Lecture Notes on Data Engineering
and Communications Technologies 75

S. Smys
Robert Bestak
Ram Palanisamy
Ivan Kotuliak *Editors*

Computer Networks and Inventive Communication Technologies

Proceedings of Fourth ICCNCT 2021

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 75

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 http://www.springer.com/series/15362

S. Smys · Robert Bestak · Ram Palanisamy ·
Ivan Kotuliak
Editors

Computer Networks and Inventive Communication Technologies

Proceedings of Fourth ICCNCT 2021

Springer

Editors

S. Smys
Department of Information Technology
RVS Technical Campus
Coimbatore, Tamil Nadu, India

Robert Bestak
Department of Telecommunication Engineering
Czech Technical University in Prague
Praha, Czech Republic

Ram Palanisamy
Gerald Schwartz School of Business
St. Francis Xavier University
Antigonish, NS, Canada

Ivan Kotuliak
Faculty of Informatics and Information Technology
Slovak University Technology
Bratislava, Slovakia

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-981-16-3727-8 ISBN 978-981-16-3728-5 (eBook)
https://doi.org/10.1007/978-981-16-3728-5

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 ICCNCT 2021 to all the participants, reviewers, and editors of ICCNCT 2021.

Preface

With a deep satisfaction, I write this Preface to the proceedings of the ICCNCT 2021 held in RVS Technical Campus, Coimbatore, Tamil Nadu, India, on April 23–24, 2021.

This conference proceedings volume contains the written versions of most of the contributions presented during the conference of ICCNCT 2021. The conference has provided a setting for discussing the recent developments in a wide variety of topics including network operations and management, QoS and resource management, wireless communications, and delay-tolerant networks. The conference has been a good opportunity for participants who came from different destinations across the globe to present and discuss the topics in their respective research areas. ICCNCT 2021 tends to collect the latest research results and applications on computer networks and next-generation communication technologies. It includes a selection of 66 papers from 282 papers submitted to the conference from universities and industries all over the world. All the accepted papers were subjected to a strict peer-reviewing purpose by 2–4 expert referees. The papers have been selected for this volume because of their quality and 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 a special thanks to all the referees for their constructive comments on all papers, and moreover, we would also like to thank the organizing committee for their hard work. Finally, we would like to thank the Springer publications for producing this volume.

Coimbatore, India	Dr. S. Smys
Praha, Czech Republic	Dr. Robert Bestak
Antigonish, Canada	Dr. Ram Palanisamy
Bratislava, Slovakia	Dr. Ivan Kotuliak

Acknowledgements

The guest editors of ICCNCT 2021 would like to acknowledge the excellent work of our conference organizing the committee, keynote speakers, and participants for their presentation on April 23–24, 2021. We also wish to acknowledge publicly the valuable services provided by the reviewers.

On behalf of the 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. We wish to acknowledge the organizers for the discussion, suggestion, and cooperation to organize the keynote speakers of this conference. We also wish to acknowledge speakers and participants who attended this conference. We are very grateful to thank all the faculty and non-faculty members and reviewers, who helped and supported this conference to become a successful event. ICCNCT 2021 would like to acknowledge the contribution made by the organization and its volunteers, conference committee members at a local, regional, and international level, who have contributed their time, energy, and knowledge.

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

Contents

Energy Efficient Clustering in Wireless Sensor Networks by Opposition-Based Initialization Bat Algorithm 1
Nebojsa Bacanin, Uros Arnaut, Miodrag Zivkovic, Timea Bezdan, and Tarik A. Rashid

Efficient Data Collection in Wireless Sensor Network 17
Meet J. Vasani and Satish Maurya

Job Scheduling in Cloud Computing Based on DGPSO 33
J. Arul Sindiya and R. Pushpalakshmi

Read–Write Decoupled Single-Ended 9T SRAM Cell for Low Power Embedded Applications 47
Amit Singh Rajput, Arpan Dwivedi, Prashant Dwivedi, Deependra Singh Rajput, and Manisha Pattanaik

Spam Detection Using Genetic Algorithm Optimized LSTM Model 59
Abhinav Sinhmar, Vinamra Malhotra, R. K. Yadav, and Manoj Kumar

Affine Recurrence Based Key Scheduling Algorithm for the Advanced Encryption Standard 73
S. Shashankh, Tavishi Kaushik, Svarnim Agarwal, and C. R. Kavitha

Simplify Your Neural Networks: An Empirical Study on Cross-Project Defect Prediction 85
Ruchika Malhotra, Abuzar Ahmed Khan, and Amrit Khera

Emotion Recognition During Social Interactions Using Peripheral Physiological Signals 99
Priyansh Gupta, S. Ashwin Balaji, Sambhav Jain, and R. K. Yadav

Phishing Detection Using Computer Vision 113
Shekhar Khandelwal and Rik Das

A Comprehensive Attention-Based Model for Image Captioning 131
Vinod Kumar, Abhishek Dahiya, Geetanjali Saini, and Sahil Sheokand

Multimedia Text Summary Generator for Visually Impaired 147
Shreya Banerjee, Prerana Sirigeri, Rachana B. Karennavar, and R. Jayashree

Keylogger Threat to the Android Mobile Banking Applications 163
Naziour Rahaman, Salauddin Rubel, and Ahmed Al Marouf

Machine Learning-Based Network Intrusion Detection System 175
Sumedha Seniaray and Rajni Jindal

BGCNN: A Computer Vision Approach to Recognize of Yellow Mosaic Disease for Black Gram 189
Rashidul Hasan Hridoy and Aniruddha Rakshit

Irrelevant Racist Tweets Identification Using Data Mining Techniques 203
Jyothirlatha Kodali, Vyshnavi Kandikatla, Princy Nagati, Veena Nerendla, and M. Sreedevi

Smart Farming System Using IoT and Cloud 215
Neha Patil and Vaishali D. Khairnar

Multipartite Verifiable Secret Sharing Based on CRT 233
Rolla Subrahmanyam, N. Rukma Rekha, and Y. V. Subba Rao

Implementation of Smart Parking Application Using IoT and Machine Learning Algorithms 247
G. Manjula, G. Govinda Rajulu, R. Anand, and J. T. Thirukrishna

Deep Face-Iris Recognition Using Robust Image Segmentation and Hyperparameter Tuning 259
Dane Brown

Text-Based Sentiment Analysis with Classification Techniques—A State-of-Art Study 277
M. S. Kalaivani and S. Jayalakshmi

Face Mask Detection Using MobileNetV2 and Implementation Using Different Face Detectors 287
Kenneth Toppo, Neeraj Kumar, Preet Kumar, and Lavi Tanwar

Image Encryption Using Diffusion and Confusion Properties of Chaotic Algorithm 305
J. N. Swaminathan, S. Umamaheshwari, O. Vignesh, P. Raziya Sulthana, A. Hima Bindu, M. Prasanna, and M. Sravani

A Sentiment Analysis of a Boycott Movement on Twitter 313
Sooraj Bhooshan, R. Praveen Pai, and R. Nandakumar

Implementing the Comparative Analysis of AES and DES Crypt Algorithm in Cloud Computing 325
R. S. Reshma, P. P. Anjusha, and G. S. Anisha

A Model for Predictive and Prescriptive Analysis for Internet of Things Edge Devices with Artificial Intelligence 333
Dinkar R. Patnaik Patnaikuni and S. N. Chamatagoudar

Comparative Analysis of SIM-Based Hybrid Modulation Schemes Over Log-Normal Channel Model 343
Siddhi Gangwar, Kavita, Subhash Burdak, and Yashna Sharma

Lesion Preprocessing Techniques in Automatic Melanoma Detection System—A Comparative Study 357
Shakti Kumar and Anuj Kumar

A Comparatıve Analysis on Three Consensus Algorithms 369
Aswathi A. Menon, T. Saranya, Sheetal Sureshbabu, and A. S. Mahesh

Issues and Challenges in the Implementation of 5G Technology 385
Mithila Bihari Sah, Abhay Bindle, and Tarun Gulati

IoT-Based Autonomous Energy-Efficient WSN Platform for Home/Office Automation Using Raspberry Pi 399
M. Chandrakala, G. Dhanalakshmi, and K. Rajesh

Detection of Early Depression Signals Using Social Media Sentiment Analysis on Big Data 413
Shruti S. Nair, Amritha Ashok, R. Divya Pai, and A. G. Hari Narayanan

Raspberry Pi-Based Heart Attack and Alcohol Alert System Over Internet of Things for Secure Transportation 423
G. Dhanalakshmi, K. Jeevana Jyothi, and B. Naveena

Automatic Classification of Music Genre Using SVM 439
Nandkishor Narkhede, Sumit Mathur, and Anand Bhaskar

Crime Rate Prediction Based on K-means Clustering and Decision Tree Algorithm 451
Jogendra Kumar, M. Sravani, Muvva Akhil, Pallapothu Sureshkumar, and Valiveti Yasaswi

Comparative Study of Optimization Algorithm in Deep CNN-Based Model for Sign Language Recognition 463
Rajesh George Rajan and P. Selvi Rajendran

Cardinal Correlated Oversampling for Detection of Malicious Web Links Using Machine Learning 473
M. Shyamala Devi, Uttam Gupta, Khomchand Sahu, Ranjan Jyoti Das, and Santhosh Veeraraghavan Ramesh

Simulation of Speckle Noise Using Image Processing Techniques 489
Noor H. Rasham, Heba Kh. Abbas, Asmaa A. Abdul Razaq, and Haidar J. Mohamad

Wi-Fi-Based Indoor Patient Location Identifier for COVID-19 503
A. Noble Mary Juliet, N. Suba Rani, S. R. Dheepiga, and R. Sam Rishi

Enabling Identity-Based Data Security with Cloud 513
Arya Sundaresan, Meghna Vinod, Sreelekshmi M. Nair, and V. R. Rajalakshmi

A Study and Review on Image Steganography 523
Trishna Paul, Sanchita Ghosh, and Anandaprova Majumder

Fault Detection in SPS Using Image Encoding and Deep Learning 533
P. Hari Prasad, N. S. Jai Aakash, T. Avinash, S. Aravind, M. Ganesan, and R. Lavanya

A Comparative Study of Information Retrieval Models for Short Document Summaries 547
Digvijay Desai, Aniruddha Ghadge, Roshan Wazare, and Jayshree Bagade

Network Attack Detection with QNNBADT in Minimal Response Times Using Minimized Features 563
S. Ramakrishnan and A. Senthil Rajan

Deep Learning-Based Approach for Satellite Image Reconstruction Using Handcrafted Prior 581
Jaya Saxena, Anubha Jain, and Pisipati Radha Krishna

CLOP Ransomware Analysis Using Machine Learning Approach 593
E. S. Aiswarya, Adheena Maria Benny, and Leena Vishnu Namboothiri

Integration of Wireless Sensors to Detect Fluid Leaks in Industries 601
N. Santhosh, V. A. Vishanth, Y. Palaniappan, V. Rohith, and M. Ganesan

Performance Analysis of Abstract-Based Classification of Medical Journals Using Machine Learning Techniques 613
A. Deepika and N. Radha

Development of Improved SoC PTS Algorithm for PAPR Reduction in OFDM Underwater Communication 627
M. Asha and T. P. Surekha

Analysis of Twitter Data for Identifying Trending Domains in Blockchain Technology 651
Sahithya Mareddy and Deepa Gupta

Enhancing the Security for Smart Card-Based Embedded Systems 673
G. Kalyana Abenanth, K. Harish, V. Sachin, A. Rushyendra, and N. Mohankumar

Implementation Mobile App for Foreign Language Acquisition Based on Structural Visual Method 687
Imad Tahini and Alex Dadykin

RETRACTED CHAPTER: Assessing Deep Neural Network and Shallow for Network Intrusion Detection Systems in Cyber Security 703
Deena Babu Mandru, M. Aruna Safali, N. Raghavendra Sai, and G. Sai Chaitanya Kumar

Leveraging Association Rules in Feature Selection to Classify Text 715
Zaher Al Aghbari and Mozamel M. Saeed

Protected Admittance E-Health Record System Using Blockchain Technology 723
Sharyu Kadam and Dilip Motwani

Using Hierarchical Transformers for Document Classification in Tamil Language 741
M. Riyaz Ahmed, Bhuvan Raghuraman, and J. Briskilal

Analysis of Hybrid MAC Protocols in Vehicular Ad Hoc Networks (VANET) for QoS Sensitive IoT Applications 753
Nadine Hasan, Arun Kumar Ray, and Ayaskanta Mishra

Programming with Natural Languages: A Survey 767
Julien Joseph Thomas, Vishnu Suresh, Muhammed Anas, Sayu Sajeev, and K. S. Sunil

An Exhaustive Exploration of Electromagnetism for Underwater Wireless Sensor Networks 781
Nebu Pulickal and C. D. Suriyakala

Two-Stage Feature Selection Pipeline for Text Classification 795
Vinod Kumar, Abhishek Sharma, Anil Bansal, and Jagnur Singh Sandhu

A Systematic Approach of Analysing Network Traffic Using Packet Sniffing with Scapy Framework 811
S. H. Brahmanand, N. Dayanand Lal, D. S. Sahana, G. S. Nijguna, and Parikshith Nayak

Detecting Ransomware Attacks Distribution Through Phishing URLs Using Machine Learning 821
B. N. Chaithanya and S. H. Brahmananda

A Framework for APT Detection Based on Host Destination and Packet—Analysis 833
R. C. Veena and S. H. Brahmananda

A Trust-Based Handover Authentication in an SDN 5G Heterogeneous Network 841
D. Sangeetha, S. Selvi, and A. Keerthana

Intrusion Detection for Vehicular Ad Hoc Network Based on Deep Belief Network 853
Rasika S. Vitalkar, Samrat S. Thorat, and Dinesh V. Rojatkar

Highly Secured Steganography Method for Image Communication using Random Byte Hiding and Confused & Diffused Encryption 867
S. Aswath, R. S. Valarmathi, C. H. Mohan Sai Kumar, and M. Pandiyarajan

An Enhanced Energy Constraint Secure Routing Protocol for Clustered Randomly Distributed MANETs Using ADAM's Algorithm 885
Bandani Anil Kumar, Makam Venkata Subamanyam, and Kodati Satya Prasad

Retraction Note to: Assessing Deep Neural Network and Shallow for Network Intrusion Detection Systems in Cyber Security C1
Deena Babu Mandru, M. Aruna Safali, N. Raghavendra Sai, and G. Sai Chaitanya Kumar

Author Index 901

About the Editors

Dr. S. Smys received his M.E. and Ph.D. degrees all in Wireless Communication and Networking from Anna University and Karunya University, India. His main area of research activity is localization and routing architecture in wireless networks. He serves as Associate Editor of *Computers and Electrical Engineering* (C&EE) Journal, Elsevier, and Guest Editor of *MONET* Journal, Springer. He is served as Reviewer for *IET*, Springer, Inderscience and Elsevier journals. He has published many research articles in refereed journals and IEEE conferences. He has been General chair, Session Chair, TPC Chair and Panelist in several conferences. He is Member of IEEE and Senior Member of IACSIT wireless research group. He has been serving as Organizing Chair and Program Chair of several International conferences and in the Program Committees of several International conferences. Currently, he is working as Professor in the Department of Information Technology at RVS technical Campus, Coimbatore, India.

Robert Bestak obtained Ph.D. degree in Computer Science from ENST Paris, France (2003), and M.Sc. degree in Telecommunications from Czech Technical University in Prague, CTU, Czech Republic (1999). Since 2004, he has been Assistant Professor at Department of Telecommunication Engineering, Faculty of Electrical Engineering, CTU. He participated in several national, EU, and third-party research projects. He is Czech Representative in the IFIP TC6 organization, and Chair of Working Group TC6 WG6.8. He annually serves as Steering and Technical Program Committee Member of numerous IEEE/IFIP conferences (Networking, WMNC, NGMAST, etc.), and he is Member of Editorial Board of several international journals (*Computers and Electrical Engineering*, *Electronic Commerce Research Journal*, etc.). His research interests include 5G networks, spectrum management and big data in mobile networks.

Dr. Ram Palanisamy is Professor of Enterprise Systems at the Gerald Schwartz School of Business, St. Francis Xavier University, Canada. He obtained his Ph.D. in information systems management from Indian Institute of Technology (IIT), New Delhi, India. He had academic positions at Wayne State University, Detroit, USA; University Telekom Malaysia and National Institute of Technology, Tiruchirappalli, India. Palanisamy's research has appeared in several peer-reviewed articles in several journals, edited books and conference proceedings.

Ivan Kotuliak received Ph.D. degree from both Versailles University and Slovak University of Technology in 2003. From that time, he joined Slovak University of Technology and worked as Researcher and Associate Professor. His research orientation is focused on network performance, including NGN architecture, wireless and mobile networking, VoIP systems and future Internet. He has been Author and Co-Author of more than sixty scientific papers and leads and participates in several international and national research projects.

Energy Efficient Clustering in Wireless Sensor Networks by Opposition-Based Initialization Bat Algorithm

Nebojsa Bacanin, Uros Arnaut, Miodrag Zivkovic, Timea Bezdan, and Tarik A. Rashid

Abstract Wireless sensor networks belong to the group of technologies that enabled emerging and fast developing of other novel technologies such as cloud computing, environmental and air pollution monitoring, and health applications. One important challenge that must be solved for any wireless sensor network is energy-efficient clustering, that is categorized as NP-hard problem. This led to a great number of novel clustering algorithms, that emerged with sole purpose to establish the proper balance in energy consumption between the sensors, and to enhance the efficiency and lifetime of the network itself. In this manuscript, a modified version of the bat algorithm, that belongs to a group of nature-inspired swarm intelligence metaheuristics, is proposed. Devised algorithm was utilized to tackle the energy-efficient clustering problems. Performance of proposed improved bat metaheuristics has been validated by conducting a comparative analysis with its original version, and also with other metaheuristics approaches that were tested for the same problem. Obtained results from conducted experiments suggest that the proposed method's performance is superior, and that it could bring valuable results in the other domains of use as well.

Keywords Lifetime optimization · Wireless sensor networks · Metaheuristics · Optimization · Bat algorithm · Enhanced algorithm · Localization

N. Bacanin (✉) · U. Arnaut · M. Zivkovic · T. Bezdan
Singidunum University, Danijelova 32, 11000 Belgrade, Serbia
e-mail: nbacanin@singidunum.ac.rs

U. Arnaut
e-mail: uarnaut@singidunum.ac.rs

M. Zivkovic
e-mail: mzivkovic@singidunum.ac.rs

T. Bezdan
e-mail: tbezdan@singidunum.ac.rs

T. A. Rashid
Department of Computer Science and Engineering, University of Kurdistan of Hewler, Erbil, KRG, Iraq
e-mail: tarik.ahmed@ukh.edu.krd

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_1

1 Introduction

Wireless sensor networks (WSNs) have grown tremendously alongside wireless communication and electronics in recent years. This is because people seek WSNs, which will operate perfectly among a wide variety of diverse and complex solutions. An indispensable part of those WSNs is sensor nodes (SN), which aim to observe nature activities, whether physical or environmental. For example, SNs are in charge of measuring humidity, temperature, and sound waves in certain areas. These measurements can become problematic if the scope is hard to approach or there are too many different types of interference. Gathering data from this kind of surroundings can be tricky and that is the main reason SNs have more than one sensor. Once data is collected, it needs to be transported to the main location, known as base station (BS). For this transmission, sensors are using analogue to digital converter (ADC), whose task is to receive information and process it further to the BS. One of the BS's main tasks is to analyze collected data and serve decision-making applications. When it comes to problems with SNs, it could be with its most important part-the power supply, as SNs have a limited battery life span, which causes networks premature depletion [22].

During the process of deployment, sensors are typically scattered randomly in a dense layout over a target area, i.e., released from a plane that flies over the volcano or hostile area. The main objective is to drop the sensors close to the target, in order to be able to obtain valid measurements and adequate coverage. From that point, sensors must operate on their own, as maintenance or human supervision is in most cases not feasible. This also means that it is not possible to replace the batteries, so the sensors must work with limited energy supplies. The sensors must cooperate and organize themselves (exact location of some individual nodes may be unknown), communicate between them wirelessly, while trying to keep the whole network alive long enough to complete its objective [39].

As to the earlier mentioned NP-hard problems, it is suggested using metaheuristics algorithms to solve them. Many complex theories find it quite challenging to solve those problems with deterministic algorithms, suggesting using heuristic algorithms. One of those theories is the complexity theory, which provides a systematic and strict theoretical framework to classify difficult problems. Deterministic algorithms can be useful to some extent, but the solution should be switched to a heuristic algorithm when it reaches that point [20].

Stochastic algorithms (SA) are sometimes tricky to understand, not because of their complexity but also their simplicity. Stochastic algorithms consist of random search guided by heuristics hints for the evaluation's next results [9].

However, they are not always precise, and there is no guarantee for an optimal solution. The reason for believing in this is that NP-hard problems run in exponential time [20].

To surpass those minor but unsolved problems using metaheuristic optimization techniques is recommended. Metaheuristic provides an optimal solution instead of giving exact results, and it consists of using iterative trial and error processes. Many of those methods are nature-inspired, and their latest development is to use metaheuristics. Moreover, metaheuristics can be used for optimization problems when it comes to incomplete information or limited capacities. Metaheuristic methods encompass different algorithms such as genetic algorithms and swarm intelligence [9, 18].

Genetic algorithm (GA) is a member of the group of adaptive search methods that are based on natural genetic selection processes. Initially, this algorithm starts generating a random set of permutations, then each of these permutations is evaluated, and at the end, the pairs of chosen individuals undergo the processes of combination and mutation to create new set of permutations [13].

Nature-inspired algorithms are behavior patterns which could be applied for solving developing problems, and they fall in two categories: based on evolutionary approach or on swarm intelligence (SI).

The evolutionary approach is based on Darwin's principle using a natural selection process. While SI approach is based on the collective intelligence and social behavior expressed by a group of small animals, for example, ant colony, firefly, gray wolf, monarch butterfly, and bee colony. As research shows, many algorithms could provide a solution for NP-hard problems. Because of their complexity, SI is remarkably the best way to find an optimal solution in an acceptable amount of time. Depending on the problem, a nature-inspired pattern could be used in various scientific fields, such as engineering, cloud computing, programming, and wireless sensor networks [19].

This paper proposes an improved implementation of a widely used bat algorithm (BA), that belongs to SI family. BA is extremely powerful optimizer, that has proven to be efficient in solving various practical problems. However, the premature converging can happen in some cases, as reported by several studies. To enhance exploitation of original BA, enhanced BA utilizes improved initialization strategy. Devised approach was adapted for tackling clustering challenge in WSNs while minimizing energy consumption. The proposed method is used to select the cluster heads in an optimal way, that is considered to be NP-hard problem.

The remainder of this paper is structured in the following way: Sect. 2 gives an overview of the SI metaheuristics and their applications, with a focus on the applications in the domain of WSN. The original and the enhanced BAs are given in Sect. 3. The results of the conducted experiments together with the discussion are presented in Sect. 4. Finally, Sect. 5 gives final remarks and suggest the possible future work.

2 Background and Related Work

People get often inspired by nature, not only for art but also for engineering, computing, and software development. Sometimes it is about the look, sometimes, it is about the mechanics, but animal behavior in nature becomes more and more popular when a solution is hard to find in software development. When we talk about nature-inspired algorithms in software development, they can be divided into two categories: based on the evolutionary approach (EA) and the other based on SI [4].

The GA are base of the evolutionary approach, which means that only the fittest will survive, as it is known that chromosomes are continually changing. According to Darwin's theory, new generations are evolving from the previous generations, and only the strongest and the fittest will have the possibility to evolve. The same is with GA, a search optimization algorithm whose elements tend to change to become more competitive. The first step in solving a problem using a genetic algorithm is defining the representation structure and covert a chromosome into a solution [12, 16, 27].

On the other hand, the SI metaheuristics inspired by the group, or to be precise, the collective behavior of animals who tend to move in swarms, such as insects, birds, fish, and other animals. This behavior is best reflected when animals are looking for food, shelter or whether trying to orient themselves in the dark. As people have almost lost their abilities to survive in nature, scientists have made an effort to develop various mathematical models that will represent animal behavior in the best possible way [4].

Many SI metaheuristics algorithms are introduced and widely used for different problems. Some of the mentioned algorithms are bat optimization algorithm (BOA) [35, 38], Harris hawk optimizer (HHO) [11], gray wolf optimizer (GWO) [3], monarch butterfly optimization (MBO) [33], moth search algorithm (MSA) [36], whale optimization algorithm (WOA) [24, 30], artificial ant colony (AAC) [10], artificial bee colony (ABC) [6], dragonfly algorithm (DA) [23] and firefly algorithm (FA) [7]. The bat optimization algorithm is built on the, as the algorithm name says, bats behavior. To be precise, when bats fly, they need something to orient themselves as they are blind. In order to get the surrounding image, they use sound reflection. Based on this reflection, bats can identify moving or immobilized objects to find flying routes or prey, as well as the distance between them and the object [8, 17].

On the other hand, HHO Algorithm observes a group of Harris's Hawks who use hunt and pursuit tactics, which is more known as a surprise pounce. The metaheuristics exploration for this algorithm finds its inspiration in the earlier mentioned Hawk's surprise pounce tactics along with their capturing strategies during the pursuit maneuvers.

One of the most used algorithms is ABC, which describes a bee swarm's behavior pattern while they are in a food quest. The scientists are familiar with two types of this algorithm, one is classic ABC, and the other one is improved ABC (iABC) [4].

Various NP-hard problems from the WSN domain have been already tackled by SI metaheuristics. The WSN energy efficiency and clustering problems have been tackled by numerous SI algorithms in the past, with the main objective to optimize the cluster head selection. For instance, PSO approach was selected in [2] to address the hot spot issue, by separating the clustering into two phases and utilization of the efficient clustering and routing approaches. This type of approach was also suggested in [37]. Authors in [14] observed the clusters as the flocks of birds and applied the algorithm that was based on the RSSI (received signal strength indicator) and were able to obtain very promising results. Localization and prolonging the WSN lifetime problems were also addressed by multiple SI algorithms with promising results, as suggested in [5, 29, 31, 32, 40]. Clustering problem was also dealt with in other

Table 1 Comparative analysis of the existing protocols

Method	Type	Energy efficiency	Limitations
LEACH	Classic	Average	Very high costs for communication
HEED	Classic	Average	Latency in network can be high
MBC	Classic	Average	Very high costs for communication
BeeSensor	SI	Satisfactory	Not cluster-based
PSO	SI	Satisfactory	Large overhead
EEABR	SI	Average	Method is proactive

recent papers, including machine learning-based approach [26] and novel cluster rotating and routing approach described in [25].

The short overview of popular methods used to address the clustering task and energy efficiency maximization is given in Table 1 [22], with short remarks about the energy efficiency and drawbacks of each method.

3 Bat Algorithm

Bat algorithm (BA) is a SI algorithm based on bat behavior. It was introduced back in 2010 by a scientist named Xin-She Yang. The algorithm's main idea is to use the bat's hunting abilities to create an optimal search process. The ability allows the bat to recognize its prey, trees, and any other object. They are using sound waves or, in other words, echolocation. The way echolocation work is based on sound waves reflection; those waves allow a bat to define a distance between him and objects in his surrounding, also it allows him to create an image of objects he is facing [8, 17].

In the first part of this section, the original BA is given. However, some deficiencies of the original approach were observed that leave space for improvements. Finally, the enhanced BA with a novel initialization strategy is described.

3.1 Original Bat Algorithm

While the optimization process is gathering data, the bat's location update is needed. The way to represent this algorithm is by using several equations to define each step in this process:

- Bat's location update
- The velocity at time
- The solution's frequency is uniformly drawn in defined range
- The current fittest solution is modified by the random walk
- Finding prey

The first step is defined by the search optimization process using the following equation:

$$x_i^t = x_i^{t-1} + v_i^t, \tag{1}$$

it allows to define bat's current location where the current solution is denoted by x_i^{t-1} and the new, updated position at iteration t of the i-th solution is denoted by x_i^t. The velocity is denoted by v_i^t.

The second step defines his velocity at time step t where it is calculated as following:

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_*) f_i, \tag{2}$$

where the current global best position is denoted by x_*, and f_i indicates the frequency of i-th bat.

In the third step frequency is uniformly allocated from earlier defined frequency range (minimum and maximum), it is evaluated as following:

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta, \tag{3}$$

where $f_{\min}$ and $f_{\max}$ are the minimum and maximum frequency, respectively and β is a random number, $\beta \in [0, 1]$.

The exploitation process of this algorithm is directed in the fourth step, where the random walk modifies the current fittest solution. It is defined as following:

$$x_{\text{new}} = x_{\text{old}} + \epsilon A^t, \tag{4}$$

where the average loudness value of all bats are denoted by A^t, ϵ is a scaling factor with a random value between 0 and 1.

In the last step, the bat's prey is found, and the loudness is updated by using following equation:

$$A_i^t = \alpha A_i^{t-1},\ r_i^t = r_i^0[1 - exp(-\gamma t)] \tag{5}$$

$$A_i^t \rightarrow 0,\ r_i^t \rightarrow r_i^0,\ \text{while } t \rightarrow \infty \tag{6}$$

where A_i^t indicates the loudness of i-th bat, at iteration t, and r is the pulse emission rate. The values of α and γ are constant.

3.2 Improved BA

According to results obtained from conducted simulations with standard benchmarks for global optimization taken from Congress on Evolutionary Computation 2006 (CEC 2006) [21] benchmark suite, as well as the previous research [35], it can be concluded that in some runs original BA converges to non-optimal regions of the search space and as a consequence worse mean values are generated. This is due to lack of exploration power, that is especially needed in early iterations.

To overcome observed drawbacks, two strategies can be employed: first, effective exploration mechanism can be adopted from other metaheuristics [8] and second efficient initialization mechanism can be incorporated, as it is proposed in this research. If initial population is generated in the promising region of the search space, logical assumption is that, assuming that the search process is guided by efficient exploitation, then the algorithm will converge toward optimum solution.

Standard initialization phase of the original BA, where solutions are generated within lower and upper bounds of the search space is replaced with opposition-based learning (OBL) mechanism. This procedure was firstly introduced by Tizhoosh in 2005 [34] and proved that it can significantly improve search process of optimization method by improving both, exploitation and exploration [1, 28].

Let x_j represents parameter j of individual x.. The opposite number x_j^o can be calculated as:

$$x_j^o = lb_j + ub_j - x_j, \tag{7}$$

where $x_j \in [lb_j, ub_j]$ and $lb_j, ub_j \in R, \forall j \in 1, 2, 3, ...D$. Notations lb_j and ub_j represent lower and upper bound of j-th parameter, respectively and D denotes the number of solution dimensions (parameters).

In the initialization phase, first N solutions are generated by using classic initialization scheme, that is given in Eq. (8). In this way standard population X is generated. Afterwards, for each solution $x \in X$, opposite individual x^o is generated by applying Eq. 7 and opposite population X^O is generated. Finally, populations X and X^O are merged together ($X \cup X^O$), and solutions in such merged population are sorted based on fitness in descending order and best N individuals are select for iterative search process of modified BA.

$$x_{i,j} = lb_j + \text{rand}(ub_j - lb_j), \tag{8}$$

where rand is a random number from the uniform or normal distribution.

The name of proposed approach, which is inspired by modified initialization phase, is opposition-based initialization BA (OBI-BA). Pseudo-code of proposed method is given in Algorithm 1. The flowchart of the proposed algorithm is given in Fig. 1.

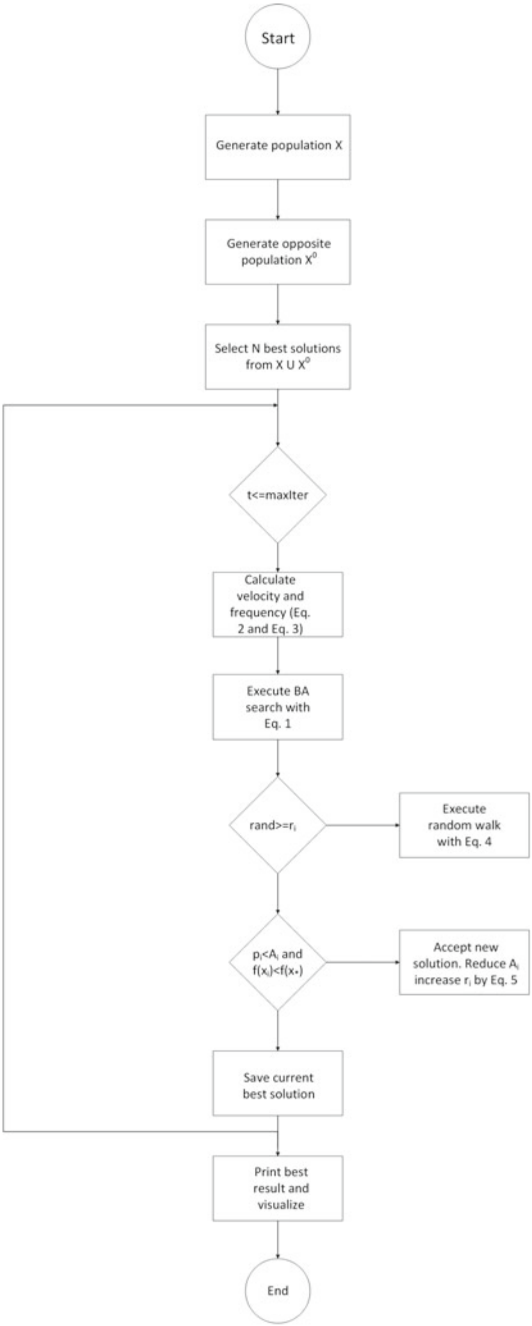

Start
Generate population X
Generate opposite population X^0
Select N best solutions from X U X^0
t<=maxIter
Calculate velocity and frequency (Eq. 2 and Eq. 3)
Execute BA search with Eq. 1
rand>=r_i
Execute random walk with Eq. 4
p_i<A_i and $f(x_i)$<$f(x_*)$
Accept new solution. Reduce A_i increase r_i by Eq. 5
Save current best solution
Print best result and visualize
End

Fig. 1 Proposed OBI-BA flowchart

Algorithm 1 Pseudocode of proposed OBI-BA

```
Objective function $f(x)$
Initialize population $X$ of $N$ bats based on Eq. (8), the values of $v_i$, $r_i$ and $A_i$, define the frequency of pulse $(f_i)$ at $x_i$, the value of the maximum iteration $(MaxIter)$, and set the iteration counter $(t)$ to 0
Generate opposition population $X^O$ by applying Eq. (7) for each solution $x$
Generate $X \cup X^O$, sort solutions based on fitness in descending order and select $N$ best individuals

while $t < MaxIter$ do
  for $i = 1$ to $N$ (each $N$ individuals in the population) do
    Calculate the velocity and frequency value by using Eq. (2) and Eq. (3)
    Perform the bat search procedure using Eq. (1)
    if $rand > r_i$ then
      Select the fittest solution
      Perform the random walk process by using Eq. (4)
    end if
    if $(p_i < A_i$ and $f(x_i) < f(x_*)$ then
      The newly generated solution is accepted
      Reduce $A_i$ and increase $r_i$ by utilizing Eq. (5)
    end if
  end for
  Find and save the current best solution $x_*$
end while
Return the best solution
```

4 Simulations and Discussion

The experiments and simulations performed in this research utilize the same experimental setup as described in [22], as it was intended to provide a valid comparative analysis between the proposed OBI-BA approach and metaheuristics described in this publication. The authors in [22] implemented the improved iABC algorithm to tackle the same research problem and compared it to the other ABC variations. The simulated deployment area is a two-dimensional $(2D)$ area with dimensions 150 m × 150 m, as stated in [22]. Sensor nodes defined by a pair of coordinates (x, y) are randomly placed over the target area with a help of a pseudo-random number generator, as shown in Fig. 2. In all conducted simulations, the BS is located inside the target area's borders. Clustering protocol chooses the cluster heads based on the residual energy of sensor nodes, as stated in [22].

Original BA, OBI-BA, and the WSN simulation framework have been developed by using the JDK (Java Development Kit) 11. All simulations were executed on the following configuration: Intel® CoreTM i7-8700K CPU with 64GB RAM and Windows 10 O.S. The parameters for the model used in the simulations were extracted from [22].

As in the referred paper [22], two WSN scenarios were implemented. The first scenario included 150–900 sensors placed in the random fashion over the target area, where the BS was located at the position (60, 120 m), in other words within the target

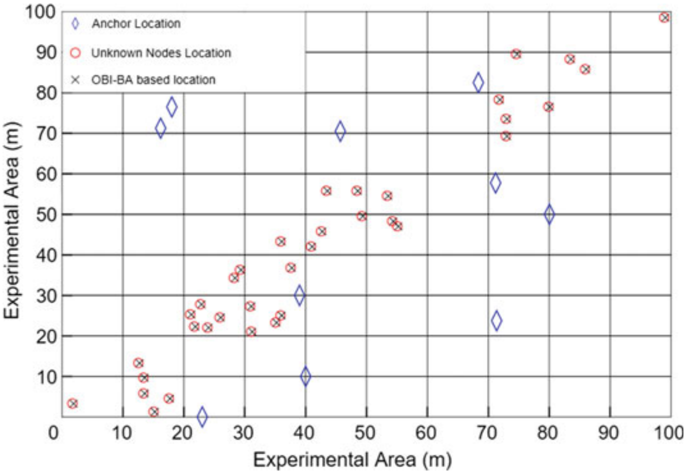


Fig. 2 Sensors deployment over the target area

area. The second scenario differs only in terms that the BS is placed at the position (100, 250 m), so practically it is outside of the target area.

The first part of the experiments determines the packet delivery ratio in the WSN for both observed scenarios, and the results are shown in Figs. 3 and 4, respectively. In the first scenario, where the base station is within the borders of the deployment area, the proposed OBI-BA approach was able to deliver the highest number of packets, slightly outperforming the EEBC metaheuristic and reaching almost 100% packets when the total number of sensors is set to 150, as shown on Fig. 3. The basic BA obtained reasonable results; however, it performed worse than EEBC. The similar results were obtained when the number of the sensor nodes was increased to 900. In case of the second scenario, where the base station is located outside the borders of the target area, OBI-BA obtained highest scores again, as shown in Fig. 4.

The simulation parameter's overview has been shown in Table 2. More details about simulation parameters can be found in the paper named "Improved artificial bee colony metaheuristic for energy-efficient clustering in wireless sensor networks" [15].

Throughput is typically used as a key performance indicator of robustness of any algorithm. The results of the conducted experiments are shown on Figs. 5 and 6, for scenarios one and two, respectively. In both scenarios, the proposed OBI-BA metaheuristics outperforms all other approaches included in the comparative analysis. On the other hand, the original BA performs worse than the EEBC method.

The results concerning the energy consumption for both observed scenarios are shown in Figs. 7 and 8. From the figures, it can be observed that the proposed OBI-BA approach slightly outperformed the EEBC metaheuristics for both scenarios, and drastically outperformed other metaheuristics included in the comparative analysis, including the basic BA.

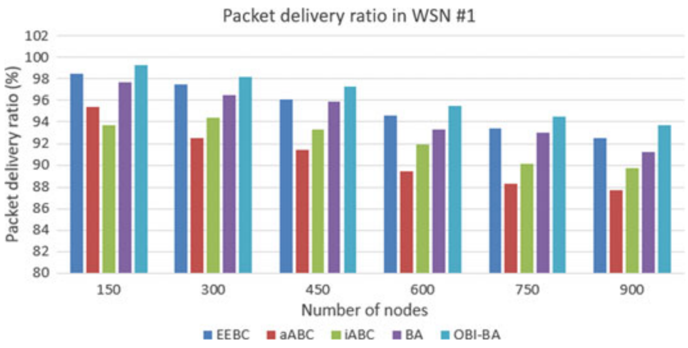


Fig. 3 Packet delivery ratio in the first scenario

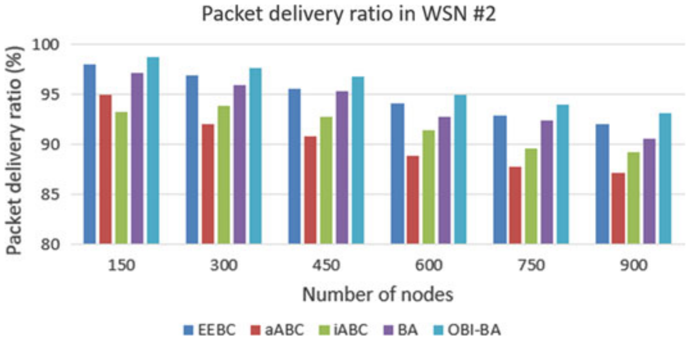


Fig. 4 Packet delivery ratio in the second scenario

Table 2 Summary of model parameters

Parameter	Value
Terrain size	150 × 150 m^2
MAC Protocol	802.11
Radio propagation	Free space
E_{fs}	6 pJ/bit/m
E_{mp}	0.00011pJ/bit/m4
Propagation limit	−111 dBm
Receiver sensitivity	−89
Data rate	2 Mbps
Packet size	3000 bits
Message size	4000 bits

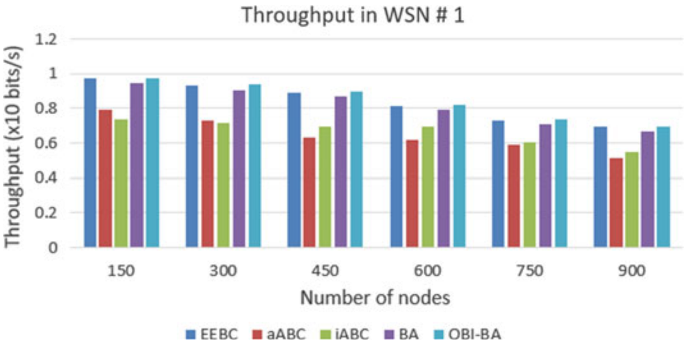


Fig. 5 Throughput in the first scenario

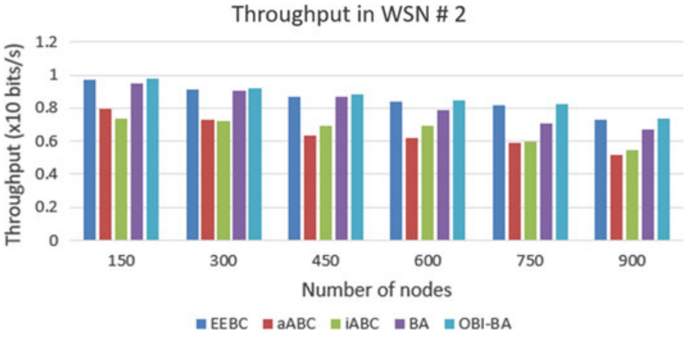


Fig. 6 Throughput in the second scenario

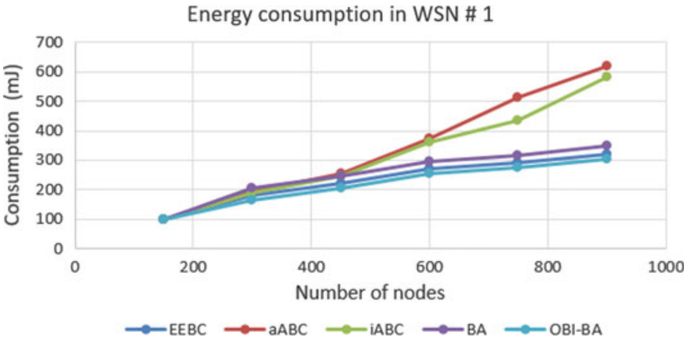


Fig. 7 Energy consumption in the first scenario

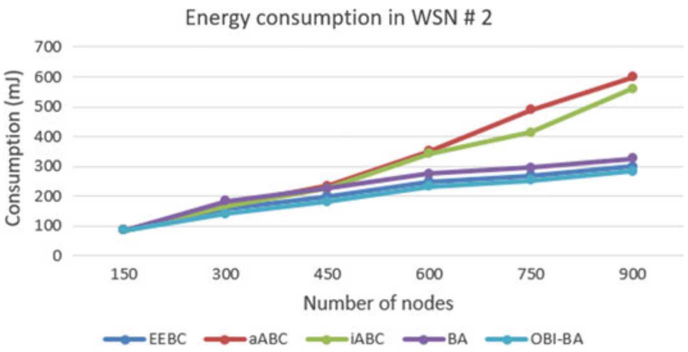


Fig. 8 Energy consumption in the second scenario

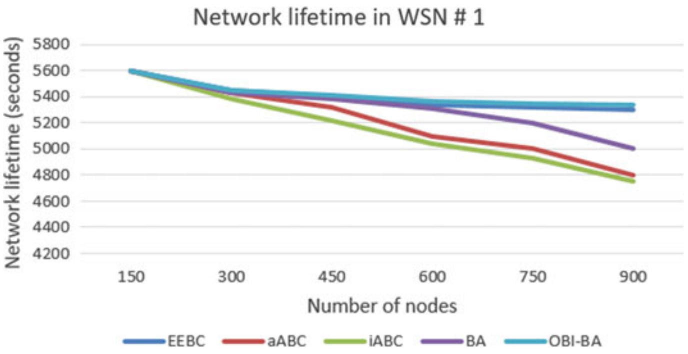


Fig. 9 Network lifetime in the first scenario

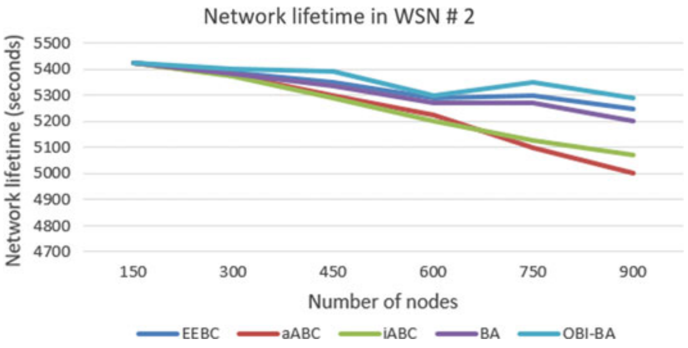


Fig. 10 Network lifetime in the second scenario

Finally, Figs. 9 and 10 show that the proposed OBI-BA metaheuristics was able to slightly extend the average WSN lifetime when compared to EEBC, and even more drastically when compared to other metaheuristics included in the research.

5 Conclusion

This manuscript presents an improved BA (OBI-BA) SI metaheuristics, that was utilized to address the clustering in WSN. The proposed method was tested through exhaustive simulations where sensors were randomly placed over a simulated area in every run. Additionally, the proposed OBI-BA was compared to the original BA, EEBC, iABC and aABC metaheuristics.

The scientific contribution of the proposed method is dual: firstly, the original BA metaheuristic was enhanced in a way to specifically target its deficiencies, and secondly, the enhanced algorithm was then applied to the clustering task in WSN, and proven to be more efficient than other methods that were included in the simulations.

The results that were obtained throughout the conducted experiments indicate that the proposed OBI-BA metaheuristics outperformed other advanced approaches. Additionally, it can be noted that the OBI-BA obtained significantly better performances compared to the original BA method. Therefore, based on the promising results of the conducted simulations, the proposed OBI-BA metaheuristics has great potential to address the WSN clustering issue. As part of the future work, OBI-BA can be adapted even further and applied to tackle other challenges from the WSN domain.

Acknowledgements The paper is supported by the Ministry of Education, Science and Technological Development of Republic of Serbia, Grant No. III-44006.

References

1. Abd Elaziz, M., Oliva, D.: Parameter estimation of solar cells diode models by an improved opposition-based whale optimization algorithm. Energy Conv. Manage. **171**, 1843–1859 (2018). https://doi.org/10.1016/j.enconman.2018.05.062, http://www.sciencedirect.com/science/article/pii/S0196890418305405
2. Azharuddin, M., Jana, P.K.: Particle swarm optimization for maximizing lifetime of wireless sensor networks. Comput. Electr. Eng. **51**, 26–42 (2016). https://doi.org/10.1016/j.compeleceng.2016.03.002, http://www.sciencedirect.com/science/article/pii/S0045790616300404
3. Bacanin, N., Bezdan, T., Tuba, E., Strumberger, I., Tuba, M., Zivkovic, M.: Task scheduling in cloud computing environment by grey wolf optimizer. In: 2019 27th Telecommunications Forum (TELFOR), pp. 1–4 (Nov 2019). https://doi.org/10.1109/TELFOR48224.2019.8971223
4. Bacanin, N., Bezdan, T., Tuba, E., Strumberger, I., Tuba, M.: Monarch butterfly optimization based convolutional neural network design. Mathematics **8**(6), 936 (2020)
5. Bacanin, N., Tuba, E., Zivkovic, M., Strumberger, I., Tuba, M.: Whale optimization algorithm with exploratory move for wireless sensor networks localization. In: International Conference on Hybrid Intelligent Systems, pp. 328–338. Springer (2019)
6. Bacanin, N., Tuba, M.: Artificial bee colony (ABC) algorithm for constrained optimization improved with genetic operators. Stud. Inf. Control **21**(2), 137–146 (2012)
7. Bacanin, N., Tuba, M.: Firefly algorithm for cardinality constrained mean-variance portfolio optimization problem with entropy diversity constraint. Sci. World J. Special Issue Com-

put. Intell. Metaheuristic Algorithms Appl. **721521**, 16 (2014). https://doi.org/10.1155/2014/721521

8. Bezdan, T., Zivkovic, M., Tuba, E., Strumberger, I., Bacanin, N., Tuba, M.: Multi-objective task scheduling in cloud computing environment by hybridized bat algorithm. In: International Conference on Intelligent and Fuzzy Systems, pp. 718–725. Springer (2020)
9. Collet, P., Rennard, J.P.: Stochastic optimization algorithms. In: Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications, pp. 1121–1137. IGI Global (2008)
10. Dorigo, M., Birattari, M.: Ant Colony Optimization. Springer (2010)
11. Heidari, A.A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., Chen, H.: Harris hawks optimization: algorithm and applications. Futur. Gener. Comput. Syst. **97**, 849–872 (2019)
12. Izadi, A., Mohammad Kimiagari, A.: Distribution network design under demand uncertainty using genetic algorithm and Monte Carlo simulation approach: a case study in pharmaceutical industry. J. Ind. Eng. Int. **10**(1), 1–9 (2014)
13. Jiang, B., Chan, W.K.: Input-based adaptive randomized test case prioritization: a local beam search approach. J. Syst. Softw. **105**, 91–106 (2015)
14. Jung, S.G., Yeom, S., Shon, M., Kim, D., Choo, H.: Clustering Wireless Sensor Networks Based on Bird Flocking Behavior, pp. 128–137, June 2015
15. Kalra, M., Singh, S.: A review of metaheuristic scheduling techniques in cloud computing. Egyptian Inform. J. **16**(3), 275–295 (2015). https://doi.org/10.1016/j.eij.2015.07.001, http://www.sciencedirect.com/science/article/pii/S1110866515000353
16. Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. Multimedia Tools Appl. 1–36 (2020)
17. Khodadadi, A., Saeidi, S.: Discovering the maximum k-clique on social networks using bat optimization algorithm. Comput. Soc. Netw. **8**(1), 1–15 (2021)
18. Kim, J.H., et al.: Meta-heuristic algorithms as tools for hydrological science. Geosci. Lett. **1**(1), 1–7 (2014)
19. Kora, P., Kalva, S.R.: Improved bat algorithm for the detection of myocardial infarction. Springerplus **4**(1), 1–18 (2015)
20. Li, W., Ding, Y., Yang, Y., Sherratt, R.S., Park, J.H., Wang, J.: Parameterized algorithms of fundamental np-hard problems: a survey. HCIS **10**(1), 1–24 (2020)
21. Liang, J., Runarsson, T.P., Mezura-Montes, E., Clerc, M., Suganthan, P., Coello, C., Deb, K.: Problem definitions and evaluation criteria for the CEC 2006 special session on constrained real-parameter optimization (2006)
22. Mann, P.S., Singh, S.: Improved artificial bee colony metaheuristic for energy-efficient clustering in wireless sensor networks. Artif. Intell. Rev. **51**(3), 329–354 (2019)
23. Mirjalili, S.: Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. Neural Comput. Appl. **27**(4), 1053–1073 (2016)
24. Mirjalili, S., Lewis, A.: The whale optimization algorithm. Adv. Eng. Softw. **95**, 51–67 (2016). https://doi.org/10.1016/j.advengsoft.2016.01.008, http://www.sciencedirect.com/science/article/pii/S0965997816300163
25. Mugunthan, S.: Novel cluster rotating and routing strategy for software defined wireless sensor networks. J. ISMAC **2**(02), 140–146 (2020)
26. Raj, J.S.: Machine learning based resourceful clustering with load optimization for wireless sensor networks. J. Ubiquit. Comput. Commun. Technol. (UCCT) **2**(01), 29–38 (2020)
27. Semnani, D., Hadjianfar, M., Aziminia, H., Sheikhzadeh, M.: Jacquard pattern optimizing in weft knitted fabrics via interactive genetic algorithm. Fashion Text. **1**(1), 1–9 (2014)
28. da Silveira, L.A., Soncco-Álvarez, J.L., de Lima, T.A., Ayala-Rincón, M.: Memetic and opposition-based learning genetic algorithms for sorting unsigned genomes by translocations. In: Pillay, N., Engelbrecht, A.P., Abraham, A., du Plessis, M.C., Snášel, V., Muda, A.K. (eds.) Advances in Nature and Biologically Inspired Computing, pp. 73–85. Springer International Publishing, Cham (2016)
29. Strumberger, I., Tuba, E., Bacanin, N., Beko, M., Tuba, M.: Monarch butterfly optimization algorithm for localization in wireless sensor networks. In: 2018 28th International Conference

Radioelektronika (RADIOELEKTRONIKA), pp. 1–6 (April 2018). https://doi.org/10.1109/RADIOELEK.2018.8376387

30. Strumberger, I., Bacanin, N., Tuba, M., Tuba, E.: Resource scheduling in cloud computing based on a hybridized whale optimization algorithm. Appl. Sci. **9**(22), 4893 (2019). https://doi.org/10.3390/app9224893
31. Strumberger, I., Beko, M., Tuba, M., Minovic, M., Bacanin, N.: Elephant herding optimization algorithm for wireless sensor network localization problem. In: Camarinha-Matos, L.M., Adu-Kankam, K.O., Julashokri, M. (eds.) Technological Innovation for Resilient Systems, pp. 175–184. Springer International Publishing, Cham (2018)
32. Strumberger, I., Minovic, M., Tuba, M., Bacanin, N.: Performance of elephant herding optimization and tree growth algorithm adapted for node localization in wireless sensor networks. Sensors **19**(11), 2515 (2019). https://doi.org/10.3390/s19112515
33. Strumberger, I., Tuba, M., Bacanin, N., Tuba, E.: Cloudlet scheduling by hybridized monarch butterfly optimization algorithm. J. Sens. Actuator Netw. **8**(3), 44 (2019). https://doi.org/10.3390/jsan8030044
34. Tizhoosh, H.R.: Opposition-based learning: a new scheme for machine intelligence. In: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), vol. 1, pp. 695–701 (2005)
35. Tuba, M., Bacanin, N.: Hybridized bat algorithm for multi-objective radio frequency identification (RFID) network planning. In: 2015 IEEE Congress on Evolutionary Computation (CEC), pp. 499–506 (May 2015). https://doi.org/10.1109/CEC.2015.7256931
36. Wang, G.G.: Moth search algorithm: a bio-inspired metaheuristic algorithm for global optimization problems. Memetic Computing (Sep 2016). https://doi.org/10.1007/s12293-016-0212-3, https://doi.org/10.1007/s12293-016-0212-3
37. Wang, J., Gao, Y., Liu, W., Sangaiah, A.K., Kim, H.J.: An improved routing schema with special clustering using PSO algorithm for heterogeneous wireless sensor network. Sensors **19**(3) (2019). https://doi.org/10.3390/s19030671
38. Yang, X.S.: A New Metaheuristic Bat-Inspired Algorithm, pp. 65–74. Springer Berlin Heidelberg, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12538-6_6, https://doi.org/10.1007/978-3-642-12538-6_6
39. Zivkovic, M., Branovic, B., Marković, D., Popović, R.: Energy efficient security architecture for wireless sensor networks. In: 2012 20th Telecommunications Forum (TELFOR), pp. 1524–1527 (Nov 2012). https://doi.org/10.1109/TELFOR.2012.6419510
40. Zivkovic, M., Zivkovic, T., Venkatachalam, K., Bacanin, N.: Enhanced dragonfly algorithm adapted for wireless sensor network lifetime optimization. In: Data Intelligence and Cognitive Informatics, pp. 803–817. Springer (2021)

Efficient Data Collection in Wireless Sensor Network

Meet J. Vasani and Satish Maurya

Abstract There has been a steady increase in climatic and man-made disasters for the past couple of years. Due to this, numerous researches on Mobile Wireless Sensor Networks (MWSN's) are attained, the sensors which are mobile and can do numerous tasks such as go to places which are dense and do not have access to the physical human body; compensation in cost, flexible, and many more which the ordinary Wireless Sensor Networks (WSN's) lack. Many ideas have been used to give the best out of the MWSN's. In every idea so far, the Mobile Sensor (MS) collects data from the sensor and return to the station which takes time, and this delays the data which was to be returned early in order to take actions, also the capacity of the storage in MS is quite low, so a number of laps are required to collect data from all sensors. In this paper, the Models proposed can eradicate such problems by using MS as a means of connectivity, which will connect sensors and the station.

Keywords Mobile device collector · Mobile sensor · Clustering · Mobility · Efficiency

1 Introduction

Natural and Man-Made disasters include Bush Fires, Earthquakes, Volcanic Eruptions and many more which can be controlled if actions are taken at a particular time frame but the sites of occurrences of these disasters are far away from the workers who are inspecting those particular sites. It would take a lot of time to know that something wrong has taken place and by then the time frame to control the situation had already been passed (Fig. 1). It is when the technology of Sensor Networks comes in sight where the workers sitting in their offices can know what is happening at the particular site at a particular moment and can take actions immediately as something goes wrong.

M. J. Vasani (✉) · S. Maurya
PP Savani University, Surat, India

S. Maurya
e-mail: satish.maurya@ppsu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_2

Volcanic Eruption
Office Building

Fig. 1 Workers are unaware of the things happening at site and cannot take action on time

Everything around is getting evolved to get the best comfortability one seeks [1], and sensors are getting in demand at a rapid rate. Sensors are elements which sense the environment [2] and store that information/data to pass it to the station, so the station can act accordingly. Wireless Sensor Networks (WSN) is said to be a communication of data between Sensor nodes placed in a particular field through wireless link [3, 4]. In single-hop transmission, data is transferred through a single sensor directly to the station in a single hop. On the contrary multi-hop transmission, data is transferred through multiple sensors that are deployed in between the sensor sending the data and the station. When the sensors come in a range of each other then and then the data can be transferred [5] (Fig. 2).

Now to collect data from places that are out of reach, a technology called Mobility/ Mobile Wireless Sensor Networks is introduced (MWSN's). In the past couple of years, there has been collaboration of robotics with the Wireless Sensor Networks which gave rise to portability [6]. In this, the sensor is able to move around in order to transfer or collect the data. When the range is too short and the distance is too long between the sensor and the station, mobility plays a vital role [7]. The sensor can be made mobile and can move towards the station until both of them are in the desired range so that the station can receive the data or vice-versa. Such different prototypes can be created based on the requirements. In single-hop transmission, there can only be two possibilities (i.e. either make the sensor mobile to go to the

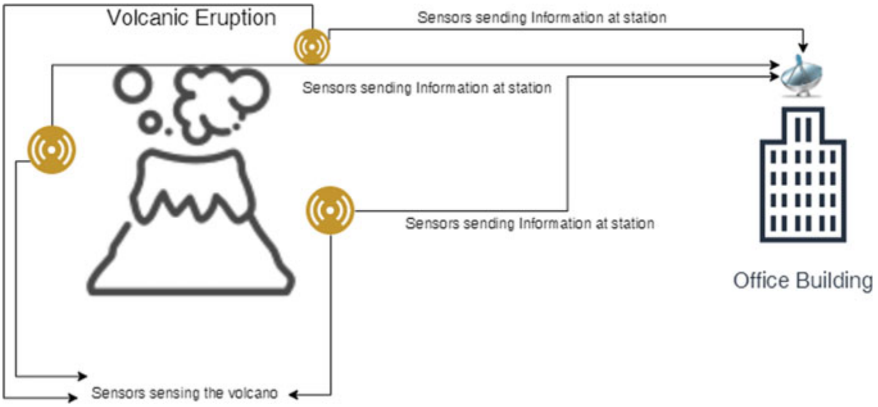

Volcanic Eruption
Sensors sending Information at station
Sensors sending Information at station
Sensors sending Information at station
Office Building
Sensors sensing the volcano

Fig. 2 Workers are aware and can take action in the particular time frame as something goes wrong

station to pass on the data or make the station move towards the sensor to collect data). In multi-hop transmission, there can be many possibilities on the passing of data to the station (i.e. various sensors can be deployed in the path of the sensor and the station which can move around to get in range and pass on the data [8] or a Cluster-based approach can be used where MS is particularly used to get data from the Cluster and come back again to the station, etc.). Multi-Hop transmissions conserve more energy than Single-Hop transmission [9]. Cluster-Based Approach: Clusters are made which consists of multiple sensors and each Cluster have a Cluster Head (CH), which transmits the data further to any of the sensor in its range [10]. Instead of using multiple sensors, one can implement mobile Sensor which will move around and do all the stuff necessary. However, due to the movement of the Mobile Sensor, it will consume more power than the regular sensor [11]. Mobility can decrease the number of hops and can give better productivity by reducing the error while transmission of the data [12].

This paper gives a brief perspective on reducing the delay of data at the station as much as possible. Further, the paper is divided into the following sections, Sect. 2 includes the main concern of proposing our Models over the previous work; Sect. 3 gives a brief of the idea to be implemented about using MS as a means of connectivity to transfer data instead of using it as a transporter. Section 4 contains Flow Diagrams for every model. Section 5 includes simulation results, which shows benefits obtained by the proposed Models; Sect. 6 states conclusion; Sect. 7 covers future work of the project idea.

2 Previous Work

The very crucial elements of the MWSN's are the following:

i. Simple Sensors: The sensor's main task is to collect the data from the surrounding [13].
ii. Sink (Station): They acquire the data collected from simple sensors sensing the surrounding and perform particular actions [13].
iii. Support Sensors: They act as an intermediate for transferring the data. They do not sense any data, but just transfers the data from one node to the other [13].

There are only two ways through which mobility can be applied:

i. Mobility of Sensor: Moving the sensor to the station.
ii. Mobility of Station: Moving the station to the particular sensor to collect the data.

Now, based on the kind of way chosen, there will be different ideas that can be/ has been introduced and every single idea will get a different output. Some of them are as follows:

i. A single sensor and a single station: Either of them can be made mobile and the data can be collected accordingly.

ii. Multiple sensors and a single station: Either sensors can be made mobile and can be sent to other sensors for collecting data and then can be sent to the station to transfer the data.

iii. Multiple sensors and a single station used as a means of mobility: The station can be made mobile and can be sent to other sensors for collecting data directly.

iv. Multiple sensors in a Cluster and a single station: In [14, 15] multiple Clusters of sensors are made. MS starts its journey from the station through its predesigned path. One by one it covers all the Clusters coming into its predesigned path. They have an algorithm set where the MS will vary its speed depending on the situation. The MS will decrease its speed while passing through any Cluster to acquire maximum data, The MS will increase its speed when it is travelling from one Cluster to the other and finally the MS will complete its journey by coming to the station. This whole journey takes place every 24 h. By chance, if there is not enough space for the MS to take data from particular Cluster, as the MS has a small amount of space, the MS will take a second lap and will have to cover all of the predesigned paths just to collect data from the particular Cluster. The MS will take as many laps as possible to collect all the data from all the Clusters. They have specified that their MS has 2 bytes of storage capacity and every 24 h 34,176 packets are generated with all the sensors sensing the environment, with each packet of 75 bytes Below Fig. 3. Is the representation of [14], they have used this same figure to illustrate their model.

v. In [13] they used the word Speed Control for the above-given idea in Fig. 3.

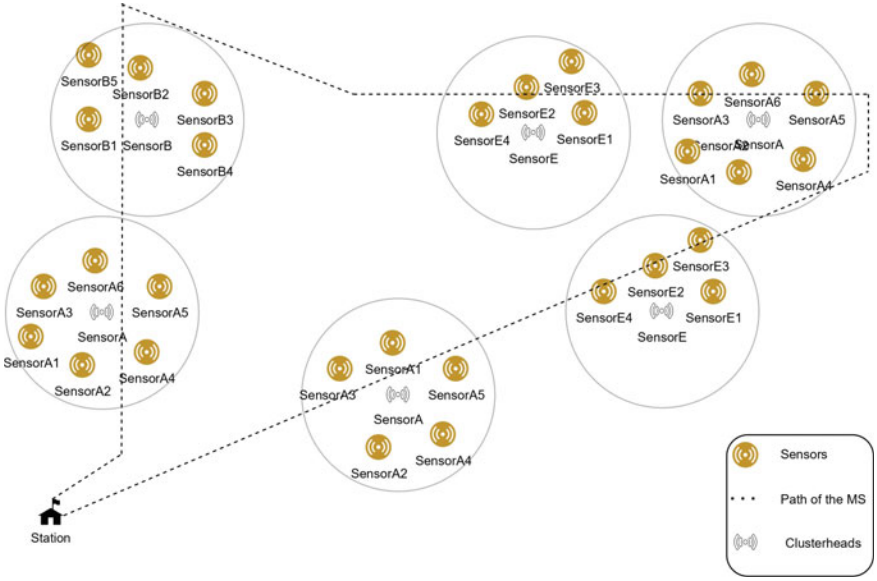


Fig. 3 Representation of [14], the complete predesigned path of the Mobile Sensor (MS)

vi. In [16], they are using the same idea of gathering the data through controlling the speed, but instead of using a simple MS to collect the data from all the Clusters, they have used a sink node to collect and store the data.

vii. In [17], multiple sink nodes are deployed to communicate and send the data from Clusters through a mobile sink node.

Now, we can see in all of the above ideas that MS is always used as a means of transporting the data from one place to another, the transportation between the sensors and the station delays the time of data to reach the station. Also, the capacity of MS is small, which makes it take laps forcefully. Also in [14], if a single Cluster wants to send data still MS will have to travel the whole predesigned path which can cause more power consumption.

To overcome these problems stated above we have designed two models which can eliminate the factors of taking multiple laps and reduce the delay of data at station.

3 Proposed Models

3.1 Model A

For Model A, we designed an idea which mainly focused on reduction of the path for MS. In [14], they have set a pre-designed path which is to be travelled by MS every time a Cluster wants to send the data. Even if a single Cluster has to send the data, then also MS has to travel an excess path to reach that particular Cluster. This causes excess paths to travel. To overcome this issue, we made a single Mega Cluster consisting of 3 to 4 Mini Clusters, such that the Clusters will send the data to the Mega Cluster and from that Mega Cluster MS can fetch the data and go back to the desired station.

In Fig. 4, the single Mega Cluster with one Cluster Head Sensor C which consists of all the 4 Mini Cluster (i.e. Cluster with Cluster Head A, Cluster with Cluster Head B, Cluster with Cluster Head E, and Cluster with Cluster Head F). Sensors sensing the environment are (Sensor A1, A2, B1, B2, E1, E2, F1, and F2), station is the sink. All the Mini Cluster Head (i.e. Cluster with Cluster Head A, Cluster with Cluster Head B, Cluster with Cluster Head E, and Cluster with Cluster Head F) are support sensors, which will not store any data, it will just act as a passer of information/data. If any of the Mini Clusters have any data, they will share directly to the Mega Cluster with one Cluster Head Sensor C and from that sensor C, MS can fetch the data and go back to the station. Here MS works both as a sink and sender. It will act as a storage when getting data from the single Mega Cluster with one Cluster Head Sensor C and a sender when giving data to the Station.

Now, suppose a single Mini Cluster with Cluster Head A wants to send the data out of all the 4 Mini Clusters, it will send the data to Mega Cluster with Cluster Head Sensor C, then MS can go to Mega Cluster directly to fetch the data, instead of going

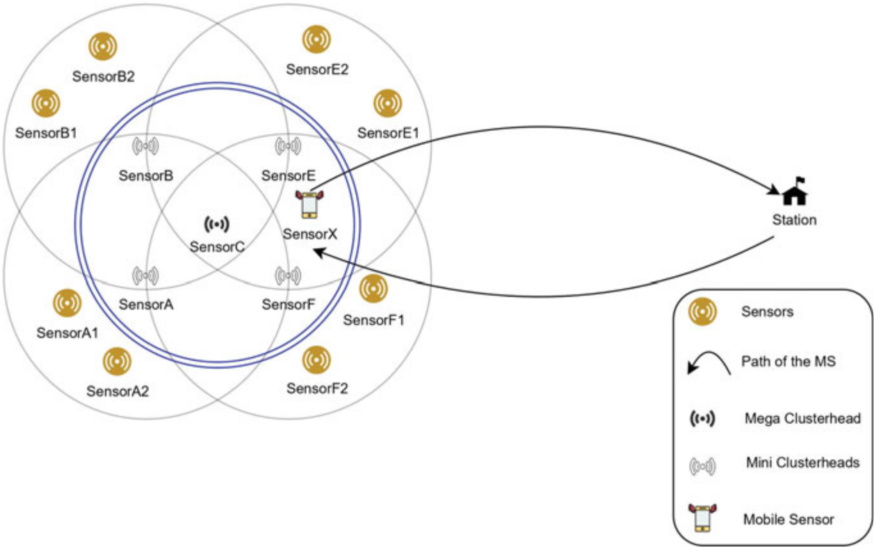


Fig. 4 Representation of Model A on how the Mobile Sensor (MS) will collect data from the Mega Cluster

to all the Mini Clusters and return back to the station as of in [14]. So, we have made Cluster of Clusters in order to remove the excess path even if a single Mini Cluster wants to send the data. We have eliminated the problem of the excess path covered, but still there will be a delay of data at the station as MS has to travel all the way back to the station to dump the data and as MS still has a small memory, indirectly, it will take a number of laps to collect all the data from all the sensors. To deal with these, Model B can be considered.

3.2 Model B

We have used the same concept from the above Model A, but instead of making MS travel all the way to the Mega Cluster and back to the station, we deployed some more sensors in the way between the station and the Mega Cluster, and a gap was left in between, whenever the Mega Cluster wants to send the data, MS will come and fill that gap and will pass whatever data any of the Clusters have to pass. In Fig. 5, single Mega Cluster with one Cluster Head Sensor C which consists of all the 4 Mini Cluster (i.e. Cluster with Cluster Head A, Cluster with Cluster Head B, Cluster with Cluster Head E, and Cluster with Cluster Head F), Sensor P, and Sensor Q are support sensors which will transfer data to the Station. Sensors sensing the environment are (Sensor A1, A2, B1, B2, E1, E2, F1, and F2), station is the sink.

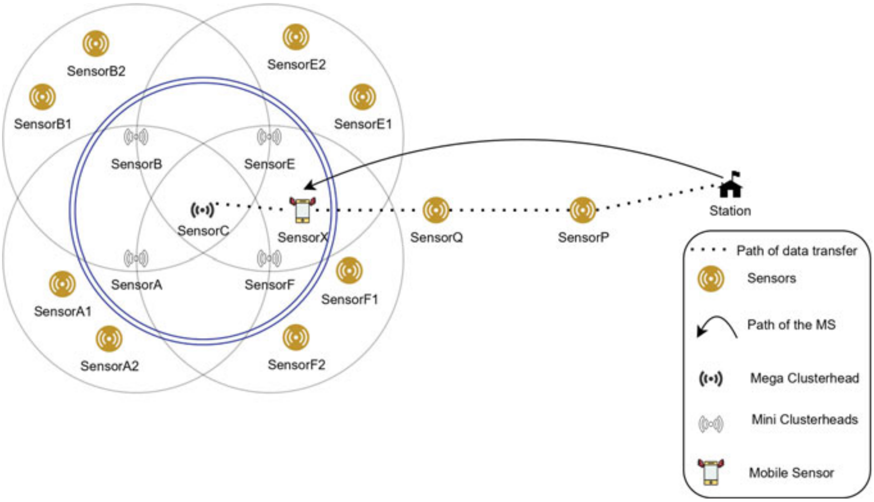


Fig. 5 Representation of Model B on how MS will collect data from the Mega Cluster

Now, here we have kept a gap between sensor Q and Mega Cluster with Cluster Head Sensor C, we made MS (sensor X) go and fill that gap so that Mega Cluster with Cluster Head Sensor C can transfer all the data directly to the Station. Here MS works as a support sensor, it will just pass the information, instead of storing.

Here we are using MS as a connector instead of a transporter, this way we will be able to reduce the number of laps. Furthermore, the storage issue will also be fixed as MS will only work as a passer of information instead of storing it. There will be no delay in data as there is minor to no transportation done in the process. A Cluster wants to send the data, MS will fill the gap and the data can be fetched directly by the station. Also, the network is made static, and so is the path of where to get connected for MS is also predesigned and for a particular time period, the Station will be getting data from a single Mega Cluster to which MS has connected to, after finishing off all the data the particular Mega Cluster wants to send, then MS will go to another Mega Cluster to collect data. So, there will be collision between Mini Clusters when sending the data and that can be controlled by applying different networking protocols, but there would be no chance of getting collision between Mega Clusters.

We could have used a sensor instead of using MS to fill the gap, but let's say if we have multiple Mega Clusters with Cluster Head (C_1, C_2, …, C_n), then we can use a single MS to fill gaps between multiple Mega Clusters and station, also the sensors which are deployed in between the station and the Mega Cluster can be used by different Mega Clusters (C_1, C_2, …, C_n) as well if they are in range, this will also reduce the number of sensors used for making the path.

In Fig. 6, MS goes to Mega Cluster with Cluster Head (C1), connects and transfers the data to the station, then goes to Mega Cluster with Cluster Head (C2), connects and transfers the data to the station, then goes to Mega Cluster with Cluster Head

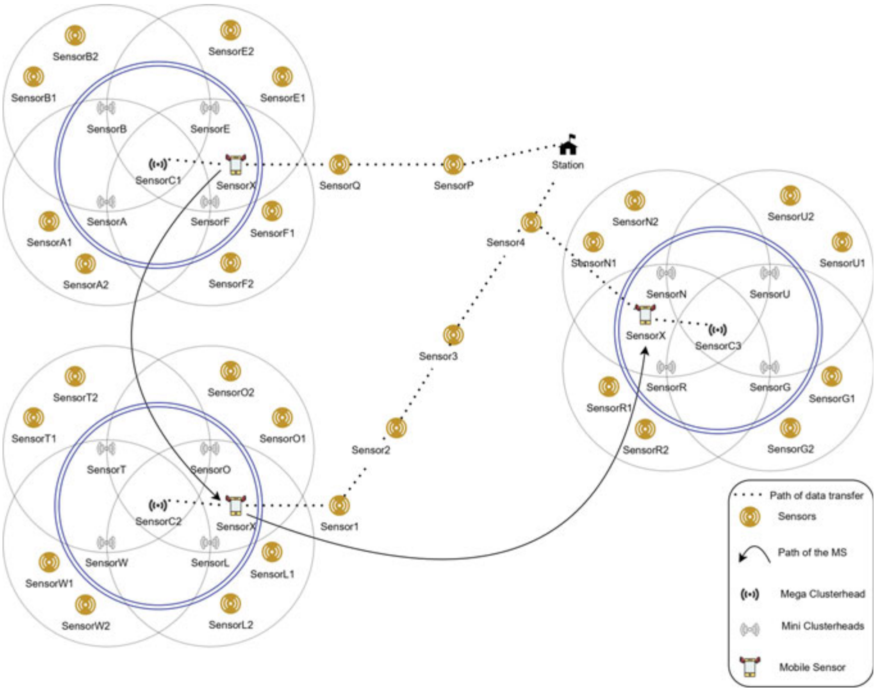


Fig. 6 Representation of Model B on how MS will collect data from the multiple mega cluster

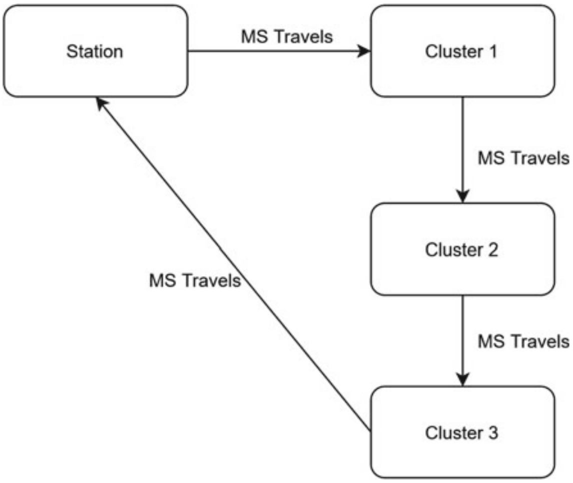
(C3), connects and transfers the data to the station. Here, in the Mega Cluster with Cluster Head (C3), we used the same deployed sensors of the Mega Cluster with Cluster Head (C2) and transfer the data, no additional sensors are used. This way the number of sensors can be reduced for cost conservation.

4 Flow Diagrams

In [14], the MS starts from the station and as it has its predesigned path set, it will travel from one Cluster to the other. Also, the MS has a capability of varying its speed, when MS is inside the range of Cluster, it will reduce its speed for maximum collection of data and while travelling from one Cluster to the other it will maximize/increase its speed. This process takes place every 24 h. Also, the MS takes multiple laps to collect all the data as MS has small memory. The Flow Diagram of [14] is given Fig. 7.

In Model A, we have a Mega Cluster which consist of multiple Clusters named Mini Clusters, these Mini Clusters consists of sensors which sense the environment and send data to their respective Mini Cluster. These Mini Clusters will send data to the Mega Cluster. Now, in order to collect data from the Mega Cluster, MS will

Fig. 7 Flow diagram of how [14] will use its MS to collect data



travel from station to the Mega Cluster, will collect the data from the Mega Cluster and will return back. The main purpose of this Model was to reduce the path which MS has to travel in [14] if a single Cluster has any data. Here all the Mini Cluster will transfer data to the Mega Cluster itself. So, no excess path will be covered, but still the number of laps will be taken by MS to collect all the data as it has a small memory. Flow Diagram of Model A is given Fig. 8.

In Model B, we have a Mega Cluster which consists of multiple Clusters named Mini Clusters, these Mini Clusters consist of sensors which sense the environment and send data to their respective Mini Cluster. This Mini Clusters will send data to the Mega Cluster. Now, the main difference between Model A and Model B, is that we have made a path between Mega Cluster and station consisting of regular support sensors and have kept a gap between them in order to connect the MS at the gap to transfer data directly to the station. The MS will travel from the station

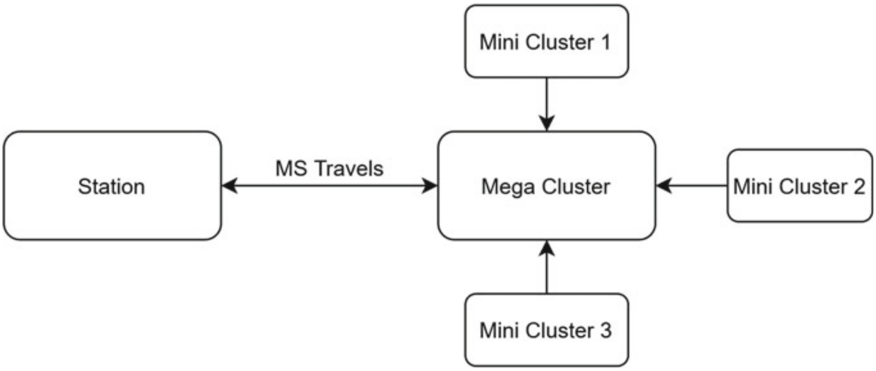


Fig. 8 Flow diagram of how Model A will use MS to collect data

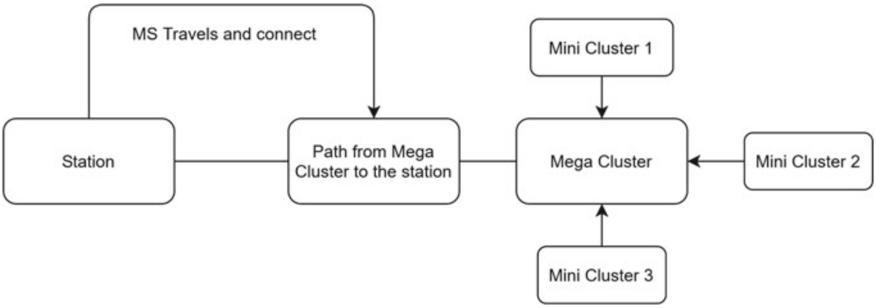


Fig. 9 Flow diagram of how Model B will use MS to collect data

and will connect at the gap, acting as a support sensor instead of storage. By this, the transportation is eliminated, and all the data will be collected in one go with elimination of multiple number of laps taken by the MS. Flow Diagram of Model B is given below Fig. 9.

5 Result Evaluation

(NOTE—All the evaluations were done on the OMNET++ Simulator. All the sensors including MS, use a radio model 802.15.4.). Below given data is obtained by implementing both the models in the OMNET++ Simulator. All the data and figures given below were fully derived from the OMNET++ simulator and no theoretical assumptions were used.

We will compare the above Model A and Model B to find out which one results in the maximum efficiency.

Table 1 states the comparison between the numbers of Packets collected per second by the MS in both the models.

Table 2 states the comparison between the numbers of Laps taken by both the models in order to gather particular data.

Table 3 states the comparison between the total numbers of time it takes to get the data to the station for both the models.

5.1 Packets Collected Per Second

As per the speed of MS, there is a variation in the number of packets collected. The graph Fig. 10, shows the comparison of packets collected in Model A and Model B whilst changing the speed of MS. In Model A, linear mobility was applied to MS for back and forth movement from the station whereas in Model B, circular mobility was applied as it has to stay connected at one place until all the data gets connected.

Table 1 Difference of data collection with respect to the MDC speed in both Model A and Model B

MS speed (ms)	Packets collected per second in (Model A)	Packets collected per second in (Model B)
0.5	131	236
1	131	223
1.5	131	221
2	131	229
2.5	131	229
3	131	230
3.5	131	230
4	131	221
5	131	221
6	131	221
7	131	221
8	131	221
9	131	221
10	132	221
50	136	230
60	114	229
100	74	233

Table 2 Number of mobile sensor (MS) rounds taken to collect the particular number of packets

MS rounds	Packets collected in (Model A)	Packets collected in (Model B)
1	1167	27,757
2	4132	30,722
3	7097	33,687
4	11,229	36,652
5	15,361	39,617
6	19,493	42,582
7	23,625	45,547
8	27,757	48,512

Table 3 Time taken to collect a particular number of packets in Model A and Model B

Time (s)	Packets collected in (Model A)	Packets collected in (Model B)
5	0	1167
18	0	4132
100	1167	7097
150	0	11,229
200	0	15,361
285	4132	19,493

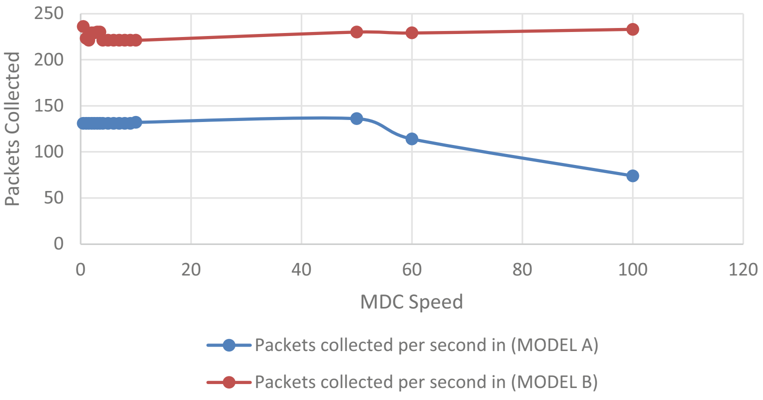


Fig. 10 Packets collected in Model A versus Model B

We can see from the Table 1, that packets collected per second is higher in Model B than in Model A.

5.2 *Number of Laps Taken by the MS*

In Model A, MS has to travel all the way to the Mega Cluster and back to the station in order to provide data to the station. Also, MS has a low capacity for storing the data, so it has to take a number of laps to collect all the data from the particular Mega Cluster. On the other hand, in Model B, MS just has to go and connect in the path in order to transfer the data to the station. MS will not store any data so it will take as many as 1 lap to collect all the data from the particular Mega Cluster. In Fig. 11, we can see that in Model A, MS takes 8 rounds in order to collect 27,757 packets, on

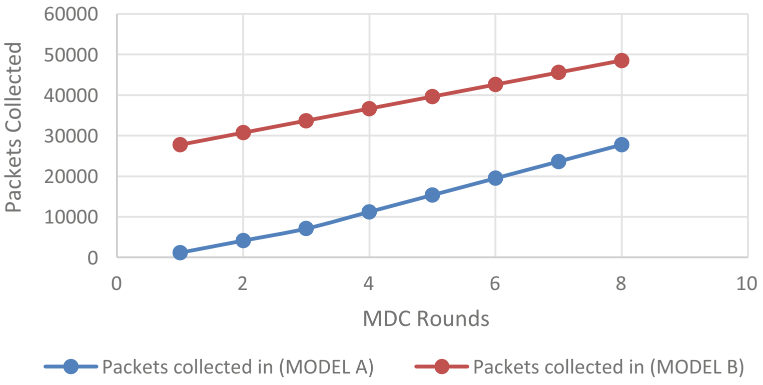


Fig. 11 Number of rounds taken by Model A versus Model B

the other hand, in Model B, MS takes only one round to collect 27,757 packets. So, Model A is taking 8 rounds to collect 27,757 packets because the MS has a small capacity, so it will not be able to collect all the data from the Clusters in one go, it will have to take multiple laps to collect all the data, on the other hand in Model B it is taking only 1 round to collect 27,757 packets as we have used the MS as a means of communicator and not transporter of data so it will just connect and pass the data directly to the station.

5.3 Time Taken to Collect the Packets

In Model A, as MS has to travel from station to the particular Mega Cluster and from that particular Mega Cluster to the station again. Due to the travelling of MS to the sensor, it will make a delay and the data will reach late at the station. On the other hand, in Model B, the data will be directly transferred to the station and there will be no delay of data at station. In Fig. 12 given below, we can see that Model A has not collected any data at 0 s and at 100 s it has collected 1167 packets, on the other hand we can see that Model B collects 1167 packets in 0 s and at 100 s it has collected 7 times more packets than Model A. As Model A will travel from Cluster to the station which will delay the data to be reached at the station and it will deliver 1167 packets at 100th second to the station on the other hand Model B will deliver 1167 packets at the 5th second to the station because we have eliminated the transportation process in order to reduce the delay of data.

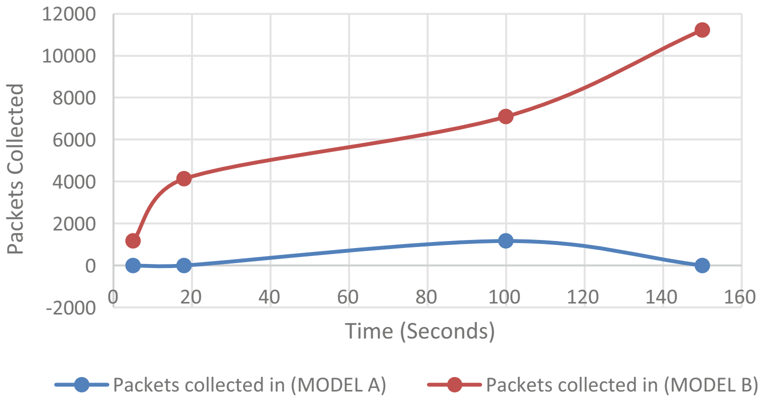


Fig. 12 Time taken to collect particular number of packets in Model A versus Model B

6 Conclusion

We can see from the above results that the number of packets collected in Model A is less than the packets collected in Model B. If we keep MS intact at one place, instead of making it travel from station to the particular Mega Cluster, we can see from the above results that the number of laps for collecting all the data is less in Model B than in Model A. As in Model A, MS stores the data, and at any instance the storage might get full, and it will have to forcefully go back to the station to dump the data and clear its storage and again go back to the particular Mega Cluster to collect the remaining data. In Model B, MS is used only to transfer the data by completing the incomplete path instead of storing the data, which on the other hand will reduce the number of laps to be taken by MS. In Model A as MS will have to travel all the way to the station in order to transfer the data to the station which it has collected from the particular Mega Cluster, due to this transportation there will be a delay of data to reach the station. On the other hand, in Model B, when we kept our MS intact, then the packets will directly reach the station without any delay.

7 Future Scope

The above Model A and Model B can be implemented both in small whereas in the large sensor networks. Modifications can be done on the final proposed solution by using power conservation techniques and to reduce the delay of data at the station even more so that the actions can be taken soon from the particular station. There should be some modifications on how to reduce the number of sensors used in making the predesigned path between the Mega Clusters and station.

References

1. Dantu, K., Rahimi, M., Shah, H., Babel, S., Dhariwal, A., Sukhatme, G.S.: Robomote: enabling mobility in sensor networks. In: IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005, pp. 404–409. IEEE (2005)
2. Amundson, I., Koutsoukos, X.D.: A survey on localization for mobile wireless sensor networks. In: International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments, pp. 235–254. Springer, Berlin, Heidelberg (2009)
3. Rawat, P., Singh, K.D., Chaouchi, H., Bonnin, J.M.: Wireless sensor networks: a survey on recent developments and potential synergies. J. Supercomput. **68**(1), 1–48 (2014)
4. Anastasi, G., Conti, M., Di Francesco, M.: Reliable and energy-efficient data collection in sparse sensor networks with mobile elements. Perform. Eval. **66**(12), 791–810 (2009)
5. Abdul-Salaam, G., Abdullah, A.H., Anisi, M.H., Gani, A., Alelaiwi, A.: A comparative analysis of energy conservation approaches in hybrid wireless sensor networks data collection protocols. Telecommun. Syst. **61**(1), 159–179 (2016)
6. Song, G., Zhou, Y., Ding, F., Song, A.: A mobile sensor network system for monitoring of unfriendly environments. Sensors **8**(11), 7259–7274 (2008)

7. Culler, D., Estrin, D., Srivastava, M.: Guest editors introduction: overview of sensor networks. Computer **37**(8), 4149 (2004)
8. Gandham, S.R., Dawande, M., Prakash, R. and Venkatesan, S.: Energy efficient schemes for wireless sensor networks with multiple mobile base stations. In: GLOBECOM'03. IEEE Global Telecommunications Conference (IEEE Cat. No. 03CH37489), vol. 1, pp. 377–381. IEEE (2003)
9. Al-Karaki, J.N., Kamal, A.E.: Routing techniques in wireless sensor networks: a survey. IEEE Wirel. Commun. **11**(6), 6–28 (2004)
10. Dhand, G., Tyagi, S.S.: Data aggregation techniques in WSN: survey. Procedia Computer Science **92**, 378–384 (2016)
11. Li, S., Shen, H., Huang, Q., Guo, L.: Optimizing the sensor movement for barrier coverage in a sink-based deployed mobile sensor network, pp. 2169–3536. IEEE (2019)
12. Zhu, C., Shu, L., Hara, T., Wang, L., Nishio, S.: Research Issues on Mobile Sensor Networks. Research Gate (2010)
13. Di Francesco, M., Das, S.K., Anastasi, G.: Data collection in wireless sensor networks with mobile elements: a survey. ACM Transactions on Sensor Networks (TOSN) **8**(1), 1–31 (2011)
14. Sayyed, A., deAraujo, G.M., Becker, L.B.: Smart data collection in large scale sparse WSNs. In: 9th IFIP Wireless and Mobile Networking Conference, July, pp. 1–8 (2016)
15. Sayyed, A., Becker, L.B.: Optimizing speed of mobile data collector in wireless sensor network. In: 2015 International Conference on Emerging Technologies (ICET), pp. 1–6. IEEE (2015)
16. Kumar, N., & Dash, D.: Maximum data gathering through speed control of path-constrained mobile sink in WSN. In: 2017 7th International Symposium on Embedded Computing and System Design (ISED), pp. 1–4. IEEE (2017)
17. Konstantopoulos, C., Vathis, N., Pantziou, G., Gavalas, D.: Efficient delay-constrained data collection in wireless sensor networks using mobile sinks. In: 2015 8th IFIP Wireless and Mobile Networking Conference (WMNC), pp. 1–8. IEEE (2015)

Job Scheduling in Cloud Computing Based on DGPSO

J. Arul Sindiya and R. Pushpalakshmi

Abstract In a Cloud Computing environment, dynamic and uncertain nature makes task scheduling problems more complex. It states the need for efficient task scheduling designed and implementation as a primary requirement for achieving QoS. A proper resource utilization enables maximum profit for the Cloud providers. The best scheduling algorithm does not consider the task set collected from the users, but it considers the resources provided by providers for operating the tasks. In this paper, we propose a Dynamic Group of Pair Scheduling and Optimization (DGPSO) algorithm. The proposed DGPSO is the performance-enhancing of AWSQP by using VM pair implementation and partition-based priority system into three levels. These three levels in the priority system such as low, medium, and high. According to the task size, the VM pairing is done. For this, the VM's parameters include communication time, system capacity, memory size, and processing speed. On the dataset, the task sizes are examined and separated according to the priority levels. On which the high priority comprises video files, the audio files under medium-level priority, and the remaining text documents, ppts, etc. included in under low priority levels. Based on the proposed task scheduling mechanism, an experiment is conducted on the aspects of computation cost, communication cost, execution time, CPU utilization, and bandwidth. The obtained results prove its achieved performance is far better than the existing approaches.

Keywords Cloud computing · Priority-based scheduling · Communication · Optimization · Task scheduling

J. Arul Sindiya (✉)
Department of CSE, CARE Group of Institutions, Trichy, India

R. Pushpalakshmi
Department of IT, PSNA College of Engineering and Technology, Dindigul, India
e-mail: push@psnacet.edu.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_3

1 Introduction

In the technological era, global computing becomes a boon for several growing companies and organizations. They deploy most of their application in the cloud and are used economically with minimum costs than earlier. Cloud on-demand services and features like pay-as-you-go payment methods are the major reason for this enormous growth. Following the advantages, there are several challenges still existing in the cloud environment. Among these, the most important challenges investigated by academia and research scholars are task scheduling and load balancing.

Scheduling is the earlier stage of balancing which schedules the arrived task in a queue for processing based on specific requirements and algorithms. The main of the scheduling algorithms is load distribution with optimal resource utilization at minimum time consumption. In the overall scheduling mechanism, Job scheduling is an important field to improve as it has a greater impact on enhancing systems flexibility and reliability. Successful scheduling should possess identifying appropriate function and resource adaptation according to adaptable time.

Traditional task schedulers do not count the task's requirements and their parameters such as the total number of tasks, their priority, waiting time, and response time. In the paper, the author develops a hybrid task scheduling algorithm using the combination of the shortest job first (SJF) and round-robin (RR) algorithms [1]. The hybrid task scheduling algorithm consists of two steps initially balancing the short and long task waiting time. The waiting time and starvation are minimized through two sub-queues such as Q1 and Q2. In both SJF and RR the calculation is done through dynamic and static quanta. Based on this evaluation the task scheduling is optimized and loads also balanced. Secondly, the processing of ready queues such as Q1 and Q2, in which Q1 states the short tasks and Q2 represents long tasks.

An efficient task scheduler analysis the environmental changes and work according to the scheduling strategy. An ant colony optimization (ACO) algorithm [2] is one of the effective scheduling algorithms for allocating the tasks in VMs. In which diversification and reinforcement are done by the slave ants. The ACO algorithm does not use a long path, which leads to incorrect ways, and it also solves ubiquitous optimization issues with the slave ants.

In cloud task scheduling, the industry-known critical issue is NP-hard problem and it can be rectified by a metaheuristic algorithm. For achieving load balancing, most of the cloud task scheduling algorithm applies the ACO algorithm.

In this research, we initially contribute our work towards reducing the makespan time of given task sets. Makespan time optimization is achieved by implementing ACO or a modified ACO algorithm. The two major tasks in cloud scheduling are scheduling and resource allocation. A combined approach of modified analytic hierarchy process (MAHP), bandwidth-aware divisible scheduling (BATS) + BAR optimization [5], longest expected process time (LEPT) preemption, and divide-and-conquer methods are used for scheduling the tasks and allocating resources. The incoming task is ranked using MAHP. With the rank list BATS + BAR, the methodology is applied for allocating the available resources. The VMs loads are continuously monitored by the LEPT method. In case the VM has a maximum load then loads are balanced through divide-and-conquer methodology. By using loads are shared to achieve optimum cloud computing performance. A detailed mechanism of the hybrid bacterial swarm optimization algorithm is analyzed for load balancing issues.

In the existing works, a combined algorithm with particle swarm optimization (PSO) [6] and bacteria foraging optimization (BFO) is used for local search. The complex factors in using this approach are bottlenecks on minimizing the user requests response time. For enhancing VM's response time and overall performance of the end user's throttled modified algorithm (TMA) is applied. TMA consists of a TMA load balancer for balancing the load using two index variables such as busy and available indices. Live migration is evolved using a new and extensible VM migration scheduler for reducing the scheduling completion time. The migration schedule applies the most appropriate migration moments and bandwidth for allocating network models. In which the scheduler decides the fast migrations with minimum completion time.

2 Related Works

In this section, various literature works dealing with load balancing, scheduling, and optimization algorithms are discussed. It contains both independent and dependent tasks in the IaaS cloud environment. We also analyze each algorithm's cloud resource model based on QoS and how these algorithms delivering the features.

Karthick et al. [1] proposed a scheduling mechanism that schedules jobs dynamically with clustering based on burst time for minimizing the starvation. It is compared to conventional methods such as SJF and FCFS. The proposed method results better by using the unused free space. It minimizes the energy consumption and maximizes the jobs taken for executions.

Shojafar et al. [2] employs FUGES a hybrid approach for load balancing. Here optimal load balancing is done with the execution time and cost. It works based on the fuzzy theory to alter the standard genetic algorithm. FUGE on scheduling works with analysis of the VM parameters like VM memory, VM bandwidth, VM processing speed, and length of the jobs. Despite several advantages still, its lack of load imbalance and to improve Inverted Ant Colony Optimization (ACO) is implemented. Asghariet et al. (2017) developed this scheme with high load balancing at minimum costs including the execution time. But it does not consider the QoS factors.

Malik et al. [3] proposed a Round Robin-based load balancing algorithm. In which task allocation is done according to the number of users, software used to cost, user type, runtime, job type, and required resources.

Sindiya and Pushpalakshmi [4] developed The Adaptive Work Size Based Queuing Process (AWSQP) allows for quick data access to virtual machines (VMs). It makes the VM available for completing the task within the deadline while keeping the cost to a minimum. Initially, for quick processing, task priority and the smallest task size are selected for the queue. Our suggested mechanism based on the size of the data, AWSQP selects the most cost-effective path. The request/response time, as well as the mean and variance of network service time, are used to calculate the data access completion time. Then AWSQP takes the best path for the task with the highest priority and repeats the process for the entire queue.

BhatuGawali et al. [5] developed an advanced Task scheduling and resource allocation scheme by combining the Modified Analytic Hierarchy Process (MAHP), Bandwidth Aware divisible Scheduling (BATS) + BAR optimization, Longest Expected Processing Time preemption (LEPT), and divide-and-conquer methods. The objective of this combined mechanism is the organization of the incoming tasks. The Analytic Hierarchy Process (AHP) considers the task its length and run time, according to which it ranks the task. This methodology is well defined for managing complex problems but suffers from restricting user's freedom of service.

Razaque et al. [6] developed a nonlinear programming divisible task scheduling algorithm for allocating the workflow tasks based on the network bandwidth availability. The disadvantage is it does not consider the VMs' energy consumptions during task allocation which may lead to an increase in time or even terminates the task without completing it. The bio-inspired algorithms such as Cuckoo Search (CS), Bees Life Algorithm (BLA), Ant Colony Optimization (ACO), genetic algorithm (GA) and Particle Swarm Optimization (PSO), etc. have an important role in scheduling the tasks to cloud nodes.

Bala and Chana [7] developed a Multilevel Priority-Based Task Scheduling Algorithm for Workflows in a Cloud Computing Environment that prioritizes workflow tasks based on instruction length. The proposed scheduling approach prioritizes cloud application tasks based on the limits established by six sigma control charts based on dynamic threshold values.

Sindiya and Pushpalakshmi [8] developed a NAERR is a novel and adaptive enhanced round-robin algorithm that computes the size and length of all requesting jobs, the capabilities of all available VMs, and the task interconnection. The jobs arrived at random time intervals with varying load conditions during the server's run time. Static or dynamic scheduling techniques are used to allocate resources. To provide efficient cloud computing, static or dynamic scheduling techniques are used to assign tasks to appropriate resources and organize the involving heterogeneous resources. As a result, user satisfaction improves.

Lakra and Yadav [8] proposed the multi-objective task scheduling algorithm. This mechanism effectively maps the tasks to the VMs through non-dominated sorting by achieving QoS. The proposed mechanism works on the principle of data center throughput and minimizes the cost without affecting the Service Level Agreement (SLA). The observation states that the proposed mechanism fails to address most of the Quality of Service factors such as awareness of VMS energy.

‘Ramezani et al. [9] presented the Multi-Objective Jswarm (MO-Jswarm) scheduling algorithm. It works is proposed for enabling optimal task distribution on each VMs even balancing with various task transferring time, task execution cost, and task execution time. The author describes the proposed Multi-Objective Jswarm (MO-Jswarm) scheduling algorithm that can improve QoS and traffic that occurred in the cloud environment. The cloud scheduler is responsible for managing the VMs and performing job executions. The best examples of cloud schedulers in the industry are the fair scheduler in Facebook, the capacity scheduler in Yahoo, and the FIFO scheduler in Hadoop MapReduce. These are schedulers that are effective in allocating the resource but fail to satisfies QoS (quality of service) constraints. In the hybrid cloud environment, QoS is more vital in real-time applications and services. This section describes several methodologies demonstrating cloud task scheduling and their issues.

3 The Proposed DGPSO

The proposed DGPSO main contribution is centered on a novel technique for task scheduling taking into considering the priority issue and aiming for better performance by optimizing execution time, makespan, resource utilization, and load balancing. So that we have revamped the performance of AWSQP [4] by adding additional parameters such as VM-pair implementation and three types of partition-based priority systems. The NAERR [8] algorithm is deployed for accomplishing the load balance of the VMs in the system.

In the AWSQP, two levels of task priority are considered such as low and high priority for forming the scheduling queue. Low priority is assigned to the larger tasks and high priority is assigned to the smaller tasks. But in the proposed DGPSO, we have added the factor of partition in the priority system into three levels such as low, medium, and high. Based on that video files come under high priority, the audio files come under medium-level priority and other text documents and ppts, etc. come under low priority. Video and audio files compose a higher file size. So that higher priority is assigned to the video files and medium priority is assigned to the audio files. By scheduling under high priority and medium priority, these will compute the task with available VMs first. In this case, suppose if it takes more processing time, automatically it can be allocated to the nearest available VM. The maximum possibility of achieving successful execution on higher-size files and medium-size files is more in this new mechanism comparing to AWSQP. Whereas the low-priority files are less in size that are computed faster with minimum time.

Overall, every task scheduling will be completed with available VMs within the deadline. This idea of improving AWSQP also achieves optimum resource utilization that results to achieve the desired results within minimum time as well as computation cost. The concept of espousing the best path for the task with high priority as defined in DGPSO is an added advantage. The pair VM implementation is applied according to the task size that executes faster than the single VM (Fig. 1).

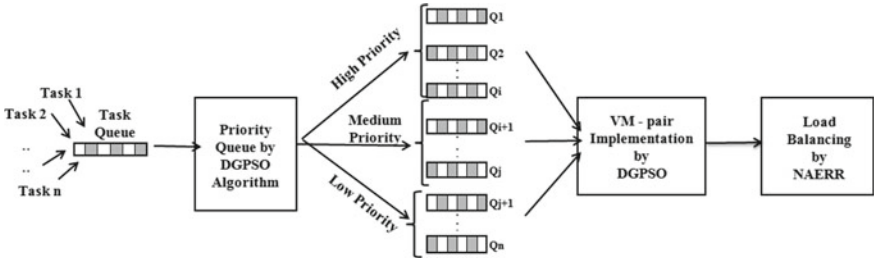


Fig. 1 Proposed DGPSO working model

- The task size is calculated by computing the task length T_{L} with processing elements PE. The task size can be expressed as below;

$$T_{\text{size}} = \sum_{i=0}^{n} \frac{T_{\mathrm{L}}}{\mathrm{PE}(\mathrm{Vm}_S)} \tag{1.1}$$

- The task length is measured in Million Instructions (MI) unit. For all the arriving tasks, the average mean (μ) and standard deviation (σ) are calculated by getting the MIPS value of each task.
- The various control limit values are set by using six sigma (6σ) X-bar control charts, i.e. upper-level control limit ($\mathrm{CL_u}$) $= \mu + 3\sigma$, A middle-level control limit ($\mathrm{CL_m}$) $= \mu$, lower-level control limit ($\mathrm{CL_l}$) $= \mu - 3\sigma$.

where,

Standard deviation (σ) is,

$$\sigma = \sqrt{\sum_{i} \frac{(T - \mu)2}{n = 1}}$$

The average mean (μ) is,

$$\mu = \sum \text{get MIPS}\,(T_i)/n$$

Here,

T_i Ready Task and $i = 1$ to n

- The arrived tasks are indexed as $T_1, T_2, T_3, \ldots T_n$ respectively. At this stage, the task queue is arranged in a FIFO manner. Then, the MIPS value is calculated for the first task in the queue. If the MIPS ($\mathrm{MI_{value}}$) value of the first task is ($\mathrm{MI_{value}}$) $\leq \mathrm{CL_l}$, then the task is classified as a smaller task (text documents and ppt, etc.) and the task is allotted to low priority queue, else if ($\mathrm{MI_{value}}$) $\geq \mathrm{CL_u}$, then the task is classified as the larger task (Video files) and the task is allotted to the higher-level priority queue. Eventually, if the MIPS value of the first task is $(\mathrm{MI}_{\mathrm{value}}) > \ =$ CL_{l} and $(\mathrm{MI}_{\mathrm{value}}) < \ = \mathrm{CL}_{\mathrm{u},}$ then the task is classified as the medium task (audio files) and the task is allotted to the medium-level priority queue. Likewise, the remaining tasks in the queue are classified to determine the priority system.

- Then, the numbers of virtual machines taking part in operation are created and indexed as Vm_1, Vm_2, Vm_3, … Vm_{n}. The virtual machines are identified based on the capacity for pairing with tasks. Here the above mechanism of priority level is applied that decides the VM based on its capacity. In this way, VM pairs with tasks are performed. The task sizes are categorized as small, medium, and high. The T_{s} denotes smaller tasks, Tm denotes medium task and T_{l} denotes larger tasks.
- Then, the three priority level queues such as low, medium, and high are formed for executing the tasks to the paired VMs.
  - Low-level priority queues (L_{q}): each queue contains the smaller tasks (text,ppt, etc.) T_{s}.
  - Medium-level priority queues (M_{q}): each queue contains the medium tasks (audio) T_{m}.
  - High-level priority queues (H_{q}): each queue contains the tasks (video) T_{l}.
- For generating the VM pairs, each VM's processor speed, memory size, and time slot are taken into consideration. Based on that three-level VM pairs are created such as low-level capacity VM pairs, medium-level capacity VM pairs, and high-level capacity VM pairs. It can be expressed as;

$$V = \sum_{i=0}^{n} (\mathrm{Vm}_1, \mathrm{Vm}_2), (\mathrm{Vm}_3, \mathrm{Vm}_4)(\mathrm{Vm}_5, \mathrm{Vm}_6) \tag{1.4}$$

the pairs are created.

- Next, the VM pair is computed based on its system capacity, processing speed, memory size, communication time. It can be expressed as below;

$$\mathrm{Vm}_s = \sum_{i=0}^{n} \frac{\mathrm{PE}}{C_t(T_{\mathrm{L}})} \tag{1.5}$$

- By combining the above equations VM pairs are created and expressed as below;

$$\begin{aligned} \mathrm{Vm}(V, T)\ T_{\mathrm{size}} &= \sum_{i=0}^{n} \frac{T_{\mathrm{L}}}{\mathrm{PE}(\mathrm{Vm}_{\mathrm{S}})} \\ &= T_l(\mathrm{Vm}_{\mathrm{s1}}, \mathrm{Vm}_{\mathrm{s2}}), T_{\mathrm{m}}(\mathrm{Vm}_{\mathrm{m3}}, \mathrm{Vm}_{\mathrm{m4}}) T_l(\mathrm{Vm}_{\mathrm{l5}}, \mathrm{V}_{\mathrm{l6}}) \end{aligned} \tag{1.6}$$

$$\mathrm{Vm\ pairs} \rightarrow (v) = (v, t); \tag{1.7}$$

Table 1 Algorithm of proposed DGPSO

S. No.	Proposed AWSQP algorithm
1	Input: no. of tasks $T = \sum_{i=0}^{n} T_1, T_2, T_3, \ldots T_n$
2	Create the number of Vm's $V = \sum_{i=0}^{n} \mathrm{Vm}_1, \mathrm{Vm}_2, \mathrm{Vm}_3, \ldots \mathrm{Vm}_n$
3	Arrange the task in queue $Q = q(t)$; //FIFO manner
	Calculate the task size $T_{\mathrm{size}} = \sum_{i=0}^{n} \frac{T_{\mathrm{L}}}{\mathrm{PE}(\mathrm{Vm}_{\mathrm{S}})}$
4	Procedure: Task priority (low, medium, high) //Calculate the MIPS value of independent task
5	**If** $(\mathrm{MI}_{\mathrm{value}}) < \; = \mathrm{CL}_{\mathrm{l}}$,**then** define the task as smaller //Low priority queue Else **if** $(\mathrm{MI}_{\mathrm{value}}) > \; = \mathrm{CL}_{\mathrm{u}}$, **then** define the task as larger //High priority queue Else **if** $(\mathrm{MI}_{\mathrm{value}}) > \; = \mathrm{CL}_{\mathrm{l}}$ and $(\mathrm{MI}_{\mathrm{value}}) < \; = \mathrm{CL}_{\mathrm{u},}$ **then** define the task as medium //Medium priority queue
6	Procedure: Set VM pair//low, medium, high-based on capacity
7	Create set Virtual machine pairs $V = \sum_{i=0}^{n} (\mathrm{Vm}_1, \mathrm{Vm}_{2)}, (\mathrm{Vm}_3, \mathrm{Vm}_4)(\mathrm{Vm}_5, \mathrm{Vm}_6) \ldots$
8	Calculate the Vm size $\mathrm{Vm}_{\mathrm{s}} = \sum_{i=0}^{n} \frac{\mathrm{PE}}{C_t(T_{\mathrm{L}})}$
9	$\mathrm{VM}(V, T)\; T_{\mathrm{size}} = \sum_{i=0}^{n} \frac{T_{\mathrm{L}}}{\mathrm{PE}(\mathrm{Vm}_{\mathrm{S}})} \;= T_{\mathrm{l}}(\mathrm{Vm}_{\mathrm{s}1}, \mathrm{Vm}_{\mathrm{s}2}), T_{\mathrm{m}}(\mathrm{Vm}_{\mathrm{m}3}, \mathrm{Vm}_{\mathrm{m}4}) T_{\mathrm{l}}(\mathrm{Vm}_{\mathrm{l}5}, V_{\mathrm{l}6})$ // Calculate the priority for each
10	Classification of VM pair (Low, Medium, High)
11	Vm pairs → $(v) = (v, t)$
12	(v, t) → call scheduling algorithm NAERR algorithm

- The (v, t) is the final VM pair which is grouped according to its priority levels such as low, medium, and high. To the paired (v, t) the NAERR algorithm is applied for scheduling. The combination of NAERR and DGPSO results in the best scheduling outputs. The results of the proposed technique are analyzed in the simulation environment by the cloudsim simulator (Table 1).

4 Experimental Results

The proposed DGPSO technique is implemented in the CloudSim simulator for analyzing their performances. For the proposed techniques, the input task that has been taken for performance consideration is 100–1000 with 100 VM's respectively.

Table 2 Parameters and quantity taken for execution

S. No.	Parameters	Qty.
1	VM RAM size	256, 312, 712, and 856 bytes
2	VM bandwidth	700, 750, 800, and 900 bits/s
3	VM manager	Xgen
4	Task length (or) instructions	500,000–100,000,000
5	PE processing capacity	174/247/355 MIPS

The results are discussed below based on the no. of task migration, task execution time, and makespan time (Table 2).

Simulation is run more than 100 times for analyzing the no. of task migration by the proposed DGPSO algorithm exploiting the space shared policy in cloudsim. Figure 2 describes the comparison result between AWSQP and NAERR algorithms with the proposed DGPSO algorithm. The comparison factor is the number of task migration against the number of virtual machines. The *X*-axis plotting represents the number of virtual machines and the *Y*-axis plotting represents the number of task migration. The attained comparison results demonstrate that while executing the tasks at the 20, 40, 60, 80 and 100 VMs, the respective NAERR task migrations are 6, 8, 10, 12 and 15 and the respective AWSQP task migrations are 5, 6, 7, 8 and 12. But simultaneously, the respective task migrations of the DGPSO algorithm are only 2, 2, 3, 3 and 4. So the combined results have manifested that among all the three techniques, the DGPSO algorithm is effective in minimizing the task migration and pick the suitable virtual machine for each task.

From Fig. 3, it is observed that the proposed DGPSO algorithm by task length has provided a faster execution time compared to other existing NAERR and AWSQP load balancing algorithms with heterogeneous tasks and heterogeneous resources.

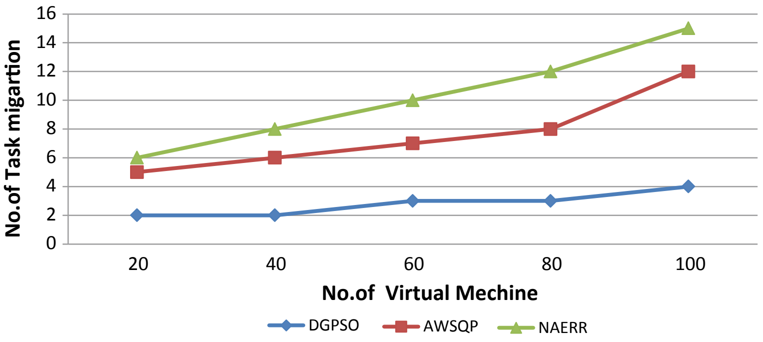


Fig. 2 Comparison of no. of VM's versus no. of task migration

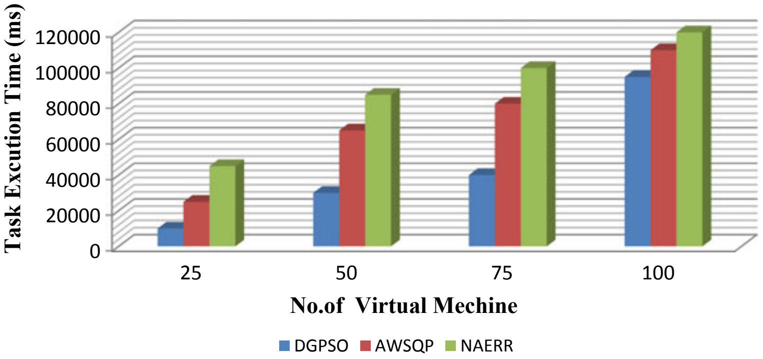


Fig. 3 Comparison of no. of VM's versus task execution time

To allocate tasks, the proposed DGPSO algorithm has computed the job length and processing ability of the heterogeneous virtual machines. Hence, the lengthy jobs are allocated to the VMs having a higher capacity in the heterogeneous environment that helps to execute the tasks in a shorter time. The proposed scheduler has considered the workload of its all configured virtual machines and its uncertain execution time of ongoing workload has been discovered. Then, the estimated execution time of arrived tasks is calculated by the scheduler in every configured VM. It includes this computed time with the execution time of existing loads on every VM. From this calculation, the least possible execution time is chosen to implement a specific task in one of the VMs. Then, the task has been allocated to that VM. The obtained results are shown that while executing the tasks at the 25, 50, 75 and 100 heterogeneous VMs, the respective execution time(ms) of the tasks are 10,000, 30,000, 40,000 and 95,000 for DGPSO, the respective execution time (ms) of the tasks are 25,000, 65,000, 80,000 and 110,000 for AWSQP and the respective execution time(ms) of the tasks are 45,000, 85,000, 100,000 and 120,000 for NAERR Thus, the proposed DGPSO algorithm is most appropriate for the heterogeneous cloud environment in term of execution time.

Figure 4 depicts the comparison result between AWSQP and NAERR algorithms with the proposed DGPSO algorithm. The comparison factor is makespan time(s) against the number of virtual machines. The *X*-axis plotting represents the number of virtual machines and the *Y*-axis plotting represents the makespan time(s). The obtained results have proven that while executing the tasks at the 10, 20, 30, 40 and 50 VMs, the respective makespan times(s) of NAERR are 1150, 1200, 1450, 1600 and 1900 and, the respective makespan times(s) of AWSQP are 850,900,1200,1315&1450. But at the same time,the respective makespan times(s) of DGPSO are only 530, 650, 815, 925 and 1250. So the combined results have demonstrated that among all the three techniques, the DGPSO algorithm is effective than NAERR and AWSQP techniques on makespan time comparison.

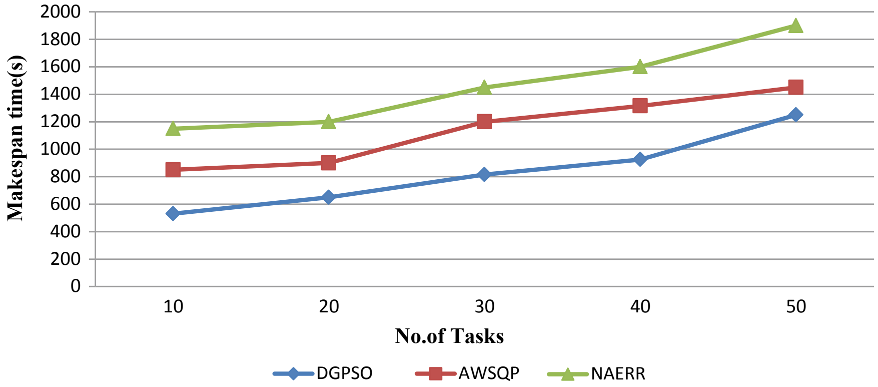


Fig. 4 Comparison of no. of VM's versus Makespan time

5 Conclusion

The concern of task scheduling in cloud computing is an emerging topic among researchers. Several methodologies are existing on the concept of task scheduling and load balancing. However, since a prominent work needs to deliver overall cloud performance efficiently. Cui et al. (2017) developed a TSS for scheduling the task in the cloud environment. TSS consists of three modules known as user module, data center module, and task scheduling module. QOT is accomplished by the user module. The infrastructure-level services such as resources allocating including hosting in series are done by data center module. The task from the QOT into memory is accomplished by the task scheduling module. These tasks are scheduled and performed according to the job scheduling algorithm with consumer's requirements. Here the optimization is achieved by implementing the GA and roulette selection mechanism. Next, the best path is chosen by using the Ant Colony Optimization algorithm, but the major issue in using the Ant Colony Optimization algorithm fails in choosing the best path on higher task size and it leads to an increase of time.

In this proposed technique, AWSQP performance is enhanced by employing VM pair implementation and partition-based priority system into three levels with DGPSO. The arriving tasks are prioritized based on their size into low, medium, and high. According to the task size, the VM pairing is carried out. For this, the VM's parameters include communication time, system capacity, memory size, and processing speed. On the dataset, the task sizes are examined and separated according to the priority levels. The high priority list comprises the video files, the audio files classified under medium-level priority, and the remaining text documents, ppts, etc. classified under low priority levels. The dataset with higher file size includes video data, social media with Facebook, Google, Amazon and the size of gb,tb. The dataset with medium size files consists of files with mb,gb, and the text document and books relevant data under small datasets size of (kb, mb). The proposed AWSQP is applied for scheduling with this priority level and allocates the available VM pair to the high-priority task initially. This mechanism increases the successful execution rate on higher-size files to a greater extent in comparison to other traditional methods. At the same time, low-priority files can be computed quickly with minimum time. This enables the successful completion of the entire dataset within the deadline. The improvement of AWSQP with VM pair implementation accomplishes optimum resource utilization with minimum cost and time. Additionally, the best path selection according to the priority levels added more credits to DGPSO. The NAERR technique is employed for balancing the workload of the VMs. Thus the implementation of DGPSO achieves the user exception on cloud computing, especially on speed and time consumption.

The various security threats such as malicious intruders, Denial of services (DoS), wrapping attacks, Man in the cloud, etc. are not considered in this proposed technique. The Reasons behind these security threats are weak cryptography, unstructured environment, topology leverage, etc. So future work can be carried out on improvising the security mechanism and focusing on these would be the most anticipated one.

References

1. Karthick, A. V., Ramaraj, E., Subramanian, R.G.: An efficient multi-queue job scheduling for cloud computing', world congress on computing and communication technologies, pp. 164–166 (2014)
2. Shojafar, M., Javanmardi, S., Abolfazli, S., Cordeschi, N.: FUGE: a joint meta-heuristic approach to cloud job scheduling algorithm using fuzzy theory and a genetic method. J. Cluster Comput. **18**(2), 829–844 (2015)
3. Malik, A., Chandra, P.: Priority-based round-robin task scheduling algorithm for load balancing in cloud computing. J. Netw. Commun. Emerg. Technol. **7**(12), 17–20 (2017)
4. Arul Sindiya, J., Pushpalakshmi, R.: Job scheduling in cloud computing based on adaptive job size based queuing process. Int. J. Adv. Sci. Technol. **28**(9), 157–168 (2019)
5. Gawalil, M.B., Shinde, S.K.: Task scheduling and resource allocation in cloud computing using a heuristic approach. J. Cloud Comput. **7**(1), 1–16 (2018)
6. Razaque, A., Vennapusa, N.R., Soni, N., Janapati, G.S.: Task scheduling in cloud computing. In: Long Island Systems, Applications and Technology Conference (LISAT), pp. 1–5 (2016)
7. Bala, Chana.: Multilevel priority-based task scheduling algorithm for workflows in cloud computing environment. In: Proceedings of International Conference on ICT for Sustainable Development, pp. 685–693 (2016)
8. Arul Sindiya, J., Pushpalakshmi, R.: Scheduling and load balancing using NAERR in cloud computing environment. Appl. Math. Inf. Sci. **13**(3), 445–451 (2019)
9. Lakra, A.V., Yadav, D.K.: Multi-objective tasks scheduling algorithm for cloud computing throughput optimization. Proc. Comput. Sci. **48**, 107–113 (2015)
10. Ramezani, F., Lu, J., Hussain, F.: Task scheduling optimization in cloud computing applying multi-objective particle swarm optimization. In: International Conference on Service-oriented Computing, pp. 237–251. Springer (2013)
11. Zhou, J., Yao, X.: Multi-objective hybrid artificial bee colony algorithm enhanced with Lévy flight and self-adaption for cloud manufacturing service composition. Appl. Intell. **47**(3), 721–742 (2017)
12. Asghari, S., Navimipour, J.N.: Cloud services composition using an inverted ant colony optimization algorithm. Int. J. Bio-Inspired Comput. **13**(4), 257–268 (2017)

13. Fang, Y., Wang, F., Ge, J.: A task scheduling algorithm based on load balancing in cloud computing. Web Inf. Syst. Mining **6318**, 271–277 (2010)
14. Lin, C.C., Liu, P., Wu, J.J.: Energy-aware virtual machine dynamic provision and scheduling for cloud computing. In: IEEE International Conference on Cloud Computing, pp. 736–737 (2011)
15. Ghanbari, S., Othman, M.: A priority-based job scheduling algorithm in cloud computing. Proc. Eng. **50**, 778–785 (2012)
16. Maguluri, S.T., Srikant, R., Ying, L.: Stochastic models of load balancing and scheduling in cloud computing clusters. In: IEEE Conference on Computer Communications, INFOCOM, pp. 702–710 (2012)
17. Gulati, A., Chopra, R.K.: Dynamic round Robin for load balancing in a cloud computing. Int. J. Comput. Sci. Mob. Comput. **2**(6), 274–278 (2013)
18. Zhu, X., Chen, C., Yang, L.T., Xiang, Y.: ANGEL: agent-based scheduling for real-time tasks in virtualized clouds. IEEE Trans. Comput. **64**(12), 3389–3403 (2015)
19. Radojevic, B., Zagar, M.: Analysis of issues with load balancing algorithms in hosted (cloud) environments. In: MIPRO Proceedings of the 34th International Convention, pp. 416–420 (2011)

Read–Write Decoupled Single-Ended 9T SRAM Cell for Low Power Embedded Applications

Amit Singh Rajput, Arpan Dwivedi, Prashant Dwivedi, Deependra Singh Rajput, and Manisha Pattanaik

Abstract Static Random-Access Memory (SRAM) is the most significant building block of embedded Systems and microprocessor. Traditional 6T cell used as a data storage element in the SRAM cell but is suffered from low stability, low process tolerance and high-power consumption issue. Technology is continuously scaling down into the nanometer regime to achieve higher integration. Minimum size cell is used to achieve higher integration density in nm technology node but it significantly increases the leakage current and decreases stability. These issues are more critical in the conventional 6T cell. This article introduces a new read/write decouple single-ended 9T cell with high stability, low process tolerance, and low static and dynamic power consumption. This 9T cell shows higher read/write stability due to read buffer and dynamic loop cutting techniques respectively. Furthermore, it shows the low leakage current due to the stack transistor technique and low dynamic power due to a single bit line (BL). In contrast to the traditional 6T SRAM cell, the proposed 9T cell has a 4.28 × higher Read Static Noise Margin (RSNM), 1.06× higher Write Static Noise Margin (WSNM), and approximately the same Hold Static Noise Margin (HSNM). The proposed 9T cell reported 0.48× lower power consumption compared to the conventional 6T cell. This 9T cell shows the half select free operation and aids bit interleaving architectures therefore it may be an appealing choice for low power embedded system.

Keywords SRAM · High static noise margin · Low power · Low leakage · Single-ended write · Single-ended read · Read–write decouple

A. S. Rajput (✉) · P. Dwivedi
Department of Microelectronics & VLSI, UTD, CSVTU, Bhilai, India

A. Dwivedi
Department of Electrical, Shri Shankaracharya Engineering College, Bhilai, India

D. S. Rajput
Government Polytechnic College, Morena, India

M. Pattanaik
ABV-Indian Institute of Information Technology & Management, Gwalior, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_4

1 Introduction

SRAM is a critical part of the embedded systems and microprocessor and works as cache memory to store and process data. More than 80% of the die area and 30% power are used by the SRAM in embedded Systems [1]. Generally, a significant part of memory (SRAM cell) remains ideal, therefore excessive leakage current is a critical problem in the SRAM cell [2]. Leakage current increases exponentially and dominates the overall power consumption below 100 nm (nm) technology node [3]. One popular method to reduce leakage current is supply voltage reduction. Supply voltage reduction reduces leakage power linearly and dynamic power quadratically [2]. Apart from higher power consumption, low read stability is also a serious problem of the conventional 6T SRAM cell. Moreover, cell stability reduces as the supply voltage decreases. Apart from that effect of process parameter variation is dominant at scaled technology node. Consequently, stability of a 6T cell is degraded to unacceptable levels [3], therefore, high stable, low power, Process tolerant SRAM cell design is an area of interest within the field of memory design.

The read current flows through the data storage node in the traditional 6T cell, so the bit line noise affects the voltage of the storage node. Moreover, read–write operations are achieved by the same access transistor in the conventional 6T SRAM cell; therefore, RSNM and WSNM cannot be optimized simultaneously. If we try to enhance read stability, its write stability is reduced. To resolve this problem, various read–write decoupled SRAM cells are proposed in the literature[4]. Z. Liu et al. presents a read–write decouple 9T cell [4]. The write operation of this cell is similar to that of a traditional 6T cell, but the read operation is done by a separate read port, resulting in improved data stability. Since write access transistors are disabled in read mode, data storage nodes are completely isolated from BL noise.[4]. However, it consumes more leakage power than a typical 6T SRAM cell due to large number of transistors. Later, Islam et al. proposed a 10T cell with low leakage [5]. The read–write operation of the Islam 10T cell is similar to the 9T cell, but due to the series-connected stack transistors, it consumes less leakage current. In Refs. [5, 6] Rajput et al. present a read–write decouple Schmitt trigger-based 10T cell. This cell shows higher read stability due to simultaneous implementation of Schmitt trigger and gate buffer technique and low leakage due to stack transistor technique; however, it suffers from half select issue because Word Line (WL) signal is connected to the entire row and turn on during the write operation. The cells where data is not to be written fell a false write operation and their store data may be disturbed, it is called half select problem. Apart from that, it consumes higher dynamic power due to differential sensing. The Bit Line Bar (BLB) and BL are precharged to VDD prior to read operations in the Rajput 10T SRAM cell, and the one-bit line is discharged to GND based on the store data. Memory has to recharge both BL before the each read operation, therefore it consumes higher dynamic power. If the number of the bit line is reduced to one, dynamic power may be reduced to half in the cell [7].

Various single-ended cells are proposed in the literature based on this idea [7, 8]. The single-ended technique is useful in power constraint applications [7]. It reduces the dynamic power and the expensive die area. S. Ahmad et al. suggested a single-ended 11T cell with enhance read stability due to the Schmitt trigger and read buffer technique but it consumes a large chip area [7]. A single-ended cell slightly increases the read/write access time, and to perform a write “1” operation is harder than a differential cell, but these issues can solve through the asymmetric SRAM cell design and write-assist techniques.

Single-ended 11T SRAM cell suffers from the multiple bit soft error [7]. To enhance the yield and reliability of the SRAM, the bit interleaving architecture is widely used. In bit interleaving architecture, multiple bit error treats as a single bit error, therefore single bit Error Correction Codes (ECC) are used to remove multiple bit error. In, Pal et al. [9] proposed a low power 9T SRAM cell. This cell is similar to the Islam at el low leakage 10T SRAM cell [5], but it doesn’t have the half select issue and can support bit interleaving architecture. It provides a significant leakage power reduction in the hold mode but consumes more dynamic power due to differential sensing. Therefore, there is a need for an SRAM cell that provides lower static and dynamic power consumption, higher read and write stability, higher process tolerance low transistor count, half select free operation, and support bit interleaving architecture at nm technology.

Figure 1 proposed gate sensing-based read, loop cutting-based write 9T SRAM cell that can address excessive power consumption, lower stability, and the higher process parameter variation issue faced by the conventional 6T cell. The noticeable features of this 9T cell are given below.

1. The proposed 9T cell shows improved read stability due to the read buffer technique and sufficient write ability because of the dynamic loop- cutting technique.

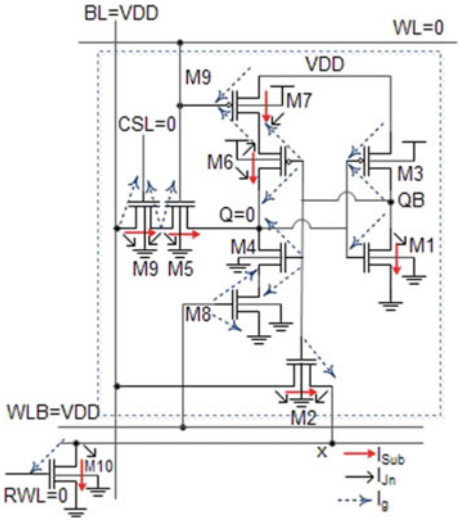


Fig. 1 Proposed single-ended 9T cell with various leakage component

2. The proposed 9T cell consumes a low leakage current and shows a substantial dynamic power consumption reduction due to a single bit line.
3. It can implement the bit interleaving architecture and ECC scheme, therefore it reduces the soft-errors in the SRAM.
4. It shows lower leakage power consumption due to series-connected stack transistors (two NMOS and two PMOS are connected in the series in one inverter) than the conventional 6T cell.

In this paper, we used the HSPICE simulation tool with 32 nm CMOS Predictive Technology Model (PTM) for minimum size transistor at operating voltage 0.9 V [10], NMOS transistors have a threshold voltage of 0.16 V, while PMOS transistors have a threshold voltage of—0.16 V. The rest of the paper was arranged as follows: Section 2 describes the proposed 9T SRAM cell and its operations. Section 3 includes a simulation setup for evaluating the proposed 9T cell's stability and leakage power consumption. The findings and discussion are summarized in Sect. 4, and the article's conclusion is presented in Sect. 5.

2 Proposed 9T Cell

In SRAM architecture, the conventional 6T cell is the most widely used cell structure. A cross-coupled inverter pair is used to store single bit data in a standard 6T cell SRAM cell, and two access transistors are used to communicate storage nodes to the bit lines on the activation of the WL signal [11]. When designing an SRAM cell, two separate design requirements must be considered. Firstly, during the read operation stored information should not be flipped but during a write operation, stored data should flip easily [12]. In the conventional 6T cell, if we try to enhance read stability its write stability reduces and vice versa. It's difficult to optimize read and write stability in the 6T SRAM cell at the same time because of this conflicting design requirement.

The proposed 9T cell is represented schematically in Fig. 1. Inverter1 (Inv-1) is formed by transistor M3 and M1, while inverter 2 (Inv-2) is made by transistor M6, M4, M7, and M8. Transistors M7/M8 are used to disconnect VDD/GND from the Inv-2 on activation of the WL and its complementary Word Line Bar (WLB) respectively. The access transistor M5 is transferred data from BL to node Q on the activation of row-based WL and column-based Column Selects Line (CSL). Table 1 displays different control signals and their values for the 9T SRAM cell in read, write and hold modes. In the write operation, WL and CSL signals are set to VDD while WWL is set to GND, it disconnects Inv-2 from power supply VDD and GND, as a result, Inv-2 become weak and the strong M5 transistor easily write data into the storage node Q. During the write mode, transistor M9 is used to eliminate the proposed cell's half select issue.

Before the start of the read operation, BL is precharged to VDD. During the read operation, RWL and WLB signals are connected to the VDD while WWL and CSL

Table 1 Proposed 9T cell signals during read, write and hold operations

Operation	BL	WL	WLB	CSL	RWL
Hold	VDD	GND	VDD	GND	GND
Read	Precharge to VDD	GND	VDD	GND	VDD
Write “0”	GND	VDD	GND	VDD	GND
Write “1”	VDD	VDD	GND	VDD	GND

signals are attached to GND. Transistor M2 becomes on or off according to voltage QB. When the QB node is high, transistor M2 turns on and precharged BL discharge to the GND through the transistors M2 and M10. However, BL remains precharge, if the node QB remains at low potential. In this cell read and write operations are decoupled, therefore, both can be optimized separately for desire stability. In the write mode, one of the inverters becomes floating therefore it overcomes the write one problem of the single BL structure. Moreover, during the hold operation signal WL, RWL and CSL are connected to GND while WLB is attached to VDD, therefore, transistors M7 and M8 become in the on-condition as a result, the proposed cell can preserve store data effectively.

3 Result and Discussion

The two significant design parameters of the SRAM cell are cell stability and power consumption, so we have selected these parameters for analysis. The stability of the SRAM cell is defined by a parameter known as “Static Noise Margin” (SNM). The Noise always has a “static” or DC nature, therefore, this parameter is named Static Noise Margin. RSNM is the maximum amount of noise that an SRAM cell can withstand before changing its state during a read process [13]. Ideally, the maximum value of RSNM is half of the supply voltage. Similarly, WSNM refers to the ability to write data in the cell via the write mode, while HSNM refers to the ability to retain stored data during a hold mode.

The butterfly method can be used to measure the stability of SRAM cells [13]. The circuit for measuring the RSNM of the 9T SRAM cell is shown in Fig. 2. To draw the Butterfly curve, first, the BL is precharged to the VDD, after that RWL and WWL are connected to the VDD, while the CSL and WL connect to the GND. In the next step, a DC voltage sweep is put on at node N1, and the QB voltage is measured to get a Voltage Transfer Curve (VTC). Similarly, Node N2 is subjected to a DC voltage sweep, and the voltage at node Q is calculated to obtain the VTC of inv-2. Subsequently, the VTC of inv-1 and the mirrored VTC of inv-2 are plotted on the same axis to form the butterfly curve (see Fig. 3). The RSNM of the cell is the arm length of the largest square that can fit inside the short wing of the butterfly curve. The method of calculating the WSNM and HSNM is the same as the RSNM calculation, but the control signal voltage will be according to Table 1 (Fig. 4).

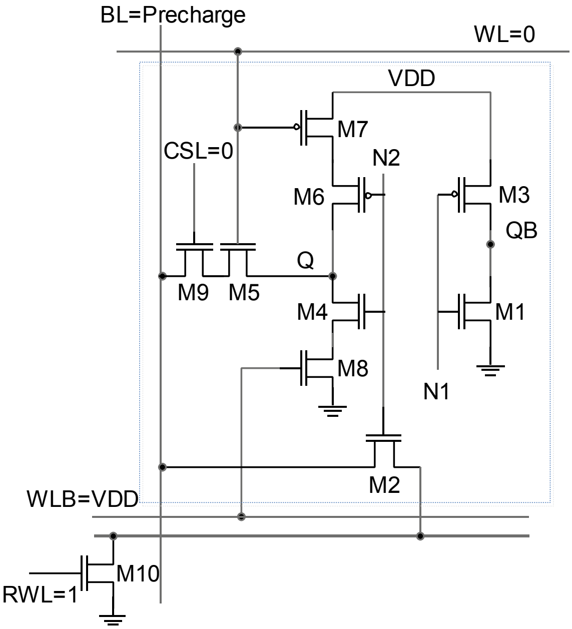


Fig. 2 Test circuit for evaluating the proposed 9T cell's read static noise margin

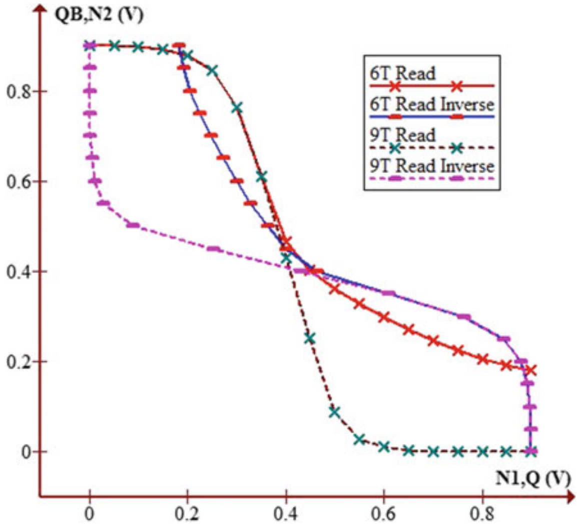


Fig. 3 Butterfly curve during the read mode in 9T and conventional 6T cell

Figure 5 shows the proposed 9T SRAM cell's RSNM, WSNM, and HSNM in contrast to a traditional 6T cell at 0.9 V. Since the read and write operations for the proposed 9T SRAM cell are decoupled, therefore both operations are optimized separately. The proposed 9T cell has an RSNM of 0.244 V, which is 4.28× higher than the regular 6T SRAM cell. This cell has a higher RSNM because of the gate sensing

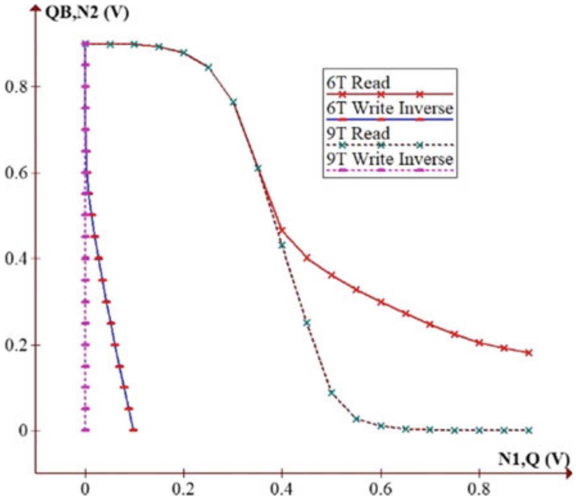
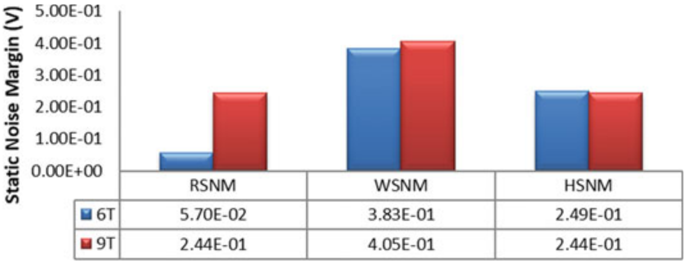


Fig. 4 Butterfly curve during write mode in 9T and conventional 6T cell



	RSNM	WSNM	HSNM
🟦 6T	5.70E-02	3.83E-01	2.49E-01
🟥 9T	2.44E-01	4.05E-01	2.44E-01

Fig. 5 Comparison of the write static noise margin, read static noise margin, hold static noise margin of proposed 9T cell and conventional 6T cell

scheme. Furthermore, since one of the inverters becomes weak during the writing process, the proposed SRAM cell shows 0.405 V WSNM, which is 1.06× higher than the traditional 6T SRAM cell. However, the reduction in HSNM is acceptable because it provides higher RSNM and WSNM compare to conventional 6T SRAM cell.

A key contributor to the standby power dissipation at the scaled technology node is a subthreshold leakage current. As the supply voltage or temperature rises, the leakage current grows exponentially. In this paper, the Leakage current was measured according to the procedure shown in Ref. [14]. To determine leakage power in hold mode, BL and WLB are connected to the VDD while WL, CSL, and RWL are tight to VDD. Figure 1 shows various Leakage current components in the proposed 9T cell. Leakage current is the sum of the leakage current(Isub between drain and source when the device is turned off), junction leakage current (Ijn between drain/source and substrate), and gate leakage current (Ig between drain/source and gate terminal), through various transistors in the SRAM cell during hold mode [5].

Equation 1 indicates the overall leakage current of the 9T cell [5, 15].

$$
\begin{aligned}
\text{Isub, PRO 9T } &= \text{Isub, M9 } + \text{Isub, M5 } + \text{Isub, M7} + \text{Isub,} \\
&\quad \text{M6} + \text{Isub, M1} + \text{Isub, M2} + \text{Isub, M10} \\
\text{Ijn, PRO 9T} &= \text{Ijnd, M9} + \text{Ijns, M9} + \text{Ijnd,} \\
&\quad \text{M5} + \text{Ijnd, M7} + \text{Ijnd, M6} + \text{Ijns, M6} + \text{Ijns,} \\
&\quad \text{M2} + \text{Ijnd, M2} + \text{Ijnd, M1} + \text{Ijnd, M10} \\
\text{Ig, PRO 9T} &= \text{Igd, M9} + \text{Igs, M9} + \text{Igd,} \\
&\quad \text{M5} + \text{Igs, M7} + \text{Igd, M7} + \text{Igs, M6} + \text{Igd,} \\
&\quad \text{M6} + \text{gd, M4} + \text{Igs, M4} + \text{Igd, M8} + \text{Igs,} \\
&\quad \text{M8} + \text{Igs, M3} + \text{Igd, M3} + \ \text{Igd, M1} + \text{Igs, M2} + \text{Igd, M10} \\
I_{\text{TOTAL\_ LEAK,PRO }9T} &= \ I_{\text{sub,PRO 9T}} \quad + \ I_{\text{jn,PRO }9T} \quad + \ I_{g,\text{PRO 9T}} \qquad (1)
\end{aligned}
$$

It seems from the above calculations that the proposed 9T cell consumes more leakage current than the 6T cell, but the result is very interesting, it consumes less leakage current because of the stacking effect. When two transistors are joined in series, if one device is attached to higher potential and the other device is connected to lower potential, the voltage of the intermediate node is increased to a value higher than voltage of the lower terminal, which reduces leakage current from the circuit, is called stacking effect [5].

(1) In standby mode, when RWL is off, the transistor stack is formed by the transistor M2 and the transistor M10, so that the intermediate node voltage grows to a positive value (629 mV in this case), consequently negative body potential raises the transistor threshold voltage (the larger body effect) and hence it reduces bit line leakage current.

(2) The intermediate node of the transistor rises to a positive value (118 mV) therefore drain to source voltage of transistor M5 and M9 decreases, it also decreases subthreshold current to some amount.

(3) One of the inverters of the 9T cell is made from four transistors M7, M6, M4, and M8, it increases the resistance between VDD and GND which helps to reduces subthreshold leakage current.

Figure 6 displays the leakage current simulation results of the 9T cell and its comparison with conventional 6T cell at 0.9 V. Despite having nine transistors, the proposed cell consumed 3.15 (4.16) n watt leakage power when data Q = VDD (0) whereas, the static power consumption was $7.51n$ W for the 6T cell in the similar situation. The proposed cell showed an average power consumption of $3.66n$ W which is 0.48× compared to a traditional 6T cell. The 9T consumes 0.42× (0.55×) leakage power compared to the 6T cell when Q = VDD (0). Due to single-ended sensing, the proposed 9T cell consumes low dynamic power in addition to low static power. Since only one-BL is needed for read or write operations in the proposed 9T cell (need to precharge BL only), its dynamic power consumption will be half that

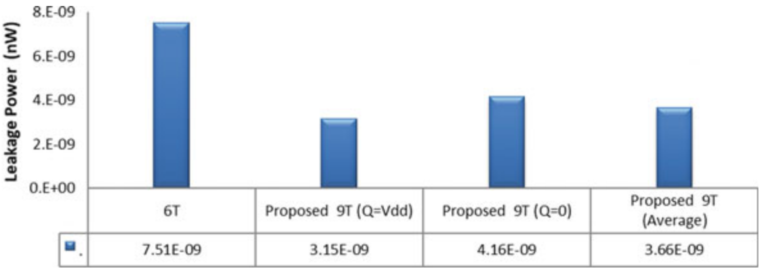


Fig. 6 The proposed 9T SRAM cell and a traditional 6T SRAM cell comparison in terms of leakage power consumption

of a differential SRAM cell [7]. As a consequence, we can say that the proposed 9T cell decreases both static and dynamic power simultaneously.

4 Bit Interleaving Architecture

The same word line is shared between the entire row in the conventional SRAM architecture, therefore during the write operation, some cells, in which do not have to perform the write operation are also get selected, and the unselected cell gets a pseudo-write operation. This problem is called a half select problem and it is undesirable in the case of memory cell design. Conventional 6T cell is suffered from half select issue, therefore, various half select free cells are proposed in the literature. One popular technique to replace single access transistors with two series-connected transistors therefore the connection between BL and data storage node is controlled by two transistors and noise from BL to data storage node is transferred only when both transistors are turned on. In the proposed cell gate of transistor, M5 is connected to the WL, that is common for the entire row, but the M9 transistor is regulated by the CSL. When CSL is low, transistor M9 becomes turn off and data storage node Q is disconnected from BL, therefore noise from BL unable to disturb the storage node voltage. Data is only written in the cell when the CSL and WL are turned on. The proposed cell has a half select free operation and can be used in bit interleaving architecture. In the bit interleaving architecture, multiple bit error can be treated like single bit error so single bit error correction codes (ECC) can be implemented effectively.

5 Conclusion

We have presented a new single-ended read/write differential 9T cell with higher stability and low power consumption. The findings of this study suggest that gate sensing (loop cutting) is an effective method to increase read (write) stability of the cell, and leakage (dynamic) power can be reduced through the stack (single-ended) technique. In contrast to the standard 6T cell, simulation results show that the proposed 9T cell offers 4.28× higher RSNM, 1.06× higher WSNM, and 0.48× lower power consumption. Finally, some potential limitations need to be considered. First, due to single-ended sensing, access time may be high, second, it requires a different SRAM architecture because of three control signals, and third, it consumes a large area due to a higher number of transistor compare to conventional 6T SRAM cell. Despite these negative aspects, we believe that at the nm technology node, the proposed 9T cell may be a better substation than the traditional 6T cell.

Acknowledgements This work was performed in the Research Hub-2, UTD, CSVTU, Bhilai under the CRP and it was sponsored by TEQIP-3.

References

1. Surana, N., Mekie, J.: Energy efficient single-ended 6-T SRAM for multimedia applications. IEEE Trans. Circuits Syst. II Express Briefs. **66**, 1023–1027 (2019). https://doi.org/10.1109/TCSII.2018.2869945
2. Sharma, V., Gopal, M., Singh, P., Vishvakarma, S.K., Chouhan, S.S.: A robust, ultra low-power, data-dependent-power-supplied 11T SRAM cell with expanded read/write stabilities for internet-of-things applications. Analog Integr. Circuits Signal Process. **98**, 331–346 (2019). https://doi.org/10.1007/s10470-018-1286-2
3. Rajput, A.S., Pattanaik, M., Tiwari, R.K.: Process invariant Schmitt trigger based static random access memory cell with high read stability for low power applications. J. Nanoelectron. Optoelectron. **14**, 746–752 (2019). https://doi.org/10.1166/jno.2019.2577
4. Liu, Z., Kursun, V.: Characterization of a novel nine-transistor SRAM cell. IEEE Trans. Very Large Scale Integr. Syst. **16**, 488–492 (2008)
5. Islam, A., Hasan, M.: Leakage characterization of 10T SRAM cell. IEEE Trans. Electron. Dev. **59**, 631–638 (2012)
6. Rajput, A.S., Pattanaik, M., Tiwari, R.K.: Stability and leakage characteristics of a Schmitt trigger-based 10T SRAM cell. In: Jain, R. (ed.) International Conference on Nanomaterials: Initiatives and Applications, pp. 88–89. Jiwaji University, Gwalior (M.P.), Gwalior (M.P.) (2018)
7. Ahmad, S., Gupta, M.K., Alam, N., Hasan, M.: Single-ended schmitt-trigger-based robust low-power SRAM cell. IEEE Trans. Very Large Scale Integr. Syst. **24**, 2634–2642 (2016)
8. Kushwah, C.B., Vishvakarma, S.K.: A single-ended with dynamic feedback control 8T subthreshold SRAM cell. IEEE Trans. Very Large Scale Integr. Syst. **24**, 373–377 (2016). https://doi.org/10.1109/TVLSI.2015.2389891
9. Pal, S., Slam, A.: 9T SRAM cell for reliable ultralow-power applications and solving multi-bit soft-error issue. IEEE Trans. Device Mater. Reliab. **16**, 172–182 (2016). https://doi.org/10.1109/TDMR.2016.2544780
10. Predictive Technology Modeling. http://www.eas.asu.edu/~ptm. Last accessed 2020/10/01
11. Upadhyay, G., Rajput, A.S., Saxena, N.: An analysis of novel 12T SRAM cell with ımproved read stability. Int. J. Innov. Res. Eng. Appl. Sci. **3** (2017). 310717/3/1-1/July
12. Gupta, R., Rajput, A.S., Saxena, N.: Improvement in read performance of 10T SRAM cell using body biasing in forward bias regime. IPASJ Int. J. Electron. Commun. **4**, 1–9 (2016)

13. Rajput, A.S., Pattanaik, M., Tiwari, R.: Estimation of static noise margin by butterfly method using curve-fitting technique. J. Act. Passiv. Electron. Dev. **13**, 1–9 (2018)
14. Chung, Y.: Stability and leakage characteristics of novel conducting PMOS based 8T SRAM cell. Int. J. Electron. **101**, 831–848 (2014). https://doi.org/10.1080/00207217.2013.805355
15. Yadav, A.S., Nakhate, S.: Low standby leakage 12T SRAM cell characterization. Int. J. Electron. **103**, 1446–1459 (2016). https://doi.org/10.1080/00207217.2015.1126859

Spam Detection Using Genetic Algorithm Optimized LSTM Model

Abhinav Sinhmar, Vinamra Malhotra, R. K. Yadav, and Manoj Kumar

Abstract The advancement in technology over the years has resulted in the increased usage of SMS which in turn has provided certain groups a chance to exploit this service by spreading spam messages to consumers making it difficult for people to receive important information and also possessing a threat to their privacy. There are numerous machine learning and deep learning techniques that have been used for spam detection and have proved to be effective. But in deep learning techniques, it is essential to fine-tune the hyperparameters which requires excessive computational power and time, making the process less feasible. The proposed work aims at reducing this computational barrier and time by using Genetic Algorithm in order to select the key hyperparameters. A randomly generated population of LSTM models was created and further generations were produced following the different stages of the genetic algorithm multiple times until the terminal condition was met, and the performance of each candidate solution was evaluated using a chosen fitness function. The most optimal configuration was obtained from the final generation which is used to classify the messages. Four metrics, namely the accuracy, precision, recall and f1-score were used to analyze the model's performance. The experimental results demonstrate that the Genetic Algorithm optimized LSTM model was able to outperform the other machine learning models.

Keywords SMS · Spam · LSTM · Evolutionary algorithm · Genetic algorithm · Tokenization · Sequencing · Padding · Fitness · Gray code · Hyperparameters.

1 Introduction

Short message service (SMS) is a technique that enables mobile users to send information to another user in the form of short text messages over a mobile network. It has been one of the significant factors for the increased usage of mobile devices. But one of the major problems faced by mobile phone users is the reception of spam SMS.

A. Sinhmar (✉) · V. Malhotra · R. K. Yadav · M. Kumar
Delhi Technological University, Delhi 110042, India
e-mail: mkumarg@dce.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_5

Spam refers to an unwanted commercial message, typically delivered to several users. A spam SMS can therefore be described as any unwanted or junk message delivered via text messages to a mobile device [1]. A spam is equally dangerous in both the forms, be it an email spam or SMS spam as it can lead to personal data breach of mobile device users. This spam SMS problem has increased to an extent that in 2012, out of all the SMS messages that were received in Asia, around 20–30% were reported as spam [2]. For an instance, in 2008, the mobile users in China got approximately 20,000 crore spam SMS messages in a week. Although, in 2011, the amount of spam messages received by a person per day in North America was about 0.1% only, still 44% of mobile device users in the US confirmed being a victim of spam SMS during a survey [3].

Over the last few years, spam has evolved and takes tremendous effort to handle it. Several methods for managing and identifying unsolicited messages have been suggested by several researchers to reduce the spam effect on mobile phone users. These consist of machine learning classification algorithms like support vector machine (SVM) [4], decision trees and others along with NLP techniques [5, 6]. Further development in the field of deep learning has resulted in the use of deep learning techniques like convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM) for spam detection and the results produced by these techniques have been satisfactory.

Neural networks, while being an undeniably powerful tool, are typically associated with a broad collection of hyperparameters that determine the topology and computational power of the network. Therefore, optimizing the configuration of the meta parameters seems like a daunting job, since it varies based on the work it performs, the dataset being considered, etc., making every problem distinct. Mostly, neural network's hyperparameters are selected through a trial and error approach. But this approach is computationally expensive and consumes a lot of time [7]. In recent times, due to the increase in the availability of computational resources, Evolutionary algorithms are being used to automatically find the optimal neural architectures. An evolutionary algorithm (EA) is an algorithm which simulates the biological phenomenons like mutation, recombination, and evolution in order to find an ideal configuration within specific constraints [8].

Genetic algorithm is a type of evolutionary algorithm and a heuristic search and an optimization method that uses an approach inspired by evolutionary mechanisms [9] such as inheritance, crossover, mutation, selection applied to the population of candidate solutions and the candidates with higher fitness values have a probability to survive more.

In the proposed work, genetic algorithm is used to optimize the hyperparameters of LSTM aimed at getting the best possible parameter set for detecting the spam SMS.

2 Related Work

Gupta et al. [1] in their research used multiple datasets and carried out comparison among eight most commonly used machine learning classifiers. There were 5574 spam and non-spam messages in English language collected from different sources in one dataset. The other dataset contained 1000 spam and ham SMS each. The classifiers to be compared were SVM, Naive Bayes, decision tree, logistic regression, random forest, AdaBoost, ANN and CNN. Four performance metrics were used to compare different classifiers, namely accuracy, precision, recall and CAP Curve with accuracy being the most important metric. From the results obtained for both the datasets, it was found that CNN had the highest accuracy in both the cases.

Mahajan and Kaur [10] in "Neural Networks using Genetic Algorithms" have analysed the concepts of genetic algorithm and neural networks and then also solved the travelling salesman problem (TSP) using the genetic algorithm. The authors have talked about no matter what the topology of the neural network is, genetic algorithm can be used to train them. They also discussed how genetic algorithm can influence the connections between the neurons. To solve the travelling salesman problem and give a maximal approximation of the problem with the reduction of cost the sequential constructive crossover technique was presented highlighting the importance of crossover operators for TSP.

The authors in their research [11] used genetic algorithm, an evolutionary algorithm, to optimize the parameters of the artificial neural network and also to modify the network weights to efficiently improve the learning of the neural network. To overcome the imperfection of backpropagation, the authors used an approach to optimize the network weights instead of fine-tuning the hyperparameters of the neural network. Experimental results showed that the hybrid algorithm implemented using the given approach gave better results when compared to the conventional artificial neural network. The experiments indicated the increase in efficiency along with a decrease in the number of false positives and false negatives by using the genetic algorithm optimized approach.

Gorgolis et al. [7] in their work proposed an approach using genetic algorithm, to optimize the hyperparameters of neural network models to alter the efficiency by not pursuing an "exhaustive search method." An int array formed from the hyperparameters selected was used to represent the chromosomes in the population for the genetic algorithm. The initial population containing the candidate solutions was formed by selecting random integer values for the hyperparameters to form the vector of integers. Parents from the population were selected and the genetic operations crossover and mutation were applied to generate new candidate solutions, and then the solutions with best values for the fitness functions were selected for the next generation. Through the experiments performed, the authors were able to present that the genetic algorithm is an effective and feasible approach for optimizing the hyperparameters to get a fine-tuned neural network.

Chung and Shin [12] in "Genetic Algorithm-Optimized Long Short-Term Memory Network for Stock Market Prediction" used the financial data available to develop a stock market prediction model. In the research, the authors proposed a hybrid approach combining the genetic algorithm with the LSTM neural network. To build an LSTM neural network, a number of hyperparameters like number of hidden layers, number of nodes per layers, need to be chosen for optimum results. This research suggested a method to decide the LSTM model's size of time window and configuration using genetic algorithm for stock market financial prediction. In all the error metrics, genetic algorithm improved LSTM Model showed improved results. The better performance attained from the hybrid model could be the result of the improved training procedure on LSTM network. The experiments indicated that adequate altering of the configuration is necessary and can lead to better results.

3 Research Methodology

3.1 Dataset Description

We have used two different datasets to train and evaluate our model.

Dataset 1: SMS Spam Collection Data Set The SMS (text) data was down-loaded from UCI Machine learning Repository. The intention of collection of the data was for SMS spam detection on cell phones. The dataset consists of 5,574 SMS messages and the messages are already labeled as either spam or ham.
The Dataset has 2 columns: Label and message, label specifies if the message is spam or ham and message contains the content of the SMS.
The dataset contains:
Ham: 4825 messages
Spam: 747 messages

Dataset 2: Spam SMS Dataset 2011-12 This dataset was collected by Yadav et al. [13] through incentivized crowdsourcing and personal contacts. The dataset consists of 2000 files distributed in two folders labelled as spam and ham containing 1000 files each. Each file contains exactly 1 message encoded in UTF-8 format.

3.2 Dataset Preprocessing

Text classification problems like sentiment analysis, spam detection, text generation and document classification generally use natural language processing (NLP). Text data can be taken as a sequence of words or characters or sentences. But it is usually considered as a sequence of words in most of the cases. We converted text into numerical representation as deep learning models do not understand raw text.

We preprocessed data in 3 steps:

- Tokenization
- Sequencing
- Padding

Tokenization Tokenization is the process of breaking sentences into tokens (words) and encoding each word into an integer value. The sentences are separated into words with the help of Tokenizer API and the words are then encoded into integer values. (Words, encoded value) is added as a key value pair in a word_index dictionary.

Sequencing Sequencing is used to convert the sentences into a sequence of numbers. The sentences are converted into integer sequences by the Tokenizer object with the help of the tokenized word_index dictionary it created.

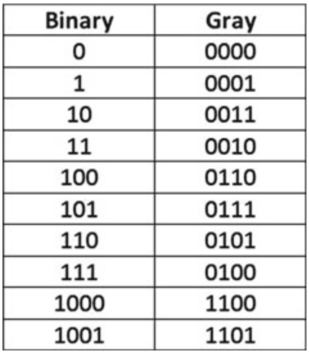
Padding Naturally, sentences present in any raw text data will be of separate lengths. But all neural networks need to have the same size inputs. Padding is used to convert sequences we obtained in the last stage into sequences of the same size. The required padding is done with the help of pad_sequences from tensor flows.

3.3 Our Approach

Genetic Algorithm to Optimize the LSTM Configuration As discussed earlier in the paper, neural network's hyperparameters are selected through trial and error approach which consumes a lot of computational power and time [7]. Hence, in order to optimize the hyperparameters of our LSTM model, we used genetic algorithm, an evolutionary algorithm, to get the best possible parameter set. The process of genetic algorithm is of the following stages: population initialization, fitness evaluation, selection, crossover, mutation, and then terminal condition check.

Before initializing the population, we need to select a representation for the candidate solution. And one of the most widely used representations for genetic algorithms is binary representation. We used bit strings to represent the genotype in this representation. The thought of using binary representation comes naturally when the solution space consists of Boolean decision variables, as the presence or absence of something can be represented by simply two states.

Problems in which we deal with integers, they can be represented by using binary representation. One issue with this type of encoding is that the significance of each bit is different, so there can be undesired implications for mutation and crossover operators. We used gray code to solve this issue to some degree, as by using gray code a shift in one bit does not have as big an impact on the solution as in binary representation because gray code being unweighted remains unaffected by a digit's positional value. Using gray code also enables us to prevent instances of walls [14]. A hamming wall can be defined as the stage where the possibility of genetic algorithm mutating in the appropriate manner to get a better fitness becomes negligible.



Binary	Gray
0	0000
1	0001
10	0011
11	0010
100	0110
101	0111
110	0101
111	0100
1000	1100
1001	1101

Fig. 1 Conversion from binary to gray code

For the conversion of binary number to its corresponding gray code encoding the following procedure is used:

1. Most Significant Bit (MSB) of the gray code and the MSB of the binary number to be converted will be exactly the same.
2. The second bit of the gray code (from the MSB) will be the result of exclusive-or (XOR) of the first and second bit of the binary number (from the MSB) taken into account.
3. Similar to the second bit, the third bit of the gray code will be the result of exclusive-or (XOR) of the second and third bit of the binary number taken into account. And the process can be repeated for the rest of the bits.

The key hyperparameters of LSTM like number of hidden units, embedding dimensions, batch size, and number of Epochs can be stored in an array represented in the form of bits and this array can be used to represent a chromosome.

Algorithm

1. population ← [list of *n* randomly generated binary arrays representing candidate solutions]
2. generation ← 1
3. max_generations ← max number of generations which will be used for the terminal condition
4. do:

   (a) evaluate_fitness ← train and calculate the fitness of each individual

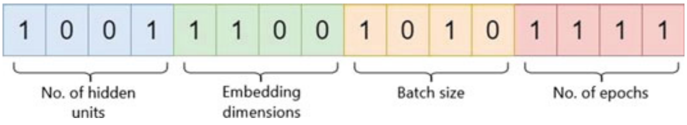


Fig. 2 Key hyperparameters of LSTM

(b) selection ← To select individuals for performing Genetic Operations
(c) crossover ← 2 individuals are selected and crossover operation is performed
(d) mutation ← 1 individual is selected and is given a chance to mutate
(e) population ← n fittest solutions are retained after every iteration
(f) generation ← generation + 1
(g) while (generation $\leq$ max_generations)

5. Obtain the hyperparameter values from the fittest solution to build the LSTM model
6. Train and predict the messages using the LSTM model obtained in the last step

Initializing Population Creating the initial population of models is the very first step of genetic algorithm. For this we randomly selected the values of hyperparameters for each model from the defined search spaces. **random**, which is a python module, was used for this purpose that follows uniform distribution.

Fitness Function The most critical part of genetic algorithm is fitness function and therefore, it has to be chosen carefully. The fitness function plays a key role in evaluating the performance of the candidate solutions generated and comparing the new generation with the old ones. The 3 candidates for a good fitness function for spam detection problems are precision, recall, f1-score. Precision is calculated as true positives divided by the number of true positives + the number of false positives. Precision is a crucial metric to evaluate whether or not the count of false positives is high. In spam classification, any message that is wrongly classified as spam is known as a false positive. Therefore, low precision can lead to loss of useful information.

After initializing the population with a number of random arrays denoting the values of hyperparameters in gray code like the one shown in Fig. 2, an LSTM model is trained using these hyperparameters and the precision is then calculated from the confusion matrix obtained by predicting the result for the testing split. This precision is used as the fitness value of each candidate solution.

Selection To select the new generation from each preceding generation, we selected a portion of that population. Candidate solutions were chosen following a fitness based criteria, where solutions with higher fitness value were preferred over the others. There are multiple selection methods. A lot of functions prefer to select the solutions with the higher value of fitness, but some weakly fitted solutions are also selected. This way the diversity of the population is preserved and weak solutions are prevented from premature convergence. Some of the most broadly used and well investigated methods for selection are Roulette wheel selection and tournament selection [15]. The method used in this paper is Roulette wheel selection.

Crossover New candidate solutions were then created from the existing population by the crossover operation. In this step, the selected superior chromosomes produced the offsprings by interchanging the gene combinations and the corresponding parts of the string. There are multiple types of crossover such as single-point crossover, multi-point crossover, etc. and the one used in our model is ordered crossover [12].

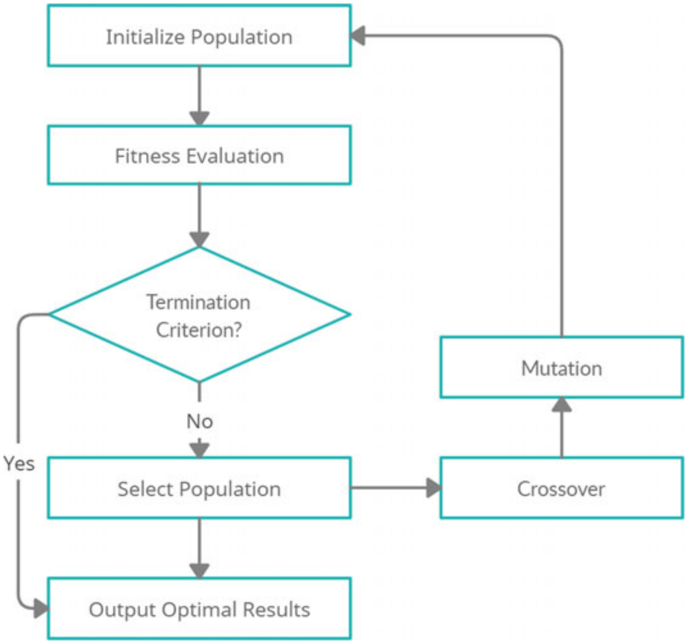


Fig. 3 Flow-diagram of the genetic algorithm

Mutation We allowed for a small chance of mutation to make sure that the individuals were not all exactly identical. Out of all the chromosomes generated in crossover step, one chromosome was selected to modify a randomly chosen bit during the mutation process. One of the limitations of the crossover process is that it cannot generate fully new information. However, modification of the corresponding bits to entirely different values through the mutation operation helps in getting rid of this problem. We have used swap mutation for our model.

Terminating Condition Upon the creation of a new generation, the process repeats again in an iterative manner, until the terminating condition is satisfied [16]. There are multiple criterion that could be used as the terminating condition such as attaining a certain fitness function value or running the genetic algorithm for a certain number of generations. The terminating condition we have used is the completion of a certain number of generations.

Training the LSTM Model The candidate solution which had the highest fitness in the last generation was taken as the best solution from the genetic algorithm and the hyperparameters extracted from the selected candidate solution were used to build our LSTM Model and it was then trained on both the datasets to classify the messages as spam or legit.

4 Results

In order to evaluate the performance of our model, we used some well known classifiers like Naive Bayes classifier, support vector machine (SVM), decision trees and artificial neural network to train on both the datasets and then compared the results obtained for each of these classifiers.

As the number of ham messages was significantly greater than the number of spam messages, accuracy alone could not be used as a metric to assess the performance of the classifier. So we also used precision, recall and F1-score along with accuracy as performance metrics for the comparison of our model with the other classifiers. All these metrics were evaluated using the following formulas:

$$\text{Accuracy}(A) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{1}$$

$$\text{Precision}(P) = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall}(R) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F1\text{−score} = 2 * \frac{P * R}{P + R} \tag{4}$$

where,

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The values of these metrics obtained for both the datasets for different classifiers are tabulated in Tables 1 and 2, respectively, and the best prediction results are in bold.

Table 1 Dataset 1: performance metrics

Classifiers	Accuracy (%)	Precision	Recall	$F1$-score
Support vector machine	98.08	0.962	0.882	0.920
Naive Bayes	98.30	0.940	0.920	0.930
Decision tree	96.41	0.888	0.826	0.856
Artificial neural network	98.56	0.974	0.925	0.949
LSTM	97.22	0.966	0.959	0.962
GA optimized LSTM (our model)	99.01	0.986	0.967	0.976

Table 2 Dataset 2: performance metrics

Classifiers	Accuracy (%)	Precision	Recall	$F1$-score
Support vector machine	94.50	0.970	0.923	0.946
Naive Bayes	95.33	0.970	0.940	0.958
Decision tree	89.33	0.905	0.887	0.895
Artificial neural network	97.75	0.985	0.970	0.977
LSTM	97.75	0.985	0.971	0.978
GA optimized LSTM (our model)	98.00	0.995	0.967	0.980

The least precision rate for Dataset 1 has been 0.888 which is obtained by decision tree classifier, and it also holds the least accuracy out of all the models. The reason for this may be its greedy approach that it uses for choosing results from the generated tree. Our model however was able to achieve a relatively higher precision rate at 0.986 along with the highest accuracy of 99.01%.

For dataset 2 also, the least precision rate is attained by decision tree classifier at 0.905 and the overall trend for all the classifiers has been almost linear. SVM and Naive Bayes classifiers bag the same precision rate at 0.970. ANN and LSTM models have shown the same accuracy of 97.75% and also the same precision rate at 0.985, whereas GA optimized LSTM model attains the highest accuracy and precision rate at 98.00% and 0.995 respectively with a comparable recall fraction at 0.967.

Machine learning algorithms are optimized using Loss functions. The loss is determined on the training and validation data, and its value depends on the performance of the model on these two data splits. Loss is the total number of errors made on each datapoint of these data splits. Whether a model performs well or poor after each pass over the entire dataset is signified by the loss value.

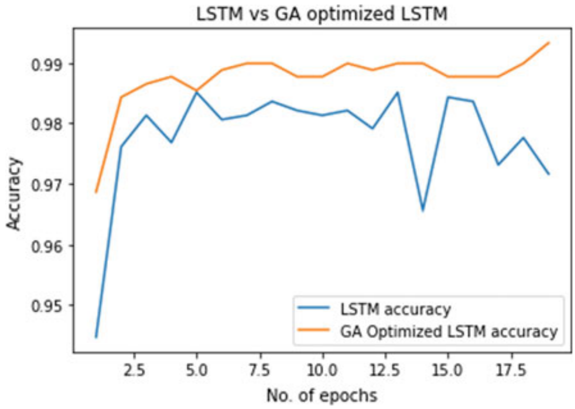


Fig. 4 Accuracy versus no. of epochs

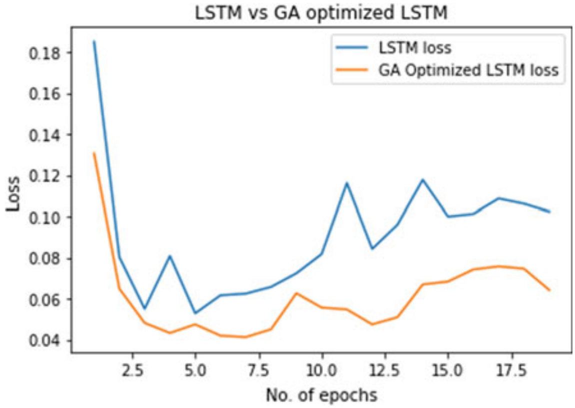


Fig. 5 Loss versus no. of epochs

The performance of an algorithm is computed using a metric called accuracy. It is generally measured as a percentage depending on the model parameters. It is a metric for how close a model's prediction is to the actual results.

Dataset 1

A comparison is shown in Fig. 4 between our model and LSTM model for dataset 1. It can be clearly seen that the accuracy increases drastically for a few initial epochs and after that the accuracy increases gradually as the number of epochs increases. And throughout this period our model has been more stable than the LSTM model as there are no drastic dips in the accuracy. Also, the accuracy of our model is higher throughout.

It can be seen in Fig. 5 that the loss reduces drastically for a few initial epochs and after that the loss reduces gradually as the number of epochs increases. And throughout this period, our model has been more stable than the LSTM model as

Fig. 6 Accuracy versus no. of epochs

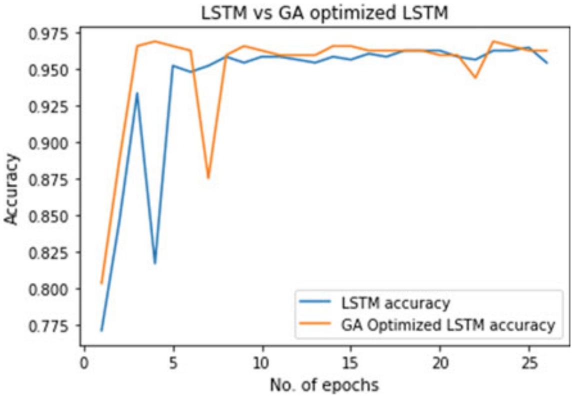
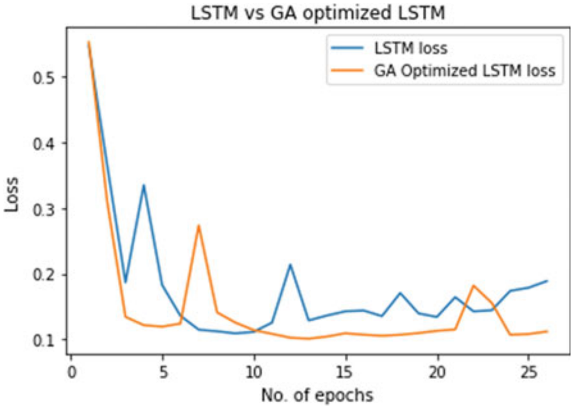


Fig. 7 Loss versus no. of epochs



there are comparatively fewer drastic spikes in the loss. Also, the loss of our model is lower throughout for dataset 1.

Dataset 2

From Fig. 6, it can be observed that for dataset 2, the accuracy of our model is quite comparable to the LSTM model as the graph for both the models intersect each other quite a number of times. But it can be seen from Table 2, that the GA optimized LSTM performs slightly better than the LSTM model for most of the performance metrics.

Similar to the accuracy graph for dataset 2, the loss graph also has both the models intersecting each other for a certain number of times. But eventually, it can be seen that our model has lower loss value than the LSTM model.

5 Conclusion and Future Work

From the results obtained for our model, genetic algorithm proves to be an efficient and practical approach in order to fine-tune a neural network as our genetic algorithm optimized LSTM model outperforms all the other classifiers in most of the performance metrics. In dataset 1, our model was able to achieve an accuracy of 99.01% along with a precision of 0.986 and the values for these metrics were 98.00% and 0.995, respectively, for dataset 2. The satisfactory performance of our model indicates that the inclusion of other evolutionary algorithms in this field can also be effective.

Hence, for future work, we propose to incorporate the other evolutionary algorithms like particle swarm optimization (PSO), firefly algorithm, etc. to optimize the neural networks in order to obtain the best possible topology of the model. Also, the parameters are not limited to the ones used in our model and fine-tuning of other parameters can also be done to obtain different results.

References

1. Gupta, M. et al.: A comparative study of spam SMS detection using machine learning classifiers. In: 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE, 2018, pp. 1–7
2. Navaney, P., Dubey, G., Rana, A., SMS spam filtering using supervised machine learning algorithms. In: 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018, pp. 43–48
3. Delany, S., Buckley, Greene, D.: SMS spam filtering: methods and data. In: Expert Systems with Applications (Feb. 2013), pp. 9899-9908. https://doi.org/10.1016/j.eswa.2012.02.053
4. Nizar Bouguila and Ola Amayri: A discrete mixture-based kernel for SVMs: application to spam and image categorization. Inf. Process. Manage. **45**(6), 631–642 (2009)
5. Bahgat, E.M., Rady, S., Gad, W.: An e-mail filtering approach using classification techniques. In: The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), Nov 28–30, 2015, pp. 321–331. Springer, Beni Suef, Egypt, 2016
6. Islam, M.S., Mahmud, A.A., Islam, M.R.: Machine learning approaches for modeling spammer behavior. In: Asia Information Retrieval Symposium, pp. 251–260. Springer, 2010
7. Gorgolis, N.: Hyperparameter optimization of LSTM network models through genetic algorithm. In: 10th International Conference on Information, Intelligence, Systems and Applications (IISA), pp 1–4. IEEE, 2019
8. Elbeltagi, E., Hegazy, T., Grierson, D.: Comparison among five evolutionary-based optimization algorithms. Adv. Eng. Inform. **19**(1), 43–53 (2005)
9. McCall, John: Genetic algorithms for modelling and optimisation. J. Comput. Appl. Math. **184**(1), 205–222 (2005)
10. Mahajan, R., Kaur, G.: Neural networks using genetic algorithms. Int. J. Comput. Appl. **77**(14) (2013)
11. Arram, A., Mousa, H., Zainal, A.: Spam detection using hybrid artificial neural network and genetic algorithm. In: 2013 13th International Conference on Intelligent Systems Design and Applications. IEEE, pp. 336–340, 2013
12. Chung , H., Shin, K.: Genetic algorithm-optimized long short-term memory network for stock market prediction. Sustainability **10**(10), 3765 (2018)

13. Yadav, K. et al.: SMSAssassin: crowdsourcing driven mobile-based system for SMS spam filtering. In: Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, 2011, pp. 1–6
14. Charbonneau, P.: An introduction to genetic algorithms for numerical optimization. In: NCAR Technical Note 74 (2002)
15. Zhong, J. et al.: Comparison of performance between different selection strategies on simple genetic algorithms. In: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), vol. 2, pp. 1115–1121. IEEE, 2005
16. Tabassum, M., Mathew, K., et al.: A genetic algorithm analysis towards optimization solutions. Int. J. Dig. Inf. Wirel. Commun. (IJDIWC) **4**(1), 124–142 (2014)

Affine Recurrence Based Key Scheduling Algorithm for the Advanced Encryption Standard

S. Shashankh, Tavishi Kaushik, Svarnim Agarwal, and C. R. Kavitha

Abstract Encryption algorithms such as Advanced Encryption Standard (AES) are known as symmetric encryption algorithms which use the same key for both encryption and decryption. These algorithms have a huge variety of applications such as for securing data and transferring files. They also have a key expansion algorithm which is used for expanding the given key. In AES the key expansion algorithm expands the given 128 bit key into 176 bytes which will then be used in the encryption process which spans for 10 rounds. This paper aims at making this process even more secure. This is implemented by including different steps to increase security in this key expansion process so that it becomes computationally infeasible to get the original key even if the adversary gets a hold of the different parts of the key. Current works suggest that there is a need for increased security in the key scheduling algorithm of the AES. It has also been tested to have inferior strict avalanche criteria in comparison with other contenders of the AES such as Serpent and Twofish. Usage of the concept of affine recurrence ensures this in the proposed model. In affine recurrence, no two outputs of the operation will have a relation between them. Another concept used is the AES Substitution Box (S-box), and this is done to ensure higher levels of confusion in the key scheduling process.

Keywords Symmetric encryption · AES · Substitution box · Affine recurrence · Bit independence criteria · Strict avalanche criterion · Related key attacks

S. Shashankh · T. Kaushik · S. Agarwal (✉) · C. R. Kavitha
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: svarnim.agarwal1805@gmail.com

S. Shashankh
e-mail: shashankh.s22@gmail.com

T. Kaushik
e-mail: tieshuekaushik@gmail.com

C. R. Kavitha
e-mail: cr_kavitha@blr.amrita.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_6

1 Introduction

Cryptography in general involves the study of methods and algorithms that are used for achieving secrecy while sharing data. These algorithms have both an encryption method and a decryption method. Symmetric Cryptography is a form of cryptography where both the encryption process and decryption process use the same key. These algorithms are widely used for transferring files over the Internet or even just viewing a webpage on a browser. These algorithms are also essential in ensuring security in the Internet of Things devices [1]. Even the Constrained Application Protocol (CoAP) has AES suite requirements [2]. There are new methodologies using which one can do computations on encrypted data. This can be used for computations among confidential information so that the data need not be decrypted at some server. This is known as homomorphic encryption [3]. These fully homomorphic encryption algorithms can provide extensive privacy and security when considering blockchain applications [4, 5].

In the process of encryption, a message or a file or any form of data is encoded such that it can be read only by an intended receiver. Encryption algorithm uses various different techniques such as scrambling and substituting so as to encode the data and all this is done with a secret key as a deciding element which will be shared with the intended receiver. The intended receiver will then use this secret key to decrypt the data which is essentially unscrambling the encoded data to get back the original or plain data. The secret key that is required by both the sender and receiver has their own procedures for being shared. There are multiple algorithms and methodologies for sharing the secret key between the sender and receiver [6]. This project is specifically using the Advanced Encryption Standard (AES) which is used in all modern applications for encryption and decryption. Most of the file transfer that takes place over the Internet makes use of the secrecy that AES is capable of providing. The AES-128 algorithm takes a 128 bit key and expands it to 176 bytes. This expansion is done using the Key Schedule Algorithm. The Key Schedule Algorithm (KSA) present in AES has recently been found to have some security-based problems. It could be attacked using related key attacks where using one of the sub-keys from expansion, other parts of the expanded key could be reverse engineered [7]. On a comparative study of key scheduling algorithms, the presence of Strict Avalanche Criterion was also evaluated to be low in the KSA of the AES [8].

This paper proposes a method so as to increase the security in the Key Scheduling Algorithm in the Advanced Encryption Standard. Based on the related works it can be concluded that the Key Scheduling Algorithm (KSA) used in the Advanced Encryption Standard (AES) compromises Strict Avalanche Criterion and Bit Independence Criterion [8]. This means that the sub-key is related to the original key. If an adversary is able to obtain a part of this expanded key there is a possibility that the adversary will be able to trace back the rest of the expanded key or even worse get the whole original key by using reverse engineering methods [7]. So, the XOR and rotation operations of the existing Key Scheduling algorithm are not sufficient for the security of the AES algorithm.

2 Related Works

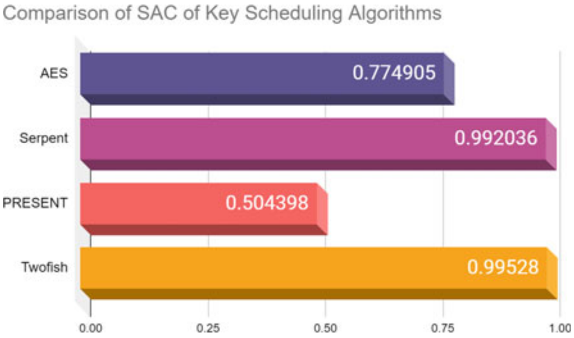
The authors of paper [7] have proposed a new variation of AES. Many different variations of this block cipher are present for providing security services. This algorithm being one of the most well-known symmetric encryption algorithms has been attacked by many adversaries by cryptanalytic processes. Therefore, it is required to strengthen the algorithm. The paper [7] points out the problems present in the key expansion process. The paper shows how related sub-keys and weak keys are being generated [9]. Many different studies have been conducted on the modern-day cryptographic algorithms but not many evaluation methodologies are present to test the key scheduling algorithms of these encryption algorithms.

The research paper [8] not only proposes different ways of judging key scheduling algorithms but also provides a comparative result. The paper has presented their results comparing the AES, Serpent, Twofish, IDEA and Present. The results show that the degree of SAC in the key scheduling algorithm of the AES is bad compared to the key scheduling algorithms of Serpent or Twofish. This paper proposes a concept of cryptographic dynamics in the key scheduling algorithms. It provides many different statistical tests to check cryptographic properties. It ultimately states that in their tests AES has performed badly when it comes to its compliance to the strict avalanche criterion.

The results from that research paper of all the different values for the degree of SAC of the Key scheduling algorithms have been compared and shown in Fig. 1. It can be noticed how AES has a lower degree of SAC compared to Serpent and Twofish which were competitors for the title of AES.

The research paper [10] is the block cipher Rijndael which was one of the proposed methods for the appointment of AES by The US National Institute of Standards and Technology (NIST). AES requested a 128-bit block cipher with a variable key length.

Fig. 1 Comparison of degree of SAC of key scheduling algorithms



It is an algorithm eventually crowned as Advanced Encryption Standard by the NIST. This algorithm is now considered to be one of the most used encryption algorithms in most web applications. The structure of the AES follows a Substitution-Permutation methodology. The AES will be elaborated on more in Sect. 3 of this paper.

In the paper [11], the authors talk about another method that was proposed for the AES competition. It came in second behind Rijndael. Serpent is very similar to Rijndael with the difference being that Rijndael is faster, while Serpent is more secure. It uses double the number of rounds to handle all currently known shortcut attacks. Serpent cipher consists of 32 rounds of encryption and also involves a different substitution box for different rounds of the algorithm. In the serpent algorithm, there are 8 substitution boxes. Each box has an input and output of 4 bits in length. Similar to the Rijndael algorithm, even serpent had key sizes of 128, 192 and 256 bits. It also has a fixed block size of 128 bits. The serpent cipher incorporates the concept of primitive polynomials. This was an inspiration for the method being proposed in this paper.

3 Advanced Encryption Standard

Arguably the most well-known and widely used symmetric encryption algorithm nowadays is the Advanced Encryption Standard [12]. The AES has immense applications in Cloud Security [13]. This is originally known as the Rijndael algorithm.

With increased computing power, attackers were capable of getting the key for DES via brute force attacks, hence 3DES was introduced. But this had a major drawback because it was found to be too slow. Hence a new algorithm was needed.

The characteristics of AES are as seen below:

- Symmetric encryption algorithm
- Uses 128/192/256-bit key
- Block size 128 bits or 16 bytes

AES has a different approach to its computations in the sense that it applies all its operations on bytes instead of on bits. Therefore, the AES will treat a block as 16 bytes instead of 128 bits. There are implementations of AES in VLSI systems as well, like many other encryption algorithms [14].

The algorithm consists of four main steps in each round, they are:

1. SubBytes which uses the AES S-box to substitute each byte in the input.
2. ShiftRows which performs a right shift to each row. The number of shifts is equal to the row number which is enumerated from 0 to 3.
3. MixColumns is a step in which matrix multiplication is performed in the field GF(2^8).

4. Add Round Key in which an XOR operation is performed between the current state of the plain text and the part of the key which is meant to be used in the respective round of the encryption process.

This Add Round Key step requires keys which are generated from the key scheduling algorithm. In the original AES key expansion, a 128-bit key is expanded into a 176-byte key which is 44 words. In this process, the originally given 128 bits are taken as the required 4 words for the first round of the encryption algorithm. The next words in the expanded key depend on the previous words of the key.

When the position of the word in the array is not a multiple of 4, the XOR of the previous word and the word four positions back is used. Otherwise, a more convoluted procedure represented as g is used which consists of the following three described steps:

1. RotWord which converts one word by applying one byte left rotate.
2. SubWord which substitutes the word by using the AES S-Box.
3. A round constant referred to as Rcon is combined with the results of the above two steps by applying an XOR function

As seen in the study of the related works, this KSA has some problems and to tackle this a new KSA is proposed by the authors called Substitution Affine Recurrence-AES (SAR-AES).

4 SAR-AES Key Scheduling

The modified version of AES that is being proposed makes use of two main steps in its key scheduling algorithm. It is the use of the AES Substitution Box (S-Box) and Affine Recurrence. A description of the aforementioned two steps is explained below with respect to AES-128. AES-128 is the variation of AES in which the initial key size is 128 bits and the expanded key is 176 bytes. This method of designing the key scheduling algorithm for AES has not been done before. The aim of the S-box is to bring in the element of confusion into key generation. This step is then followed by affine recurrence which has never before been applied in the key scheduling algorithm for AES. In other solutions to creating a new key schedule algorithm for AES, there is some element that usually causes an increase in the time consumed for computing the expanded key. But this method performs similar to the original AES, a comparison of the time consumed has also been shown in the results section.

The important steps to be considered while implementing this algorithm are shown in Fig. 2. It is a more concise representation of the complete procedure being followed in SAR-AES.

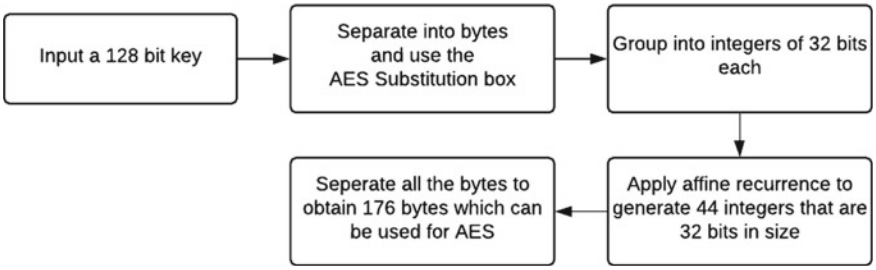
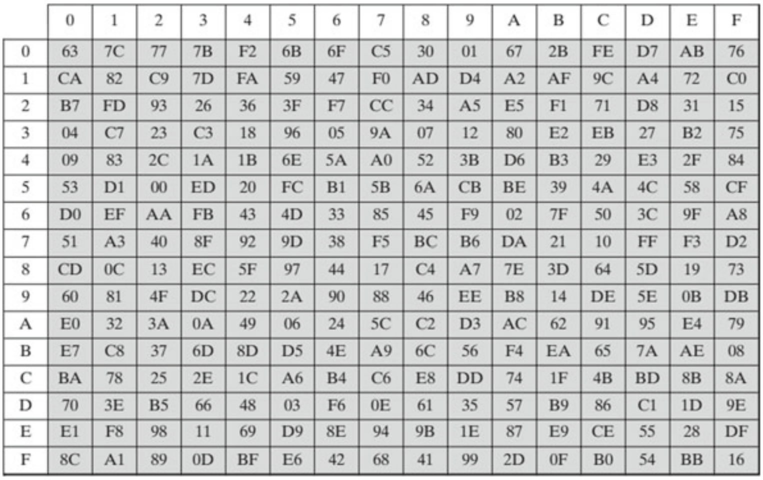

Input a 128 bit key
Separate into bytes and use the AES Substitution box
Group into integers of 32 bits each
Apply affine recurrence to generate 44 integers that are 32 bits in size
Seperate all the bytes to obtain 176 bytes which can be used for AES

Fig. 2 Flow diagram

4.1 AES Substitution Box

Within the modified key schedule algorithm, the first step is using the substitution box. The AES substitution box is designed for creating confusion. It is essentially a 16 × 16 matrix which maps two hexadecimal digits to another two hexadecimal digits [10]. In this algorithm, we give an input of 128 bits in total to the algorithm. This program will divide this into pairs of hexadecimal digits and then apply the AES substitution box on the inputs. This will produce a 128-bit output. This step guarantees high compliance to the Strict Avalanche Criterion and Bit Independence Criterion. The original AES S-box is shown in Fig. 3.



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	63	7C	77	7B	F2	6B	6F	C5	30	01	67	2B	FE	D7	AB	76
1	CA	82	C9	7D	FA	59	47	F0	AD	D4	A2	AF	9C	A4	72	C0
2	B7	FD	93	26	36	3F	F7	CC	34	A5	E5	F1	71	D8	31	15
3	04	C7	23	C3	18	96	05	9A	07	12	80	E2	EB	27	B2	75
4	09	83	2C	1A	1B	6E	5A	A0	52	3B	D6	B3	29	E3	2F	84
5	53	D1	00	ED	20	FC	B1	5B	6A	CB	BE	39	4A	4C	58	CF
6	D0	EF	AA	FB	43	4D	33	85	45	F9	02	7F	50	3C	9F	A8
7	51	A3	40	8F	92	9D	38	F5	BC	B6	DA	21	10	FF	F3	D2
8	CD	0C	13	EC	5F	97	44	17	C4	A7	7E	3D	64	5D	19	73
9	60	81	4F	DC	22	2A	90	88	46	EE	B8	14	DE	5E	0B	DB
A	E0	32	3A	0A	49	06	24	5C	C2	D3	AC	62	91	95	E4	79
B	E7	C8	37	6D	8D	D5	4E	A9	6C	56	F4	EA	65	7A	AE	08
C	BA	78	25	2E	1C	A6	B4	C6	E8	DD	74	1F	4B	BD	8B	8A
D	70	3E	B5	66	48	03	F6	0E	61	35	57	B9	86	C1	1D	9E
E	E1	F8	98	11	69	D9	8E	94	9B	1E	87	E9	CE	55	28	DF
F	8C	A1	89	0D	BF	E6	42	68	41	99	2D	0F	B0	54	BB	16

Fig. 3 AES substitution box

4.2 Affine Recurrence

This step of the key scheduling process involves the execution of the below expression.

$$w_i := (w_{i-8} \oplus w_{i-5} \oplus w_{i-3} \oplus w_{i-1} \oplus \emptyset \oplus i) \ll 11$$

where Ø is the fractional part of the golden ratio $\left(\sqrt{5}+1\right)/2$.

The above seen expression can be converted to its underlying polynomial $x^8 + x^7 + x^5 + x^3 + 1$. This polynomial is known to be primitive [11].

A primitive polynomial is one which is irreducible [15]. This implies no two different outputs of this recurrence will produce results that are related in any way. Hence this leads to removal of both weak keys and related keys. The concept behind a primitive polynomial is that it will create more extensions to a base field. It is also considered to be an irreducible polynomial because it cannot be represented as a multiplication of other polynomials. Looking at this from the perspective of cryptography, it implies that there cannot be any other relation that can be formed using other polynomials to relate the results of the affine recurrence formula. In the way it's used in the key scheduling algorithm each sub-key is an output of affine recurrence. Hence this means that there can't be any possible polynomial to represent the relation between the outputs of the key scheduling algorithm.

In this step, the 128 bits that have been obtained from the S-Box will be considered as pre-keys. This is divided into 4 parts of 32 bits and copied once. So now there are 8 integers that are 32 bits in length. These 8 pre-keys will then be used in the affine recurrence [11]. Each iteration of the affine recurrence will provide us with 8 bits of the expanded key. Next this process will be iterated 44 times to obtain 44 integers that are 32 bits each. This output of 44 integers will count to a total of 176 bytes in size. This 176-byte long integer will be returned as the output of the key scheduling process. This is the expanded key size for the AES-128 created from an original key of 128 bits.

Figure 4 shows a flowchart describing the steps involved in the SAR-AES. It shows how exactly both the AES S-box and Affine Recurrence have been fitted into the key scheduling process. It also shows what needs to be taken care of with respect to the representation of the data, such as representing the sub-keys as 32-bit integers. This is for the ease of computing the steps that are to follow in the procedure.

5 Methodology

The aim was to implement the proposed variation of the AES. This was to be done by creating a file encryption/decryption application with a GUI. The goal was to implement it such that it did not matter what format or type of file was given as input.

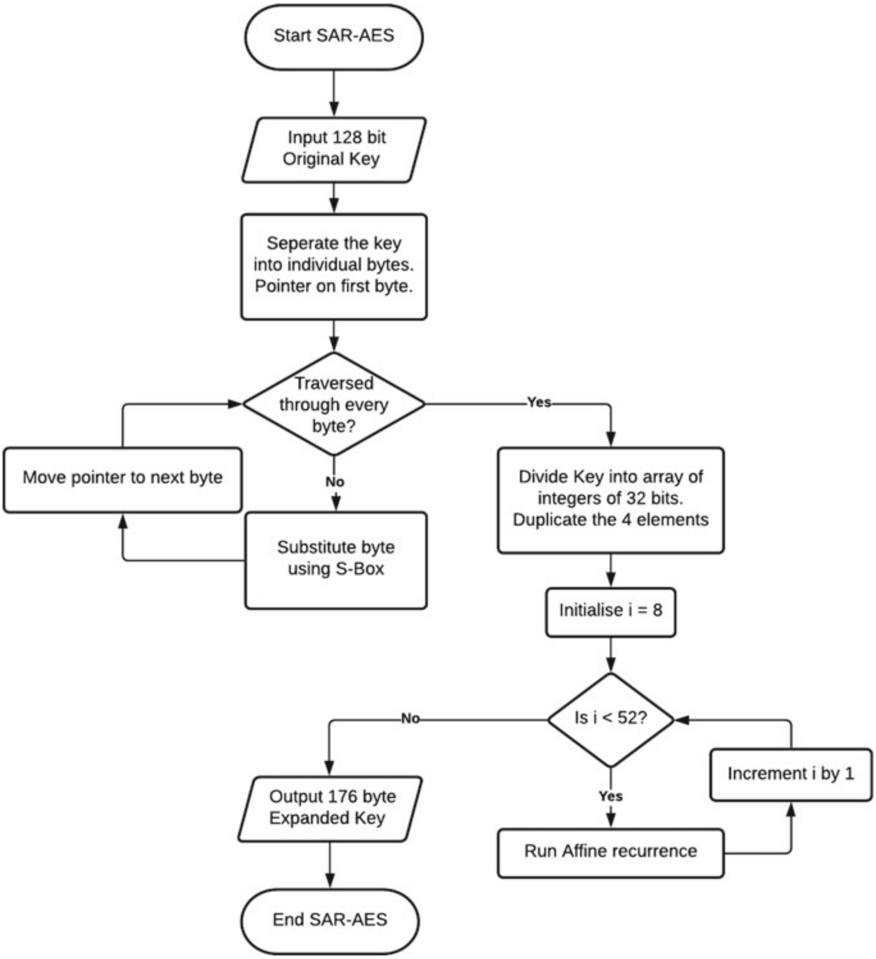

Start SAR-AES
Input 128 bit Original Key
Seperate the key into individual bytes. Pointer on first byte.
Traversed through every byte?
Yes
No
Move pointer to next byte
Substitute byte using S-Box
Divide Key into array of integers of 32 bits. Duplicate the 4 elements
Initialise i = 8
Is i < 52?
No
Yes
Increment i by 1
Run Affine recurrence
Output 176 byte Expanded Key
End SAR-AES

Fig. 4 Flowchart for SAR-AES

Irrespective of which the algorithm should be executed successfully. The proposed modified version of AES was implemented using Python 3.8. The first step in doing so was to implement it just to convert some of the input plain text to cipher text and then to convert it back into the original plain text. After doing this the implementation was tweaked to accept a text file and this process was repeated again. Next finally the implementation was completed by including the encryption and decryption process by using the modified version of AES for all file formats. This was done by taking the file input as a binary file. Then the data was read from the file 16 bytes at a time and the encryption procedure was called. The same process was followed while decrypting the file as well. This new data is put into another file in the same location as the input file. The complete detailed flowchart showing the complete algorithm for SAR-AES has been shown in Fig. 4.

6 Results

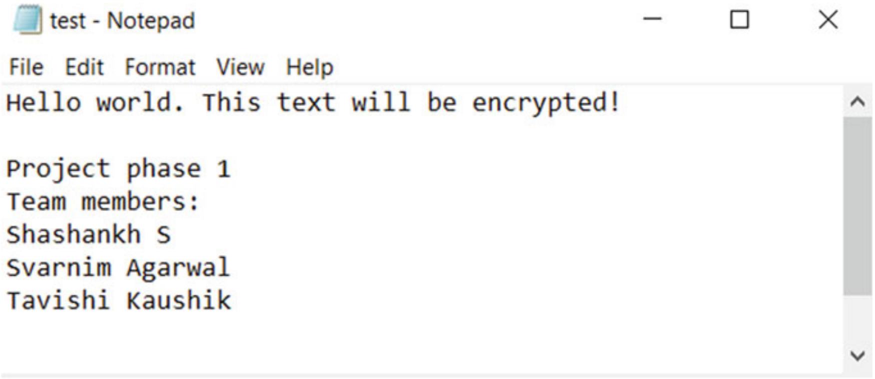
6.1 Sample Execution on Text File

Figure 5 is the image of the input plain text. In the shown example a text file has been used. A text file has been used for the example so that it can be visualized as to what happens when the encryption process is completed. In the encryption process, the whole file is broken down into blocks so it won't matter what the format of the file is. But in the case of the file being an image file or even a PDF, it is to be noted that the intermediate/encrypted form of the file cannot be seen. This is mainly because the software that is being used to view the image file or PDF is unable to decode the encrypted data and hence show it. Hence for this result, we are showing the example in a text file so that even the encrypted version of the file can be visualized.

The text data being used for the encryption is—"Hello world. This text will be encrypted! Project phase 1 Team members: Shashankh S Svarnim Agarwal Tavishi Kaushik".

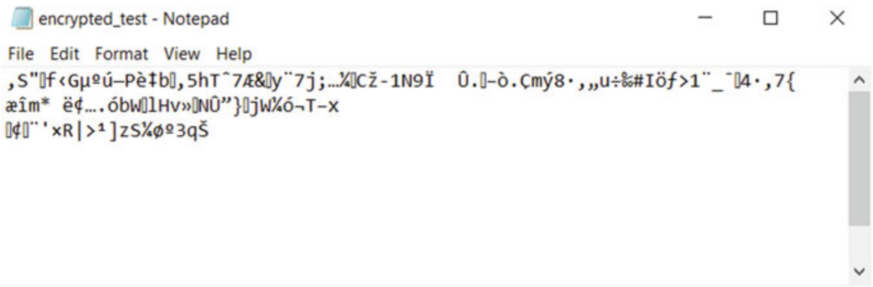
Figure 6 shows the encrypted file. For this example, 'secretkey123' has been used as the secret key for the encryption and decryption process. After encryption the text is converted to a set of random symbols and to make sense of this text, we need to decrypt the file using the same secret key. Figure 7 shows the decrypted file after using 'secretkey123' as the secret key again.

Figure 7 is a screenshot of the text file after decrypting the file from Fig. 6. As seen it is identical to the original text file in Fig. 5. This assures that the encryption and decryption procedures are working with perfect integrity. The same process was



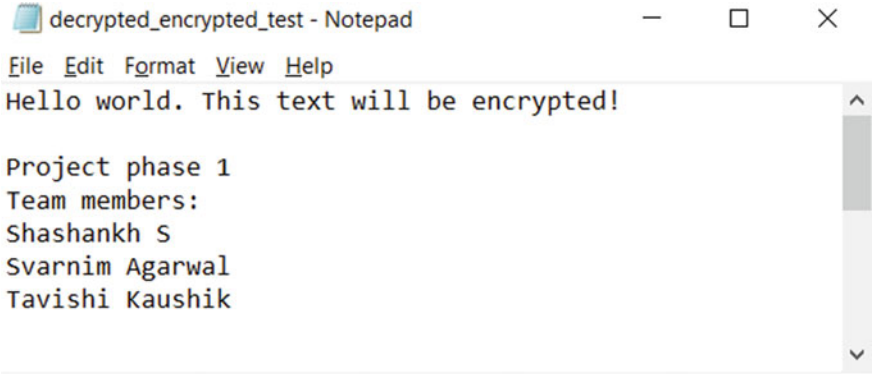
```
Hello world. This text will be encrypted!

Project phase 1
Team members:
Shashankh S
Svarnim Agarwal
Tavishi Kaushik
```

Fig. 5 Sample text file



```
,S"f‹Gµºú–Pè‡b,5hTˆ7Æ&y¨7j;…¼Cž-1N9Ï  Û.–ò.Çmý8·,„u÷‰#Iöƒ>1¨_¯4·,7{
æîm* ë¢….óbWlHv»NÛ”}jW¼ó¬T-x
¢¨'×R|>¹]zS¼øº3qŠ
```

Fig. 6 Encrypted text file



```
Hello world. This text will be encrypted!

Project phase 1
Team members:
Shashankh S
Svarnim Agarwal
Tavishi Kaushik
```

Fig. 7 Decrypted text file

also conducted on image and PDF files. The encryption and decryption process were successful for those file formats as well. The other formats on which this algorithm was tested were image files and PDF files.

6.2 Time Comparison with AES

A comparison of the execution time of the original AES and the proposed variant of AES was conducted. It was noticed that on a varying set of different sized files the amount of time taken for the execution of both the algorithms were similar with negligible differences. When this test was conducted on a 10,000-byte file the original AES took 0.4551 s and the SAR-AES took 0.4168 s for the encryption procedure. More tests that were conducted to compare the execution time of AES and SAR-AES for files of sizes 1000 bytes and 16 KB. The results can be seen both numerically and graphically in Fig. 8.

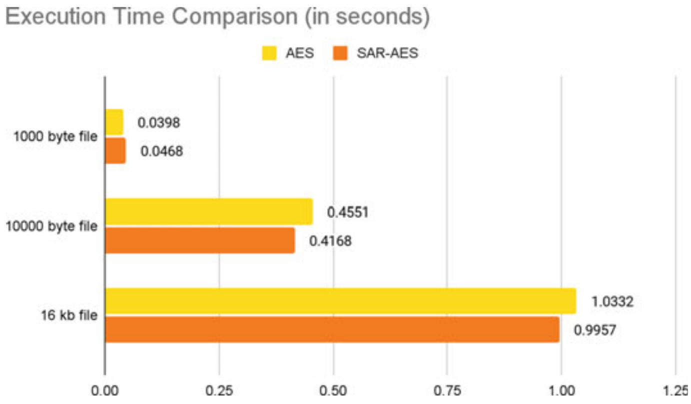


Fig. 8 Execution time comparison between AES and SAR-AES

7 Conclusion

AES is a well-known symmetric encryption algorithm used in a variety of applications. There are however some limitations to this algorithm as well. The limitation being addressed in this paper is associated with how it has a weak key scheduling algorithm. It was noted from existing works that the strict avalanche criterion of the key scheduling algorithm of the AES is inferior in comparison with key scheduling algorithms that are used in ciphers such as Serpent and Twofish [8]. The cause for this was identified to be the methods used in the expansion process. The process lacked the cryptographic concept of confusion [7]. It was also noticed that there were a lot of possibilities for related key attacks in the existing model. This was because using some of the sub-keys generated, an adversary could calculate some of the other sub-keys. This can lead to the possibility of an adversary reverse engineering all the way to get the initial secret key that is being used to encrypt and decrypt the data between the sender and receiver.

The approach in the proposed solution is using two major steps, AES substitution box and Affine Recurrence. The AES substitution box is present in the solution to create the concept of confusion in the algorithm. Affine recurrence is the main and distinguishing step of this proposed model. Affine recurrence is eliminating the risk of related key attacks. This is because it is a primitive polynomial. No two outputs of this mathematical equation can be related to each other. Hence the proposed model is addressing and eliminating the likelihood of related key attacks.

There are potential future works related to SAR-AES. As of now this work is currently proposed, implemented and tested only on AES-128. This is a variation of AES in which the original key size is 128 bits. A possible future work could relate to how it can be scaled to accommodate the key scheduling in AES-192 and AES-256 which are the variations of AES with key sizes 192 and 256 bits.

References

1. Pallavi, G.S., Anantha Narayanan, V.: An overview of practical attacks on BLE based IOT devices and their security. In: 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 694–698 (2019)
2. Arvind, S., Anantha Narayanan, V.: An overview of security in CoAP: attack and analysis. In: 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019. Institute of Electrical and Electronics Engineers Inc., pp. 655–660 (2019)
3. Alkady, Y., Farouk, F., Rizk, R.: Fully homomorphic encryption with AES in cloud computing security. In: International Conference on Advanced Intelligent Systems and Informatics (AISI), vol. 845. Springer, Cham (2018)
4. Shakya, S.: Efficient security and privacy mechanism for block chain application. J. Inf. Technol. **1**(02), 58–67 (2019)
5. Suma, V.: Security and privacy mechanism using blockchain. J. Ubiquitous Comput. Commun. Technol. (UCCT) **1**(01), 45–54 (2019)
6. Sungheetha, A., Sharma, R.: Novel shared key transfer protocol for secure data transmission in distributed wireless networks. J. Trends Comput. Sci. Smart Technol. (TCSST) **2**(02), 98–108 (2020)
7. Rahul Saha, G.G., Kumar, G., Kim, T.: An ımproved version of AES using a new key generation process with random keys. Hindawi Secur. Commun. Netw. **2018**, 9802475 (2018)
8. Afzal, S., Yousaf, M., Afzal, H., Alharbe, N., Mufti, M.R.: Cryptographic strength evaluation of key schedule algorithms. Hindawi Secur. Commun. Netw. **2020**, 3189601 (2020)
9. Vimmadisetti, D., Sharma, T., Bhaskar, A., Kavitha C.R.: Data security on cloud: A survey, research issues and challenges. IJAER Int. J. Appl. Eng. Res. **10**(11), 28875–28887. ISSN 0973-4562 (2015)
10. Daemen J., Rijmen V.: The block cipher Rijndael. In: Quisquater, J.J., Schneier, B. (eds.) Smart Card Research and Applications. CARDIS 1998. Lecture Notes in Computer Science, vol. 1820. Springer, Berlin, Heidelberg (2000)
11. Anderson, R., Biham, E., Lars Knudsen, S.: A Proposal for the Advanced Encryption Standard, Finalist in the Advanced Encryption Standard (AES) Contest and it is in the Public Domain (1998)
12. Farhan, A.A., Kavitha C.R.: End-to-end encryption scheme for IoT devices using two cryptographic symmetric keys. IJCTA Int. J. Control Theor. Appl. **9**(20), 43–49 (2016)
13. Akhil, K.M, Praveen Kumar, M., Pushpa, B.R.: Enhanced cloud data security using AES algorithm. In: International Conference on Intelligent Computing and Control (2017)
14. Parvathy, P., Remya Ajai, A.S.: VLSI ımplementation of blowfish algorithm for secure ımage data transmission. In: International Conference on Communication and Signal Processing (ICCSP) (2020)
15. Weisstein, E.W.: "Primitive Polynomial." From MathWorld—A Wolfram Web Resource

Simplify Your Neural Networks: An Empirical Study on Cross-Project Defect Prediction

Ruchika Malhotra, Abuzar Ahmed Khan, and Amrit Khera

Abstract Ensuring software quality, when every industry depends on software, is of utmost importance. Software bugs can have grave consequences and thus identifying and fixing them becomes imperative for developers. Software defect prediction (SDP) focuses on identifying defect-prone areas so that testing resources can be allocated judiciously. Sourcing historical data is not easy, especially in the case of new software, and this is where cross-project defect prediction (CPDP) comes in. SDP, and specifically CPDP, have both attracted the attention of the research community. Simultaneously, the versatility of neural networks (NN) has pushed researchers to study the applications of NNs to defect prediction. In most research, the architecture of a NN is arrived at through trial-and-error. This requires effort, which can be reduced if there is a general idea about what kind of architecture works well in a particular field. In this paper, we tackle this problem by testing six different NNs on a dataset of twenty software from the PROMISE repository in a strict CPDP setting. We then compare the best architecture to three proposed methods for CPDP, which cover a wide range of scenarios. During our research, we found that the simplest NN with dropout layers (NN-SD) performed the best and was also statistically significantly better than the CPDP methods it was compared with. We used the area under the curve for receiver operating characteristics (AUC-ROC) as the performance metric, and for testing statistical significance, we use the Friedman chi-squared test and the Wilcoxon signed-rank test.

Keywords Neural networks · Machine learning · Software defect prediction · Cross-project defect prediction · Software quality.

R. Malhotra · A. A. Khan (✉) · A. Khera
Delhi Technological University, Main Bawana Road, Delhi 110042, India
e-mail: ruchikamalhotra@dtu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_7

1 Introduction

As the industrial applications of software continue to increase, it is becoming crucial to deliver fault-free software to avoid human and financial losses. Exhaustive testing of software is an effort-intensive task, and due to the limited availability of testing resources, it is imperative to cautiously allocate the resources [1]. Software defect prediction (SDP) is the process of predicting modules containing defects in software based on historical data. It aims to identify the defect-prone areas in software where effort and resources should be expended to deliver fault-free software. SDP helps in carefully managing the limited resources by directing them to the right areas to maintain software quality and testing efficiency. Given the dependency of today's world on software in areas as diverse as medicine and economics, making sure that said software is bug-free also ensures minimal probability of suffering losses. The source of data used to determine the defect-prone areas bifurcates SDP into two forms: within project defect prediction (WPDP), wherein the data is local to the software, and cross-project defect prediction (CPDP), wherein the data is external to the software. It has been observed that WPDP leads to better results than CPDP as it utilizes historical data of the same software [2]. However, the availability of sufficient labeled historical data is scarce, especially for new or in development software [3]. Hence, there has been an increased interest to employ cross-project models for defect prediction [4–6].

Given the versatility of neural networks (NN) in applications of machine learning, they have been used extensively for SDP as well [7–10]. A major question posed to researchers is that of the architecture of the NN to be used. Often, this architecture is obtained on a trial-and-error basis [7, 11]. An insight on the right architecture to be used will not only help in saving the effort expended but might also improve the results obtained. With this in mind, in this paper, we conducted an empirical study between differently sized NNs to analyze their efficiency and efficacy of predicting software defects in the cross-project domain. We conducted the study on six NNs with varying depths and architectures. To the extent of our knowledge, such a study has not yet been conducted, and hence, we aim to bridge this research gap.

We conducted the study on six NNs with varying depths and architectures. The results were obtained over a comprehensive dataset comprising twenty software from the PROMISE repository [12]. For the context of this study, we used area under the curve for receiver operating characteristics (AUC-ROC) as the performance metric to compare the results of various NN architectures and SDP models. For testing the statistical significance of our results, we used the Friedman chi-square test and the Wilcoxon signed-rank test. The study was conducted in two phases, corresponding to the following research questions.

- RQ1: Which NN architecture performs the best for CPDP?

In the first phase, we compared and ranked the six NNs to identify the best-performing architecture. The networks were of two types, three with and three without dropout layers. In each type, the three NNs were of varying depths from shallow and sparse to deep and dense networks. The architectures are further expounded upon in Sect. 3.

- RQ2: How does the best NN architecture compare to existing CPDP methods?

In the second phase, we compared the best-performing network with three methods proposed for CPDP. This is done to verify the validity of the comparison performed in answering RQ1 and to make sure that the networks chosen for said comparison are competent. The selected methods for comparison were chosen to incorporate a wide variety of approaches with Panichella et al.'s combined defect predictor (CODEP) being a combined classifier, Nam et al.'s CLAMI incorporating clustering for unlabelled data, and Ryu et al.'s hybrid instance selection using nearest neighbor (HISNN) utilizing local and global knowledge through a relevancy filter [1, 13, 14]. Each of these methods has produced results that deem them as competent methods to compare with, and as such, performing comparably or better than these methods would amount to the NN being a competent classifier for CPDP.

In this study, we found that small-sized NNs performed better than their larger counterparts, and the NNs with dropout layers performed better than those without. As such, the smallest NN with the dropouts (NN-SD) was the best-performing network. When this was compared with the three methods for CPDP, NN-SD once again achieved the highest rank with statistical significance. This indicates that simpler NNs perform better than complicated ones in general and are a good fit for CPDP.

Further sections in this paper will elaborate on our work. Section 2 covers the literature review and background work. Section 3 explains the research methodology in detail. Section 4 comprises the results of the two phases of the study, and finally, we conclude the paper in Sect. 5.

2 Literate Review

SDP refers to the process of classifying software modules as defective or non-defective based on historical data. Depending on the source of this historical training data, several divisions can be made. Most notably, if the data is sourced from the history of the same project as that which is being tested, SDP, in this case, is termed as WPDP. Likewise, if the data is sourced from software other than (and usually similar to) the one being tested, it is termed as CPDP. Whether the historical data is labeled or not dictates if supervised or unsupervised classification methods are to be used.

It has been observed that WPDP tends to outperform CPDP, as corroborated by Hosseini et al. [5]. This observation also seems intuitive given that one will seldom find historical data of other software representing the defects in a software better than its own historical data. However, several studies have also proven the utility of CPDP, such as increasing the probability of detecting defects while allowing higher false-positive rates [2, 15, 16] In addition, the fact that in the case of new software and pilot versions, historical data is nearly absent, CPDP demands the attention of the research community.

Typically, the SDP pipeline comprises four stages: data acquisition, data preprocessing, training, and testing. Each of these stages has seen massive strides that have streamlined the process. Several archives and repositories that are now publicly available for researchers to use Chidamber and Kemerer have proposed a reliable set of software metrics to test SDP models on, and the last couple of decades have seen tides of SDP methods being proposed which cover a wide variety of scenarios encountered in this field [12, 17, 18]. The first stage, as mentioned before, also serves as a means to divide SDP into two broad categories: WPDP and CPDP. As previously mentioned, due to better representation of the testing data in the training data, WPDP tends to give better results than CPDP, which suffers from the heterogeneity of data among others [2]. However, when dealing with software that has insufficient historical data, WPDP is unable to function well, and a need for CPDP arises [3].

Lately, CPDP has been picking up steam among researchers due to increasing difficulty in obtaining historical data for new software. Hosseini et al. conducted a thorough literature review of CPDP in a five-pronged research tackling all aspects of CPDP [5]. First, they evaluated the independent variables used for training and testing CPDP models. Second, they studied the various modeling techniques used in CPDP, such as naive Bayes and logistic regression. Third, they looked at performance measures used for comparison. Fourth, they looked at a variety of CPDP approaches, such as Ryu et al.'s TCSBoost and Chen et al.'s DTB [19, 20]. And lastly, they compared the performance of CPDP models with WPDP models. As such, in their review, Hosseini et al.'s covered the advances and current capabilities of CPDP as a whole [5]. Taking a further look at the width of CPDP approaches, some are discussed in this section.

Nam et al. proposed a clustering-based method titled CLAMI to work with unlabelled datasets, which was found to work well with several classifiers such as logistic regression and decision trees [13]. Ryu et al. introduced another approach that utilized clustering to utilize both local and global knowledge for CPDP in their model HISNN [1]. Panichella et al. proposed a combined defect prediction classifier called CODEP, which used a conglomeration of six different classifiers [14]. CODEP has since been used in CPDP benchmarking studies several times. He et al., in their work, introduced a method to better select datasets, focusing on distribution similarities between testing and training data, as opposed to the brute force method of finding appropriate datasets [21].

The previously mentioned versatility of NNs has not gone unnoticed by the community either. Several methods have been put forward, including those that utilize transfer learning, cost-sensitive learning, as well as convolutional NNs [7, 9, 20]. Qiao et al. tested a deep learning model and its efficiency at predicting software defects and produced positive results, with their model outperforming state-of-the-art approaches [22]. Zheng proposed three methods employing AdaBoost in combination with a CSNN involving threshold moving and weight updation [10]. Ryu et al.

proposed utilizing a cost-sensitive boosting algorithm along with transfer learning and SMOTE to handle class imbalances [20].

This brings us to the problem statement we are trying to tackle in this paper. Arar et al., much like other authors of papers on NNs, even outside the domain of SDP, mentioned that they arrived at the architecture for their NN through trial-and-error [7, 11]. If there were a general direction that can be taken when designing NNs for research, it can be of help to researchers and save valuable time and effort. In this spirit, we conduct an empirical study to compare six different NNs with varying structures and depths. The NNs are tested on an extensive dataset of 20 software to ensure the greater generalizability of our results. Through the tests, we find that the simplest NN with dropout layers performs the best among the six. These findings are in keeping with Zhou et al.'s conclusion that simpler classifiers tend to work as well as or even better than complex ones. Further, we compared this NN-SD with Panichella et al.'s CODEP, Ryu et al.'s HISNN, and Nam et al.'s CLAMI to find that NN-SD performs statistically significantly better than all three, thus corroborating the conclusion [1, 13, 14].

3 Research Methodology

3.1 A. Datasets and Preprocessing

The study comprises several datasets from the widely used PROMISE repository [12]. The software were chosen after filtering out the datasets containing more than 50% defective modules as proposed by Tantithamthavorn et al. [23]. We also filtered out datasets containing less than 5% defective modules. From the filtered data, twenty software were chosen to carry out the experiments. Table 1 contains the characteristics of the twenty software utilized in this study. The software systems are represented by the metrics suggested by Jureczko and Madeyski [24]. Each module is represented by 20 numerical features, and a boolean variable is used to represent whether the module is defective (represented by 1) or not (represented by 0). A large dataset was utilized to minimize the external threats to validity and increase the generalizability of our results. The datasets were subjected to the following preprocessing:

3.1.1 Normalization

The metrics used to represent the modules have different ranges. These differences in the range of the metrics can adversely affect the prediction results [25]. To prevent this, normalization was applied to the metrics, which transforms the data and reduces the differences in the ranges. In this paper, L2 normalization was applied, which normalizes the metrics by dividing each value of the metric by the sum of squares of all the values of that metric. This transforms the range of the metrics to [0–1].

Table 1 Dataset description

Dataset	Version	# Module	# Defect	Defect %
e-learning	1.0	64	5	7.81
tomcat	6.0	858	77	8.97
ivy	2.0	352	40	11.36
arc	1.0	234	27	11.54
poi	2.0	314	37	11.78
systemdata	1.0	65	9	13.85
xalan	2.4	723	110	15.21
xerces	1.3	453	69	15.23
redaktor	1.0	176	27	15.34
camel	1.6	965	188	19.48
pbeans	2.0	51	10	19.61
serapion	1.0	45	9	20.00
skarbonka	1.0	45	9	20.00
ant	1.7	745	166	22.28
log4j	1.0	135	34	25.19
jedit	4.1	312	79	25.32
termoproject	1.0	42	13	30.95
synapse	1.2	256	86	33.59
velocity	1.6	229	78	34.06
berek	1.0	43	16	37.21

3.1.2 Balancing

SDP often struggles with the issue of imbalanced data since the number of non-defective classes often outweighs the defective ones, as is visible in the Defect% column in Table 1. The imbalance of classes in the data can result in the model becoming biased toward the majority class (non-defective class) [26]. Hence, to mitigate this, we treated the data with the synthetic minority oversampling technique (SMOTE) to balance the distribution. SMOTE is a data augmentation technique which, unlike random oversampling methods, synthesizes an artificial point between a randomly chosen minority point and one of its k-nearest neighbors [27].

3.2 Experiment Overview

In this section, we present an overview of the experiments carried out to answer the two RQs. Figure 1 shows the experiment methodology in detail. The study was conducted on twenty software systems from the PROMISE repository [12]. Nor-

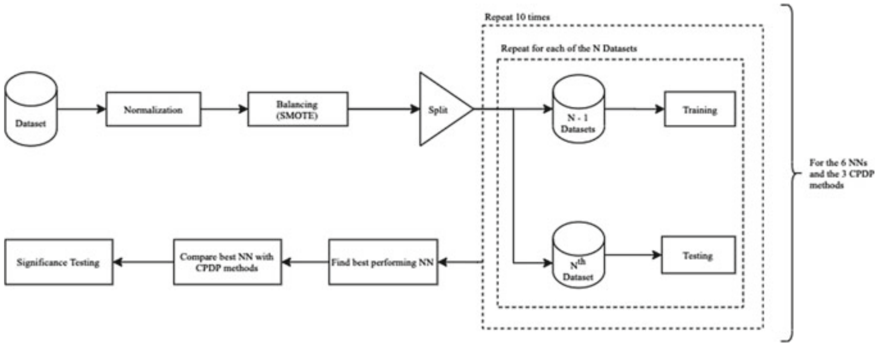


Fig. 1 Experiment methodology

malization and the data balancing technique SMOTE were applied to preprocess the data. Following the convention of strict CPDP, each model was tested on each of the datasets keeping the rest as training data [28]. This means that in the case of N datasets, $N - 1$ are used for training, and the model was then tested on the Nth dataset. This was repeated N times so that each of the N datasets is tested upon once. The experiments were conducted ten times on the preprocessed data to obtain robust results, and the mean AUC-ROC values were reported. On the basis of the mean AUC-ROC scores, the best-performing NN was identified. The best-performing NN (NN-SD) was then compared with the CPDP methods on the basis of the mean AUC-ROC scores. We performed statistical tests to validate the results. For this purpose, we employed the Friedman chi-square Test and Friedman mean ranks. Further, for the comparison of the best architecture with the CPDP methods, we also employed the Wilcoxon signed-rank test. The study was conducted in the following two phases:

3.2.1 Comparison of NN Models

To answer RQ1, we compared six NNs with different architectures. The six NNs comprised the following two types of layers:

- **Dense Layer**: This layer is the basic layer that builds up the NN. It simply consolidates all the outputs from the previously connected layers in each of its units, applies the activation function on it, and passes the output to each unit of the succeeding layer.
- **Dropout Layer**: NNs may suffer from overfitting due to the densely connected layers and a large number of weights between them. Dropouts help minimize the overfitting by shutting off a fraction of units (dropout percentage) in a layer. That is, temporarily, those fractions of units are dropped out of the network. Dropouts are considered a good method to reduce overfitting in NNs [29].

  Out of the six NNs, three comprised only the dense layers. These three were the small (NN-S), the medium (NN-M), and the large (NN-L)-sized NNs without

Table 2 Model architectures

Models	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
NN-S	20	16	4	1	–	–	–	–	–	–	–
NN-M	20	64	32	16	4	1	–	–	–	–	–
NN-L	20	256	128	64	32	16	4	1	–	–	–
NN-SD	20	16	D(0.5)	4	1	–	–	–	–	–	–
NN-MD	20	64	D(0.5)	32	D(0.5)	16	4	1	–	–	–
NN-LD	20	256	D(0.5)	128	D(0.5)	64	D(0.5)	32	16	4	1

dropouts. Whereas the remaining three comprised both dense and dropout layers. These three were the small (NN-SD), the medium (NN-MD), and the large (NN-LD)-sized NNs with dropouts. Table 2 summarizes the six NN architectures based on the ordering of the said layers. The values in the cells represent the number of units in the respective layers. The dropout layers have been represented as $D(x)$, where 'x' signifies the dropout percentage. The cells of the output layers have been bolded. The leaky ReLU activation function is employed in each of the intermediate layers. That is, the output of these layers is passed through the leaky ReLU activation function before being passed as input to the subsequent layers. Similarly, the Sigmoid activation function is employed for the output layer to transform the output to the [0–1] range. From the table, it is clear that NN-S is less dense and smaller than NN-M, which is in turn less dense and smaller than NN-L. Similarly, for the NNs with dropouts, NN-SD is less dense and smaller than NN-MD, which is less dense and smaller than NN-LD. As mentioned, the NNs were tested in keeping with strict CPDP conventions, a total of ten times to reduce the effects of stochastic elements, and the AUC-ROC values were averaged over the ten runs before comparison. Based on the experimental results, NN-SD was found to be the best-performing model.

3.2.2 NN-SD Versus CPDP Methods

To evaluate the performance of the best NN architecture, we compare it against Panichella et al.'s CODEPBayes, referred to as simply CODEP, Nam et al.'s CLAMI in conjunction with logistic regression (LR), and Ryu et al.'s HISNN [1, 13, 14]. CODEP is a good example of ensemble classifiers being employed for CPDP, CLAMI represents the use of unsupervised learning toward CPDP, and HISNN is a method that employs a combination of both local and global knowledge. These methods were hence chosen to incorporate the diversity of CPDP methods and test the quality of the results obtained in answering RQ1. NN-SD was found to outperform each of these methods with statistical significance, which was tested using a combination of Friedman chi-square and Wilcoxon signed-rank tests.

4 Experiment Results

4.1 *RQ1: Which NN Architecture Performs the Best for CPDP?*

Table 3 shows the mean AUC-ROC performance of the six NN models obtained over the ten runs. The best results for each dataset have been marked in boldface. It can be seen that NN-SD performs the best out of all the architectures achieving the best result on eight of the datasets. The p-value obtained from the Friedman chi-square test is 7.52×10^{-5}, which makes the difference significant. The Friedman mean ranks are presented in Table 4, which gives us two levels of insights into the performance of the six NNs. First, looking at the two groups of NNs with and without dropouts, we observe that the smallest NNs perform the best, followed by the medium-sized, and finally the large ones. Second, when comparing NNs of each size with their counterparts, i.e., NN-S with NN-SD, and so on, we observe that the NNs with dropouts consistently perform better than the NNs without dropouts. In summary, it can be seen that shallow NNs perform better than deeper ones, and NNs with

Table 3 AUC-ROC values (NN Models)

Dataset	NN-S	NN-M	NN-L	NN-SD	NN-MD	NN-LD
e-learning-1.0	0.7207	0.7471	**0.8169**	0.7892	0.7119	0.7098
tomcat-6.0	0.8184	0.8190	0.8129	0.8241	**0.8252**	0.8218
ivy-2.0	**0.8257**	0.7720	0.7736	0.8252	0.8210	0.8100
arc-1.0	0.6951	0.6858	0.6815	0.7149	**0.7234**	0.7128
poi-2.0	0.7295	0.7271	0.7110	0.7403	**0.7456**	0.7401
systemdata-1.0	**0.8258**	0.8036	0.7639	0.7937	0.7956	0.8123
xalan-2.4	0.7423	0.7032	0.6801	**0.7533**	0.7453	0.6862
xerces-1.3	0.7465	0.7084	0.6984	**0.7513**	0.7354	0.7222
redaktor-1.0	0.4665	0.4201	0.3485	**0.4800**	0.4734	0.4728
camel-1.6	0.5972	0.5991	0.6072	0.6004	0.5988	**0.6090**
pbeans-2.0	0.7600	0.7439	0.7195	**0.7663**	0.7439	0.7176
serapion-1.0	0.7599	0.7543	0.7444	**0.7617**	0.7519	0.7599
skarbonka-1.0	0.7716	0.7623	0.7284	**0.7827**	0.7741	0.7691
ant-1.7	0.8074	0.7888	0.7889	**0.8118**	0.8059	0.7964
log4j-1.0	0.7880	0.7858	0.7691	0.7803	0.7924	**0.8008**
jedit-4.1	**0.8184**	0.7982	0.7718	0.8121	0.8120	0.8122
termoproject-1.0	0.8934	0.9056	0.8695	0.8960	**0.9093**	0.9003
synapse-1.2	0.7391	0.6856	0.6254	**0.7416**	0.7377	0.7278
velocity-1.6	0.7020	0.6957	0.6677	0.6985	0.6888	**0.7042**
berek-1.0	0.9907	**0.9926**	0.9889	0.9898	0.9912	0.9875

Table 4 Friedman mean ranks (NN models)

Model	Mean rank	Relative rank
NN-SD	4.75	1
NN-MD	4.1	2
NN-S	4.05	3
NN-LD	3.6	4
NN-M	2.85	5
NN-L	1.65	6

dropouts perform better than those without them. Hence, it is natural that NN-SD, the shallowest network with dropouts, performs the best and achieves the best mean rank.

4.2 RQ2: How Does the Best NN Architecture Compare to Existing CPDP Methods?

Table 5 contains the AUC-ROC scores of NN-SD compared with the chosen CPDP methods. The boldface signifies the best value obtained for each dataset. We can see that NN-SD achieves the maximum AUC-ROC score on seventeen of the datasets. Table 6 shows the Friedman means ranks, and the 1.76×10^{-7} p-value of the Friedman chi-square test corroborates that the difference is significant. We can see that NN-SD achieves the best mean rank of 3.75, making it the better method for CPDP. Further, post-hoc analysis results are carried out via the Wilcoxon signed-rank test, and the results are present in Table 7. From the table, it is evident that NN-SD performs significantly better than the CPDP methods it is compared to.

The boxplot in Fig. 2 further corroborates the superiority of NN-SD when compared to the CPDP methods in questions while also validating the competency of the neural networks in comparison with RQ1. It is clear that both measures of central tendency, the median and the mean (represented by the ’X’), are higher for NN-SD compared to the other three. Further, the minimum, first quartile, third quartile, and maximum are also all higher for NN-SD than the remaining three.

5 Threats to Validity

This section describes three kinds of threats to validity with respect to the study conducted in this paper.

First, internal threats to validity, which cover the errors present in the experiments conducted. The experiments were conducted with the utmost care by utilizing reliable

Table 5 AUC-ROC values (NN-SD versus CPDP methods

Dataset	NN-SD	CODEP	CLAMI	HISNN
e-learning-1.0	0.7892	0.5000	**0.9051**	0.4831
tomcat-6.0	**0.8241**	0.7176	0.7174	0.6741
ivy-2.0	**0.8252**	0.7574	0.7313	0.6189
arc-1.0	**0.7149**	0.5805	0.6611	0.5821
poi-2.0	**0.7403**	0.5648	0.6066	0.6377
systemdata-1.0	**0.7937**	0.6022	0.7312	0.6042
xalan-2.4	**0.7533**	0.6639	0.7108	0.5907
xerces-1.3	**0.7513**	0.6151	0.6842	0.5988
redaktor-1.0	0.4800	0.5523	0.5121	**0.5781**
camel-1.6	0.6004	0.5227	**0.6213**	0.5413
pbeans-2.0	**0.7663**	0.6378	0.7341	0.7012
serapion-1.0	**0.7617**	0.6111	0.6111	0.5972
skarbonka-1.0	**0.7827**	0.5139	0.7778	0.4167
ant-1.7	**0.8118**	0.6836	0.6964	0.6408
log4j-1.0	**0.7803**	0.6372	0.7148	0.6169
jedit-4.1	**0.8121**	0.7374	0.7030	0.6183
termoproject-1.0	**0.8960**	0.6538	0.8011	0.4655
synapse-1.2	**0.7416**	0.6796	0.6630	0.5386
velocity-1.6	**0.6985**	0.5442	0.6852	0.6223
berek-1.0	**0.9898**	0.9063	0.8519	0.7141

Table 6 Friedman mean ranks (NN-SD versus CPDP methods)

Model	Mean rank	Relative rank
NN-SD	3.75	1
CLAMI	2.7	2
CODEP	2	3
HISNN	1.5	4

Table 7 Wilcoxon signed-rank test (NN-SD versus CPDP methods)

Classifier	*p*-value	#*p*-value
CODEP	0.00014	0.00028
CLAMI	0.00249	0.00499
HISNN	0.00016	0.00032

#*p*-value denotes the Bonferroni corrected *p*-value for 2 tests

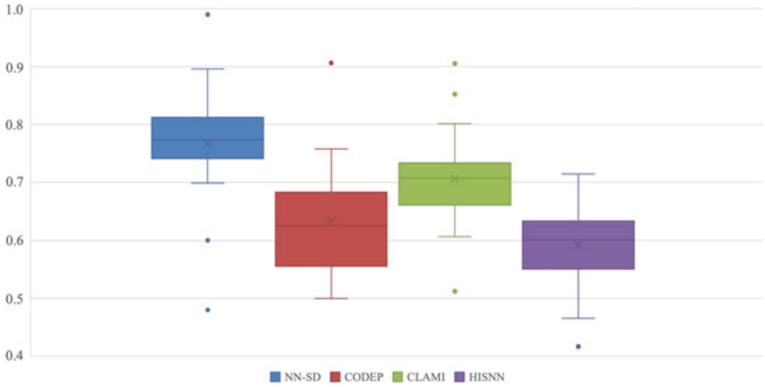


Fig. 2 NN-SD versus CPDP methods

and standard tools and software packages. In spite of this, there may still be some errors that were not identified. The reliable Keras and scikit-learn python packages were utilized for data preprocessing, implementation, and comparison.

Second, external threats to validity, which cover the generalizability of the study and its results. To increase the generalizability of the results, we analyzed the experiments on a comprehensive dataset comprising 20 software containing a total of 6107 instances from the PROMISE repository. However, this poses a threat as the results may reflect the characteristics and biases of the dataset. It is natural that the requirements for each instance of SDP will vary and so will the architecture of the network which fits the case best. Nevertheless, this study provides a reliable start-line for researchers to build their network from. In the future, this threat can be further minimized by analyzing the experiments on a more diverse dataset sourced from several repositories.

Third, threats to construct validity, which refer to the quality of datasets and the performance metrics used to draw conclusions in the study. The dataset has been sourced from the reliable PROMISE repository. This repository is widely trusted and has been adopted by several researchers in the defect prediction domain [14, 20, 26]. Nevertheless, there could still be some flaws due to the choice of the dataset. The widespread metric, AUC-ROC, was utilized as the only performance metric. In the future, the threat due to the use of a single metric can be minimized by utilizing several performance metrics in conjunction to solidify the conclusion.

6 Conclusion

Maintaining software quality is becoming ever more difficult with an increase in size and complexity of software. Moreover, releasing software containing bugs may cause system breaks and lead to major losses. Hence, it is imperative to promptly identify and fix defects in software. The purpose of SDP is to streamline the process of software testing to identify defects. It aids in the judicious allocation of the limited testing resources toward the defect-prone areas. In this dynamic industry, there are many instances when projects lack sufficient historical data for defect prediction. In such cases, CPDP becomes the only viable option. Over the years, there has been increased interest in the applications of NNs for defect prediction. However, there is still little clarity over the right architectures to be used for SDP. Hence, in this paper, we try to bridge this gap by analyzing different NN architectures for CPDP.

In this paper, we compare and analyze six NN architectures to identify the characteristics of a NN that is viable for CPDP with datasets with class imbalance. To make the results more generalizable, we performed the experiments on twenty datasets from the PROMISE repository. We performed two steps of data preprocessing, namely normalization and balancing, using L2 normalization and SMOTE, respectively. For increased reliability of results, we performed the experiments ten times and utilized the mean AUC-ROC scores for the comparison. Further, to verify the competency of the best-performing NN model, we compared it with a diverse group of CPDP methods, namely CODEP, CLAMI, and HISNN. The results indicate that shorter networks outperform denser ones and that networks containing dropouts outperform those without them. In line with this, NN-SD is found to be the best-performing model. Moreover, NN-SD performs better than the CPDP methods it was compared to, thus verifying the quality of the results produced. The results were validated by the Friedman chi-square test and Wilcoxon signed-rank test and were found to be statistically significant.

In the future, we plan to extend the study by comparing a larger number of different architectures. We further plan to analyze the results on a more diverse dataset sourced from several reliable repositories. Finally, we also plan to examine the effect of feature selection techniques on the performance of different network architectures.

References

1. Ryu, D., Jang, J.I., Baik, J.: A hybrid instance selection using nearest-neighbor for cross-project defect prediction. J. Comput. Sci. Technol. **30**(5), 969–980 (2015)
2. Turhan, B., et al.: On the relative value of cross-company and within-company data for defect prediction. Empir. Softw. Eng. **14**(5), 540–578 (2009)
3. Zimmermann, T., et al.: Cross-project defect prediction: a large scale experiment on data versus domain versus process. In: Proceedings of the 7th Joint Meeting of the European Software Engineering Conference and the ACM Sigsoft Symposium on the Foundations of Software Engineering, 2009, pp. 91–100
4. Herbold, S., Trautsch, A., Grabowski, J.: A comparative study to benchmark cross-project defect prediction approaches. IEEE Trans. Softw. Eng. **44**(9), 811–833 (2017)
5. Hosseini, S., Turhan, B., Gunarathna, D.: A systematic literature review and meta-analysis on cross project defect pre- diction. IEEE Trans. Software Eng. **45**(2), 111–147 (2017)
6. Menzies, T., et al.: Defect prediction from static code features: current results, limitations, new approaches. Autom. Softw. Eng. **17**(4), 375-407 (2010)
7. Arar, Ö.F., Ayan, K.: Software defect prediction using cost-sensitive neural network. Appl. Soft Comput. Software defect prediction using cost-sensitive neural network. **33**, 263–277 (2015)

8. Jindal, R., Malhotra, R., Jain, A.: Software defect prediction using neural networks. In: Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization, pp. 1–6. IEEE (2014)
9. Liu, M., Miao, L., Zhang, D.: Two-stage cost-sensitive learning for software defect prediction. IEEE Trans. Reliab. **63**(2), 676–686 (2014)
10. Zheng, J.: Cost-sensitive boosting neural networks for software defect prediction. Expert Syst. Appl. **37**(6), 4537–4543 (2010)
11. Wang, S.: Training deep neural networks on imbalanced data sets. Int. Joint Conf. Neural Netw. (IJCNN). **2016**, 4368–4374 (2016)
12. Menzies, T., et al.: The promise repository of empirical software engineering data. West Virginia University, Department of Computer Science (2012)
13. Nam , J., Kim, S.: Clami: Defect prediction on unlabeled datasets (t). In: 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 452–463. IEEE (2015)
14. Panichella, A., Oliveto, R., Lucia, A.D.: Cross-project defect prediction models: L'union fait la force. In: 2014 Software Evolution Week-IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering (CSMR-WCRE), pp. 164–173. IEEE (2014)
15. He, Z., et al.: An investigation on the feasibility of cross-project defect prediction. Autom. Softw. Eng. **19**(2), 167–199 (2012)
16. Kocaguneli, E., et al.: When to use data from other projects for effort estimation. In: Proceedings of the IEEE/ACM International Conference on Automated Software Engineering, pp. 321–324 (2010)
17. Chidamber, S.R., Kemerer, C.F.: A metrics suite for object oriented design. IEEE Trans. Softw. Eng. **20**(6), 476–493 (1994)
18. Wu, R., et al.: Relink: recovering links between bugs and changes. In: Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering, 2011, pp. 15–25
19. Chen, L., et al.: Negative samples reduction in cross-company software defects prediction. Inf. Softw. Technol. **62**, 67–77 (2015)
20. Ryu, D., Jang, J.-I., Baik, J.: A transfer cost-sensitive boosting approach for cross-project defect prediction. Software Qual. J. **25**(1), 235–272 (2017)
21. He, Z.: Learning from open-source projects: an empirical study on defect prediction. In: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 45–54. IEEE (2013)
22. Qiao, L., et al.: Deep learning based software defect prediction. Neurocomputing **385**, 100–110 (2020)
23. Tantithamthavorn, C., et al. Automated parameter optimization of classification techniques for defect prediction models. In: Proceedings of the 38th International Conference on Software Engineering, pp. 321–332 (2016)
24. Jureczko, M., Madeyski, L.: Towards identifying software project clusters with regard to defect prediction. In: Proceedings of the 6th International Conference on Predictive Models in Software Engineering., pp. 1–10 (2010)
25. Cruz, A.E.C., Ochimizu, K.: Towards logistic regression models for predicting fault-prone code across software projects. In: 3rd International Symposium on Empirical Software Engineering and Measurement, pp. 460–463. IEEE (2009)
26. Bennin, K.E., et al.: Mahakil: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. IEEE Trans. Softw. Eng. **44**(6), pp. 534–550 (2017)
27. Chawla, N.V., et al.: SMOTE: synthetic minority over-sampling technique. J. Artif Intell Res **16**, 321–357 (2002)
28. Qiu , S., et al.: An investigation of imbalanced ensemble learning methods for cross-project defect prediction. Int. J. Pattern Recogn. Artif. Intell. **33**(12), 1959037 (2019)
29. Goodfellow, I., et al.: Deep Learning, vol. 1.2. MIT Press Cambridge (2016)

Emotion Recognition During Social Interactions Using Peripheral Physiological Signals

Priyansh Gupta, S. Ashwin Balaji, Sambhav Jain, and R. K. Yadav

Abstract This research aims to present a method for emotion recognition using the K-Emocon dataset (Park et al. in Sci Data 7(1):1–16 [8]) for use in the healthcare sector as well as to enhance computer–human interaction. In the following work, we use peripheral physiological signals to recognize emotion using classifier models with multidimensional emotion space models. These signals are collected using IoT-based wireless wearable devices. Emotions are measured in terms of arousal and valence by using physiological signals obtained from these devices. Several machine learning models were used for emotion recognition. Thirty-eight input features were extracted from a variety of physiological signals present in the dataset for analysis. Best accuracy achieved for valence and arousal in our experiment was 91.12% and 62.19%, respectively. This study targets recognition and classification of emotions during naturalistic conversations between people using peripheral physiological signals. It is shown that it is viable to recognize emotions using these signals.

Keywords K-Emocon · Emotion recognition · Machine learning · Neural network · IoT · Wearable sensors

1 Introduction

Emotions play an important role in the way we perceive things and also in the way we behave. They can influence the decisions we make in our everyday life to a very large extent. Emotions have three components: subjective (how we experience emotion), physiological (how our body reacts to it) and expressive component (how we behave). A person may be exposed to the same emotion through different experiences, and some people may be weak with their expressiveness for emotion but we can observe a general trend of physiological signals with emotions. Over the past two decades, there has been a large increase in the studies of emotion processing in the area of affective computing. Emotions recognition can help us with a better human–computer

P. Gupta (✉) · S. A. Balaji · S. Jain · R. K. Yadav
Delhi Technological University, Main Bawana Road, Delhi 110042, India
e-mail: rkyadav@dtu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_8

interaction as the computer would be more sensible in their interaction as per the emotions of the human which promotes the user experience. Apple, for example, has been using emotion recognition AI to enhance the emotion intelligence of its products. Other than this, it could help patients who cannot express feelings like autism spectrum disorder patients and similar such applications for technologies are rapidly growing in the medical sector. Screening for conditions such as dementia, sleep apnoea and epilepsy has previously been tested for variability in heart rate, movement and light exposure. Emotions are reflected in our words, in our voice, body language, facial expressions, acoustic characteristics, physiological signals, etc. While other factors can be faked/bluffed, it is extremely hard to control physiological signals of our body. Companies such as Emotient, Snapchat have been utilizing emotions to predict attitudes and future actions of people. There has been much research for emotion recognition using facial expressions but because of its unreliable nature. For example, even when in a negative emotional state, people can smile on a formal social occasion. Hence, the focus has been shifted to emotion recognition using physiological signals as these are controlled by the nervous system. All methods have their own accuracy and limitation to detect emotion.

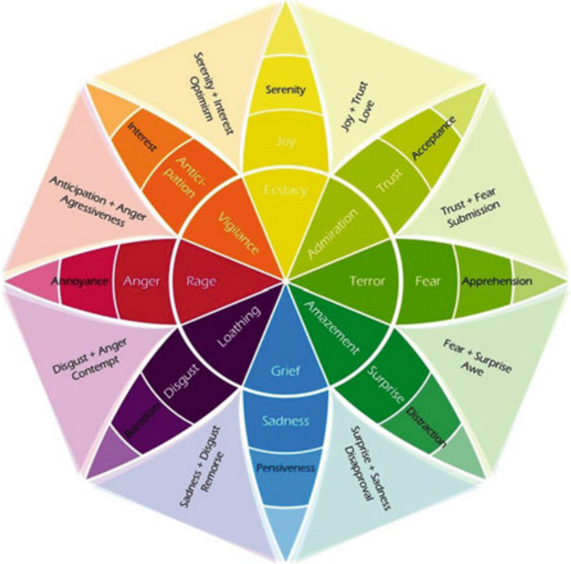
2 Theory of Emotion Recognition

Common physiological signals such as electroencephalography (EEG), heart rate variance (HRV), electrodermal activity (EDA), respiration (RSP), and skin temperature (SKT) can be used for recognizing emotions. Nowadays, IoT technology makes physiological data available along with activity data. In order to track their health, such as heart rate, blood pressure, number of calories burned and evaluate their movements, people are interested in buying connected items which are generally connected to their phones. We can find a lot of work for healthcare applications, but at the same time, this technology can be used to identify emotions that take into account people's emotional state (stress, happiness, sorrow, anger) and evaluate that using the appropriate AI tools to detect emotions, categorize them and then analyse their effect on heart disease. These physiological signals can be measured using devices like wristbands, chest patches, fitbands, and smartwatches. As compared to the clinical setup, these devices are less accurate and efficient, but at the same time provides us advantage of being powerful, comfortable, cheaper, easily available and helps us get out of the clinical environment as using smart wearables, knowledge of activity and psychology can be obtained in a passive way, without disturbing the user's everyday life. Each iteration of these devices is better than their previous version. Today, wristbands have the ability to measure PPG/BVP, EDA, GSR, gyrometer, accelerometer and SaO2 and future devices seem to provide us everything we can get from a clinical setup using large machines. These devices are already being used for monitoring everyday sleep and activities of the people, and their results seem promising for everyday use. The data obtained by these devices can be transferred wirelessly to our mobiles and computers in real time for further computations and fetching the final result. A few devices available for the purpose are Apple, Empatica E4, Microsoft Band 2, Garmin Venu Sq, Samsung Galaxy Fit-2, Polar Grit X, Muse Cue, fitBit Charge 4, OuraRing and VitalPatch. Of these devices, Empatica E4 is widely used for research purposes while Apple watches are widely used for activity monitoring purposes [12].

According to a survey in 2014, India has the highest number of smartwatch users globally and its market is growing rapidly [11].

For classification of emotions, two types of models are used, discrete emotion model and multidimensional emotion space model. In the discrete emotion model, we consider a small number of core emotions and consider that emotions at all times faced by a person can be classified into one of these categories. For this model, we have Ekman's basic emotions [4] and also Plutchik's wheel of emotions (Fig. 1) [9] which defines a set of emotions for the classification. In the multidimensional emotion space model, emotions are associated with a collection of dimensions. The model is defined in a space of two dimensions (valence and arousal) as shown in Fig. 2 or three dimensions (valence, arousal and dominance). The arousal represents the excitement level, the valence represents the positive and negative emotion and the control that a person has over emotion is represented by dominance. On the basis of the values of physiological state in the two or three dimensions, we can decide upon the emotions [14]. These dimensions of emotions can capture definitions of subtle emotions that vary little from the general categories of emotions. A dimensional emotional model represents a useful representation that captures all related emotions and provides a way to measure emotional similarity. Thus, we can say emotion is amalgamation of valence and arousal.

Fig. 1 Plutchik's wheel of emotion



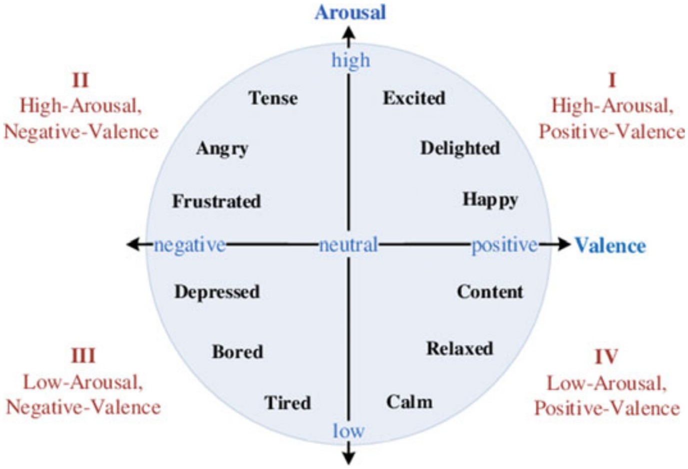


Fig. 2 2D Emotion space model

The classification systems themselves can be either user-dependent or independent. User-based systems require specific calibration for each user before triage, while standalone systems can identify feelings of unfamiliar users without individual configuration. User-dependent systems with 6–8 people tend to achieve greater accuracy in those subjects but lack an overall accuracy for a generic audience. Having a larger audience is admirable for a user-independent system, and this approach generally achieves less precision and requires more complex algorithms to accurately interpret emotions. So, it is a trade-off situation with no one solution that fits all problems, we can either have a mode that is more accurate to a select audience or a model that works better on all types of audiences but is less accurate overall.

This research presents an evaluation of physiological signals for emotion classification using the K-Emocon dataset which is one of the most recently published datasets in this field. The recognition rate depends upon the features used, the classifier or emotion model in consideration and also on the dataset itself. In the datasets that were published before K-Emocon, emotions were evoked using specifically selected pictures, videos, audios or even combinations of both. This method has a disadvantage in that different people are evoked differently to the same resource in both sign and intensity. K-Emocon proposed a more reliable method of collecting a dataset where they simulated natural conversation between participants and recorded emotions and physiological signals during the conversation. Five machine learning models are compared in the upcoming research work which includes K-nearest neighbours (KNN), Gaussian naive Bayes (GNB), decision tree (DT), support vector machine (SVM) and a custom deep learning-based neural network classifier. Using the following models on physiological signals, valence and arousal are classified into five categories and then their results are compared to get the accuracy and F1 scores. The results obtained here will act as a baseline for further research work.

3 Previous Work

Emotions are very important in our daily activities. Over the past few decades, there has been significant research for emotion classification using facial expressions and physiological signals like HRA, skin temperature, and EDA. The physiological signals reflect the status of our nervous system and hence carry information about our emotion states of our inner body. Among these stated signals, the HRA, GSR and EEG turn out to be the most useful for the classification as they have shown strong variation with the emotions.

In [10], the researchers used a decision tree for classification of emotions. The features used are EDA, HRV and SKT for which they achieved an accuracy of 65%. Reference [3] used multiple classification models including linear discriminant analysis, quadratic discriminant analysis and decision tree using HRA and GSR and achieved an accuracy of 70%. The authors of [5], using values of EMG, ECG, respiration and EDA on SVM and ANFIS models, achieved an accuracy of 79.3% and 76.7% for the two models, respectively. Reference [15] achieved an average accuracy of 87.3 for bi-classification using the DKMO model in deep learning on EEG, EMG, GSR, MEG, EOG and RES signals. Reference [1] on the DEAP dataset, achieved best accuracy of 73.08% for arousal and 72.18% accuracy for valence using random forest, SVM and LR on signals like RES, PPG and SKT. An accuracy of 57% arousal accuracy and 62.7% valence accuracy using decision fusion model on EEG signals was achieved in [6]. Reference [2] achieved an accuracy of 66 and 75% for SVM and RSVM classification models using HRV, SKT, BVP and EEG signals.

4 Our Methodology

4.1 Dataset-K-Emocon

K-EmoCon is a multimodal dataset containing different types of emotion annotations taken in a continuous manner [8]. It contains emotional annotations of the three available viewpoints, distinct from previous datasets: the self, the arguing partner and the outside observer. This dataset includes multimodal measurements, including audiovisual images, electroencephalograms and physiological peripheral signals, obtained from 16 sessions, including a couple of approx. ten minutes of exchanges on a social topic. The annotators noted emotional manifestations every 5 seconds in terms of arousal value and 18 more categorical emotions while watching the debate video. K-EmoCon is the first freely accessible emotional dataset to host more emotional evidence during social experiences. The participants selected for the process were in the age group of 19 to 36 and Asian. The dataset aims to provide an additional aspect of external reviewers and the opposite participant's perspective on emotion to improve the classification. The debate was such that each participant spoke after one another in a not completely structured fashion. The topic selected for debate was Yemeni refugees on Jeju Island, South Korea. The participants were seated in a well-lit room and were made to wear a Polar H7 Heart Rate sensor to detect ECG biosignals.Empatica E4 wristband in order to measure triaxial acceleration, heart rate indirectly via the PPG signals, galvanic skin response and temperature. Total of 172.92 min of data was obtained. The five external raters selected for rating participants' emotions were in the age group of 22–27 years with 3 male and 2 female. The data obtained was preprocessed by synchronizing it time-wise. The EEG sensors are negatively influenced by noises, and errors are also possible in other devices. There are four different types of emotion annotations (target values) in the K-Emocon dataset.

1. Self-Annotations—These are the annotations done by the participants themselves through self-evaluation of emotions.
2. Partner Annotations—In this method, the person in conversation with the first participant is evaluating and annotating their emotions.
3. Self-Partner Mean Annotations—This is the mean of the first two annotations.
4. Aggregate External Annotations—Three external observers observing the 2 talking participants have also evaluated their emotions and then their evaluations have been combined into a single type of annotation.

4.2 Physiological Signals Used and Feature Extraction

Selecting optimal input parameters is a crucial part of the machine learning process. For our analysis of emotions, these parameters are the physiological signals that are exhibited by the body. In this paper, we have used all the physiological signals present in the K-Emocon dataset that is captured by the Empatica E4 and the Polar H7 Heart Rate Sensor, namely blood volume pressure (BVP) (64 Hz), heart rate (HR) (1 Hz), skin temperature (SKT) (4 Hz), electrocardiograph (ECG) (1 Hz), electrodermal activity (EDA) or galvanic skin response (GSR) (4 Hz) and triaxial acceleration (32 Hz). Along with optimal input parameter selection, proper feature extraction is also extremely crucial for a fair and extensive analysis of the problem. For feature extraction, we have used the open-source library called PyTeap which is a Python implementation of Toolbox for emotion analysis using physiological signals (TEAP) and extracted all possible features of the signals present. We split the physiological signals up and tested them in several different rolling window sizes and extracted several feature combinations according to the window. From features like ECG, heart rate and triaxial acceleration common features like mean and standard deviation have been extracted. Higher degree features have also been extracted from some physiological signals. From BVP, features such as heart rate variability, interbeat interval, multi-scale entropy at five levels, spectral power of different frequency bands, spectral power ratio between 0.0–0.08 Hz and 0.15–0.5 Hz bands, Tachogram's low-, medium- and high-frequency spectral values and Tachogram's energy ratio have been extracted. From GSR, we have extracted the number of peaks

Table 1 Features extracted and their corresponding signals

Signals used	No. of features extracted	Features extracted
Blood volume pressure (BVP) (64 Hz)	17	Heart rate variability (HRV), Interbeat interval (IBI), Multiscale entropy (MSE), Spectral power of low/medium/high frequency bands, spectral power ratio, Tachogram's spectral values and ratio, Mean, Std. Deviation
Electrodermal activity (EDA) (4 Hz)	5	Number of peaks per second, average amplitude of peaks, peak rise time, Mean, Std. Deviation
Skin temperature (SKT) (4 Hz)	6	Kurtosis, skewness, spectral power, Mean, Std. Deviation
Heart rate (HR) (1 Hz)	2	Mean and Std. Deviation
Electrocardiograph (ECG) (1 Hz)	2	Mean and Std. Deviation
Triaxial acceleration (32 Hz)	6	Mean and Std. Deviation

per second, their average amplitude and their rise times. From SKT, complex features like kurtosis and skewness were extracted. After this process, we are left with 38 features per subject which we have used for classification. Both temporal and frequency domain features were extracted . In order to properly capture signal information, each physiological signal was divided into smaller windows. Sub-signal lengths of 30 seconds were used to extract the optimal result. Features as specified in Table 1 were extracted from each window for each participant to form the feature vector.

4.3 Classification

Historically many different classification methods have been suggested by previous studies for emotion recognition and classification. A Gaussian naive Bayes classification method was used in [6] to classify emotions. Along with accuracy, they also used the $F1$ scores to check for classification performance of the model as well as a leave one subject out (LOSO) scheme for cross-validation of the model. Several other studies have shown the use of different types of classifiers such as K-nearest neighbours (KNN) [13], support vector machine (SVM) [1, 2], logistic regression (LR) [1], decision trees (DT) [10] [13], random forest classifier (RFC) [1, 13] and multiLayer perceptron (MLP) [7]. We observed that SVM is a particularly popular classifier in this field as it tends to give the best results on physiological singal-based datasets.

The K-Emocon dataset [8] is relatively newer as compared to other emotion recognition datasets. It has a different approach to emotion elicitation than before, this dataset captures emotions of participants during naturalistic conversations which induce a real-life social interaction type of environment as opposed to previous methods of showing audio/video clips as done in DEAP [6]. For this reason, we have chosen the following models for optimum emotion classification as well as comparison purposes:

1. Gaussian Naive Bayes (GNB)—The GNB classifier is a classification algorithm based on the Bayes theorem. This algorithm is based on finding the probability of an event occurring given that another event has already occurred. This algorithm assumes that all data points in the dataset have an independent and equal contribution to the output, which is theoretically ideal for our dataset.

2. K-nearest neighbours (KNN)—KNN is a supervised machine learning algorithm that is widely used for classification problems. KNN functions by finding the distances between a specific data point and rest of the points in the data, selecting the pre-defined number of points (K) closest to the query, then votes for the most frequently occurring label. After hit- and trial-based testing, we have taken $K = 7$ for our classification as we found this to produce the best result.

3. Decision Tree (DT)—DT is a flowchart-like structure in which each internal node represents an attribute query, each branch represents the test result, and the resultant class is represented by each leaf node. The classification laws reflect the paths from root to leaf. The decision tree approach is one of the essential learning strategies that offer an adequate representation of the grouping of laws to accurately differentiate the chosen characteristics. In this process, the most robust characteristics for the initial splitting of the input data were found through the construction of a tree-shaped model.

4. Support Vector Machine (SVM)—One of the main statistical learning methods is a SVM capable of distinguishing invisible knowledge by deriving selected features and creating a high-dimensional hyperplane to divide the data points into two groups to create a model. Since SVM has the ability to handle high-dimensional data using limited function preparation, it has recently become very common in medical applications to extract physiological data. ECG, heart rate and SpO_2, most of which are used in a short-term and annotated fashion, are typical health parameters considered by SVM methods.

5. Neural Network (NN)—An artificial intelligence solution is a NN that is commonly used for classification and prediction. By learning the known classification of the classes and matching it with expected nodes of the records, the NN approach designs the train data in a way to adjust the network weights for the next iterations of learning. For our experiments, we have made a custom configuration of NN to best suit our purpose. Our neural network consists of 1 input layer for the 38

features, a hidden dense layer with 16 nodes and the "ReLU" activation function, followed by a dropout layer to prevent overfitting of training data, another hidden dense layer with the same parameters as the first one, and a final output layer powered by the "softmax" activation function. For the compilation of the model, the Adam optimizer is used and sparse categorical cross-entropy is used for the calculation of loss. The model keeps on training as long as we have a continuous decrease in loss with each epoch, until a maximum of 50 epochs.

We have calculated the accuracy values for all five of these classifiers and their respective $F1$ scores to measure confidence. We aim to find the best classifier that can work with all the physiological signals present in the K-Emocon dataset.

4.4 Platform of Implementation

All performed experiments (preprocessing, feature extraction, training and testing of various ML models) have been performed using the Python Language (v3.8.5) on the open-source Jupyter environment. The Graphic Card Nvidia GeForce RTX 2060 (mobile) was used for training. The following packages were used: OS, Glob, Numpy, and Pandas for basic calculations, data retrieval, cleaning, processing and visualization; SciKit learn for importing machine learning models; Keras for building an artificial neural network; PyTeap for feature extraction, selection and preprocessing; and finally matplotlib for plotting bar graphs.

5 Results Obtained and Analysis

In order to assess the performance of our proposed emotion classification models, we have implemented leave one subject out (LOSO) approach for cross-validation. The data is divided into 21 parts in the LOSO method, where each part contains the data of a single individual subject. Out of those 21 parts, 20 are used for training the model and the last part is used for testing the model. This method is repeated 21 times, each time considering one individual's data to be the test set. To calculate the accuracy of the emotion classification models, accuracies from 21 experiments are averaged. LOSO-based analysis was carried out for both arousal and valence.

As previously mentioned, we have used five machine learning models including a custom configured deep learning model for a proper classification analysis. Accuracy and respective $F1$ score for confidence measure are obtained after each test run of every model for both target values, namely arousal and valence. Moreover, K-Emocon provides us with five different types of target values, namely "self", "partner", "self-partner mean" and "aggregate external", based on the annotator of those target values. These target values are integers and lie in the closed range of 1–5 for both arousal and valence. Finally using the LOSO approach for cross-validation means, we have

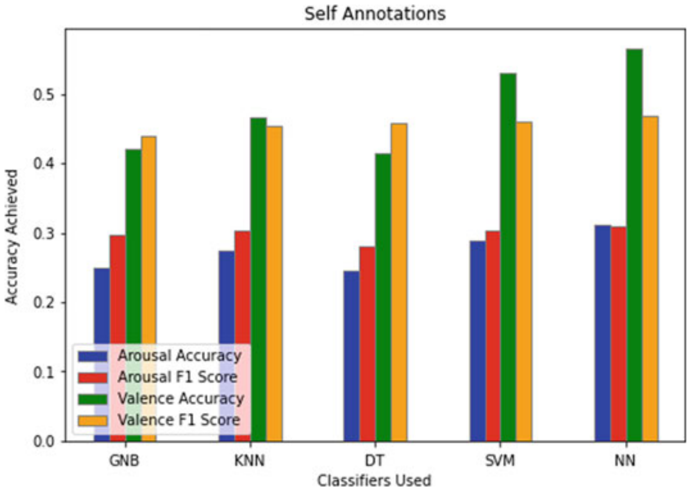


Fig. 3 Performance of ML models on self-annotations

21 output values for each configuration of test, 1 for each participant. This results in a final array of total 840 values for accuracy and 840 respective $F1$ scores for every participant. The mean of accuracies of all 21 participants is taken at the end of a single test run and that is labelled as the average accuracy of that particular model in that test run. This approach of LOSO cross-validation and calculating mean accuracy allows us to precisely calculate the performance of every model in each configuration without any unfair bias. The performance of said models in these configurations is shown in the following figures.

Figure 3 shows the performance of all five models on self-annotation target values. The neural network achieved the best accuracy for both valence and arousal at 56.63 and 31.03%.

Figure 4 shows the performance of all five models on partner annotation target values. The neural network achieved the best accuracy for both valence and arousal at 54.57 and 32.95%.

Figure 5 shows the performance of all five models on self-partner mean annotation target values. The neural network achieved the best accuracy for both valence and arousal at 60.05 and 55.22%. These annotations give the best accuracy for arousal.

Figure 6 shows the performance of all five models on aggregate external annotation target values. SVM achieved the best accuracy for valence at 91.12% with the neural network behind it by 0.08%. Best accuracy for arousal was achieved by neural network at 45.42%.

Out of the five models used, we observe that the mean accuracy of our custom configured neural network almost always exceeds that of the other four models. The NN model outperforms GNB by 9.16%, KNN by 6.63%, DT by a massive 11.36% and SVM by a marginal 2.38% on average in terms of accuracy. There was only a single case where the average accuracy of the NN model was lower than one of the

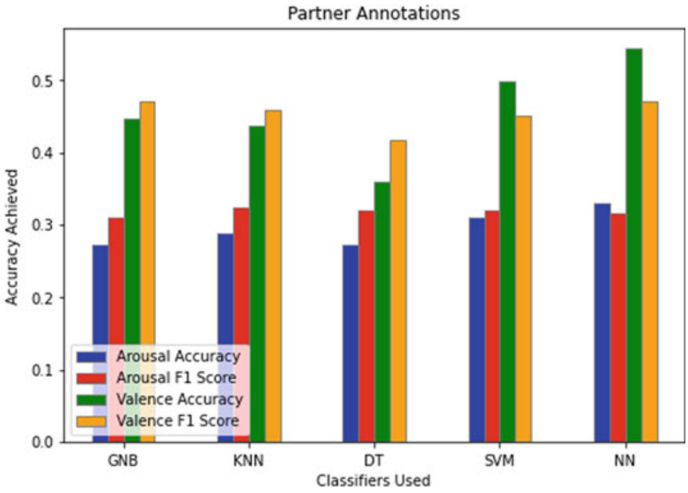


Fig. 4 Performance of ML models on partner annotations

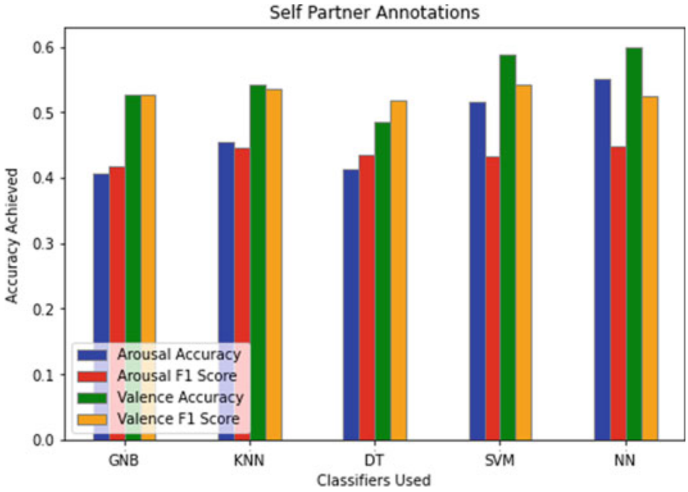


Fig. 5 Performance of ML models on self-partner mean annotations

other models, in external annotations, the average valence accuracy of SVM model beat the neural network by a value of 0.08%. The best accuracy achieved for valence was 91.04% by NN with an $F1$ score of 87.62 and 91.12% by SVM with an $F1$ Score of 87.47%. The best accuracy achieved for arousal on the other hand was 55.22% by NN with an $F1$ Score of 44.86%. This is the accuracy score obtained when all physiological signals and their 38 extracted features were used for classification. We also tested all different combinations of physiological signals to observe their effect on accuracy and found in one case that using only heart rate signal resulted in

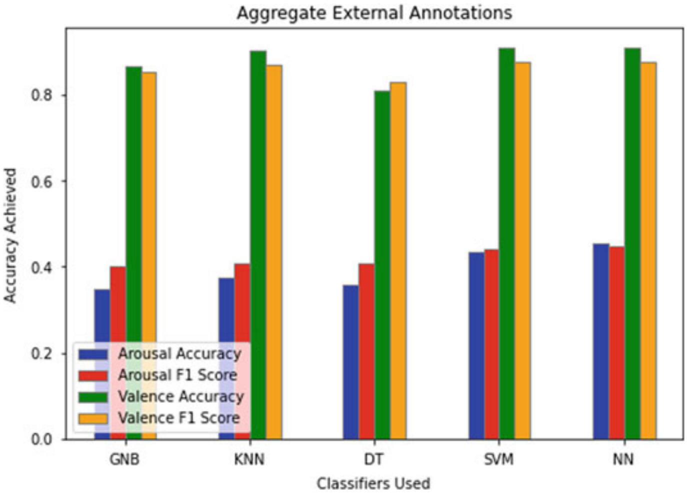


Fig. 6 Performance of ML models on aggregate external annotations

an improved arousal accuracy of 62.19% in self-partner mean annotations but this significantly worsened performance on valence accuracy.

Just like any other system, our proposed work also has few limitations. K-Emocon the dataset used collected data from a limited number of participants (21 total). Hence, judging the generalisability of the proposed model is difficult. The features extracted from each physiological signal play a crucial role on the accuracy obtained, and hence, there may be a better set of features providing better accuracy.

6 Conclusion and Discussion

This study was aimed at recognition and classification of emotions during naturalistic conversations between people using peripheral physiological signals.

This approach of using physiological signals has the advantage that it shows us the true emotion that the person is feeling as these signals cannot be voluntarily modified by the person unlike other factors like facial expressions and voice. To achieve the best possible recognition rate, we tested and compared the performance of four standard machine learning models, namely GNB, KNN, DT and SVM, along with one advanced deep learning model, the custom neural network configuration, on the K-Emocon dataset. This allows us to find the best possible classifier without any bias. A total of 38 input features were extracted from a variety of physiological signals present in the dataset for this analysis. It was observed that the best accuracy was achieved by the custom NN model in all cases except 1, where it was behind SVM by a minimal 0.08%. The best accuracies achieved for valence and arousal in our experiment were 91.12% and 62.19%, respectively. As valence describes the type of emotion which is easier to identify than arousal which describes the intensity of emotion, we believe this reason can contribute to the relative disparity of accuracies that is seen between the two target values. Due to the nature of our multi-class classification as well as the fact that K-Emocon is a relatively newer dataset at the time of writing this paper, it is not possible for us to directly compare these results to other work. However, the results obtained with this dataset are definitely an improvement over results obtained in previous works with older datasets like DEAP. Overall it is shown that it is viable to recognize emotions using peripheral physiological signals.

7 Future Work

A limitation of our emotion recognition model is that it is a user-independent model; i.e. it is not calibrated according to any specific user but instead is a generic model designed to work with anyone. This leads to less accurate overall results when compared to a user-based model which is calibrated and trained to work on a single type of user as specific calibration tends to give better accuracy. Another limitation we faced was that there were only 21 test subjects on which the data was available in the dataset, having more test subjects makes the model more generalizable. Future work in this field can work on overcoming these limitations as well as different feature extraction and selection techniques or the use of even more advanced deep learning techniques for better emotion recognition.

References

1. Ayata, D., Yaslan, Y., Kamasak, M.E.: Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. J. Med. Biol. Eng. 1–9 (2020)
2. Dai, Y.: Reputation-driven multimodal emotion recognition in wearable biosensor network. IEEE Int. Instrument. Measur. Technology Conf. (I2MTC) Proc. **2015**, 1747–1752 (2015)
3. Dobbins, C. et al.: A lifelogging platform towards detecting negative emotions in everyday life using wearable devices. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, 2018, pp. 306–311
4. Ekman , P.E., Davidson, R.J.: The Nature of Emotion: Fundamental Questions. Oxford University Press (1994)
5. Katsis, C.D., et al.: Toward emotion recognition in car-racing drivers: a biosignal processing approach. IEEE Trans. Syst. Man Cybern. Part A Syst. Human **38**(3), 502–512 (2008)
6. Koelstra, S., et al.: Deap: A database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2011)
7. Lee, C.: Using neural network to recognize human emotions from heart rate variability and skin resistance. IEEE Eng. Med. Biol. 27th Ann. Conf. **2006**, 5523–5525 (2005)
8. Park, C.Y., et al.: K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci. Data **7**(1), 1–16 (2020)
9. Plutchik, R.: The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. **89**(4), 344–350 (2001)

10. Pollreisz, D., TaheriNejad, N.: A simple algorithm for emotion recognition, using physiological signals of a smart watch. In: 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 2353–2356 (2017)
11. PTI: India has the highest number of smartwatch owners globally: Survey. 2014. urlhttps://www.bgr.in/news/india-has-the-highest-number-of-smartwatch-owners-globally-survey-330358/, visited on 10/20/2014
12. Saganowski, S.L., et al.: Review of Consumer Wearables in Emotion, Stress, Meditation, Sleep, and Activity Detection and Analysis. arXiv:2005.00093 (2020)
13. Shu, L., et al.: Wearable emotion recognition using heart rate data from a smart bracelet. Sensors **20**(3), 718 (2020)
14. Sreeja, P.S., Mahalakshmi, G.S.: Emotion models: a review. Int. J. Control Theor. Appl. **10**(8), 651–657 (2017)
15. Zhang, X., et al.: Spatial-temporal joint optimization network on co-variance manifolds of electroencephalography for fatigue detection. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 893–900. IEEE, 2020

Phishing Detection Using Computer Vision

Shekhar Khandelwal and Rik Das

Abstract Phishing is a cyber-crime wherein innocent web users are trapped into a counterfeit website, which is visually similar to its legitimate counterpart, but in reality, it is fake. Initially, users are redirected to phishing websites through various social and technical routing techniques. Users being ignorant about the illegitimacy of the website, provide their personal information such as user id, password, credit card details, bank account details to name a few. These details are stolen by the phishers and later used for either financial gains, or to tarnish a brand image or even more grave crimes like identity theft. Many phishing detection and prevention techniques are proposed in the literature; however, there is much scope in the cyber-security world with the advent of smart machine learning and deep learning methods. In this research, we explored computer vision techniques and build deep learning and machine learning classifiers to detect phishing website and their brands. Some of the experiments include Transfer Learning and Representation Learning techniques by utilizing various off-the-shelf Convolutional Neural Network (CNN) architectures to extract image features. It is observed that DenseNet201 outperforms all experiments conducted as well as the existing state-of-the-art on the dataset used, proving the hypothesis that Convolutional neural networks are an effective solution for extracting relevant features from phishing webpages for phishing detection classification.

Keywords Phishing detection · Deep learning in cyber-security · Anti-phishing mechanism · Website brand prediction · Image classification · Transfer learning · Representation learning · Computer vision

1 Introduction

Phishing is a term derived from fishing, by replacing "f" with "ph", but contextually mean the same [1]. Like fishes are being trapped into the fishing net, innocent

S. Khandelwal (✉)
IBM India Software Labs, Bangalore, India

R. Das
Xavier Institute of Social Service, Ranchi, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_9

web users are being trapped into phishing websites. Phishing websites are basically counterfeit websites which are visually similar to their legitimate counterparts.

Phishing detection mechanisms are broadly categorized into 4 groups–

1. List (whitelist/blacklist) based
2. Heuristics (pre-defined rules) based
3. Visual similarity based
4. AI/ML based

In list based anti-phishing mechanisms, a whitelist and blacklist of the URLs are created, and a suspicious website URL is compared against these lists to conclude whether the website under scrutiny is a phishing website or a legitimate one [2] [3].

There are various limitations with list-based approach, namely.

(1) It is dependent on a third-party service provider that captures and maintains such lists like Google safe browsing API [4].
(2) Since listing a newly deployed phishing website in the white/blacklist is a process that takes time. First, it has to identified, and then it has to be listed. Since the average lifetime of a phishing website is 24–32 h, hence zero-day phishing attacks [5] cannot be prevented by list-based mechanisms.

In heuristic-based approaches, various website features like image, text, URL, and DNS records are extracted and used to build a rule-based engine or a machine learning based classifier to classify a given website as phishing or legitimate. Heuristic-based approaches are quite effective anti-phishing mechanism, however [6] addresses some drawbacks in heuristic-based approaches like.

(1) The required time and computational resources for training is too high
(2) Cannot be used as a browser plugin
(3) Would be ineffective once the key features are discovered by the scammers as they will find a way to bypass the rules.

Visual Similarity Based techniques are very useful to detect phishing since phishing websites look similar to their legitimate counterparts. Visual similarity-based technique uses visual features like text-content, text-format, DOM features, CSS features, website images, etc. to detect phishing. In visual similarity-based techniques, DOM, CSS, HTML tags and pixel-based features are compared to their legitimate counterparts in order to make a decision.

As depicted in Fig. 1, within pixel-based techniques, there are two broad categories through which phishing detection is achieved. One approach is through comparison of visual signature of suspicious website images with the stored visual signatures of legitimate websites. For example, hand-crafted image features like SIFT (Scale Invariant Feature Transform) [8], SURF (Speeded Up Robust Features) [9], HOG (Histogram of Oriented Gradient) [10], LBP (Local Binary Patterns) [11], DAISY [12], MPEG7 [13] are extracted from the legitimate websites and stored in a local datastore. Further, similar features from the website under scrutiny are extracted and compared with the stored baseline, to classify whether the website is legitimate or phishing.

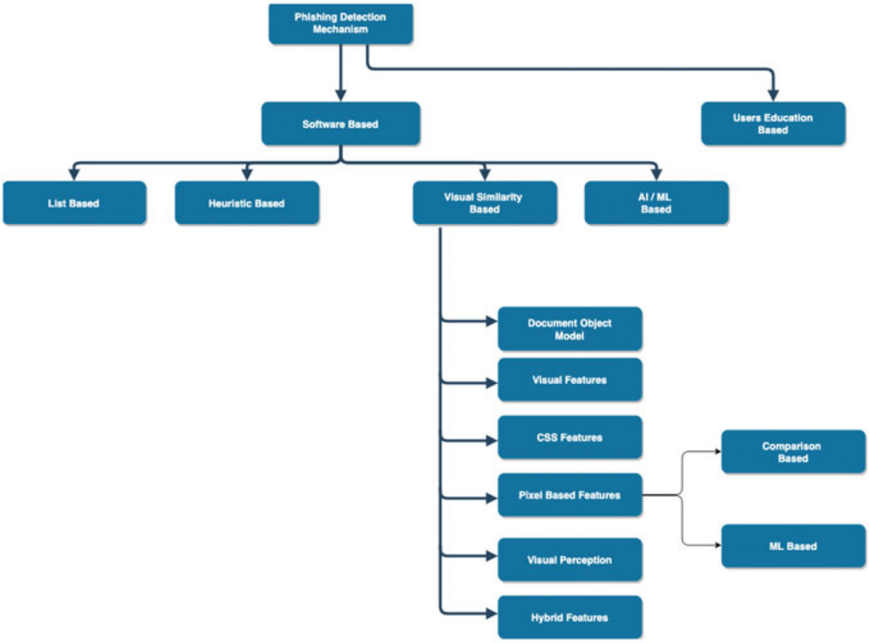


Fig. 1 Phishing detection mechanisms [7]

There are two approaches that have not been explored extensively within pixel-based techniques, which is experimented in this research. First, extracting features from images using Convolutional Neural Networks. All the previous work that has been done on feature extraction is mainly done on extracting handcrafted features like SIFT, SURF, HOG, etc. Second, once the features are extracted, instead of comparing the features with a stored baseline, build a machine learning classifier to learn the pattern of phishing webpage images, for classification.

In this research, we used transfer learning approach to extract image features using various state-of-the-art CNN architectures. Further, various machine learning, and deep learning classifiers are trained using those extracted features. In the phishing detection literature, this approach has not been implemented so far, and this is the first time Transfer Learning and Representation learning approaches are implemented on a publicly available dataset for phishing detection using website images.

The research is broadly classified into 4 sections.

1. Perform thorough literature review on the topic of Phishing and their detection mechanisms that previous researchers have worked upon. Thoroughly study the use of machine learning and deep learning approach in previous research on the topic of phishing detection. Additionally, explore datasets available for research on this topic.

2. Based on the gaps in the previous research works that utilize machine learning and deep learning techniques for phishing detection, define various mechanisms and approaches that are carried out in this research.
3. Perform data pre-processing like feature extraction, label encoding, data-imbalance treatment, dimensionality reduction and feature fusion. Further, implement the defined approaches and capture the results.
4. Finally, analyze the results, and compare the performance of the best performing model of this research with the state-of-the-art performance on the chosen dataset.

2 Literature Survey

One of the noticeable heuristic-based anti-phishing mechanism that also uses image of the webpage is GoldPhish [14] which uses OCR technology to extract the text from the logo of the webpage. Once brand name is extracted, it uses Google Page Rank algorithm to get the top 5 webpages using the extracted text from the logo. Then the suspicious web page domain is compared with the top 5 domains retrieved using google search api, and if the suspicious website domain does not match with any of those 5 domains, then it's declared a phishing website.

Like GoldPhish extracts logo using OCR technology, Verilogo [15] uses SIFT image descriptor [8] to extract logo of the given suspicious website. Verilogo, not only extracts logo using SIFT descriptor, but it also deploys multi-level phishing detection mechanism. If logo of the extracted website does not match any logo in the local datastore, it gives the website a benefit of doubt. Since there is no delta to compare with, and hence user is allowed to continue on the website. This limits the mechanism detection capability of Verilogo only to the brands whose legitimate logos are stored in the local datastore for comparison.

One such solution is described in this paper [16] where computer vision technique called SURF descriptors [9] were used to extract discriminative key point descriptors of both legitimate and suspicious website and based on similarity threshold, displays a warning for further inspection of the suspicious website.

Another comparison based phishing detection application utilizes Contrast Context Histograms (CCH) descriptors [17] to compare similarity degree between suspicious and authentic web pages.

On similar lines, where image descriptors of the suspicious websites are compared with some locally stored legitimate website descriptors is the one which utilizes HOG image descriptors to detect phishing [18].

Now, the next few literatures are the Machine Learning based classification mechanism for phishing detection. Similar to comparison-based approach, features are extracted from the website images, but unlike comparison-based approaches, wherein those features are compared with the locally stores legitimate features, here patterns are extracted from those features using some machine learning algorithms and a classification model is built, which is used for phishing detection in real time.

One such approach is, PhishIRIS [19] which utilizes MPEG7 [13] and MPEG7 like Compact Visual Descriptors (CVD) like Scalable Colour descriptor (SCD), Colour Layout Descriptor (CLD), CEDD, Fuzzy Colour and Textual Histogram (FCTH) and Joint Composite Descriptors (JCD). The aforementioned image descriptors were extracted from the website snapshot and transformed into feature vectors and fed into machine learning algorithms like Support Vector Machine (SVM) and Random Forest (RF) for classification purposes to classify the image between legitimate or phishing website.

Eroğlu et al. [20] explores GIST and LBP features to determine the brand of the phishing webpages. In this study, GIST and LBP features are extracted and ingested into machine learning algorithms like SVM, RF and XGB for the classification task. Similarly, [21] explores SIFT [22] and DAISY [12] features to determine the brand of the phishing webpages.

In conclusion, previous research on this topic, where website images are used for phishing classification, are mostly comparison based, in which features extracted from the trusted website image samples are stored in a local datastore, and then similar features are extracted from the unseen website images, and then compared with the stored features, and based on the comparison result, classification is done.

Various research has been done on this topic where brand of the given website image is extracted, and then domain of the extracted brand is compared with the legitimate domain of that brand. Legitimate domains are either stored in the local datastore or are extracted using search engine api services like google api. In both, the above approaches, either a local datastore is required, or a third-party service is required for an end-to-end classification of the website images as phishing or legitimate.

In this research, we focus on using the image pixel values and the features extracted using those values to build a machine learning and deep learning classifier. Such application classifies based on the features extracted from the static images of the website, eliminating the need for any local datastore to store legitimate website features, as well as any third-party services to validate the legitimacy of the domain of the website in question. Figure 2 depicts various mechanisms for phishing detection which are pixel based, as well as the mechanism proposed in this research.

Additionally, all the previous research where image features are extracted for classification, only hand-crafted featured have been extracted and explored. Hence this study is significant in terms of exploiting convolutional neural networks to extract image features from various layers of various off-the-shelf CNN architectures. Also, since the experiments are conducted using Transfer Learning techniques, hence classification model can be trained on less images with less computational requirements.

Hence, various significant approaches experimented in this research are -

(1) Utilize various off-the-shelf CNN applications for transfer learning by replacing the final classification layer.
(2) Extract features from website images using Representation learning utilizing various off-the-shelf CNN applications like VGGNet [23], GoogLeNet [24],

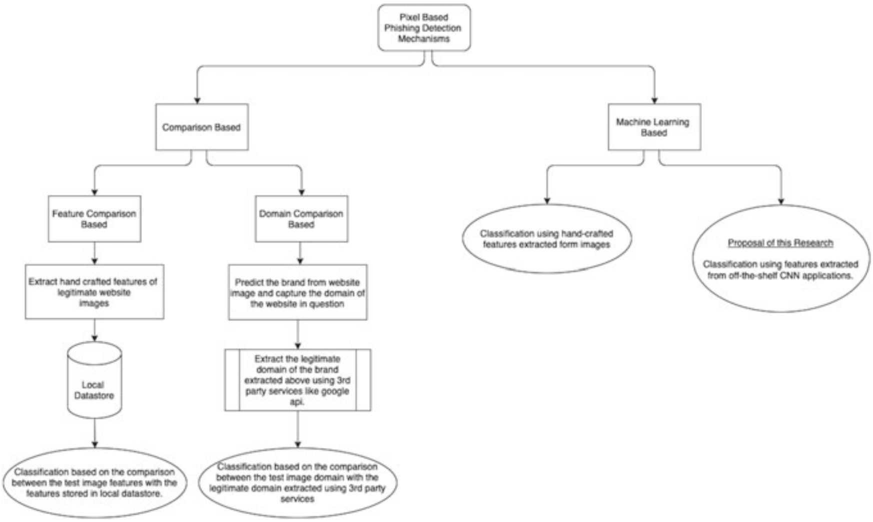

Pixel Based Phishing Detection Mechanisms
Comparison Based
Machine Learning Based
Feature Comparison Based
Domain Comparison Based
Classification using hand-crafted features extracted form images
Proposal of this Research
Classification using features extracted from off-the-shelf CNN applications.
Extract hand crafted features of legitimate website images
Predict the brand from website image and capture the domain of the website in question
Local Datastore
Extract the legitimate domain of the brand extracted above using 3rd party services like google api.
Classification based on the comparison between the test image features with the features stored in local datastore.
Classification based on the comparison between the test image domain with the legitimate domain extracted using 3rd party services

Fig. 2 Various image processing based work and proposed solution

Residual Network [25], DenseNet [26] etc. Use the extracted features to train a machine learning classifier to classify phishing websites and their respective brands.

(3) Apply dimensionality reduction (PCA) technique on the features extracted from different CNN applications and build a machine learning classifier.

(4) Fuse features of same image extracted from different techniques and use the fused features to build a machine learning classifier, as discussed in fusing local and global image descriptors [27], combining multiple global descriptors [28] or combining various image features [29].

3 Datasets

Dataset used for this research is generated by [19] and made publicly available at [30] for future research purposes. It is a labelled dataset readily available for research work. Dataset comprises of 2852 screenshots of websites which contains screenshots of 14 highly phished brands in their respective folders and one folder which contains legitimate webpages screenshots. So overall it is a multiclass dataset with (14 + 1) classes. The data is collected from March–May 2018.

Table 1 shows the list of brands whose phishing webpage images are there in the dataset along with the information of number of images of each brand.

The “other” folder contains the images of all the legitimate websites. Since this is not a comparison-based approach, hence other folder contains images of brands other than the 14 brands chosen for phishing website images. Also, since most of the websites on internet are legitimate websites, hence with the addition of this

Table 1 PHISH-IRIS dataset details

Brand name	Training samples	Testing samples
Adobe	43	27
Alibaba	50	26
Amazon	18	11
Apple	49	15
Bank of America	81	35
Chase Bank	74	37
Dhl	67	42
Dropbox	75	40
Facebook	87	57
Linkedin	24	14
Microsoft	65	53
Paypal	121	93
Wellsfargo	89	45
Yahoo	70	44
Other	400	1000
Total	1313	1539

"other" folder makes this dataset an open set since the images in this folder have no common colour scheme, edge structure that characterizes its own class. There are 1313 training and 1539 testing webpages in the dataset, which are already separated in two different folders "train" and "val" as depicted in Fig. 3.

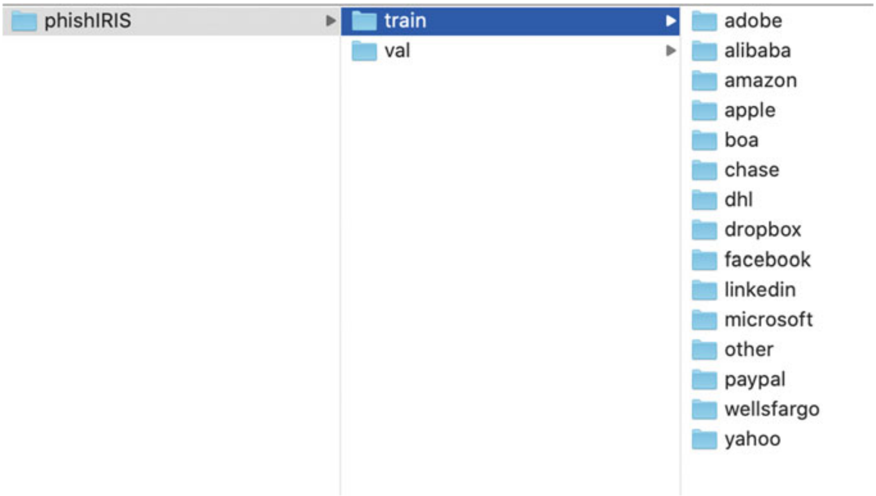


Fig. 3 Folder structure of the dataset

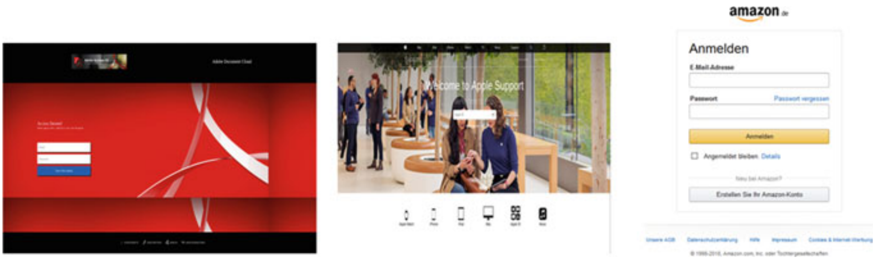


Fig. 4 Phishing website images of adobe, apple and amazon

As depicted in Fig. 3, the images are labelled since images of individual brands are stored in the folder named by its brand. The folder structure is same for both training and validation data. Each folder contains phishing website images of the respective brand.

Figure 4 shows some phishing website images from the Phish-IRIS dataset. The phishing website images of the respective brands seem similar to their legitimate counterparts to naked eyes. Hence, there is a need for a machine to intelligently learn the patterns in the images of phishing websites, if any. Hence, by training a machine learning model, we expect that the model to identify the patterns of the phishing webpages from the sample images provided for training the model.

4 Proposed Techniques and Tools

As proposed in Fig. 2, image features are extracted using Convolution Neural Networks for building a classification model, to detect phishing websites as well as brand of the phishing website. The phishing detection task in this research is an image-based multi-class classification task.

The number of images available in Phish-IRIS dataset, that we will use in this research, contains 1513 images in training dataset. This is not a considerable number of images to train a CNN model from scratch. Hence, in this research, experiments are conducted with various off-the shelf CNN architectures using transfer learning and representation learning approach. Various CNN applications are utilized to extract relevant image features, and then use those features are used to build a machine learning and deep learning multi-class classification model.

List of all the CNN architectures used for experiments in this research are listed in Table 2.

On a high level, experiments are conducted utilizing traditional CNN, transfer learning and representation learning approaches, as shown in Table 3.

Off-the-shelf pre-trained CNN applications as listed in Table 2, are used to extract features from images in Phish-IRIS dataset [30] utilizing transfer learning techniques. All the CNN applications listed in Table 2 are utilized for representation learning

Table 2 CNN architectures used for transfer learning

Xception
VGG16, VGG19
ResNet50V2, ResNet101V2, ResNet152V2
InceptionV3, InceptionResNetV2
MobileNet, MobileNetV2
DenseNet121, DenseNet169, DenseNet201

Table 3 Description of all the classification approaches conducted

Approaches	Description
Approach 1	Traditional CNN with image size of (128, 128, 3) and (256, 256, 3)
Approach 2	Transfer the learning till last fully connected layer
Approach 3	Transfer the learning till last convolution layer, extract features, and SMOTE and implement a machine learning classifier
Approach 4	Horizontal fusion of the features extracted in Approach 3 in a combination of 2 and 3 CNN applications
Approach 5	PCA of the features extracted in Approach 3 and implement a machine learning classifier
Approach 6	Horizontal fusion of the features extracted in Approach 5 in a combination of 2 and 3 CNN applications
Approach 7	Vertical fusion of the features extracted in Approach 5 in a combination of 2 and 3 CNN applications
Approach 8	Apply RF and SVM on the best performing model from all the above approaches for an image size of (512, 512, 3)

and extracted the features of the images from the Phish-Iris dataset. Transfer learning approach was used to build a classifier by using the pre-trained weights of all the aforementioned CNN pre-built models using Keras library, by just removing the final classification layer and replacing it with 15 units based softmax classifier as depicted in Fig. 5.

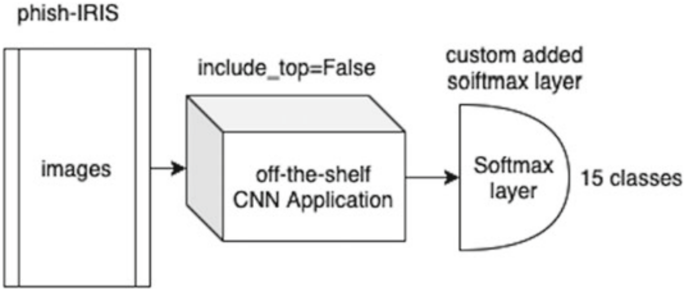


Fig. 5 Approach 2

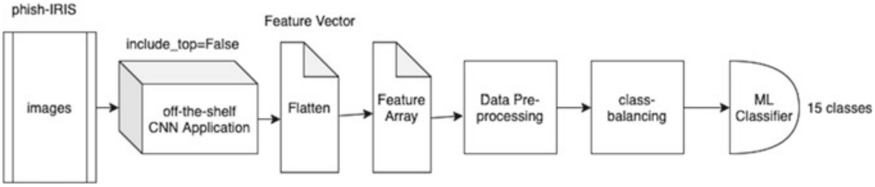


Fig. 6 Approach 3

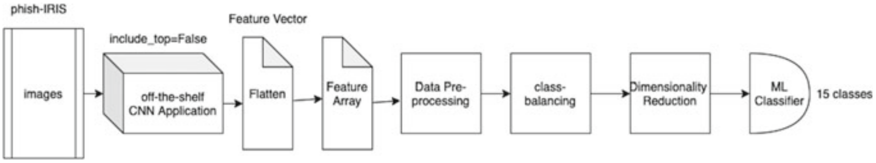


Fig. 7 Approach 4

Additionally, features were extracted from the dataset utilizing representation learning techniques using CNN architectures listed in Table 2. The features are extracted from resizing the images in three different sizes (128, 128, 3), (256, 256, 3) and (512, 512, 3). These individual feature vectors were then treated using SMOTE technique to treat the class imbalance. The final dataset is used to train a machine learning classifier. We utilized Random Forest and Support Vector machine classifiers throughout the experiments, as depicted in Fig. 6.

Further, we experimented with the dimensionality reduction of the class imbalanced data using Principal Component Analysis (PCA) technique, as depicted in Fig. 7.

Finally, we fused the features extracted from various CNN models in a combination of 2 CNN architectures and a combination of 3 CNN architectures, using horizontal feature fusion technique. We used the class imbalanced data as well as PCA'd data for the fusion technique. Fusion of features extracted from different CNN architectures can only go through Horizontal fusion because the dimensions of the features extracted from different applications vary. Hence, initially only horizontal fusion of features were conducted and experimented as shown in Fig. 8.

Further, after dimensionality reduction, all features from all CNN architectures are brought to same dimensions, hence vertical fusion of features could be achieved.

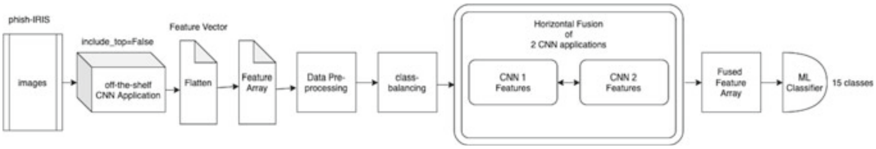


Fig. 8 Horizontal fusion with feature fusion

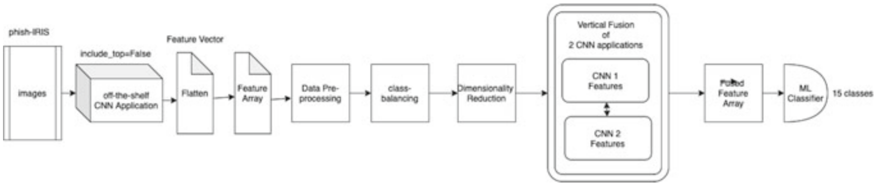


Fig 9 Vertical fusion with feature fusion

Hence, vertical fusion of features were conducted after PCA is applied, as shown in Fig. 9.

The research methodology for feature extraction as well as various experiments conducted in this research is shown in Fig. 10.

5 Results and Discussion

Images in the phish-IRIS dataset are resized into three different dimensions—(128, 128, 3), (256, 256, 3) and (512, 512, 3). Features are extracted utilizing CNN applications like VGGNet, DenseNet, etc. for all the image sizes. Additionally, dimensionality reduction and feature fusion of the extracted features are done to build a numerous classification model. Table 4 lists all the best performing models within each approach listed in Table 3, which is experimented in this. The table is sorted by recall (TPR) value against each experiment.

As it can be seen in Table 4, classifier built using features extracted using DenseNet201 for the image size of (512, 512, 3) performed exceptionally well when trained using SVM classification algorithm. For a multi-class classification model, the performance of the DenseNet201 based model performance against each class can be seen in Fig. 4 and the confusion matrix can be seen in Fig. 11.

As it can be seen in Fig. 11, the precision, recall and $F1$-score of individual classes, and summarized the overall metrics for the entire dataset.

Precision in the multi-class context would mean that number of webpages detected as a particular class, they actually belong to that class. No other class is detected for that phishing brand. This would not be a good measure in case of detecting phishing website because it may miss to detect a lot of phishing website as "phish", but still would show a high value. For example, in the confusion matrix depicted in Fig. 12, there are 31 webpages that were predicted as adobe, but out of those 31, only 24 webpages actually classified correctly. Hence, the precision of the model for adobe class is 77%.

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

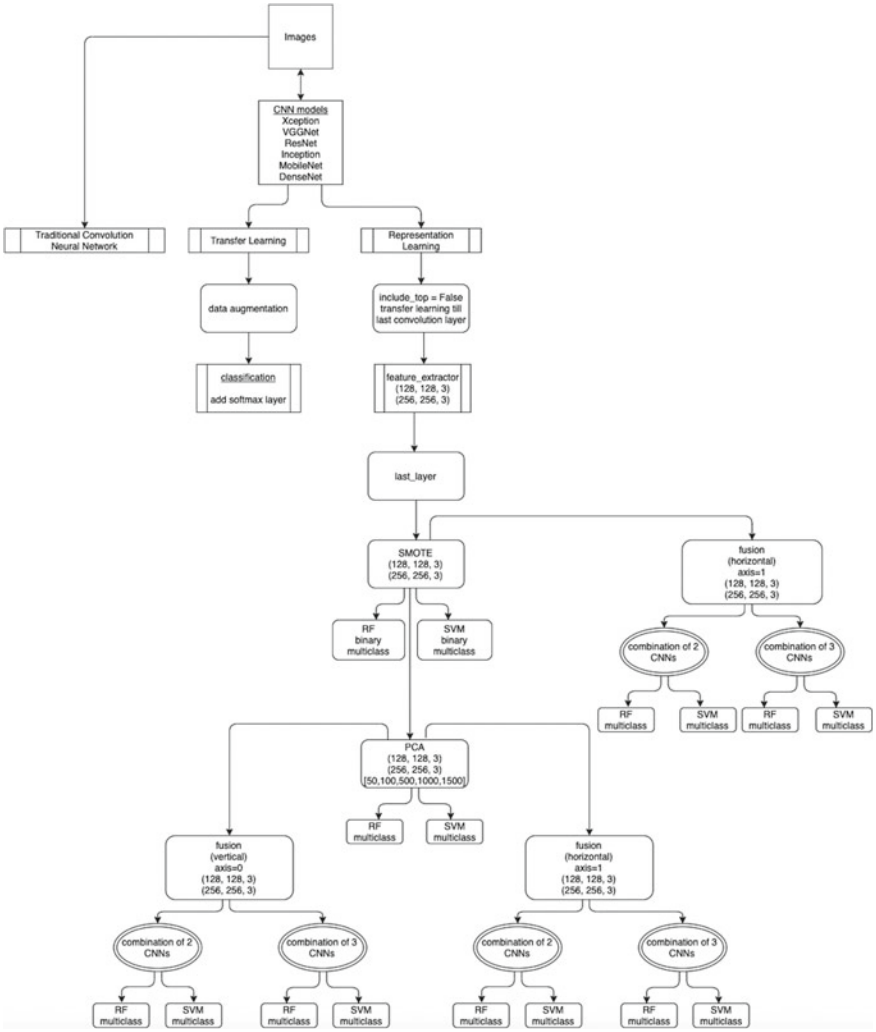

Images
CNN models
Xception
VGGNet
ResNet
Inception
MobileNet
DenseNet
Traditional Convolution Neural Network
Transfer Learning
Representation Learning
data augmentation
include_top = False transfer learning till last convolution layer
classification
add softmax layer
feature_extractor (128, 128, 3) (256, 256, 3)
last_layer
SMOTE (128, 128, 3) (256, 256, 3)
fusion (horizontal) axis=1 (128, 128, 3) (256, 256, 3)
RF binary multiclass
SVM binary multiclass
combination of 2 CNNs
combination of 3 CNNs
RF multiclass
SVM multiclass
PCA (128, 128, 3) (256, 256, 3) [50,100,500,1000,1500]
fusion (vertical) axis=0 (128, 128, 3) (256, 256, 3)
fusion (horizontal) axis=1 (128, 128, 3) (256, 256, 3)

Fig. 10 Experiments conducted in this research

Recall, on the other hand actually tells that for a given brand webpages, how many of them are detected correctly. For example, in the confusion matrix depicted in Fig. 12, out of 27 adobe webpages, 24 webpages are classified correctly. Hence, the recall of the model for adobe class is 88%.

$$\mathrm{REC} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP}} \mathrm{REC} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{FN} + \mathrm{TP}}$$

Table 4 Comparison of best performing models within different approaches

Approach	Model	Recall	$F1$	Image size	Classifier
8	**DenseNet201**	**0.908**	**0.909**	**512**	**SVM**
3	DenseNet201	0.903	0.903	256	SVM
3	DenseNet201	0.897	0.89	256	RF
4	VGG19_ResNet152V2_DenseNet201	0.897	0.893	256	RF
5	DenseNet201_500	0.897	0.896	256	SVM
4	Xception_DenseNet201	0.895	0.891	256	RF
3	MobileNet	0.891	0.89	128	SVM
4	ResNet101V2_MobileNetV2_DenseNet201	0.89	0.888	128	RF
3	VGG16	0.885	0.88	128	RF
5	MobileNet_500	0.884	0.88	128	SVM
4	VGG19_DenseNet169	0.884	0.883	128	RF
5	DenseNet169_100	0.883	0.88	256	RF
6	MobileNetV2_DenseNet169_DenseNet201_100	0.88	0.881	256	RF
5	MobileNet_100	0.877	0.874	128	RF
7	VGG16_DenseNet169_100	0.875	0.871	256	RF
6	VGG16_DenseNet169_100	0.875	0.871	256	RF
6	MobileNet_MobileNetV2_DenseNet201_100	0.873	0.873	128	RF
7	VGG16_MobileNetV2_DenseNet169_100	0.869	0.864	256	RF
2	DenseNet169	0.869	0.884	256	NN
6	VGG16_MobileNetV2_100	0.855	0.853	128	RF
7	VGG16_MobileNetV2_100	0.855	0.853	128	RF
7	VGG16_MobileNetV2_DenseNet169_100	0.85	0.847	128	RF
2	DenseNet169	0.791	0.825	128	NN

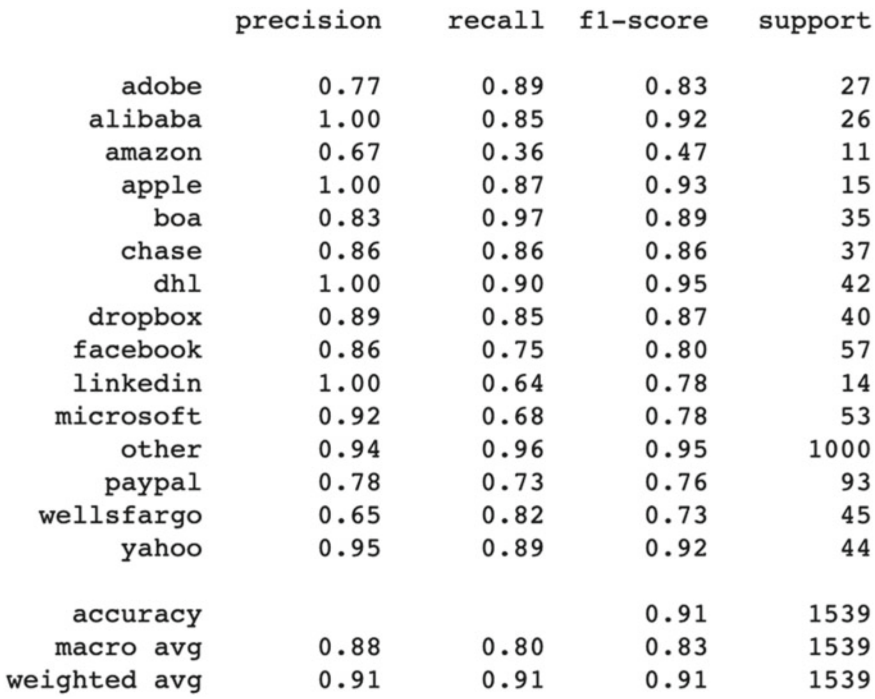
$F1$-score is the harmonic mean of Recall (REC) and Precision (PRE), and provide a more accurate performance of a model.

$$F1 = 2.\frac{\text{PRE.REC}}{\text{PRE} + \text{REC}} F1 = 2.\frac{\text{PRE.REC}}{\text{PRE} + \text{REC}}$$

Comparison of the Best Performing Model with the State-of-the-Art

State-of-the-art performance of a machine learning based classifier for phish-IRIS dataset as stated in their official site [30] is used to compare the best performing model implemented in this research as shown in Table 5.

Performance metrics of all the approaches conducted in this study as per Table 4, along with State-of-the-art results against True Positive Rate (Recall) values can be seen in Fig. 13.



```
              precision    recall  f1-score   support

       adobe       0.77      0.89      0.83        27
     alibaba       1.00      0.85      0.92        26
      amazon       0.67      0.36      0.47        11
       apple       1.00      0.87      0.93        15
         boa       0.83      0.97      0.89        35
       chase       0.86      0.86      0.86        37
         dhl       1.00      0.90      0.95        42
     dropbox       0.89      0.85      0.87        40
    facebook       0.86      0.75      0.80        57
    linkedin       1.00      0.64      0.78        14
   microsoft       0.92      0.68      0.78        53
       other       0.94      0.96      0.95      1000
      paypal       0.78      0.73      0.76        93
  wellsfargo       0.65      0.82      0.73        45
       yahoo       0.95      0.89      0.92        44

    accuracy                           0.91      1539
   macro avg       0.88      0.80      0.83      1539
weighted avg       0.91      0.91      0.91      1539
```

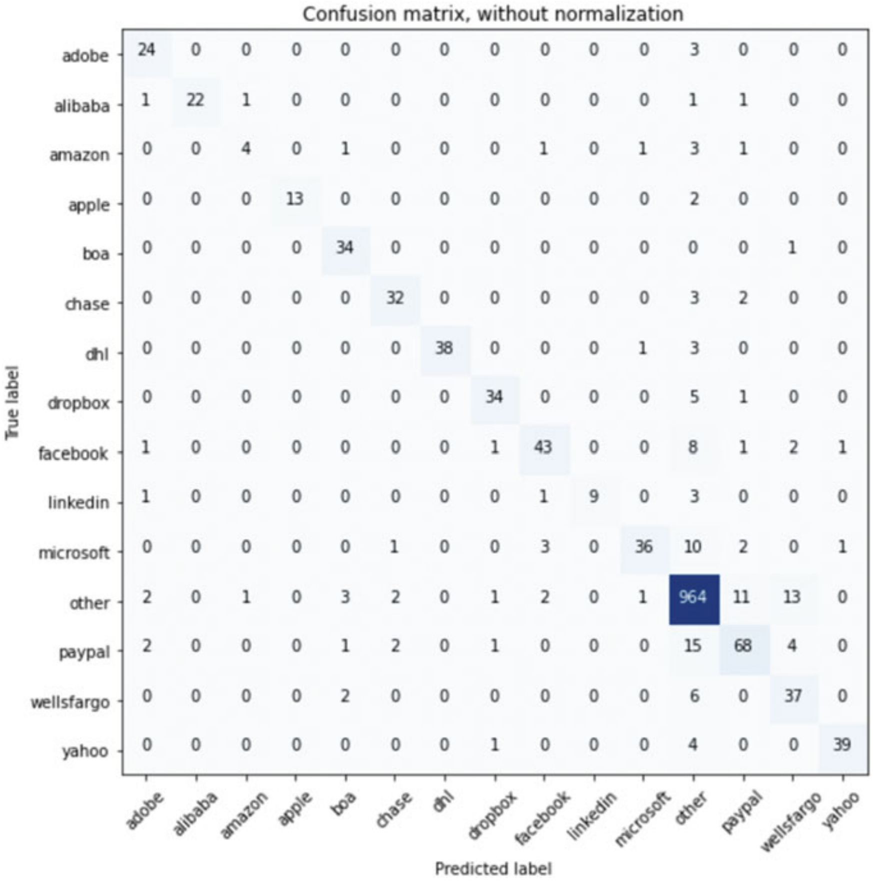
Fig. 11 performance metrics for DenseNet201 against each class

Performance metrics of all the approaches conducted in this study as per Table 4, along with State-of-the-art results against $F1$ score values can be seen in Fig. 14.

In Figs. 13 and 14, it can be seen how the top-performing models built using various approaches defined in Table 3 performed in terms of Recall and $F1$ score metrics. As per Approach 8 from Table 3, features are extracted using image size of (512 * 512) utilizing DenseNet201 CNN architecture, and ingested the representation of the images in SVM machine learning classifier. The results using this model have exceeded the state-of-the-art performance on this dataset as seen in Table 5.

6 Conclusion

To conclude, phishing attacks are an ever-increasing threat that cyberworld is facing today. Anti-phishers come up with their unique solutions to protect the users against this threat. However, phishers find a way to break every security layer, and finally, bring users on the phishing websites. This research explores various deep learning methods utilizing computer vision techniques and provides an additional security layer to the cyber-users. Therefore, if phishers somehow manage to land the user on the phishing websites, bypassing all the list-based or rule-based security layers; with



	adobe	alibaba	amazon	apple	boa	chase	dhl	dropbox	facebook	linkedin	microsoft	other	paypal	wellsfargo	yahoo
adobe	24	0	0	0	0	0	0	0	0	0	0	3	0	0	0
alibaba	1	22	1	0	0	0	0	0	0	0	0	1	1	0	0
amazon	0	0	4	0	1	0	0	0	1	0	1	3	1	0	0
apple	0	0	0	13	0	0	0	0	0	0	0	2	0	0	0
boa	0	0	0	0	34	0	0	0	0	0	0	0	0	1	0
chase	0	0	0	0	0	32	0	0	0	0	0	3	2	0	0
dhl	0	0	0	0	0	0	38	0	0	0	1	3	0	0	0
dropbox	0	0	0	0	0	0	0	34	0	0	0	5	1	0	0
facebook	1	0	0	0	0	0	0	1	43	0	0	8	1	2	1
linkedin	1	0	0	0	0	0	0	0	1	9	0	3	0	0	0
microsoft	0	0	0	0	0	1	0	0	3	0	36	10	2	0	1
other	2	0	1	0	3	2	0	1	2	0	1	964	11	13	0
paypal	2	0	0	0	1	2	0	1	0	0	0	15	68	4	0
wellsfargo	0	0	0	0	2	0	0	0	0	0	0	6	0	37	0
yahoo	0	0	0	0	0	0	0	1	0	0	0	4	0	0	39

Fig. 12 Confusion matrix for DenseNet201

Table 5 Comparison of best model with SOTA

Metrics	Phish-IRIS SOTA (%)	DenseNet201-SVM-512 (%)
True positive rate	90.60	90.80
False positive rate	8.50	1.20
$F1$-measure	90.50	90.60

the proposed deep learning methods, quick prediction of a legitimate or a phishing page is recommended by the proposed method in this research.

A detailed literature review is performed comparing the existing phishing detection methods, and it is observed that most of the state-of-the-art phishing detection methods utilize hand-crafted feature extraction techniques. Hence, there is a need for deep learning methods that can automate the feature extraction process. Hence, this

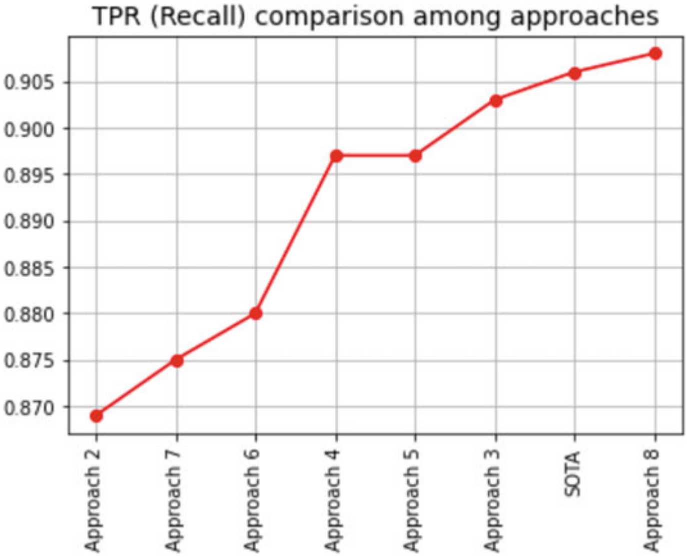


Fig. 13 Comparison based on recall among all approaches and SOTA

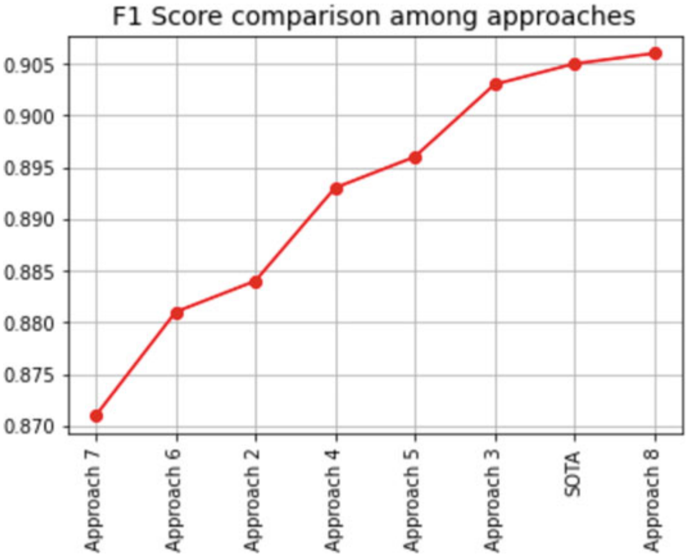


Fig. 14 Comparison based on $F1$ score among all approaches and SOTA

research provides a computationally efficient solution based on Transfer Learning techniques. Various state-of-the-art CNN architectures are trained on the standard public dataset that includes both phishing and legitimate webpages. It is observed that DenseNet201 outperforms all the existing state-of-the-art CNN architectures experimented in this research, with a True Positive Rate of 90.8%.

In this research, experiments are performed with traditional CNNs, transfer learning, and representation learning approaches. The features are extracted from the convolutional layers of various off-the-shelf CNN-based models and ingested in a machine learning and deep learning-based classifiers. This research is a First of a Kind (FOAK), as the aforementioned techniques are not utilized in the phishing detection domain that utilizes computer vision methodologies.

In this research, we experimented with the image sizes of (128 * 128), (256 * 256) and only for few iterations we utilized the image size of (512 * 512). Hence, for future work, images of bigger sizes can be utilized to extract features and experiment. Also, we utilized SVM and RF algorithms to build the machine learning classifier. Hence, for future works, other machine learning algorithms like XGBoost, etc. can be experimented with to see if they show any significant improvement in the performance metrics. Finally, in this research, we utilized CNN architectures that have support from python Keras library like VGG, ResNet, Inception, DenseNet, etc. For future work, I recommend experimenting with more CNN state-of-the-art architectures like Efficient Net, etc. to extract image features for experiments.

References

1. Phishing Definition & Meaning | What is Phishing?. https://www.webopedia.com/definitions/phishing-meaning/. Accessed Jan 30, 2021
2. Jain, A.K., Gupta, B.B.: A novel approach to protect against phishing attacks at client side using auto-updated white-list. Eurasip J. Inf. Secur. **1**, 2016 (2016). https://doi.org/10.1186/s13635-016-0034-3
3. Prakash, P., Kumar, M., Kompella, R.R., Gupta, M.: PhishNet: Predictive blacklisting to detect phishing attacks. In: 2010 Proceedings IEEE INFOCOM, Mar 2010, pp. 1–5. https://doi.org/10.1109/INFCOM.2010.5462216
4. Google Safe Browsing|Google Developers. https://developers.google.com/safe-browsing. Accessed Jan 30, 2021
5. Zero-day (computing)—Wikipedia. https://en.wikipedia.org/wiki/Zero-day_(computing). Accessed Jan 30, 2021
6. Varshney, G., Misra, M., Atrey, P.K.: A survey and classification of web phishing detection schemes. Secur. Commun. Netw. **9**(18), 6266–6284 (2016). https://doi.org/10.1002/sec.1674
7. Jain, A.K., Gupta, B.B.: Phishing detection: analysis of visual similarity based approaches. Secur. Commun. Netw. **2017**(i), 1–20 (2017). https://doi.org/10.1155/2017/5421046
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
9. Bay, H., Tuytelaars, T., Van Gool, L.: LNCS 3951—SURF: speeded up robust features. Comput. Vis. ECCV 2006, 404–417 (2006) [Online]. Available https://doi.org/10.1007/11744023_32
10. Li, B., Cheng, K., Yu, Z.: Histogram of oriented gradient based GIST feature for building recognition. Comput. Intell. Neurosci. **2016** (2016). https://doi.org/10.1155/2016/6749325

11. Nhat, H.T.M., & Hoang, V.T.: Feature fusion by using LBP, HOG, GIST descriptors and canonical correlation analysis for face recognition. In: 2019 26th International Conference on Telecommunication ICT 2019, pp. 371–375, 2019.https://doi.org/10.1109/ICT.2019.8798816
12. Tola, E., Lepetit, V., Fua, P.: DAISY: An efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 815–830 (2010). https://doi.org/10.1109/TPAMI.2009.77
13. Rayar, F.: ImageNet MPEG-7 Visual Descriptors—Technical Report, pp. 21–23, 2017 [Online]. Available http://arxiv.org/abs/1702.00187
14. Dunlop, M., Groat, S., Shelly, D.: GoldPhish: using images for content-based phishing analysis. In: 5th International Conference on Internet Monitoring Protection ICIMP 2010, pp. 123–128, 2010https://doi.org/10.1109/ICIMP.2010.24
15. Wang, G.: Verilogo: Proactive Phishing Detection Via Logo Recognition. Jan 2010, pp. 1–20 (2010) [Online]. Available http://escholarship.org/uc/item/6m26d488.pdf
16. Rao, R.S., Ali, S.T.: A computer vision technique to detect phishing attacks. In: 2015 Fifth International Conference on Communication Systems and Network Technologies, Apr 2015, pp. 596–601. https://doi.org/10.1109/CSNT.2015.68
17. Chen, K.T., Chen, J.Y., Huang, C.R., Chen, C.S.: Fighting phishing with discriminative keypoint features. IEEE Internet Comput. **13**(3), 56–63 (2009). https://doi.org/10.1109/MIC.2009.59
18. Bozkir, A.S., Sezer, E.A.: Use of HOG descriptors in phishing detection. In: 2016 4th International Symposium on Digital Forensic and Security (ISDFS), Apr 2016, no. 2013, pp. 148–153. https://doi.org/10.1109/ISDFS.2016.7473534
19. Dalgic, F.C., Bozkir, A.S., Aydos, M.: Phish-IRIS: a new approach for vision based brand prediction of phishing web pages via compact visual descriptors. In: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Oct 2018, pp. 1–8. https://doi.org/10.1109/ISMSIT.2018.8567299
20. Eroğlu, E., Bozkir, A.S., Aydos, M.: Brand recognition of phishing web pages via global image descriptors. Eur. J. Sci. Technol. 436–443. https://doi.org/10.31590/ejosat.638397
21. Bozkir, A.S., Aydos, M.: Local image descriptor based phishing web page recognition as an open-set problem. Eur. J. Sci. Technol. 444–451 (2019). https://doi.org/10.31590/ejosat.638404
22. Wu, J., Cui, Z., Sheng, V.S., Zhao, P., Su, D., Gong, S.: A comparative study of SIFT and its variants. Meas. Sci. Rev. **13**(3), 122–131 (2013). https://doi.org/10.2478/msr-2013-0021
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representation ICLR 2015—Conference on Track Proceedings, pp. 1–14, 2015
24. Szegedy, C. et al.: Going deeper with convolutions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07–12-June, pp. 1–9, 2015. https://doi.org/10.1109/CVPR.2015.7298594
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computing Society Conference on Computer and Vision Pattern Recognition, 2016-Dec, pp. 770–778, 2016. https://doi.org/10.1109/CVPR.2016.90
26. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of 30th IEEE Conference on Computer and Vision Pattern Recognition, CVPR 2017, 2017-Jan, pp. 2261–2269, 2017. https://doi.org/10.1109/CVPR.2017.243
27. Wilson, J., Arif, M.: Scene Recognition by Combining Local and Global Image Descriptors, 2017 [Online]. Available http://arxiv.org/abs/1702.06850
28. Jun, H., Ko, B., Kim, Y., Kim, I., Kim, J.: Combination of Multiple Global Descriptors for Image Retrieval, 2019 [Online]. Available http://arxiv.org/abs/1903.10663
29. Gao, H., Chen, W.: Image classification based on the fusion of complementary features. J. Beijing Inst. Technol. English Ed. **26**(2), 197–205 (2017). https://doi.org/10.15918/j.jbit1004-0579.201726.0208
30. Phish-IRIS Dataset—A Small Scale Multi-Class Phishing Web Page Screenshots Archive. https://web.cs.hacettepe.edu.tr/~selman/phish-iris-dataset/. Accessed Nov 29, 2020

A Comprehensive Attention-Based Model for Image Captioning

Vinod Kumar, Abhishek Dahiya, Geetanjali Saini, and Sahil Sheokand

Abstract Image captioning/automatic image annotation is referred to as description of image in text according to the contents and properties observed in a picture. It has numerous implementations such as its utilisation in virtual assistants for people with visual impairment, for social media and several other applications in computer vision and deep learning. Another interesting application is that a video can be explained frame by frame by image captioning (considering it to be carousel of images). In this paper, we have used an encoder–decoder architecture along with attention mechanism for captioning the images. We have used layers of CNN in the form of an encoder and that of RNN as decoder. We used Adam optimiser which gave the best results for our architecture. We have used Beam Search and Greedy Search for evaluating the captions. BLEU score was calculated to estimate the proximity of the generated captions to the real captions.

Keywords Attention mechanism · Encoder–decoder architecture · Automatic image annotation.

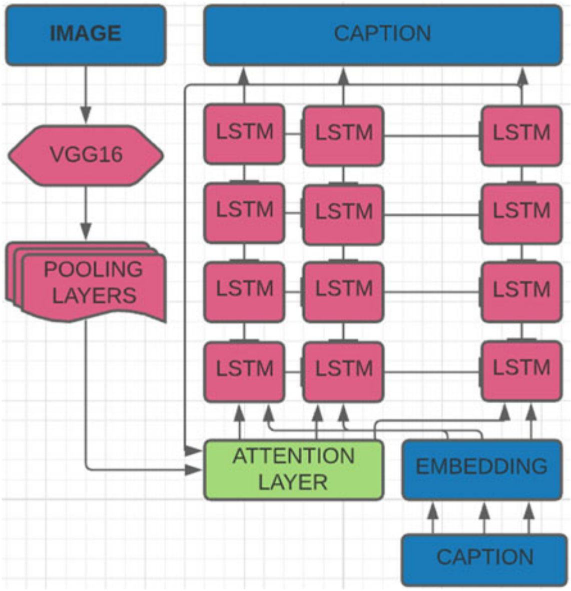
1 Introduction

Image captioning [1–3] involves generation of a human decipherable text form description of an image. Contrary to human understanding, it is quite exigent for a machine such as a computer as it requires recognition of what constitutes an image as well as how to transfer this interpretation into human understandable textual language. With time, deep learning models have overcome statistical methods and are able to produce state of the art results for the problem in almost all of the topics of research in present time. We aim to understand how deep learning methods can be utilised to generate textual explanation of images. An encoder–decoder architecture [1, 4], which is used for solving the image description creation problem, involves two components [1]:

V. Kumar · A. Dahiya · G. Saini (✉) · S. Sheokand
Delhi Technological University, Delhi 110042, India
e-mail: vinod_k@dtu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_10

Fig. 1 Working of encoder–decoder with a feature vector



1. Encoder: A NN model that inputs the given image and combines the components of that image into a vector, of predefined constant length, called the feature vector.
2. Decoder: A NN model that is given the feature vector and has to generate captions for it.

CNN [5, 6] is used for extraction of components of the images converting them into feature vectors. A RNN [7, 8] such as a LSTM [9, 10] is used to generate the current timestamp word in the sequence given the generated feature vector and the text sequence predicted until the current timestamp. The encoder–decoder neural architecture is inadequate to catch the substance of the whole image all at once [2]. The word it produces is just describing a fragment of the image. It becomes an issue. Attention mechanism helps solve this problem by determining proper weightage to different sections of the input image and, thus, selects most useful elements from the larger input image data (Fig. 1).

2 Related Work

Different techniques are implemented in various research papers to tackle image captioning. Reference [11] follows an encoder–decoder configuration which uses abstract image feature vectors as input . A new object relation transformer model is introduced which incorporates data about the connection between input detected articles through geometric attention. Quantitative as well as qualitative results illustrate the need for such geometric attention for developing captions of images, directing to improvements on all common captioning metrics on the MS-COCO dataset. In [12], a unique sequence transduction model algorithms are used which are based on RNNs and CNNs. An attention mechanism layer is sandwiched between the layer of RNN and CNN for better performance as with the attention model, and the architecture is time efficient to train. With 165,000,000 parameters, the model scores 27.5 % BLEU on English-to-German translations. On English-to-French, the model outperformed the current best model by 0.7 BLEU, while achieving a BLEU score of 41.1. In [13], multilayer convolutional neural network (CNN) to generate vocabulary is used for image description and a long short-term memory (LSTM) to precisely structure purposeful description using the produced words. A sequential API of Keras was implemented with Tensorflow as a backend to use the deep learning architecture to achieve a BLEU score of 0.683 for this model. Reference [14] tells us that with the availability of millions of images on the Internet, we can create more sophisticated and vigorous models to organise the images and ensure maximum interaction with them. Hence, they introduce a robust new DB called "ImageNet", a large-scale ontology of pictures which is created upon the structure of the WordNet. ImageNet strives to populate the greater part of the 80k Synonym rings of WordNet with more than five hundred clear and high resolution images. Organization of such a large number of images will happen by the connotation hierarchy of WordNet. This paper provided an in-depth analysis of ImageNet, whose current state comprises 12 subtrees with 5247 Synonym rings and 3.2 million images in total. The database ImageNet being large in size and having more diversity came out to be more precise than the image datasets which are used. Reference [15] states that since most of the caption generating models use encoder–decoder architecture for this problem, it introduces a new methodology which integrates "policy-network" and a "value network" to produce picture captions. The policy network predicts the next state according to current state, whereas the value network estimates all the possible extensions of current state. An actor critic reinforcement (ACR) learning model is used to train the network. Different evaluation techniques are subsequently used to estimate the precision of the architecture. Reference [16] proposes that since we know that human methodologies of evaluation of a model take ample amount of time and can be ineffective and impractical, therefore we need a machine translation evaluation which is effective and inexpensive and can be reused an indefinite number of times. Thus, the paper proposed the use of a numerical metric for translation. BLEU score can be implemented for evaluation which is a much smarter way to evaluate.

3 Method

The proposed architecture of our model comprises mainly three parts. The first part is a feature extractor that is a CNN, the second part is the attention layer that is responsible for giving proper weightage to the extracted features and the final third part being RNN that is responsible for generation of captions using the partial caption generated so far along with the weighted extracted features.

3.1 CNN

CNN is used for extraction of components of the images converting them into feature vectors. We used a pre-trained CNN model—VGG16 on ImageNet dataset that has already learned to classify images to encode our photograph [17]. This process of using a pre-trained model is called transfer learning [18]. VGG16 (also called OxfordNet) is a convolutional neural network architecture. In VGG16, there are 16 layers having tunable parameters of which 13 are convolutional layers and 3 are fully connected.CNNs when used in image classifications will determine the components of the input image and try to classify it under one of the given classes (for example, ball, dog, human). The input image is interpreted by the computer as an array of pixels. The 13 image will be a 3D tensor in case of RGB with a value between 0 and 255 determining the intensity of the colour. CNN consists of a combination of convolution layers with filters (Kernels), Pooling (Fig. 2).

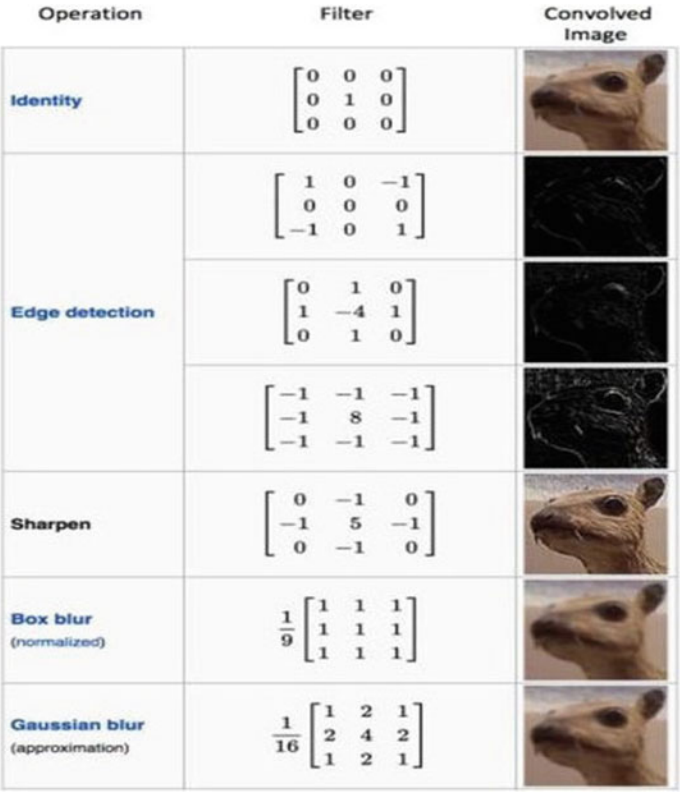
Pooling layers provide an approach to down sampling feature maps by summarising the presence of features in patches of the feature map. Two common pooling methods are average pooling and max pooling that summarise the average presence of a feature and the most activated presence of a feature, respectively.

1. Max Pooling—taking the largest value from each patch of the feature map.
2. Average Pooling—taking the average of all the values a patch of the feature map (Fig. 3).

3.2 RNN

RNN is a neural network that has a memory. RNNs are recursive in nature because they use the same procedure for each time stamp data input, while the output of the present timestamp input is determined on the previous timestamp data. The output is fed back into the recurrent network as the output of the last timestamp. Unlike multi-layered perceptrons, RNNs can utilise their cell state (memory) to process input sequences which makes them relevant to a distinct domain of tasks (such as speech recognition or unsegmented, connected handwriting recognition).

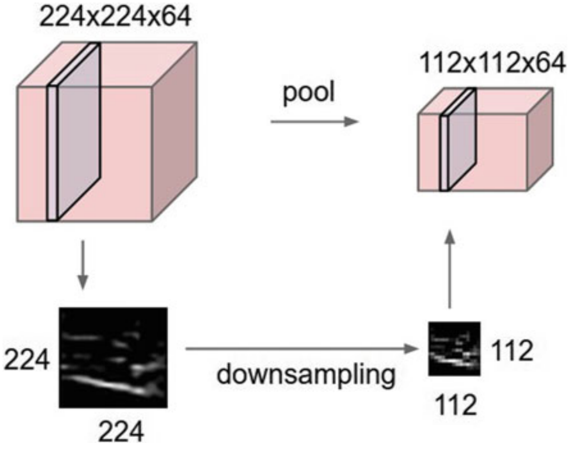
The problem with RNNs is their short-term memory. Therefore, transferring information from initial timestamps to subsequent ones is tough if the sequence is lengthy enough. While processing a paragraph, RNNs could miss out on important information from the initial parts. RNNs have the vanishing gradient problem during backpropagation. Gradient values are used to upgrade a neural network's weights.



Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

Fig. 2 The following image shows different filters or kernels used for different types of feature extraction such as edge detection, curve detection, sharpen, etc. The kernel is a matrix with values 0s and 1s which when multiplied by the input image matrix gives information about different features of the image

Fig. 3 Pooling in CNN



When the gradient shrinks during its propagation through time, it does not contribute to much learning as gradient value becomes extremely small and the product of small numbers(less than 1) tends to zero. So in RNNs, layers which get a small gradient upgrade stop learning (generally the earlier layers). Therefore, RNNs can fail to remember, thus having a short-term memory. This could be a problem in long sequences.

The working of LSTM [19] depends on the cell state, input gate, forget gate and output gate (Fig. 4). The cell state transmits confined knowledge all the way down the sequence chain. Details from the previous time steps go to later time steps, thus diminishing the repercussions of short-term memory. As the information flows through the cell state, data gets summed up or removed via gates. The gates are distinct neural networks which can determine what data is allowed to stay in the cell state and what is discarded. With the help of backpropagation, the gates are imparted the knowledge of what data is needed in the cell state and which data can be ignored.

We start by giving an initial hidden state from the encoder(NIL) to our LSTM along with the start marker (< START>). On each step, the LSTM outputs a probability distribution for the next word, over the entire vocabulary. We pick the highest probability word, add it to the caption, and feed it back into the LSTM. This is repeated until the LSTM generates the end marker (Fig. 5).

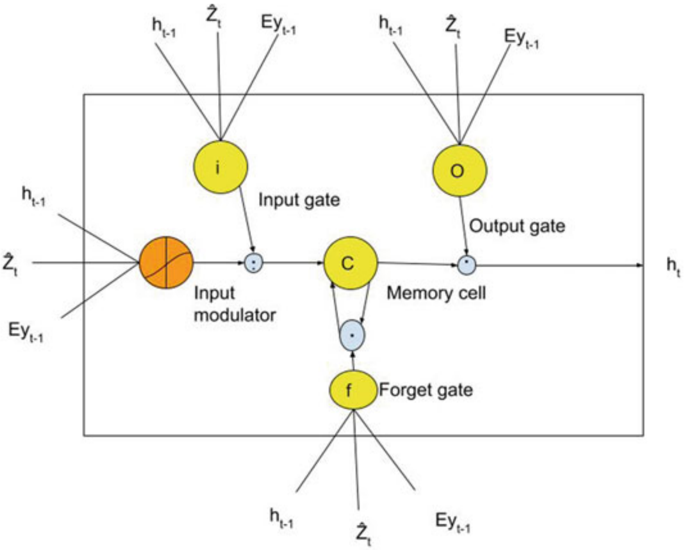


Fig. 4 Working of LSTM

3.3 Attention Mechanism

The encoder–decoder model is inadequate to catch the essence of the entire input image all at once [2]. The word it produces is just describing a fragment of the image. It becomes an issue. By using the attention mechanism, the image is first dissected into n parts and CNN feature extraction of each part is computed, say, $h_1, h_2, ..., h_L$ [2, 20]. When any new word is generated by the RNN, the attention mechanism is focussed on the relevant part of the image, and therefore, the decoder uses that fragment of image only [21], thus making better predictions about the image (Fig. 6). It can be said that the architecture involving only an encoder and a decoder did not do the translation task the way human beings are designed to do as human beings do not read the whole paragraph, understand it and then translate it to another language. Rather, they read a portion of the sentence and translate it simultaneously, which is more like the attention mechanism.

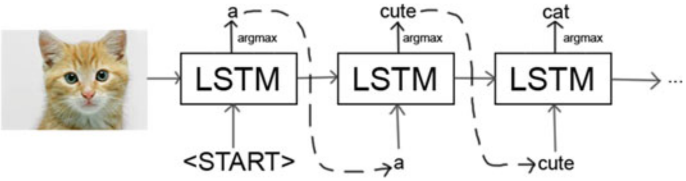


Fig. 5 Process of prediction of words using LSTM

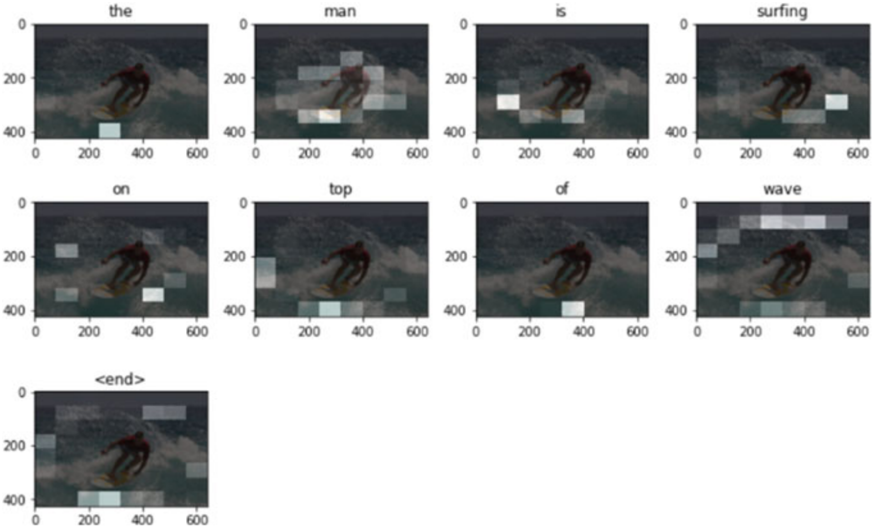


Fig. 6 An example of attention mechanism. The white areas/pixels have higher attention for the corresponding word in the caption

The attention weights for fragmants of an image are calculated by the equations given below. In the following equation, z_t the context vector and a_j $(a1, a2, a3\ldots)$ are the features extracted by the encoder. The context vector z_t is a dynamic representation of the relevant part of the input image at time t. We define a mechanism ϕ that computes z_t from the annotation vectors a_i , i = 1, . . . , L corresponding to the features extracted at different image locations. For each location i, the mechanism generates a positive weight α which can be interpreted either as the probability that location i is the right place to focus for producing the next word (the "hard" but stochastic attention mechanism), or as the relative importance to give to location i in blending the a_i's together. The weight α_i of each annotation vector a_i is computed by an attention model f_{att} for which we use a multilayer perceptron conditioned on the previous hidden state h_{t-1} [1]

$$e_{ti} = f_{\text{att}}(a_i, h_{t-1}) \tag{1}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \tag{2}$$

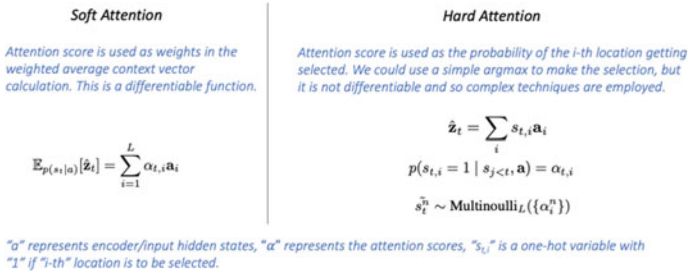
Once the weights (which sum to one) are computed, the context vector z_t is computed by:

$$z_t = \phi(\{a_i\}, \{\alpha_i\}) \tag{3}$$

where ϕ is a function that returns a single vector given the set of annotation vectors and their corresponding weights. The details of ϕ function depends upon the type of attention we used. Attention can be either hard or soft.

3.3.1 Hard Versus Soft Attention

Soft attention is when we calculate the context vector as a weighted sum of the feature vector. Hard attention is when, instead of weighted average of all feature vector, we use attention scores to select a single feature vector.



Soft Attention

Attention score is used as weights in the weighted average context vector calculation. This is a differentiable function.

$$\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i}\mathbf{a}_i$$

Hard Attention

Attention score is used as the probability of the i-th location getting selected. We could use a simple argmax to make the selection, but it is not differentiable and so complex techniques are employed.

$$\hat{\mathbf{z}}_t = \sum_i s_{t,i}\mathbf{a}_i$$

$$p(s_{t,i} = 1 \mid s_{j<t}, \mathbf{a}) = \alpha_{t,i}$$

$$\tilde{s}_t^n \sim \text{Multinoulli}_L(\{\alpha_i^n\})$$

"a" represents encoder/input hidden states, "α" represents the attention scores, "$s_{t,i}$" is a one-hot variable with "1" if "i-th" location is to be selected.

3.4 *Algorithm*

3.4.1 Data Preprocessing

The image size is reshaped to 224 * 224 * 3 since we are using VGG16 model. For the purpose of fast computation on the dataset, we had to split it in batches of size 64 with buffer size of 1000 and embedding dimension of 256. The captions are preprocessed by adding "<start>" and "<end>" tags to each and every caption in our dataframe. This helps our machine learning model to understand when the caption starts and when it ends. The preprocessing steps on the captions are as follows:

1. First step is to tokenize the captions. Tokenization is done to create a vocabulary of all the distinct words and to get rid of the redundant words or characters.
2. Next, the vocabulary was restricted to top 5000 words. We replaced all the other words with the token <unk> (for words not in vocabulary). The restriction and optimisation can be done according to the use cases.
3. Finally, we generated a word to index mapping (word->index) and vice versa along with converting words into numerical data using word2vec [22].
4. Since we need all the sequences to be the same length, padding is done on the shorter sequences to make them to be the same length as the longest one.

3.4.2 Training Step

1. The output of CNN, i.e. VGG16, initial timestamp hidden state (0 at the beginning) and the decoder input (i.e. <start> character showing start of the caption) are fed to the decoder.
2. The decoder outputs the next character and its present timestamp hidden state.
3. The decoder's previous timestamp hidden state is fed back to the neural network again, and the projections are then used to compute the metrics.
4. According to the teacher forcing phenomenon, the target word serves as the next input to the RNN decoder. It is used to acquire an understanding of the correct sequence from the sequence, quickly and with accuracy.
5. The last step is the calculation of the gradient. Then it is applied to the optimizer, and backpropagation is done.

3.4.3 Optimiser

Adam optimiser gave the best results for our architecture. Adaptive moment estimation (Adam) works with 1st and 2nd momentums [23]. The motivation behind the optimiser is that we strive to carry out a careful search which shall give better results and try not jumping over the minimum. Other optimisers which can be used are stochastic gradient descent, Adam, Adagrad.

3.4.4 Testing Step

This step is similar to what is done in the training step except the gradients are not updated. The predicted output is given as an input to our decoder (RNN cell) at subsequent timestamps. One of the main purposes of the test step is also to determine if the proposed architecture is overfitting or not.

4 Experiment

4.1 Dataset

Multiple open-source datasets are available for such problems. For example, Flickr 8k, Flickr 30k, MS-COCO, nocaps, etc. To accomplish the aim of this project, we have made use of the Flickr 8k dataset and COCO dataset. Flickr 8k dataset comprises 8000 images, each having five captions. Our training set will comprise of 6000 images with 1000 images in the Dev set and 1000 images in the test set, respectively.
Flickr8k_Dataset: Consisting of all the images
Flickr8k_text: Consisting of all the captions
The MS-COCO dataset comprises 180k images, each having five captions. We used 120k images for training, 10k images in the dev set and 50k images for testing.
MS-COCO_Dataset: Consisting of all the images
MS-COCO_text: Consisting of all the captions
Inorder to analyse and estimate the performance of our model, we held back a section of data from the given dataset while training our model. This data is known as Dev set or as the validation set is used during hyper-parameter tuning, i.e. to optimise the model in the development process. The Dev set or the validation set can be very useful in finding problems like overfitting or underfitting in our model.

4.2 Metrics

Bilingual evaluation understudy score (BLEU) is an algorithm to assess a generated sentence to a reference sentence [24]. It evaluates machine-generated texts. A score of 1.0 means a perfect match. In contrast, 0.0 means a perfect mismatch. The score was designed for estimating the outputs of the machine translation systems. BLEU score is quick and inexpensive, quite easy to understand, language independent, agrees with human evaluations and is broadly accepted. BLEU score approach performs by taking into account the matching n-grams in the candidate translation to n-grams in the reference text. Unigram means a token. Bigram means each word pair. The differentiation is made without regard to the order of the words. The Python NLTK (Natural Language Toolkit library) provides a function to calculate BLEU score that can be used to assess the machine-generated text against a reference text. There are two types of BLEU score:

1. Sentence BLEU Score—sentence_bleu() assesses a candidate sentence against single or multiple reference verdicts.
2. Corpus BLEU Score—corpus_bleu() evaluates the BLEU score for more than one sentence like a paragraph or a document.

4.3 Platform of Implementation

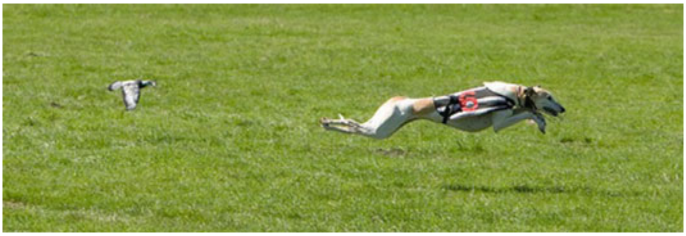
The whole process of this experiment (preprocessing the data, feature extraction, the training step, implementing the attention mechanism and the RNN decoder, optimising and the testing step) have been carried out using Python Language(v3.8.3) on the open-source Jupyter environment. We used a number of Python packages such as Numpy, Pandas, Keras, SciKit Learn, Tensorflow, OS, Matplotlib, Python Imaging Library, pickle and Natural Language Toolkit (NLTK).

5 Results

We analysed the captions generated by our model according to their BLEU scores. We used two approaches to evaluate the captions:

1. Greedy Approach: It chooses the word which has the maximum probability. That is why it is called 'Greedy'. It works on the principle of maximum likelihood estimation (MLE). MLE is a statistical mode to obtain the values of the parameters which gives the most probable output/result.
2. Beam Search: The best 'k' states are given to the model as input. Beam Search processes every part, and every time the best translation (the highest probability, as determined by the model) is returned. The algorithm keeps on doing this till it reaches the translational limit or reaches the defined <end>. The Beam Search algorithm has been found out to be very effective for machine translational systems

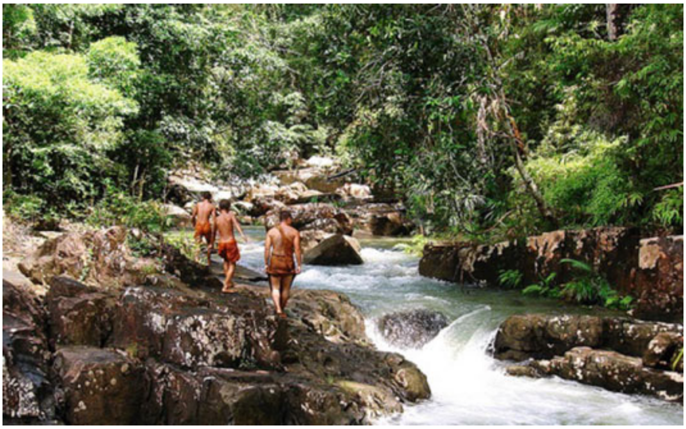
Some Captions generated with BLEU scores:



REAL CAPTION: White and tan dog leaps through the air **CAPTIONS GENERATED**:

- Greedy Search: Dog in harness is running through the grass (BLEU = 20.96)
- Beam Search (*b* = 3): Dog is leaping through the grass (BLEU = 68.46)
- Beam Search (*b* = 7): Dog is leaping through field (BLEU = 73.19)
- Beam Search (*b* = 10): A dog is running through a grassy field (BLEU = 50.0)

OBSERVATION: Beam Search (*b* = 7) is more meaningful here.



REAL CAPTION: Three people walking along a stream in a jungle
CAPTIONS GENERATED:

- Greedy Search: Three people are walking along a rocky mountain (BLEU = 45.67)
- Beam Search (*b* = 3): Four people are walking on a mountain(BLEU = 49.19)
- Beam Search (*b* = 7): Three people are walking along a rocky mountain (BLEU = 45.67)
- Beam Search (*b* = 10): Three people are walking across a river (BLEU = 23.19)

OBSERVATION: Beam Search (*b* = 7) and Greedy Search are more meaningful here

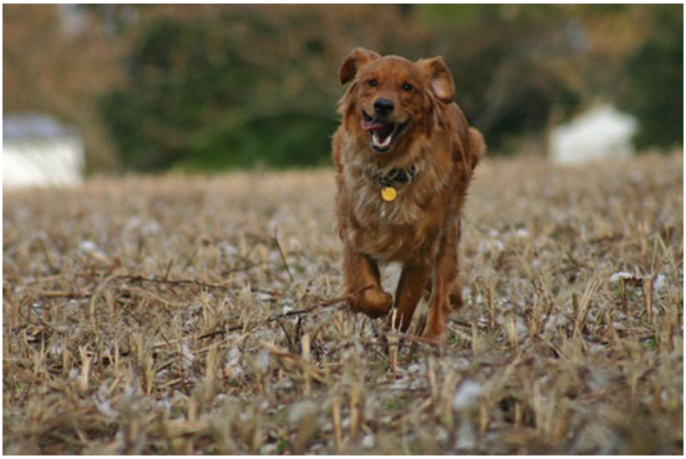


Table 1 With proper hyper-parameter tuning, our architecture gave the following BLEU Scores on COCO and flickr8k dataset (with beam size = 3)

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Flickr 8k	Soft attention	66.6	44.7	29	19.3
	Hard attention	66.6	45	32	21.2
COCO	Soft attention	70.3	53.6	39.8	29.5
	Hard attention	71	50.1	35.5	24.9

REAL CAPTION: Brown dog with his tongue out as he runs through the field
CAPTIONS GENERATED:

- Greedy Search: Brown dog is running through the camera (BLEU = 15.09)
- Beam Search (*b* = 3): Brown dog is running through the grass (BLEU = 62.8)
- Beam Search (*b* = 7): Brown dog is running through the grass with his tongue out of grass with an orange toy in the dead leaves (BLEU = 37.7)
- Beam Search (*b* = 10): Brown dog is running through the grass (BLEU = 62.8)

OBSERVATION: Beam Search (*b* = 10) is more meaningful here.
Figure 7 depicts how the loss while training our model on MS-COCO dataset is reduced after each epoch. The loss is due to the wrong predictions made by our decoder in comparison with the labelled caption word that was expected to be predicted. These losses are backpropagated over time into the model so that the model can learn from its mistakes, adjust the weights, gradients of layers in our model and make better predictions next time. As seen in the graph, the loss is constantly

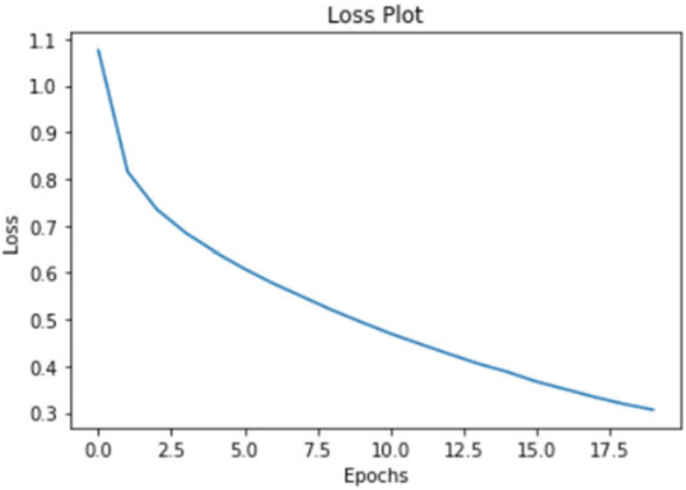


Fig. 7 The loss versus epochs graph for COCO dataset

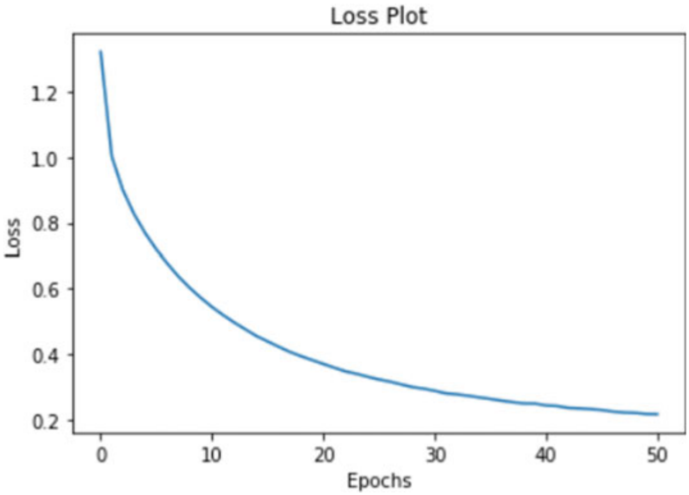


Fig. 8 The loss versus epochs graph for flickr8k dataset

decreasing which shows that the model is understanding how to correctly describe images on it's own.

Similarly, Fig. 8 shows this phenomenon of backpropagation of loss over time and the model learning on how to correctly describe an image on Flickr8k dataset as our training dataset

6 Conclusion and Future Work

The model has done a fairly good job in creating captions from images (Table 1). Beam Search helped in obtaining better outcomes than Greedy Search. The images used during testing and the images used during training of the model should be similar. Like, if the model is coached using the images of different animals like cats, horses, cows, etc., it should not be tested on the images of fruits, mountains, cars, etc. For example, training the model on images of different fruits and testing on images of different cars; in such cases, it is hard to design a ML model which could give a satisfactory result.

For future work, we can try to further improve our model by adding batch normalisation layer which will standardises the inputs to a layer for each mini batch.This stabilises the learning step and reduces the number of epochs used to train the architecture. To avoid the issue of overfitting, we can tune our model with different dropout. A dropout makes the model to learn more robust features by ignoring some units(neurons) during the training step. We can do more hyper-parameter tuning (learning rate, number of layers, number of units, batch size, etc.). Another integral part of computer vision is object detection and image segmentation [24, 25]. Object detection aids in pose estimation, vehicle detection, surveillance, etc. Thus, we will improve upon our model's ability to differentiate and detect all different objects in the image faster and in a more robust way thus improving decoder's ability to predict the correct word with these robust features. Therefore, algorithms like R-CNN, YOLO, etc., can be used. Along with attention mechanism, we can make use of the difference techniques like XAI, IML which can improve the model's explanation. Interactive ML (IML) adds the component of human expertise to AI/ML processes by enabling them to re-enact and retrace AI/ML results, while explainable AI (XML) studies transparency and traceability of opaque AI/ML model thus making the results more reliable. We can further try to integrate the generated captions with an audio mechanism using deep learning. An image to speech system can generate a spoken description of a photograph directly, without first generating text or generating text if we need.

References

1. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057. PMLR, 2015
2. Vinyals, O., et al.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164
3. Lu, J., et al.: Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 375–383
4. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
5. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. IEEE, 2017
6. O'Shea, K., Nash, R.: An introduction to convolutional neural networks. arXiv:1511.08458 (2015)
7. Schuster, M., Paliwal, K.K.: Networks bidirectional reccurent neural. IEEE Trans. Signal Process. **45**, 2673–2681 (1997)
8. Sundermeyer, M., et al.: Comparison of feedforward and recurrent neural network language models. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8430–8434. IEEE, 2013
9. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. J. Mach. Learning Res. 115–143 (2002)
10. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing LSTM language models. arXiv:1708.02182 (2017)
11. Herdade, S., et al.: Image captioning: transforming objects into words. arXiv:1906.05963 (2019)
12. Vaswani, A., et al.: Attention is all you need. arXiv:1706.03762 (2017)
13. Papineni, K., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, 2002
14. Lakshminarasimhan Srinivasan, D.S., Amutha, A.L.: Image captioning–a deep learning approach. Int. J. Appl. Eng. Res. **13**(9), 7239–7242 (2018)

15. Deng, J.: Imagenet: a large-scale hierarchical image database. IEEE Conf. Comput. Vis. Pattern Recogn. **2009**, 248–255 (2009)
16. Ren, Z., et al.: Deep reinforcement learning-based image captioning with embedding reward. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 290–298
17. Qassim, H., Verma, A., Feinzimer, D.: Compressed residual-VGG16 CNN model for big data places image recognition. In: IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp. 169–175. IEEE (2018)
18. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big data **3**(1), 1–40 (2016)
19. Harika, G., et al.: Building an Image Captioning System Using CNNs and LSTMs. Int. Res. J. Mod. Eng. Technol. Sci. **2**(6) (2020)
20. Fukui, H., et al.: Attention branch network: Learning of attention mechanism for visual explanation. In: Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition, 2019, pp. 10705–10714
21. Li, L., et al.: Image caption with global-local attention. Proc. AAAI Conf. Artif. Intell. **31**(1) (2017)
22. Kottur, S., et al.: Visual word2vec (vis-w2v): learning visually grounded word embeddings using abstract scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4985–4994
23. Zakir Hossain, M.D., et al.: A comprehensive survey of deep learning for im- age captioning. ACM Comput. Surv. (CsUR) **51**(6), 1–36 (2019)
24. Manoharan, S.: Performance analysis of clustering based image seg- mentation techniques. J. Innov. Image Process. (JIIP) **2**(01), 14–24 (2020)
25. Chawan, A.C., Kakade, V.K., Jadhav, J.K.: Automatic detection of flood using remote sensing images. J. Inf. Technol. **2**(01), 11–26 (2020)

Multimedia Text Summary Generator for Visually Impaired

Shreya Banerjee, Prerana Sirigeri, Rachana B. Karennavar, and R. Jayashree

Abstract With the advancing methodologies in the field of NLP, text summarization has become an important application and is still under research.Taking into consideration today's busy world, time is a most important factor for everyone and people do not bother to make time to listen to long audio news or read long news articles, and as the visually impaired are an important part of our society, it becomes still difficult for them to read although there is braille, but it is inconvenient for them every time to read through braille. The aim is to make use of advancing technology to make their lives easier. This gave us inspiration that led to the idea of generating concise and short summaries for the visually impaired. Our system makes use of various APIs like speechrecognition, pyspeech, Google Cloud Speech API, etc., to extract text and then use summarisation techniques to present the most accurate summary and convert it back to audio format so that the news is more accessible for the visually impaired.

Keywords Summarisation · NLP · API's · Speechrecognition · Google cloud speech API · BERT · BART · GPT-2 · ROUGE

1 Introduction

Summarization of text has gained a lot of importance in recent times because of the excess of data available online and the lack of time in people's lives to give enough attention to it. This is resulting in increasing demand of more ca pable text summarizers. As the volume of data being generated and consumed is growing day by day, representing information in the smallest possible form, while preserving its meaning has become an important task. It is a tedious task to read long articles to get all the relevant information, instead, reading well- formed summaries to get the maximum information is way more convenient.

S. Banerjee (✉) · P. Sirigeri · R. B. Karennavar · R. Jayashree
Department of CSE, PES University, Bangalore, India
e-mail: jayashree@pes.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_11

A huge number of visually impaired individuals cannot read news reports like sighted individuals, and they need to peruse braille with their fingerprints, but it is an extremely inefficient methodology because it takes a lot of time and needs several steps to be followed. Our goal is to summarize speeches or articles in audio or text format by applying various NLP techniques and convert them into audio format so that the information is conveyed effectively to the visually impaired.

Summarization can be defined as an activity of compressing a text document into a shorter version while keeping its key contents. Even the text is shorter, it should mean the same. As summarizing with help of human takes a lot of time and effort, automising this task would gain more popularity and therefore makes this an interesting subject for research.

Text summarization has applications in various NLP related tasks like text classification, question answering, legal texts summarization, news summarize- tion, and headline generation.

Text summarization helps to reduce the length of a document. The generated summaries are often used as a part of these bigger systems.

Text summarization techniques are mainly divided into two groups:

1. Extractive summarization: It involves choosing tokens and sentences from the original document and including them in the summary.
   Ex: Can be compared to a process where we use highlighter when reading something and using those highlighted sentences in the summary
2. Abstractive summarization is an efficient sort of summarization as it in- volves a way of constructing entirely new phrases and sentences to with hold the meaning of the original document.

Ex: Can be compared to an examination where we understand and write new content based on understanding, i.e., human-written summaries

The goal is to summarize articles using various natural language summarizing techniques which includes various methods like BERT, BART, GPT-2, etc. Different methods of summarisation were applied on BBC news articles dataset, and these methods were evaluated using ROUGE metric to decide upon the best method for summarizing the articles.

The goal is to also present multi-document summarisation where similar topic articles from different sources are collected and applying summarisation on those articles. Ex: Covid-19 can be taken as a topic and different articles on this can be collected from different newspapers or magazines and can be passed to summariser for the summarization.

In the following sections, we discuss about the literary work done for this project, along with the detailed description of the methods, datasets, and the metrics used.

2 Problem Statement

It is not possible for the blind to read huge articles and speeches and is a time-consuming task for a sighted individuals. The lack of ability of those individuals to browse text incorporates an immense impact on their lives. They do have some techniques available like Braille but its highly inconvenient for them and is time-consuming and needs several steps to be followed.

We are aiming to build "Multimedia Text Summary Generator For Visually Impaired."

The objective is to save time and effort by automating the process of generating accurate and short summaries. Our goal is to summarize speeches and convert them into audio format so that the information is conveyed effectively to the visually impaired. The input to our model can be either the speeches or articles in text or audio format, and with this input in order to present the summary, we use models like BERT, BART, and GPT-2 to summarize the text to a precise and short summary and converting it back to audio format to make it easily accessible to the visually impaired. The audio form of input is mainly focused on so that it is convenient for the visually impaired to get information quickly. They can save time and listen to the short version any time while traveling or doing some other work

Our system could take input in both text and audio format. The system is easily accessible by both normal and visually impaired. The output summary is in the audio format. This makes it convenient for the visually impaired to get information quickly. They can save time and listen to the short version any time while traveling or doing some other work.

3 Related Work

3.1 Summarization System for the Visually Impaired [1]

This section deals with the challenges related to summary generation for the blind and discusses methods to produce better and accurate summaries for them.

The summaries generated by current summarization systems focus on the quality of the content. Usually, these summaries are produced for people who have the gift of sight. The factors considered for generating such summaries do not include other factors important for summaries generated for the blind. Apart from content quality and fluency, length is another important factor to be considered while generating summaries for them. The blind reads with the help of their fingers. Shorter summary reduces the effort required to be put in by them. The challenge is to create short summaries while maintaining the content quality and fluency.

The traditional summarization tasks focus on improving the content quality of a summary. But while developing a summary generator for the blind, the task should aim at producing a short summary. This summary might be translated from a summary with a given length. The solution to this new summarization task might include two steps,

1. Apply an existing summarization algorithm to produce a summary.
2. Translate the summary into a braille summary.

But, the challenge is that two sentences with the same content may translate into two braille sentences with different lengths. Therefore, the braille length of each sentence should be considered in the new solution.

The summaries are evaluated from two aspects,

1. Content quality: This is evaluated by measuring the intersection of content among the human generated summaries and the reference summaries along with the ROGUE metric.
2. Braille length: This is calculated by adding the braille lengths of all the sentences in the summary.

3.2 Critical Evaluation of Neural Text Summarization [2]

This section deals with critical evaluation of the current research setup for text summarization which mainly includes the datasets, the evaluation metrics, and the models.

In most of the summarization experiments conducted till now, the majority of datasets deal with the news domain. Some of the most common news datasets used are CNN/Daily Mail, XSum. Outside the news domain, TIFU (collection of posts scraped from reddit, TL; DR summary) is one of the most popular datasets.

The problem with the current datasets used is:

1. Summarization is mostly done as an under constrained task. Assessing important information depends on prior knowledge of the user. A study was conducted to demonstrate the difficulty and ambiguity of content selection. A constrained summary is more precise, whereas an underconstrained summary is more verbose.
2. Data scraped from the web might contain noise. The quality is highly dependent on the pre-processing steps. Manual inspection to check the noise is tough.

The most popularly used evaluation tool for summaries is the ROGUE package. The basis of automatic metrics offered by this package is the grammatical overlap between candidate and reference summaries and is based on exact token matches. But, the issue is that the overlap between phrases that are grammatically different but actually mean the same is not supported. Many extensions of the ROGUE package are available such as ParaEval, ROGUE-WE, ROGUE 2.0, and ROGUE-G.

The datasets require additional constraints for the creation of good summaries. Layout bias is one of the major factors that affect the result of current methods in a negative way. The current evaluation protocol is weakly correlated to human judgments and fails to evaluate important features such as factual correctness of the summary.

3.3 Pre-trained Language Models [3]

The section discusses detailed implementation of a pre-trained language model to the application of text summarization.

Pre-trained language models can be fine-tuned to do various task-specific jobs. The aim is to use a pre-trained language model to compress a document into a shorter form while retaining most of the important information. The section deals with three main things: document encoding, ways to effectively employ pretrained models, and proposed models that can be used for the task of summarization.

Pretrained language models expand the concept of words embeddings by learning contextual representations. BERT is a advanced representational language model which is trained along with masked language modeling. The general architecture can be described as follows: Input text is given as input through three embeddings, First one being Token embeddings, which indicates the meaning of every token. Second one is segmentation embeddings which discriminates between two sentences. The last one is position embeddings that indicates the position of each token within the text sequence. Single input vector is formed by adding these embeddings and is fed to a bi-directional transformer. The final output is the output vector for each token with contextual information.

ROGUE packages R-1 and R-2 are used to assess instructiveness, whereas R-L(LCS) is used to assess fluency. BERT-based models outperform other models across all datasets.

3.4 Learning to Summarize from Human Feedback [4]

The usage of language model pretraining has led to significantly better performance. However, there is still a misalignment between this fine-tuning objective-maximizing the likelihood of human-written text and generating high-quality outputs as determined by humans. The approach to optimize for quality might just be the way to overcome these problems.

In the technique explained in this section, a dataset of human preferences between pairs of summaries is first collected, and then a reward model is trained via supervised learning to predict a summary better suited to human judgement. Finally, a policy is trained to maximize the score given by the reward model. The policy generates a token of text at each step and is updated using the PPO algorithm based on the reward model reward given to the generated summary. The main contributions are as follows:

1. Training with human feedback significantly outperforms very strong baselines on English summarization.
2. Human feedback models generalize much better to new domains than supervised models.
3. Conduction of extensive empirical analyses of the policy and reward model.

The initial policy is to fine-tune the model via supervised learning on the Reddit TL; DR summarization dataset. The process then consists of three steps that can be repeated iteratively.

1. Comparisons from the samples are sent to humans, who are tasked with selecting the best summary of a given Reddit post. Given a post and a candidate summary, learn a reward model from human comparisons.
2. A reward model is trained to predict the log odds that this summary is the better one, as judged by the labelers.
3. The logit output of the reward model is treated as a reward. It is optimized using reinforcement learning, specifically with the PPO algorithm.

Models

1. Pretrained models These models are used as baselines by padding the context with examples of high-quality summaries from the dataset.
2. Supervised baselines These models are fine-tuned via supervised learning to predict summaries from the dataset. These models are used to sample initial summaries for collecting comparisons, to initialize policy and reward models, and as baselines for evaluation.
3. Reward models To train the reward models, it is started from a supervised baseline and then add a randomly initialized linear head that outputs a scalar value. This model is trained to predict which summary is better as judged by a human.
4. Human feedback policies

The reward model trained above is used to train a policy that generates higher-quality outputs as judged by humans. This is primarily done using reinforcement learning, by treating the output of the reward model as a reward for the entire summary that is maximized with the PPO algorithm.

The aspect that makes our project differ from the existing ones is that, we have included the audio summarisation along with multi document summarisation as the next steps of the project, which will summarize the same article from different sources for a better summary.

4 Dataset

The dataset we chose for this project was the BBC NEWS SUMMARY dataset for text summarization which included 417 articles from BBC news covering between 2004 and 2005. There are five different domains of articles which include business, sports, entertainment, politics, and technology. It contains articles in a folder called news articles, and for each article, there are summaries for reference along with their summaries for these 5 domains.

This data was constructed using one other dataset used for data categorization which had 2225 different articles from BBC website correlating to different stories from 5 areas in 2004–2005 which was presented in paper "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," Proc. ICML 2006 by D. Greene and P. Cunningham whose rights are protected under BBC.

Dataset source: https://www.kaggle.com/pariza/bbc-news-summary

5 Summarization Techniques

5.1 GPT-2: Generative Pre-trained Transformer

NLP covers variety of actions like QA, classifying the documents, summarizing the text, and assessing similarity in semantics.

GPT-2 was bought into picture by an institute named OpenAI. This was trained to generate upcoming tokens in a series of tokens in an unsupervised way by using transformer architecture.

This was successor of GPT whose main reason for its decline was that it was pre-trained using traditional language modeling which was used in prediction of the following word in a sentence, which contradicted BERT which was trained using masked language modeling which meant guessing missing words given its predecessor and successor. BERT was bi-directional which helped it in giving a rich presentation and finer performance among other techniques.

It was formed as an successor of GPT with additions to number of parameters and size of the dataset. GPT-2 has around 1.5 billion attributes and was trained on data covering 8 million sites. Its goal was to come up with a sequence of tokens during the sample and predict next likely word. It can also build complete sentences and paragraphs by anticipating the additional words which will then give us meaningful sentences.

Being trained under casual language modeling's purpose, it was better in anticipating following token in a sequence. Making use of this characteristic allows GPT-2 to come up with text which are syntactically coherent.

It has unique features like answering questions, summarization of documents, and also translation between languages.

5.2 BERT

BERT or bi-directional encoder representations from transformers makes use of a transformer. In its basic form, a transformer contains two different mechanisms—(1)

an encoder to read the input text and a (2) decoder to produce a prediction for the task. It is a kind of mechanism that learns contextual relations between words in the given text.

Fine-tuning BERT for Extractive Summarization In this type of summarization, the aim is to identify the most important sentences in a text document and form a summary using those sentences. The neural encoder is used to create representations of sentences, and the classifier is to predict the important sentences that could be used in the summary. This can be seen as a task in which every sentence is labeled as 0 or 1, indicating whether to include it in the summary. To capture document-level features, several transformer layers are put together on top of BERT output. The final output layer is a sigmoid classifier. The model is used to obtain the score of each sentence. The Top-3 highest scored sentences formed the summary. Trigram blocking is used to reduce redundancy. It involves skipping a candidate sentence if there exists a trigram overlapping between that sentence and the summary.

Fine-tuning BERT for Abstractive Summarization In this type of summarization, the input and output both are sequences. Sequences of tokens are mapped to a sequence of representations by the encoder. The decoder generates target summary token-by-token. The encoder is pre-trained, whereas the decoder is a layered transformer. The inconcsistency between the encoder and the decoder might cause instability. Separate optimizers of encoder and decoder can be used. A two-stage fine-tuning approach can be used where the encoder is first fine-tuned on an extractive summarization task and then on an abstractive summarization.

5.3 BART: Bi-Directional and Auto-Regressive Transformers

BART is a form of summarisation technique which is mainly used in pretraining seq to seq models.

The process of training in this method is done in two ways:

1. An arbitrary noise function is used to corrupt the text.
2. Reconstruct the first text by learning a model.

The architecture mainly used in this method is the transformer-based neural MT architecture which also has bi-directional encoder and left to right decoder like BERT and GPT, respectively. This gives us the information that the mask which is fully visible is encoders attention mask, and hence, the decoder mask is casual.

In pretraining tasks, BART acts as a denoising autoencoder, and the pretraining task is done by the order of first sentences being randomly shuffled and a new in-filling scheme where the text spans are replaced with one mask token. BART also works for comprehensive tasks but not that effective when compared to efficiency for text generation tasks when fine-tuned.

BART having a autoregressive decoder, some of tasks such as summarization, abstract Q&A are fine-tuned most importantly for sequence generation tasks. For both of the above-mentioned tasks, the data is same as that of input, but it is also altered which relates to the objective of denoising the pretraining. This is the reason for the input of encoder being input sequence, and hence, the output generated for it by the decoder is done in autoregressive manner.

6 Metrics

Recall-oriented understudy for gisting evaluation or more popularly known as ROGUE is a set of metrics used to evaluate automatic summarization. The idea is to compare a candidate summary with a reference summary usually written by humans to evaluate the system-generated candidate summary. This sort of comparison often leads to results that are highly dependent on the reference summary. For example, a generated summary might not be identical to a given human-written summary in words but might have the same meaning. This factor is often not considered in ROGUE evaluation. There are a number of variations available for this type of evaluation:

1. ROUGE-N: evaluates a summary by checking the overlap of N-grams between the candidate and reference summaries. N-gram is primarily a set of occurring words within a given window. For example:
   ROUGE-1: evaluates a summary by checking the overlap of unigram between the candidate and reference summaries. Unigram considers each word as a single entity and hence compares the two summaries word by word.
   ROUGE-2: evaluates a summary by checking the overlap of bigram between the candidate and reference summaries. Bigrams consider two adjacent words as a single entity and hence compares the two summaries by taking two words at a time.
2. ROUGE-L: This is a longest common subsequence (LCS)-based statistics. It takes sentence level structure similarity under consideration and measures the longest matching sequence of words. The basis for this metric is that longer the LCS of two summary sentences is, more similar they are. For example, let us consider two sentences,
   S1: P Q R S T U V W
   S2: P A B Q D R S C
   Here, the LCS is PQRS.
3. ROUGE-W: This is a weighted LCS-based statistics that favors consecutive longest common subsequences. Even though ROGUE-L is a good method for evaluation, it does not consider the distance between the words in LCS. For example, let us consider three sentences,
   S1: P Q R S T U V
   S2: P A B Q U R S

S3: P Q R S A B C
Now, LCS between S1 and S2 is PQRS and LCS between S1 and S3 is also PQRS. Both will be considered equally good by ROGUE-L. However, S3 is a better summary because the distance between the words in the LCS from S3 is lesser. This factor is considered by ROGUE-W, and hence, it is considered better than ROGUE-L sometimes.

4. ROUGE-S: This is a skip-bigram-based co-occurrence statistics. For example, let us consider the following sentence,
S1: P Q R S T U
Possible skip-bigrams are: (P, Q), (P, R), (P, S), (P, T), (P, U), (Q, R), (Q, S), (Q, T), (Q, U), (R, S), (R, T), (R, U), (R, T), (S, T), (S, U), (T, U). This method of summary evaluation involves counting of common skip-bigrams between two summaries. A limit can be put on the maximum distance between the two words considered for skip-bigram to avoid spurious results. For example, if we take the limit of the maximum distance between the two words considered for skip-bigram to be 2, the possible skip-bigrams would be reduced to: (P, Q), (P, R), (Q, R), (Q, S), (R, S), (R, T), (S, T), (S, U), (T, U).
5. ROUGE-SU: This is an extension of ROGUE-S. ROGUE-S is extended with the addition of unigram as a counting unit. For example, let us consider three sentences,
S1: P Q R S
S2: S R Q P
S3: A B C D
S1 and S2 do not have any skip-bigram in common but S2 is more similar to S1 than a sentence that has nothing in common with S1 (for example-S3), and ROGUE-S will consider S2 and S3 to be equally good which should not be the case. Hence, ROGUE-SU considers this factor and takes unigram as a counting unit.

Precision and recall are computed using the overlap.

1. Recall: This refers to what proportion of the reference summary is covered by the candidate summary. When considering only individual words, it can be defined as
Recall = No. of overlapping words/Total words in reference summary
Candidate summaries are often lengthy, containing almost all the words present in the reference summary. Many of these words in the candidate summary could be purposeless, making the summary extremely long which is why we use another metric, precision.
2. Precision: This refers to what proportion of the candidate summary was actually required. It basically measures the relevance of the words used in the summary. It can be defined as,
Precision = No.of overlapping words/Total words in candidate summary

This aspect is very important when length of the summary is an important factor, i.e., the aim is to generate concise summaries. It is often considered best to report F-measure.

7 Proposed Approach

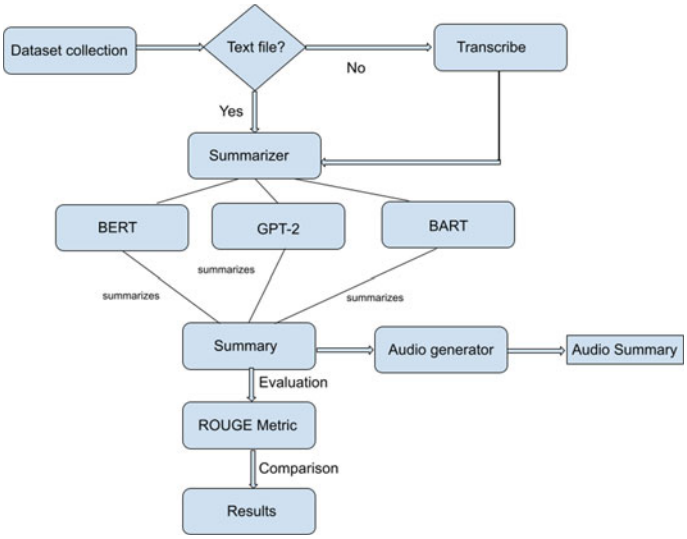


Fig. 1 The workflow diagram

8 Results

We verify the accuracy of the system-generated summary using the ROUGE metrics which compares our automated summary to the reference summary available in the dataset.The ROUGE has different versions like rouge-1, 2, n, l, etc., which basically checks the overlap between the two summaries, and based on the overlap, we get the results on precision, recall, and f-score.

Recall will refer to how much of the reference summary the system summary is capturing. Precision refers to how much of the system summary was in fact relevant. It is always best to compute both the precision and recall and then report the F-measure.

All three models were applied on ten articles from five different domains and were evaluated using Rouge-1 and ROGUE-L metrics. The results are as follows.

8.1 BERT Results

The results on summaries generated by the BERT model from domain business, entertainment, sports, politics, and tech are given in Tables 1, 2, 3, 4, and 5, respectively. It can be observed that precision goes upto 0.77 in ROGUE-1 evaluation metrics for articles in entertainment domain, while the recall values mostly revolves around 0.45 for every domain. The maximum *F*-score obtained is 0.58 in ROGUE-1 for Entertainment domain (Tables 1, 2, 3, 4, and 5).

Table 1 Business

Metric	*F*-score	Recall	Precision
Rouge-1	0.49	0.41	0.59
Rouge-L	0.42	0.37	0.50

Table 2 Entertainment

Metric	*F*-score	Recall	Precision
Rouge-1	0.58	0.47	0.77
Rouge-L	0.54	0.45	0.69

Table 3 Sports

Metric	*F*-score	Recall	Precision
Rouge-1	0.51	0.44	0.64
Rouge-L	0.47	0.41	0.56

Table 4 Politics

Metric	*F*-score	Recall	Precision
Rouge-1	0.51	0.47	0.57
Rouge-L	0.43	0.40	0.47

Table 5 Tech

Metric	*F*-score	Recall	Precision
Rouge-1	0.50	0.46	0.58
Rouge-L	0.42	0.39	0.49

8.2 BART Results

The results on summaries generated by the BART model from domain business, entertainment, sports, politics, and tech are given in Tables 6, 7, 8, 9, and 10, respectively. It can be observed that precision goes upto 0.70 in ROGUE-1 evaluation metrics for articles in entertainment domain, while the recall values mostly revolves around 0.30 for every domain. The maximum F-score obtained is 0.42 in ROGUE-1 for entertainment and tech domain (Tables 6, 7, 8, 9 and 10).

Table 6 Business

Metric	*F*-score	Recall	Precision
Rouge-1	0.35	0.25	0.57
Rouge-L	0.29	0.22	0.44

Table 7 Entertainment

Metric	*F*-score	Recall	Precision
Rouge-1	0.42	0.30	0.70
Rouge-L	0.40	0.30	0.61

Table 8 Sports

Metric	*F*-score	Recall	Precision
Rouge-1	0.39	0.30	0.60
Rouge-L	0.32	0.26	0.46

Table 9 Politics

Metric	*F*-score	Recall	Precision
Rouge-1	0.38	0.30	0.50
Rouge-L	0.30	0.25	0.40

Table 10 Tech

Metric	*F*-score	Recall	Precision
Rouge-1	0.42	0.37	0.52
Rouge-L	0.33	0.30	0.39

8.3 GPT-2 Results

The results on summaries generated by the GPT-2 model from domain business, entertainment, sports, politics, and tech are given in Tables 11, 12, 13, 14, and 15, respectively. It can be observed that precision goes upto 0.76 in ROGUE-1 evaluation metrics for articles in entertainment domain, while the recall values mostly revolve around 0.45 for every domain. The maximum *F*-score obtained is 0.58 in ROGUE-1 for entertainment domain.

Table 11 Business

Metric	*F*-score	Recall	Precision
Rouge-1	0.48	0.41	0.59
Rouge-L	0.42	0.36	0.50

Table 12 Entertainment

Metric	*F*-score	Recall	Precision
Rouge-1	0.58	0.47	0.76
Rouge-L	0.54	0.45	0.68

Table 13 Sports

Metric	*F*-score	Recall	Precision
Rouge-1	0.51	0.44	0.64
Rouge-L	0.46	0.40	0.55

Table 14 Politics

Metric	*F*-score	Recall	Precision
Rouge-1	0.51	0.47	0.56
Rouge-L	0.43	0.40	0.47

Table 15 Tech

Metric	*F*-score	Recall	Precision
Rouge-1	0.50	0.46	0.59
Rouge-L	0.43	0.40	0.48

9 Conclusion

In this paper, we have summarized articles from different domains from BBC news dataset using three summarization techniques, i.e., BERT, GPT-2, and BART. The evaluation is done using ROGUE-1 and ROGUE-L evauation metrics.

Among the three models used, GPT-2 has given better results than the other two models in all the five domains. The next best results are given by BERT. The GPT-based model takes advantage of tranfer learning.

Since the results obtained from ROGUE metrics are highly dependent on the reference summaries, it cannot be considered as an absolute measure of the quality of a summary. However, among the available evaluation metrics, these are the most accurate.

References

1. Wan, X., Hu, Y.: BrailleSUM: A News Summarization System for the Blind and Visually Impaired People
2. Kryscinski, W., Keskar, N.S., McCann, B.: Neural Text Summarization: A Critical Evaluation
3. Liu, Y., Lapata, M.: Text Summarization with Pretrained Encoders
4. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodeia, D., Christiano, P.: Learning to Summarize From Human Feedback

Keylogger Threat to the Android Mobile Banking Applications

Naziour Rahaman, Salauddin Rubel, and Ahmed Al Marouf

Abstract Android is presently the world's most prevalent operating system, reaching more mobile customers than any other operating system to date by providing numerous services via smartphone and various android devices to make our life easy. Most of the android applications are developed by third-party android developers, android provides them an enormous platform to build their application. Modern cyber attackers are highly interested in this platform to access user's sensitive information; with their own build malicious application or take amenities of other android developer's application to spy on user's activity. We have found that keyloggers can thieve personal information from users, such as credit card information or login pin/password from their typed keystroke in social networking and mobile banking apps. In case of mobile banking generally the mobile devices such as smartphones, tablets are being used for financial communications with the banks or financial institutions, by allowing clients and users to conduct a variety of transactions. In android app store (Google Play) keylogger apps are initially blocked but using some vulnerabilities in app permission it can be installed with benign and trusted apps. Both expert and maladroit android smartphone users use the mobile banking application, inexpert users are unable to find the vulnerabilities and attacker's use this as an advantage to place an attack. The security android has provided for all the application is not sufficient for the sensitive application such as mobile banking application. In our paper, we discuss how attackers steal mobile banking app users sensitive information for their financial gain and also proposed a method to avoid keylogger attacks on android mobile banking apps.

N. Rahaman · S. Rubel · A. A. Marouf (✉)
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
e-mail: marouf.cse@diu.edu.bd3

N. Rahaman
e-mail: naziour.cse@diu.edu.bd

S. Rubel
e-mail: salauddin15-7033@diu.edu.bd

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_12

Keywords Keylogger · Prevalent OS · Mobile banking · Mobile security · Financial Institute · Maladroit

1 Introduction

Cyberattacks on financial services firms have risen by 72% globally between 2014 and 2018. According to The Cost of Cyber Crime Study [1], the average cost per company increased around $1.3 million between 2017 to 2018. With over 3.7 billion smartphone users worldwide, the growth of the mobile app industry is unsurprising. More than 2.8 million android applications are in Google Play Store. About 25.75% of android user have used finance categories app in September 2019 [2].

Users, developers, and cybercriminals are all drawn to the versatility and functionality that Android has to deliver. Because of its ease of use and ability to cover high mobility, mobile banking has become a norm in the banking industry. Individuals who use mobile banking should be aware of the potential for cybercrime to affect their banking statements. Keylogger attack can be one of the criminal techniques that may occur on the mobile banking application.

Keylogger can embed themselves into PCs, Macs, Androids and iPhones in the same way like other malwares does and these types of malware called 'rootkit' viruses. Only in 2019, 9.9 billion malware attacks are reported along with mobile malware [3] which convey a security threat to the mobile banking application. Google Play Store has taken a wise step by removing all the keylogging application. As a result, the only option that mount a keylogger program is to do so remotely. Since we all know, Android asks for system permissions before running any app; however, often people disregard certain approvals. Since the user is ignorant to the application's secret permissions, keyloggers take advantage of the situation. Keeping these approaches in mind, using social media analytics and insights from social media such as recommendation engines [4, 5], personality trait polls [6, 7] or the stylometric features [8] can be utilized for such attacks.

Different methods may be used to prevent such keylogger attack. Now mobile baking applications are used by all kind of people, most of the people do not have enough knowledge about security. Keylogger attack can be mostly prevented by using antivirus or antimalware software. But most of these software are paid, users have to spend money for the service. Users from poor or developing country are not interested of using paid antivirus software but these countries have the most app user.

2 Literature Review

Keylogger attack takes place from the client-side because it steals a user's confidential and personal information from the user's input channel. The openness of the android platform has brought out a huge privacy risk, especially on transaction-based activity.

Quite a number of works have been done on keyloggers but most of them are focused on computers' keylogger. Recently, a few keylogging attacks on mobile devices have been studied. The method of stealing data from a mobile banking app has been identified by Prayogo Kuncoro [9]. However, their research did not include any mobile banking apps and did not provide any solutions for preventing keylogging attacks. Many attacks from third-party keyboards were introduced by Fadi Mohsen in [10]. Their research mostly focused on analyzing current keyboard authorization which can be still abused. However, this information is not sufficient to prevent keylogger attack. Junsung Cho's study analyzes the security of the third-party keyboards on android system, they have proposed to use a trustworthy keyboard for the bank Web site but keyboard design is not specified [11].

To avoid and prevent keylogging attack several models has been proposed by [12–14]. These prevention methods are applicable for malicious behavior of keylogger but with advanced android, permission keylogger can still place an attack. A study of Dr. Manisha M. More has shown the current scenario of cyberattacks in online banking [15]. Dr. N. Bhalaji has shown a very useful method for secured mobile experience by replicating data method [16]. To detect suspicious activity, Dr. Joy Iong have used hybrid deep learning which is capable of detecting keylogging [17]. Their technique can detect harmful activity but prevention is no specified. We, on the other hand, showed different ways of tracking the mobile banking apps by keylogger through an intense test with a potential solution.

3 Keylogger for Mobile Banking Application

3.1 Frequently Used Permissions in Mobile Banking Application

Permissions are used by an application for getting authorized access to different components of android. Developers declare all the required permission in the AndroidManifest.xml file and the permission are granted by the user while installation. In recent android devices (Android 6.0 and above) have a runtime permission system [18]. Permissions are granted when the applications are running and also have an enough context on why permission is required. We have installed 50 mobile banking applications from 15 different countries. Total 16 permissions are requested by these applications, 50% of them require only 5 or 6 permissions. There are two types of mobile banking applications: (i) only transaction-based and (ii) transaction and lifestyle-based. Only transaction-based applications provide money transfer, ATM and balance information. On the other hand, transaction and lifestyle-based applications provide many services including mobile recharge, bill payment, shopping offers, etc. This second kind of mobile banking application requires many permissions which makes them more vulnerable. For user personalization, the developer needs to access more user's data, thus make the application a good source of user

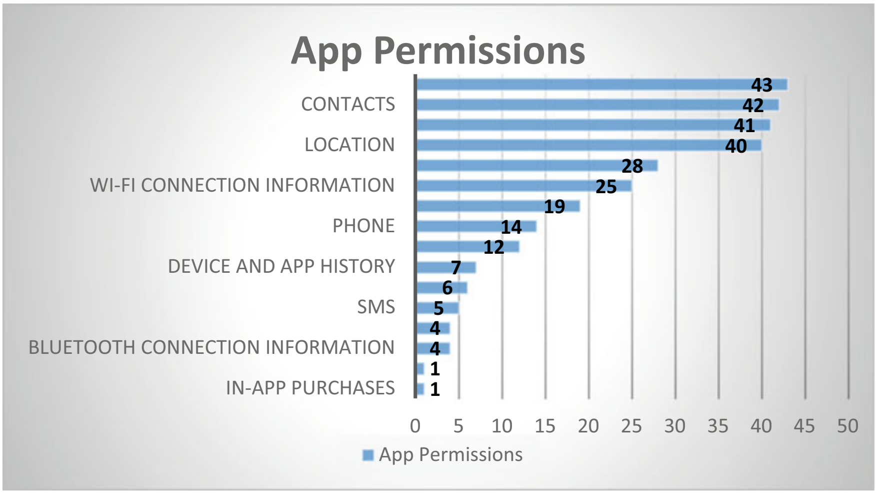


Fig. 1 Frequently used app permissions

information and others application try to take this as advantage. Problems happened when other applications having less permission, uses these sensitive permissions by corrupting ICC (Inter Component Connection) [19, 20]. Broadcast theft, activity hijacking, service hijacking, intent spoofing, privilege escalation and application collusion attack are some of the possible ways to do this [21] (Fig. 1).

3.2 How Keylogger Gets Installed

In android, keylogger app can be installed in various way with the app's apk. Consciously a user will not install a keylogger apps can get installed by either with a pop-up ad or pleasant app. After being installed, the apps began to log the user's keystrokes. The attacker saves keystroke data in a file in local storage, which is then submitted to a remote server. Advance keylogger can record and send the keystrokes to a database that is updated frequently. We demonstrate the attack scenario of a keylogger application (Fig. 2).

From the following figure, we can find three ways to install a keylogger application in android mobile.

- The attacker can create a benign application such as a calculator application with hidden keylogger malware, and it will be very difficult for a user to identify the keylogger. Even attackers' benign applications can incline users to install other keylogger application from other resources. Stalkerware is a process of installing an app without user interaction, this model can be used by attackers to install an unwanted keylogger app [22].

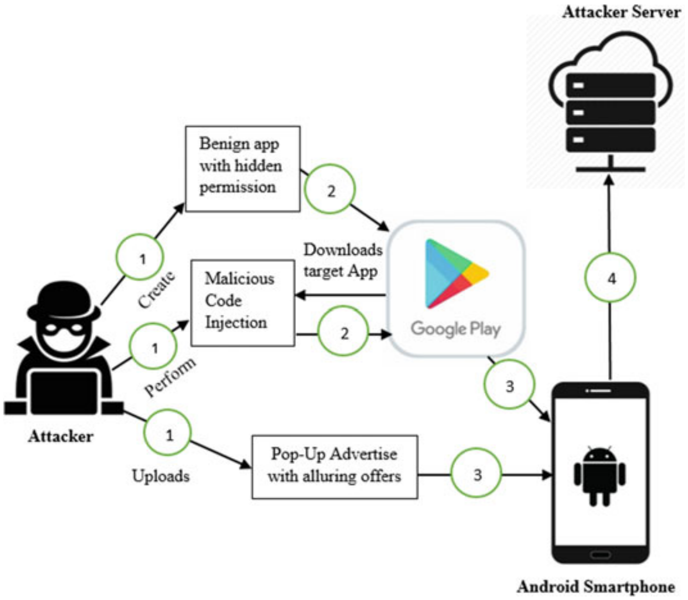


Fig. 2 Keylogger attack scenarios

- Attackers target most popular application from play store, with reverse engineering they can retrieve the code and inject their malicious code into the application [23]. Most of the time it happened for the paid application. Attackers provide the paid application for free with malicious code inside it.
- A keylogger can be installed from pop-up advertisement, attackers can provide an advertising with an alluring offer like, install this application and get 10$. Once a user installs that kind of application, constantly user gives access to any permission, thus keylogger can get installed.

4 Proposed Mechanism to Prevent Threat

We know that sensitive banking information is saved as an encrypted cipher. The client clicked the unencrypted value while entering their pin in the mobile banking app for logging. As a result, a keylogger application can easily monitor the value (Fig. 3).

Keylogger records typed value according to the sequence of typing. Each value in a keyboard layout design has two parts: labeled value and codecs value [24]. The label value is what we can see on keyboard but the value of codecs isn't really available in user interface. The ASCII value of a character is the codecs value of a key. Every key has a unique ASCII value, but every keyboard has a certain key-label and key-codecs value. The keylogger extracts the key's label value and stores it in a file. In the

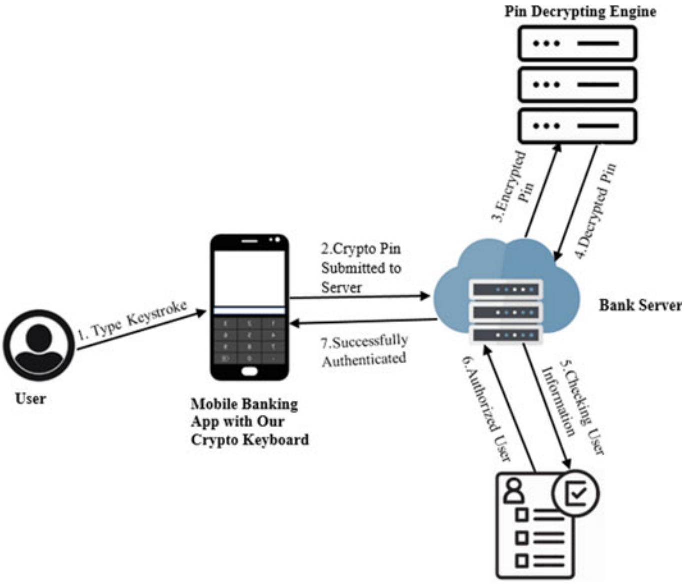

Pin Decrypting Engine
3.Encrypted Pin
4.Decrypted Pin
2.Crypto Pin Submitted to Server
Bank Server
1. Type Keystroke
7.Successfully Authenticated
User
Mobile Banking App with Our Crypto Keyboard
6.Authorized User
5.Checking User Information

Fig. 3 Our proposed model of encryption

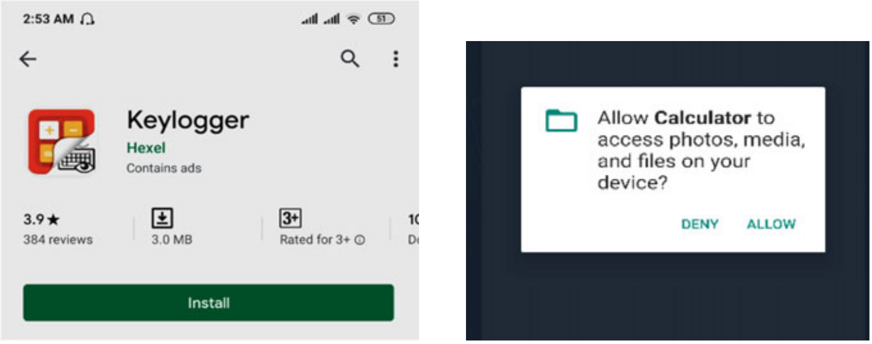
model we've suggested, the key label will be similar to the default keyboard layout. On the other hand, key codec value will be diverse. User's typed keystroke will be different from the labeled key value. The actual labeled value will not be recorded by the keylogger because of varied codec value of the key. As an Example: keylogger will store 'm' instead of recording '3' for the labeled key of '3' as we assign the codec value of '3' to the ASCII value of 'm'. The keylogger will unable to record the original key value and store encrypted value. The decryption engine will decode the encrypted value of the key and the original value will be sent to server. The encryption process will be started from the user end while typing the Pin/Password. Encrypted pin will be sent to banking server where a pin decrypting engine will decrypt the pin. Then the decrypted information will be used to authenticate user. After successful authentication user will be able to login.

5 Experimental Analysis

5.1 Proof of Concept

For our experiment, we downloaded a keylogger application from the Google Play store. Though keylogger is illegal in most of the terms but some keylogger application still available in Google Play Store. Before publishing any application in play store Google has a review process [25]. So, attackers designed their application in such way that it can be able to pass the review process. We downloaded the application named ‘calculator’ developed by Hexel. During installation, this application only requires ‘Photos and Media’ permission. Recently Google has updated the permission system, permissions are allowed by user during using the application. But not any effective warning messages were displayed for using the permission of WRITE_EXTERNAL_STORAGE, most users are unlikely to notice the significant threat of this approval and will proceed with the installation. But this application needs to enable ‘accessibility’ for the keylogging process. After enabling accessibility keylogger to start keylogging. It records all the keystrokes typed by the user and saved into a file (Fig. 4).

We run our experiment on an app named ‘Rocket’ (An application of Dutch Bangla Bank Limited) [26]. We login into the mobile banking application and after some activities we logged out of the application. In this period of time, our keylogger enabled us to store the login pin of the application in a text file which was later manually checked. Most of the mobile application has same pin as the pin for the bank account. Once the application pin accessed by an unauthorized person, the account will be in danger (Fig. 5).

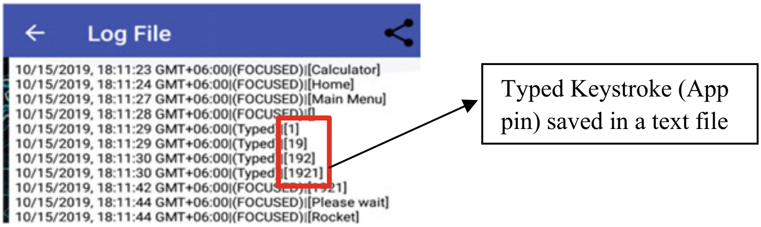


(a) App Information

Allow **Calculator** to access photos, media, and files on your device?

(b) Storage Permission

Fig. 4 No effective warning message to user mobile



```
10/15/2019, 18:11:23 GMT+06:00|(FOCUSED)|[Calculator]
10/15/2019, 18:11:24 GMT+06:00|(FOCUSED)|[Home]
10/15/2019, 18:11:27 GMT+06:00|(FOCUSED)|[Main Menu]
10/15/2019, 18:11:28 GMT+06:00|(FOCUSED)|[]
10/15/2019, 18:11:29 GMT+06:00|(Typed)|[1]
10/15/2019, 18:11:29 GMT+06:00|(Typed)|[19]
10/15/2019, 18:11:30 GMT+06:00|(Typed)|[192]
10/15/2019, 18:11:30 GMT+06:00|(Typed)|[1921]
10/15/2019, 18:11:42 GMT+06:00|(FOCUSED)|[1921]
10/15/2019, 18:11:44 GMT+06:00|(FOCUSED)|[Please wait]
10/15/2019, 18:11:44 GMT+06:00|(FOCUSED)|[Rocket]
```

Typed Keystroke (App pin) saved in a text file

Fig. 5 Keystroke typed by Android built-in keyboard (Gboard)

From the above figure, we can find the recorded keystrokes. Keystrokes are also tokenized in the segment so that the application login pin can be easily found.

5.2 Applying Our Mechanism

From our proposed mechanism (Sect. 4) we design a prevention method to avoid keylogging attack in mobile banking applications. We have implemented our mechanism with the same mobile banking application and keylogger application. We face some problems in implementing single digit encryption. In our present android application programming, we cannot implement a function in key codes. That's why encryption from the key codecs cannot apply. As android is controlled by Google, so changes in android are dependent on them. In this scenario, we just change the codecs value with encrypted value and record the keystroke (Fig. 6).

To prevent keylogging attack, we have designed an encryption method for the keyboard. For implementing our mechanism of encrypting keystroke digit from the user end this model can be used. This algorithm will encrypt the typed keystroke in a single digit, our proposed methodology is Single Digit Key Encryption (SDKE) which is an application of AES (Advanced Encryption Standard) encryption algorithm [27]. User will type pin/password for login into his account, this pin will be encrypted as AES cipher. Each digit will be encrypted and save into a stack. From the encrypted cipher, a digit will be choosen randomly and this random digit will be the codec value for the digit. If any keylogger able to steal the keystroke, actually it will record this random digit. When user press the submit button for login, all the

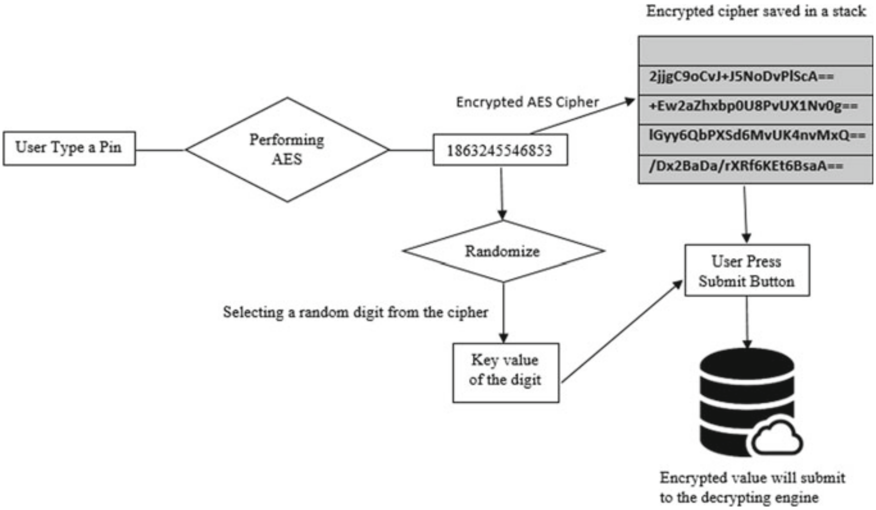
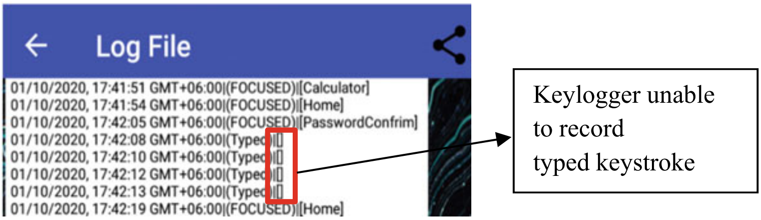


Fig. 6 Applying Pin-Crypto method to avoid keylogger attack



```
01/10/2020, 17:41:51 GMT+06:00|(FOCUSED)|[Calculator]
01/10/2020, 17:41:54 GMT+06:00|(FOCUSED)|[Home]
01/10/2020, 17:42:05 GMT+06:00|(FOCUSED)|[PasswordConfrim]
01/10/2020, 17:42:08 GMT+06:00|(Typed)|[]
01/10/2020, 17:42:10 GMT+06:00|(Typed)|[]
01/10/2020, 17:42:12 GMT+06:00|(Typed)|[]
01/10/2020, 17:42:13 GMT+06:00|(Typed)|[]
01/10/2020, 17:42:19 GMT+06:00|(FOCUSED)|[Home]
```

Keylogger unable
to record
typed keystroke

Fig. 7 Keystroke typed by our designed keyboard with our recommended settings

encrypted cipher from the stack will be submitted to the decrypting engine along with the key value. Then the decrypting engine will decrypt the cipher and process for the next authentication task. Using this model keylogger unable to record the actual value. Suppose user type the pin '5678' for login after performing 128-bit AES with the secret key value of '1,863,245,546,853,697' this will return some cipher for '5' like '2jjgC9oCvJ+J5NoDvPlScA==' [28] and from this cipher a random digit 'j' will be choose. So if the user type '5' keylogger will receive 'j' for the typed keystrokes (Fig. 7).

From the above figure, we can find that the keylogger application unable to store the keystrokes. The keylogger record the codecs value which we have changed into an encrypted value. So, the value is untraceable by the keylogger. If any keylogger is able to record the typed keystroke, it cannot find the original key value. Because the value is already encrypted. In that case, it can record the encrypted value. Example: user type a pin '1234', the keylogger record encrypted value like 'h&)*'. Its method ensures more integrity of a mobile banking application.

6 Recommendation

Application login Pin length should be more than 4 Digit: Password length is a very important issue in the security of any online account. Having password length equal to or less than 4 digits lays a platform for brute-force attack. In mobile banking application, there are many applications which are using only 4 digits for application login. Though a small password is easy to remember but it is less secure. Using 5 digits is 6 times (30,240) more and using 6 digits is 30 times (151,200) more secure than 4 digits pin in a mobile banking application.

Password field input type change from number pin to text password: Existing mobile banking applications' password field is designed for the number input type. We are talking about digit encryption, when we want to encrypt any digit in numberPassword field, the encrypted digit will be also a number digit. Encrypting any number digit to another number digit does not effective as there is a chance of having same input and encrypted digit. Using text type password field will provide access to a large number of Unicode characters. Then the encryption combination will be huge. In a password field, the user cannot see the typed digit so the changes may not affect the user.

Disable Copy-Paste option in the password field: To avoid collecting the application pin, this function can be effective. Most of the mobile banking application is using this function.

Mobile Banking application should have their own keyboard layout: Using the android's default or third-party keyboard for mobile banking application can be dangerous. Mobile banking application should use their own keyboard layout for their application.

7 Conclusion and Future Work

The paper presented the potential keylogger threat to Android Mobile Banking App. Keylogger attack may not happen if a user is conscious while using their smartphone or have a little security knowledge. The Trojan that delivers keylogger can drop more malware such as adware, spyware, ransomware or even a legacy virus on the system. So, it is not only about keylogger attack but also other malware that can take place in an android phone. In our proposed method we use AES algorithm which may slow down the overall process of user logging. We need to find more efficient and optimized algorithm for this process and specifically for digit encryption. This method can be only implemented when we can able to use functions in android xml file which provides the codecs value for the key.

For our future work, we will expand our keyboard model and build a fully cryptographic keyboard to ensure proper security. During our experimental analysis, we have found that keylogger can also read messages from Facebook, Gmail and WhatsApp. We will find more efficient algorithm and design for the keyboard that can be used to stop or prevent keylogging attack.

References

1. Help Net Security: Financial services firms most adept at making balanced security investments—Help Net Security, 2020 [Online]. https://www.helpnetsecurity.com/2018/02/14/financial-services-security-investments
2. Statistica: Leading Android App Categories Worldwide 2019, 2020 [Online]. https://www.statista.com/statistics/200855/favourite-smartphone-app-categories-by-share-of-smartphone-users/
3. Securitymagazine.com, 2020 [Online]. https://www.securitymagazine.com/articles/91660-more-than-99-billion-malware-attacks-recorded-in-2019
4. Marouf, A.A., Ajwad, R., Tanbin Rahid Kyser, M.: Community recommendation approach for social networking sites based on mining rules. In: 2nd IEEE International Conference on Electrical and Information and Communication Technology (iCEEiCT), Jahangirnagar University, Bangladesh, 21–23 June, 2015

5. Mehedi Hasan, M., Shaon, N.H., Marouf, A.A., Kamrul Hasan, M., Mahmud, H., Mohiuddin Khan, M.: Friend recommendation framework for social networking sites using user's online behavior. In: 18th IEEE International Conference on Computer and Information Technology (ICCIT), MIST, Bangladesh, 21–23 December, 2015
6. Marouf, A.A., Kamrul Hasan, M., Mahmud, H.: Comparative analysis of feature selection algorithms for computational personality prediction from social media. IEEE Trans. Comput. Soc. Syst. **7**(3), 587–599 (2020)
7. Marouf, A.A., Kamrul Hasan, M., Mahmud, H.: Identifying neuroticism from user generated content of social media based on psycholinguistic cues. In: 2019 2nd IEEE Conference on Electrical, Computer and Communication Engineering (ECCE 2019), CUET, 7–9 Feb, 2019
8. Hossain, R., Marouf, A.A.: BanglaMusicStylo: a stylometric dataset of bangla music lyrics. In: 1st IEEE International Conference on Bangla Speech and Language Processing (ICBSLP), SUST, 21–22 Sept 2018
9. Kuncoro, A., Kusuma, B.: Keylogger ıs a hacking technique that allows threatening ınformation on mobile banking user. In: 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE), 2018
10. Mohsen, F., Shehab, M.: Android keylogging threat. In: Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2013
11. Cho, J., Cho, G., Kim, H.: Keyboard or keylogger?: a security analysis of third-party keyboards on android. In: 2015 13th Annual Conference on Privacy, Security and Trust (PST), 2015
12. Enck, W., et al.: TaintDroid. ACM Trans. Comput. Syst. **32**(2), 1–29 (2014)
13. Nauman, M., Khan, S., Zhang, X.: Apex. In: Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security—ASIACCS'10, 2010
14. Pearce, P., Felt, A., Nunez, G., Wagner, D.: AdDroid. In: Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security—ASIACCS'12, 2012
15. More, D.M.M., Nalawade, M.P.J.D.K.: Online banking and cyber attacks: the current scenario. Int. J. Adv. Res. Comput. Sci. Softw. Eng. Res. Paper, 2015
16. Bhalaji, N.: Effıcıent and secure data utilization in mobıle edge computing by data replication. J. ISMAC **2**(1), 1–12 (2020)
17. Chen, D., Smys, S.: Social multimedia security and suspicious activity detection in SDN using hybrid deep learning technique, vol. 2, no. 2, pp. 108–115 (2020)
18. Google Play|Android Developers: Android Developers, 2020 [Online]. https://developer.android.com/distribute/best-practices/develop/runtime-permissions
19. Li, L., Bartel, A., Klein, J., Traon, Y.: Automatically exploiting potential component leaks in android applications. In: 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications, 2014
20. Schartner, P., Bürger, S.: Attacking Android's Intent Processing and First Steps Towards Protecting it. Technical Report TR-syssec-12-01, Universität Klagenfurt, 2012
21. Wang, J., Wu, H.: Android Inter-App Communication Threats, Solutions, and Challenges. arXiv:1803.05039, 2018
22. Baraniuk, C.: The rise of stalkerware. New Scientist **244**(3257), 20–21 (2019)
23. RSAC: Reverse-Engineering an Android App in Five Minutes. PCMAG, 2020 [Online]. Available https://www.pcmag.com/news/rsac-reverse-engineering-an-android-app-in-five-minutes
24. Keyboard|Android Developers: Android Developers, 2020 [Online]. https://developer.android.com/reference/android/inputmethodservice/Keyboard
25. Google Play|Android Developers: Android Developers, 2020 [Online]. https://developer.android.com/distribute/best-practices/launch/launch-checklist
26. Play.google.com, 2021 [Online]. https://play.google.com/store/apps/details?id=com.dbbl.mbs.apps.main&hl=en&gl=US

27. Search Security: What is Advanced Encryption Standard (AES)? Definition from WhatIs.com, 2020 [Online]. https://searchsecurity.techtarget.com/definition/Advanced-Encryption-Standard
28. Online Tool for AES Encryption and Decryption. devglan, 2020 [Online]. https://www.devglan.com/online-tools/aes-encryption-decryption

Machine Learning-Based Network Intrusion Detection System

Sumedha Seniaray and Rajni Jindal

Abstract As the network is dramatically extended, security has become a significant issue. Various attacks like DoS, R2L, U2R are significantly increasing to affect these networks. Thus, detecting such intrusions or attacks is a major concern. Intrusions are the activities that breach the system's security policy. The paper's objective is to detect malicious network traffic using machine learning techniques by developing an intrusion detection system in order to provide a more secure network. This paper intends to highlight the performance comparison of various machine learning algorithms like SVM, K-Means Clustering, KNN, Decision tree, Logistic Regression, and Random Forest for the detection of malicious attacks based on their detection accuracies and precision score. A detailed analysis of the network traffic features and the experimental results reveal that Logistic Regression provides the most accurate results.

Keywords Intrusion · Intrusion detection · Network-based intrusion detection system · Network security · Machine learning · Network traffic · KDD · Feature selection

1 Introduction

An attack is any kind of action that threatens the integrity, confidentiality; attempting to achieve unauthorized access to the sensitive information of a network system is known as an attack. An Intrusion Detection System is a system that helps detect a variety of malicious or abnormal network traffic and computer usage that is not feasible to detect with the help of a conventional firewall or is unknown to the user. This comprises of network attacks against all the services that are vulnerable, data-driven attacks on applications, host-based attacks like unauthorized system or

S. Seniaray · R. Jindal (✉)
Delhi Technological University, New Delhi, India
e-mail: rajnijindal@dce.ac.in

S. Seniaray
e-mail: sumedhaseniaray@dtu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_13

software login, privilege escalation, and access to personal/sensitive user files and data, and malware (viruses, worms, and trojan horses). Intrusion detection systems and firewalls both are a part of network security, but they differ from each other as firewall looks on the outside for intrusions so that intrusions can be stopped before happening. Firewalls forbid access between networks so that intrusion can be prevented. If an attack is within the network, then it does not signal. In contrast, an Intrusion Detection System (IDS) detects a suspicious intrusion once it has occurred and then signals an alarm to notify that an intrusion has been detected. Firewalls are like barriers that protect the system from the outside threats and signals the system if unauthorized or forceful attempts are made from the outside. In contrast, an Intrusion Detection System signals the system when it detects such malicious activity. The main agenda of IDS is to protect the host or the network from any malicious or unusual activity that can enter the system and compromise the data. Thus, the aim is to detect an intrusion before the hackers get to the information and damages or corrupt it. Intrusion detection can be performed for various application areas, for instance, in Digital forensics, in IoT [1] for detection of intrusions in the network, wireless sensor networks (WSN) [2], social media networks [3], real-time security systems, and also in combination with firewalls for additional security of the network as well as the host system (Fig. 1).

Intrusion Detection Systems are of two types:

- **Network-based Intrusion Detection System (NIDS)**

NIDS [4] detects any threat or intrusion like denial of service (DoS), etc., introduced in the network by keeping track of the network traffic. A network-based intrusion detection system resides on the network monitoring the network traffic flows, that is, the inbound and outbound traffic to and fro from all the devices connected in the network.

- **Host-based Intrusion Detection System (HIDS)**

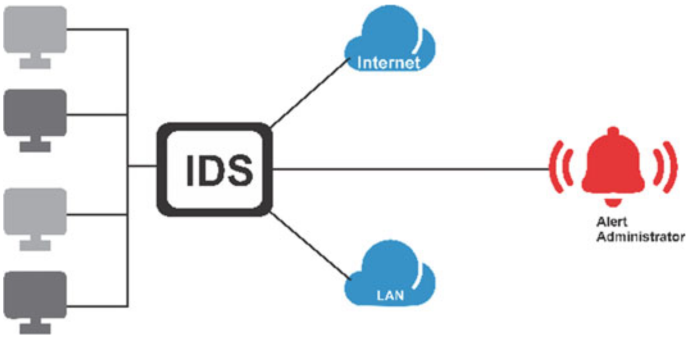


Fig. 1 Intrusion detection system

HIDS is installed on the client's system and helps detect any threats which are introduced in a specific host via the network packet is received, or which host accessed the data or whether any unauthorized access has been done, etc.

On the basis of methods of detection, Intrusion Detection System (IDS) can be categorized as:

- **Signature-based IDS**

Signature-based IDS [5] provides significant intrusion detection results for well-known, specified attacks. Thus, they are not very capable of detecting unfamiliar or new attacks.

- **Anomaly-based IDS**

An anomaly from the perspective of security is an event that is suspicious. Anomaly-based IDS [4, 6] have the potential to detect previously unknown, unseen intrusion events or attacks. Therefore, anomaly-based IDS can detect both known and unknown intrusions. Thus anomaly-based systems have a higher rate of detection capability than signature-based IDS.

We aim to work on the anomaly-based network intrusion detection system. Thus, monitoring the network traffic flows in order to detect not just the known but also unknown or abnormal network traffic flow. An anomaly-based intrusion detection system monitors the network traffic and compares it to the normal traffic flows, and if it detects some unusual pattern or anomalies on the network, it alarms the signal indicating a potential threat. Based on this comparison, the network traffic flow is categorized as “normal” or “abnormal or malicious".

There are various Machine Learning techniques that are incorporated into the intrusion detection procedure to decrease the false alarm rates. Machine Learning is also used to automate the building of an analytical model. As Machine Learning is a part of Artificial Intelligence, which prevails on the concept that a system gets trained, makes decisions, and learns to diagnose patterns with fewer human intervention; thus, it is determined to build a model that enhances its performance on the basis of previous results. For this purpose, various Machine Learning techniques are Support Vector Machine (SVM), KNN, Random Forest, Logistic Regression, Decision tree, and Naïve Bayes, etc.

To give a detailed analysis of these algorithms for detection of intrusion and to establish our anomaly-based network intrusion detection system, we collected the non-malicious network traffic, that is, KDD'99 data set and malicious network traffic, to train these machine learning classifiers on the basis of the collected network traffic data. The main contributions of this paper are summarized as follows:

1. Collection of both non-malicious or normal data and malicious data.
2. After analyzing the network traffic, feature extraction is performed, and 14 traffic features are extracted.
3. Feature selection is performed using the Feature Importance technique on these 14 extracted features to inherit the most significant ones, on which Machine Learning techniques are performed.

4. Machine learning classifiers are then trained on individual features to identify the intrusion detection accuracy and precision.
5. A combination of network traffic features is done, which lie above 50% on the Feature Importance Scale.

2 Related Work

In this modern era of Machine Learning, network intrusion detection system has become a vital component in network security. In today's period, it is vital for the organization and individual to secure their computer and network data as once the network is compromised, it can cause a lot of information damage. Various machine learning algorithms are applied to intrusion detection systems, such as decision tree [7–10], Logistic Regression [11–13], Support Vector Machine (SVM) [14–16], and Random Forest [15, 17]. In [17], a Random Forest-based intrusion detection system model was developed where the effectiveness of the Random Forest-based intrusion detection system model was tested on the NSL-KDD dataset, and it was noticed that the Random Forest performance was slow for real-time predictions when the number of trees was increased. Their results exhibited a detection rate of 99.67% in comparison with J48. In [15], detailed analysis and comparison are drawn on different machine learning algorithms, namely Random Forest, Support Vector Machine (SVM), and Extreme Learning Machine (EML), to find out the algorithms which give a better result when to amount of data to be analyzed is increased. And it was realized that Extreme Learning Machine (EML) gives the best results when the entire data is taken into consideration. When half of the data was considered, SVM performs better than the other two. In [16], the overall performance of SVM is improved by accelerating the convergence of the algorithm and increasing its training speed. A new function was created with the intention that the error rate of the SVM is reduced. Repalle and Kolluru [18] discovered that it is crucial to obtain a well-labelled dataset in order to provide efficient results. K-Nearest Neighbour (KNN) was found to be the best working algorithm; for the analysis, the values assigned to the variable 'K' is of importance. Fayyad et al. [19] discusses a comprehensive analysis of cybersecurity with the help of intrusion detection using machine learning (ML) and data mining (DM) methods, where performances of both ML and DM techniques are addressed to analyze accuracies of each of these techniques, which contributes to the field of cybersecurity. Tao et al. [15] proposed the FWP-SVM-GA algorithm, an intrusion detection algorithm that is based on the characteristics of the Support Vector Machine (SVM) and Genetic algorithm (GA) algorithm where the FWP-SVM-GA algorithm performs feature selection, parameter optimization of SVM based on GA. This reduces the SVM error rate and enhances the true positive rate. Finally, an optimal feature subset is used on the feature weights and SVM parameters in order to optimize them. As a result, classification time, error rates decrease, and the true positive rate increases. According to [20], intrusion detection is considered as a multi-class and two-class classification. This is performed using the SVM machine learning algorithm. SVM acts as a decision-making model throughout the training phase in the proposed SVM-based intrusion detection model. They performed three kinds of experiments on the 1999 KDD dataset, wherein they performed the experiments on 41 features set and presented a comparison of SVM IDS with KDD 1992 contest winner and concluded that SVM IDS is more effective when it comes to intrusion detection. In [21], an intrusion detection framework using SVM along with feature augmentation is performed on NSL-KDD dataset. Feature augmentation was done in order to provide a more concise and high-quality training data set for the SVM classifier, which helped improve the efficiency of the SVM-based proposed model. As a result of the experiment, the proposed model achieved a high detection accuracy of 99.18%.

3 Methodology and Implementation

This section describes the way the machine learning classifiers are implemented on the network traffic features to design an anomaly-based intrusion detection model. The implementation is summarized in four phases, represented in Fig. 2, namely: (1) Network traffic collection (2) Data pre-processing (3) Feature Extraction and Feature Selection (4) Implementation of Machine Learning (ML) techniques for the proposed intrusion detection system.

3.1 Network Traffic Collection

As the estimation and analysis of the machine learning techniques are performed on the network traffic, we need two sets of network traffic data, that is, malicious or intrusive traffic data and non-malicious or normal traffic data. The normal traffic data via an extensive network traffic analyzer software Wireshark [22] is used to capture them and converted into TCP and UDP flow conversations. The other set of data used is KDDCUP’99 [19, 23] dataset, which includes the malicious or intrusive network traffic data. KDD training dataset consists of approximately 4,900,000 records, each of which contains 41 features, labelled as normal or an attack. This data is in “pcap” format, which can be further analyzed by Wireshark. Wireshark is also used to analyze

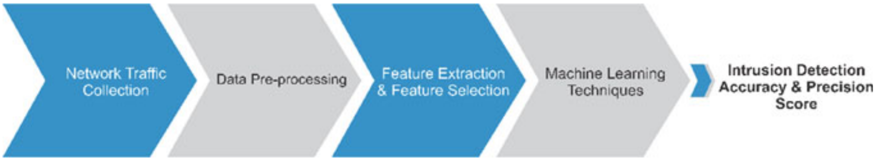

Network Traffic Collection
Data Pre-processing
Feature Extraction & Feature Selection
Machine Learning Techniques
Intrusion Detection Accuracy & Precision Score

Fig. 2 Proposed ıntrusion detection system model

network data, and then the data can be classified as normal and abnormal or intrusive data. Wireshark presents the data packets in a human-readable format, making it easier to understand the data better.

3.2 Data Pre-processing

The dataset is pre-processed, and in order to do that, the KDDCUP'99 data set is cleaned, and the redundant data is eliminated. The combination of both the datasets together is then divided into two sets of data, that is, training and testing datasets in the ratio of 70:30. These datasets, normal and intrusive datasets, before dividing them are labeled as "normal" and "attack" which helps distinguish the type of network traffic data.

3.3 Feature Extraction and Feature Selection

Based on [24, 25] and the survey conducted on the previous related work on the network traffic features pertaining to intrusion detection, techniques like PCA, LDA, etc. were used, and we found that 14 network traffic features extracted from the TCP and UDP flow conversations were of at most importance, represented in Table 1, in order to analyze the collected network traffic features for normal samples.

Out of these 14 traffic features, we had to obtain an optimal set of features to perform the machine learning techniques for detecting the intrusion. For this purpose, we used the feature selection technique, Feature Importance. Feature selection is a technique that reduces the amount of data to be analyzed. This is accomplished by identifying the most important features (or attributes) of a data set and discarding the less important ones. Feature importance renders a score for each of the features of the

Table 1 Network traffic features extracted

Network traffic features	
Average packet size	Average time between packets received
Average packet size received	Average time between packets sent
Average packet size sent	Ratio of ıncoming to outgoing bytes
Average flow duration	Ratio of ıncoming to outgoing packets
Number of packets	Bytes received
Rel start	Bytes sent
Duration	Number of bytes

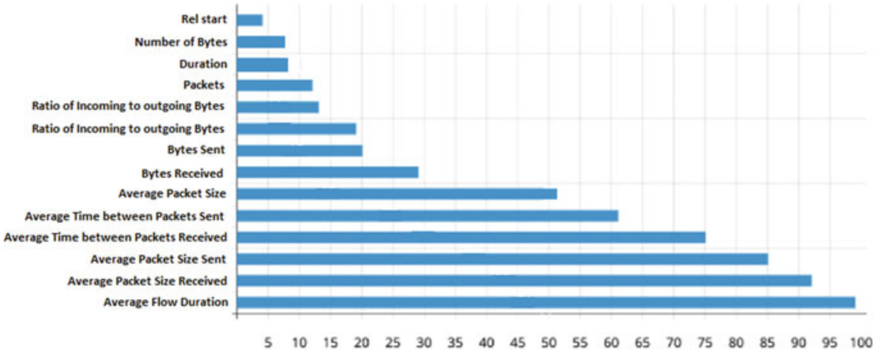


Fig. 3 Feature selection score graph

Table 2 Optimal network traffic features selected

Label	Network traffic features
F1	Average flow duration
F2	Average packet size received
F3	Average packet size sent
F4	Average time between packets received
F5	Average time between packets sent
F6	Average packet size

network traffic data; the higher the score more relevant or important is the feature. As we have implemented the intrusion detection model in Python, the importance selection is used, which is an in-built class.

Figure 3 depicts the Feature Selection score graph, which presents the least important to the most important network traffic features based on the importance score evaluated. Based on this feature selection score graph, we have set a threshold to 50; that is, we will select the most significant features that lie above the set threshold. So, all the features having a score greater than 50 are considered the most optimal feature set, represented in Table 2.

3.4 ML for Intrusion Detection System

After selecting the most optimal network traffic feature dataset, various machine learning techniques like Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, Decision trees, K-Nearest Neighbour (KNN), and Random Forest are applied to these feature set in order to detect the intrusive traffic from the normal.

Support Vector Machine (SVM)

Support Vector Machine(SVM) - A Support Vector Machine(SVM) model is a supervised machine learning algorithm that analyzes the data used for the purpose of regression and classification. This algorithm aims to discover a hyper plain in N-dimensional space (where N is the number of features) that separately performs the data points classification. Support Vector Machine(SVM) [8] can perform linear and non-linear classification, implicitly mapping the inputs in the high-dimensional feature space. All flat affine subspaces are called hyperplanes. SVM Kernel is used to add more dimension to low dimensional space making it easier to segregate the data; it converts the inseparable problem to a separable problem by adding more dimension using kernel tree.

Decision Tree

A decision tree, as the name suggests, is a tree-like graph that has internal nodes that represent the test done on the attributes/features, and the branches show the decision rules of the test, leaf nodes which represent the outcome. Decision tree, which is a supervised machine learning algorithm, is used in making the classification and regression models.

Logistic Regression

Logistic Regression [26], being a supervised machine learning algorithm, is used in classification analysis, which helps in the prediction of variable data set probability. It assesses the interrelation between the dependent (Label) and the independent (Features) variable. Sigmoid function is used in the logistic function as a cost function. This logistic function helps map predictions to probabilities, and by fitting the data to this function, the probability of occurrence of an event can be predicted.

Naïve Bayes

The Naïve Bayes classifier is a probabilistic classifier that imposes a strong independence assumption [27], which suggests that the probability of an attribute doesn't affect the probability of the other. The dataset is converted into frequency tables, and further, a new table is generated on the basis of the evaluated probabilities of the features/attributes under consideration. For an n attributes series, the naïve Bayes classifier produces $2n!$ independent assumptions. Nevertheless, the Naïve Bayes classifier often provides correct results.

K-Nearest Neighbour (KNN)

K-Nearest Neighbour(KNN) is a supervised classifier, machine learning algorithm. KNN stores all the values present in the data set and classifies them into a new data point based on the character similarities. K-Nearest Neighbour (KNN) assumes that things with similar characteristics are near each other; that is, similar things exist in close proximity. The position where the target variable will be placed is predicted by finding the k closest neighbour, by calculating the Euclidean Distance.

Random Forest

Random Forest algorithm is mainly utilized for the purpose of classification analysis but can also be used for regression analysis. Different decision trees are created on various data samples, the prediction is taken from each of the decision trees, and then the voting is done to get the final prediction. Higher accuracy will be achieved by including a higher number of trees in the model.

4 Results and Discussions

In this section, we have implemented the Machine Learning (ML) algorithms stated in the previous section on the selected optimal set of network traffic features. A detailed analysis and comparison of these features on the basis of the ML algorithms have been drawn to depict their corresponding accuracy and precision.

4.1 ML Detection Accuracy on Individual Network Traffic Features

Table 3 presents the evaluated accuracy of ML techniques on individual network traffic features. İt is observed that when all the considered ML algorithms are applied to the individual features, it was observed that Random Forest has the highest detection accuracy, that is, an average accuracy of 97.85%.

Table 3 Detection accuracy (%) for ındividual features

Algorithm		$F1$	$F2$	$F3$	$F4$	$F5$	$F6$
Decision tree		88.72	89.23	89.74	91.73	81.02	87.72
Naïve Bayes		79.51	79.18	83.82	88.34	80.50	89.87
Random forest		88.89	99.10	99.85	99.96	99.68	97.45
SVM		85.29	95.88	94.43	90.22	99.98	97.45
Logistic regression		86.86	95.13	89.31	87.77	98.90	93.86
KNN	$K = 5$	84.85	99.35	99.30	99.51	99.89	98.85
	$K = 10$	85.37	99.34	99.30	99.50	99.87	98.59

4.2 ML Detection Accuracy on Combined Network Traffic Features

We now evaluate the detection accuracy for the combination of all the network traffic features, which is summarized in Table 4, representing the detection results for all six traffic feature combinations. İt is observed that when the ML algorithms are applied to the combination of features, Logistic Regression has the highest detection accuracy, that is, an average accuracy of 99.048%.

We observed that combining the optimal network traffic features leads to better intrusion detection accuracy. These results are concluded based on the traffic features we had selected. At the same time, if we include a traffic feature that has a network selection score less than (<) 50, we observed that the detection accuracy for the combined feature set of 7 features, that is including the feature *Bytes Received*, the intrusion detection accuracy is reduced in comparison to the detection accuracy of the combination of top 6 features, summarized in Table 5.

Table 4 Detection accuracy (%) for 6 combined features

Algorithm		$F1$ and $F2$ and $F3$ and $F4$ and $F5$ and $F6$
Decision tree		92.389
Naïve Bayes		88.487
Random forest		97.881
SVM		91.636
Logistic regression		99.048
KNN	$K = 5$	98.593
	$K = 10$	98.472

Table 5 Detection accuracy (%) for 7 combined features

Algorithm		$F1$ and $F2$ and $F3$ and $F4$ and $F5$ and $F6$ and $F7$
Decision tree		91.781
Naïve Bayes		88.394
Random forest		96.126
SVM		90.208
Logistic regression		99.012
KNN	$K = 5$	97.71
	$K = 10$	97.23

Table 6 Precision score on individual features

Algorithm		$F1$	$F2$	$F3$	$F4$	$F5$	$F6$
Decision tree		0.64	0.89	0.64	0.85	0.69	0.47
Naïve Bayes		0.89	0.73	0.64	0.92	0.66	0.43
Random forest		0.66	0.95	0.97	0.99	0.97	0.99
SVM		0.66	0.96	0.97	0.95	0.94	0.97
Logistic regression		0.91	0.87	0.98	0.96	0.93	0.99
KNN	$K = 5$	0.67	0.98	0.99	0.99	0.99	0.98
	$K = 10$	0.70	0.98	0.99	0.99	0.99	0.98

Table 7 Precision score on combined features

Algorithm		$F1$ and $F2$ and $F3$ and $F4$ and $F5$ and $F6$
Decision tree		0.86
Naïve Bayes		0.92
Random forest		0.97
SVM		0.96
Logistic regression		0.99
KNN	$K = 5$	0.98
	$K = 10$	0.98

4.3 ML Detection Precision on Individual Network Traffic Features

Comparison of Precision evaluated with the help of the stated ML algorithms on all the individual features are represented in Table 6.

Table 7 display the evaluated precision values when the ML algorithms were implemented on the combination of 6 features optimal feature set.

On performing the analysis, we observed that Logistic Regression had the highest average testing precision score of 0.94 when ML algorithms were implemented on individual features, and also, Logistic Regression outperformed the other ML algorithms when a combination of all the features was considered, with the highest precision score of 0.99. Once again, we can observe that we get a better precision score on combining the features than the individual features precision score.

5 Conclusion

In this paper, we compared the Machine Learning (ML) techniques on the normal and the intrusive network traffic dataset based on the detection accuracy and precision score. Once we extracted the network traffic features, we first selected the optimal set of features by performing the Feature Selection technique, Feature Importance. This project aims to find the optimal feature set, which would, in turn, provide us better detection results for the anomaly-based Network Intrusion Detection System with the help of the Machine Learning algorithms. On experimenting, we observed that the ML techniques performance was improved when implemented on the combination of network traffic feature set, that is, rather than implemented on the individual features. The highest accuracy of 99.048% and the highest precision score were achieved using the Logistic Regression technique when applied to the combined feature set. A detailed analysis of all the ML algorithms is also presented in our work. To the best of our knowledge, none of the existing work focuses on multiple supervised and unsupervised ML techniques. Also, the experiments in the existing work are performed on the standard dataset like KDD'99 or NSL-KDD datasets. We intended to select the most optimum feature set on the collected real-time normal dataset and perform the supervised and unsupervised ML techniques on them for effective intrusion detection. For our future work, we look forward to considering a larger, more extensive network traffic feature set in order to find a more optimal feature set for the improvement of intrusion detection. Perform intrusion detection based on the types of attacks involved in the network as our proposed work does not involve network attack classification, and also perform intrusion detection using a Deep Learning-based model.

References

1. Smys, S., Basar, A., Wang, H.: Hybrid intrusion detection system for internet of things (IoT). J. ISMAC **02**(04), 190–199 (2020)
2. Baraneetharan, E.: Role of machine learning algorithms intrusion detection in WSNs: a survey. J. Inf. Technol. Dig. World **02**(03), 161–173 (2020)
3. Sathesh, A.: Enhanced soft computing approaches for intrusion detection schemes in social media networks. J. Soft Comput. Paradigm (JSCP) **1**(02), 69–79 (2019)
4. Vengatesan, K., Kumar, A., Naik, R., Verma, D.K.: Anomaly based novel intrusion detection system for network traffic reduction. In: 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), pp. 688–690, Palladam, India (2018)
5. Gao, W., Morris, T.: On cyber attacks and signature based intrusion detection for modbus based industrial control systems. J. Dig. Forensics Secur. Law **9**(1), 37–56 (2014)
6. Jyothsna, V., Rama Prasad, V.V., Munivara Prasad, K.: A review of anomaly based intrusion detection systems. Int. J. Comput. Appl. **28**(7), 26–35 (2011)
7. Sinclair, C., Pierce, L., Matzner, S.: An application of machine learning to network intrusion detection. In: 15th Annual Computer Security Applications Conference (ACSAC'99), pp. 371–377, Phoenix (1999)
8. Mulay, S.A., Devale, P.R., Garje, G.V.: Intrusion detection system using support vector machine and decision tree. Int. J. Comput. Appl. **3**(3), 40–43 (2010)
9. Eesa, A.S., Orman, Z., Brifcani, A.M.A.: A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Syst. Appl. **42**(5), 2670–2679 (2015)
10. Kim, G., Lee, S., Kim, S.: A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Syst. Appl. **41**(4), 1690–1700 (2014)

11. Dreiseitl, S., Ohno-Machado, L.: Logistic regression and artificial neural network classification models: a methodology review. J. Biomed. Inform. **35**(5–6), 352–359 (2002)
12. Ghosh, P., Mitra, R.: Proposed GA-BFSS and logistic regression based intrusion detection system. In: 3rd International Conference on Computer, Communication, Control and Information Technology (C3IT), pp. 1–6, Hooghly (2015)
13. Bapat, R., Mandya, A., Liu, X., Abraham, B., Brown, D.E., Kang, H., Veeraraghavan, M.: Identifying malicious botnet traffic using logistic regression. In: Systems and Information Engineering Design Symposium (SIEDS), pp. 266–271, Charlottesville, VA (2018)
14. Bamakan, S.M.H., Wang, H., Tian, Y., Shi, Y.: An effective intrusion detection framework based on mclp/svm optimized by time-varying chaos particle swarm optimization. Neurocomputing **199**, 90–102 (2016)
15. Ahmad, I., Basheri, M., Iqbal, M.J., Rahim, A.: Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access **6**, 33789–33795 (2018)
16. Tao, P., Sun, Z., Sun, Z.: An improved intrusion detection algorithm based on GA and SVM. IEEE Access **6**, 13624–13631 (2018)
17. Farnaaz, N., Jabbar, M.: Random forest modeling for network intrusion detection system. Proc. Comput. Sci. **89**(1), 213–217 (2016)
18. Repalle, S.A., Kolluru, V.R.: Intrusion detection system using ai and machine learning algorithm. Int. Res. J. Eng. Technol. (IRJET) **4**(12), 1709–1715 (2017)
19. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: Knowledge discovery and data mining: towards a unifying framework. KDD **96**, 82–88 (1996)
20. Kim, D.S., Park, J.S.: Network-based intrusion detection with support vector machines. In: International Conference on Information Networking ICOIN 2003, Lecture Notes in Computer Science, pp. 747–756, Korea (2003)
21. Wang, H., Jie, Gu., Wang, S.: An effective intrusion detection framework based on SVM with feature augmentation. Knowl.-Based Syst. **136**, 130–139 (2017)
22. Gupta, S., Mamtora, R.: Intrusion detection system using wireshark. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **2**(11), 358–363 (2012)
23. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the kdd cup 99 data set. In: IEEE Symposium on Computational İntelligence for Security and Defense Applications, pp. 1–6, Otawa (2009)
24. Arora, A., Peddoju, S.K.: Minimizing network traffic features for Android mobile malware detection. In: 18th ACM International Conference on Distributed Computing and Networking ICDCN'17, no. 32, pp. 1–10, India (2017)
25. Arora, A., Peddoju, S.K.: Malware detection using network traffic analysis in android based mobile devices. In: 8th International conference on Next Generation Mobile Apps, Services and Technologies, pp. 66–71, India (2014)
26. Böhning, D.: Multinomial logistic regression algorithm. Annal. Inst. Stat. Math. **44**(1), 197–200 (1992)
27. Al-Sharafat, W.S., Naoum, R.: Development of genetic-based machine learning for network intrusion detection. Int. J. Comput. Inf. Eng. **3**(7), 1677–1681 (2009)

BGCNN: A Computer Vision Approach to Recognize of Yellow Mosaic Disease for Black Gram

Rashidul Hasan Hridoy and Aniruddha Rakshit

Abstract The yellow mosaic disease is a common black gram leaf disease that causes severe economic losses to local farmers and a hindrance to healthy production which can be prevented by computer vision based fast and accurate recognition system. In this paper, Black Gram Convolutional Neural Network (BGCNN) has been proposed for the recognition of this disease, and the performance of BGCNN has compared with the state-of-the-art deep learning models such as AlexNet, VGG16, and Inception V3. All the models have trained with original dataset having 2830 images and expanded dataset generated with image augmentation having 16,980 images that increase test accuracy of all the models significantly. BGCNN realizes accuracy of 82.67% and 97.11% for the original and expanded dataset, respectively. While, AlexNet, VGG16, and Inception V3 have achieved 93.78%, 95.49%, and 96.67% accuracy for the expanded dataset, respectively. The obtained results validate that BGCNN can recognize yellow mosaic disease efficiently.

Keyworks Yellow mosaic disease · Leaf disease · Black gram · Disease recognition · Computer vision · Convolutional neural networks · Deep learning · Transfer learning

1 Introduction

Black gram (scientific name: *Vigna mungo*) is widely used in Southeast Asian cuisine which is a remarkably prized pulse of this region. It is a type of bean, also known as black lentil. It is very nutritious, from ancient times, black gram has been cultivated. A remarkable number of south Asian farmers such as Bangladesh, India, and Nepal are engaged with black gram cultivation. Besides these countries, it is also cultivated

R. H. Hridoy (✉) · A. Rakshit
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
e-mail: rashidul15-8596@diu.edu.bd

A. Rakshit
e-mail: aniruddha.cse@diu.edu.bd

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_14

in the Caribbean, Fiji, Mauritius, Myanmar, and Africa. In medieval times, black gram is also used in the construction of the medieval crucible [1]. Cultivation of black gram is very easy and also provides a high investment return. Black gram has tremendous numbers of health benefits. It has an amazing ability to boost energy and improve immunity, also very helpful for psoriasis patients, as it reduces inflammation. For centuries, the black gram was remarkably used in skin and hair treatment which is the powerhouse of antibacterial properties that reduces acne. Moreover, it helps to control sugar levels, improves digestion and blood circulation, and contains various types of minerals that improve bone mineral density, also helpful for arthritis and osteoporosis patients. Black gram provides instant energy that strengths the nervous system and contains potassium that helps to control high blood pressure by preventing the constriction of blood vessels. Besides these health benefits, it also shows significant performance in both weight loss and gains, black gram oil is a traditional remedy for joint pain. Table 1 shows all the nutrition values of black gram after drying in the sun per 100 g [2].

Plants of a black gram can be affected by different kinds of diseases such as leaf spot, yellow mosaic disease, bacterial leaf blight, anthracnose, powdery mildew, rust, etc. Among these, yellow mosaic disease is the most common disease found in various areas of Bangladesh. A study conducted in Bangladesh found that 63% of farmers continuing black gram cultivation for higher yields and income. Besides 33% of farmers want continuing cultivation as its cultivation is very easy and not costly [3]. Initial symptoms of the yellow mosaic disease appear on young leaves of black gram with light scattered yellow spots. Day by day size of spots increases size, then some leaves completely turn yellow. Necrotic symptoms are also found in affected leaves. Affected plants become stunted and take more time to mature. These plants produce very few flowers. Pods of diseased plants turn in yellow [4]. Yellow mosaic is the most vulnerable disease than other diseases which decreased pod size and quality, affected plants contain fewer and smaller seeds also. In recent times, computer vision has shown surprising performance in various disease recognition related tasks such

Table 1 Nutritional values of raw black gram

Constituents	Approximate composition	Constituents	Approximate composition
Energy	341 kcal	Thiamin	0.27 mg
Carbohydrates	58.99 g	Zinc	3.35 mg
Protein	25.21 g	Sodium	38 mg
Total fat	1.64 g	Potassium	983 mg
Dietary fiber	18.30 g	Calcium	138 mg
Folates	216 mg	Copper	0.98 mg
Niacin	1.45 mg	Iron	7.57 mg
Pantothenic acid	0.91 mg	Magnesium	267 mg
Pyridoxine	0.28 mg	Phosphorus	379 mg

as leaf disease, fruit disease, crop disease, skin disease, etc. Convolutional Neural Network (CNN) is now widely used for recognizing various leaf diseases such as jute, cucumber, rice, wheat, grape, pumpkin, and tomato, etc. To recognize the yellow mosaic disease of black gram, CNN is used in this study. With minimal error, the proposed BGCNN recognizes yellow mosaic disease by classifying black gram leaf images. According to the experimental results, the BGCNN model has achieved 97.11% test accuracy, which is better than other classic models. In addition, after data augmentation, using a dataset of 16,980 images of black gram leaves, the accuracy increases by 14.44%. The performance of BGCNN is compared with state-of-the-art CNN architectures such as AlexNet, VGG16, and Inception V3. We have developed a variant of CNN architecture that consists of convolution, max pooling, and fully connected layers from scratch with fewer parameters which has achieved greater accuracy than other state-of-the-art CNN models. The major contributions are:

- Our proposed CNN architecture, namely, BGCCN has been developed from scratch to recognize the yellow mosaic disease of black gram.
- A new dataset of black gram leaves has been used in this study.
- Moreover, existing CNN models such as AlexNet, VGG16, and Inception V3 have also been used with the transfer learning approach.

The remainder of this study is organized as follows. Section 2 discusses the literature review. The dataset, deep neural network architectures, and experiment are given in Sect. 3. The results obtained in the study are given and discussed in Sect. 4. Finally, the study is concluded with Sect. 5.

2 Literature Review

A remarkable number of researchers have made enormous efforts to detect diseases of the leaf to minimize the damage of diseases. In the paper, Mia et al. [5] have used the support vector machine (SVM) to classify four diseases of the mango leaf and achieved an average of 80% accuracy. K-means clustering has been used for extracting the interesting region of leaf from L*A*B color space. Gray-Level Cooccurrence Matrix (GLCM) method has been used for extracting 13 features from the diseases affected region. To recognize two diseases of the coffee leaf, Sorte et al. [6] have compared the performance of Texture Based Disease Recognition (TBDR), and Deep Learning Disease Recognition (DLDR) approach. In TBDR, texture attribute vectors have been used and Patternnet feedforward neural network for training and testing. To extract statistical attributes GCLM has been used. They have obtained the best result in TBDR using the Local Binary Pattern (LBP). A modified AlexNet neural network has been used directly to the sample images in DLDR. This approach has performed better than TBDR, the Kappa coefficient is 0.970, and sensitivity is 0.980. Han and Watchareeruetai [7] have used AlexNet, VGG16, Inception, Xception, ResNet50, MobileNet, and MobileNetV2 to classify six nutrient deficiencies in Black Gram, and ResNet50 has performed best generalization performance with data augmentation. Precision, recall, *f*-measure, and test accuracy of ResNet50 are 68.01%, 64.39%, 66.15%, and 65.44%, respectively. Liu et al. [8] have proposed a CNN model named Dense Inception Convolutional Neural Network (DICNN) to diagnose six diseases of the grape leaf and achieved an accuracy of 97.22%. The performance of both Adam and SGD optimization algorithms has been analyzed, and with the same learning rate, SGD has performed better. With the dense connection, DICNN has performed better than without the dense connection. In this research, recognition performance of pre-trained models VGG16, GoogLeNet, ResNet34, DenseNet169, UnitedModel, and AFGDC is 88.96%, 94.25%, 94.67%, 94.89%, 96.58%, and 92.33%, respectively. Atila et al. [9] have analyzed the performance of EfficientNet, AlexNet, ResNet50, VGG16, and Inception V3 models to classify 26 leaf diseases of 14 plants. EfficientNet group consists of 8 models, these are B0, B1, B2, B3, B4, B5, B6, and B7. The input size of B4 and B5 of EfficientNet architecture model is 380 × 380 and 456 × 456 pixels, respectively. In the original dataset B5 model has achieved the highest accuracy of 99.91%. But B4 model has achieved the highest accuracy of 99.97% in the augmented dataset. In both types of datasets AlexNet has achieved the lowest accuracy, but while training it has taken the lowest time per epoch. Using shape, color, and texture features of leaves, Rao and Kulkarni [10] have achieved 93.18% accuracy. But using only the shape feature, their classifier has gain 91.74%. Using the GLCM feature, Gabor feature, and Curvelet feature extraction, a combined feature extraction model has been proposed in their study. For classification, the neuro-fuzzy classifier has been used to classify leaf diseases in their research. Based on multiple linear regression, Sun et al. [11] have proposed a disease recognition system for plant disease recognition. An improved histogram segmentation method has been introduced that can accurately calculate the threshold automatically and multiple linear regression and image feature extractions have been utilized in their research. In another research work, Mohanty et al. have used deep learning to detect plant diseases based on leaf image using 38 classes [12]. Their proposed model has achieved 99.35% accuracy on the test set, it can identify 26 diseases of 14 crop species. GoogLeNet architecture has performed better than AlexNet in their study. Three versions of the dataset have been used, and those are color, gray-scale, and segmented. Chandy has used deep learning for pest infestation identification [13]. Shakya have used SVM, KNN, random forest, and discriminant analysis to analyze image classification techniques based on artificial intelligence [14].

As a matter of fact, we come to know that not a single work has been introduced for yellow mosaic disease recognition using any computer vision approach and this is the first attempt to recognize this disease using state-of-the-art CNN models. Hence, it is essential to recognize yellow mosaic disease through a computer vision approach that will assist farmers to maintain the standard level of nutrition and enhance production by taking precautions for affected leaves of black gram.

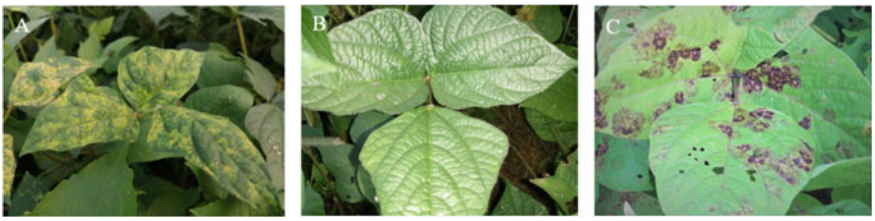


Fig. 1 Examples of each class of black gram leaf dataset, **a** yellow mosaic disease, **b** healthy leaf, and **c** miscellaneous

3 Materials and Methods

3.1 Black Gram Leaf Dataset

From five black gram plantation farms, images of the dataset were acquired with smartphone camera. Three representative images of the black gram leaf dataset belonging to each class are shown in Fig. 1. A total of 2830 images have obtained belonging to 3 classes, 43.46% of those were affected with the yellow mosaic disease. Dimensions of captured images were 4160 × 3120 pixels, both horizontal and vertical resolution was 96 dpi, bit depth was 24, ISO speed was ISO-50, focal length was 4 mm, and max aperture was 1.69. By using a library of Python 3, namely, Pillow all images have been reshaped for the purpose of this study.

3.2 Data Augmentation

To attain satisfactory performance in deep learning, a large amount of data is needed to train CNNs. In the training stage of these networks, overfitting is a common problem. It can be defeated using data augmentation. When a CNN network fits too well with the training set, then it becomes very difficult to generalize new data by the model that was not in the training set, then overfitting happens. Details on both original and expanded black gram leaf datasets are presented in Table 2.

Rotation, cropping, flipping, shearing, zooming, and changing the brightness level of image are the most commonly used operations of data augmentation. Images are rotated clockwise by a given number of degrees from 0 to 360 in the rotation augmentation technique, 50° rotation has been used in this study. The horizontal flip and vertical flip have been used which is an extension of rotation, in which the rows or columns of pixels are reversed. The width shift and height shift have been used to make shift-invariance to the images, the range value of 0.2 has been used in both width shift and height shift. The image generation process of the expanded dataset illustrates in Fig. 2.

Table 2 Summary of black gram leaf original and expanded dataset

Dataset name	Class name	Training images	Validation images	Testing images	Total number
Original	Yellow mosaic disease	762	234	234	1230
	Healthy leaf	642	214	214	1070
	Miscellaneous	318	106	106	530
Expanded	Yellow mosaic disease	4428	1476	1476	7380
	Healthy leaf	3852	1284	1284	6420
	Miscellaneous	1908	636	636	3180

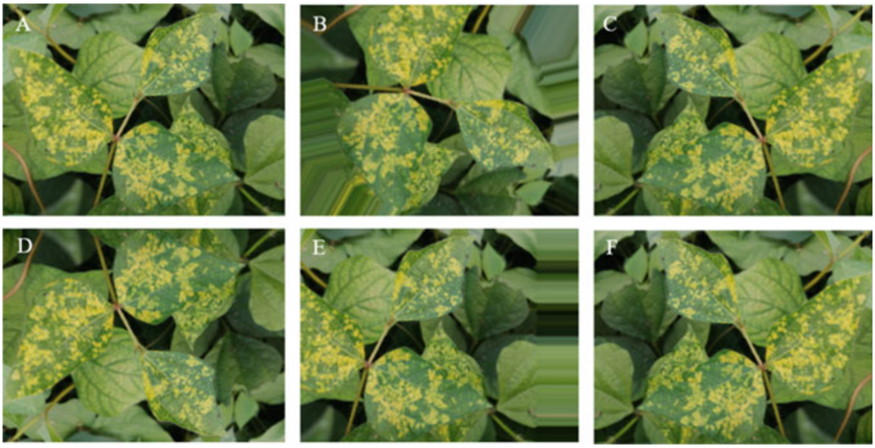


Fig. 2 Image augmentation of yellow mosaic disease image, **a** the original image, **b** rotation, **c** horizontal flip, **d** vertical flip, **e** width shift, and **f** height shift

3.3 CNN Based Models

CNNs have been used in this study to build classifiers for the yellow mosaic disease of black gram. The performance of the proposed CNN named BGCNN is compared with state-of-the-art CNN architectures such as AlexNet, VGG16, and Inception V3. The architecture of AlexNet consists of 5 convolutional layers with rectified linear unit (ReLU) activation function, 3 fully connected layers (FC), finally the Softmax layer and contains approximately 61 million parameters [15]. VGG16 contains five convolution blocks with 3 × 3 filters with stride 1 and same padding, maximum pooling layers that use 2 × 2 filters with stride 2, and three FC layers with approximately 138 million parameters [16]. To decrease the number of connections and parameters without losing the efficiency of the architecture, factorization is used in Inception V3 that consists of 42 layers. Symmetrical and asymmetrical building blocks of this architecture containing convolutions, max pooling, average pooling, concats, dropouts, and FC layers [17].

In the proposed BGCNN architecture, three convolution and maxpooling layer have been used. The convolution layer is the key element of CNNs, processes images using convolution filters. Using a set of convolution filters, the raw input image is directly applied to this layer. 3 × 3 kernel has been utilized by CNN for convolving the whole raw input image as well as the intermediate feature maps in the convolution layers. Activation functions decide whether the information received by a neuron is relevant to the given information or should be ignored. To make CNN capable of learning and performing complex tasks, these functions are applied to the input. In the proposed BGCNN architecture, ReLU has been used as it does not activate all neurons at the same time. To learn global features stacking of many convolution layers are needed, as first convolution layers extract edges, lines, corner, and other low-level features. So, three convolution layers have been used in this study. To reduce the spatial size of the representation, in CNN the pooling layer is used in between successive convolutional layers. In the network, it controls overfitting by reducing the number of parameters and computation. On the number of filters, polling has no effects [18]. By eliminating non-maximal values maxpooling layer reduces computation for upper layers. The FC layers are the last few layers in CNNs. All the features are extracted from the previous convolutional and subsampling layers are combined by this layer. In the last FC layer, the number or neurons is the same as the number of classes used to train the architecture. The size of the output layer of the proposed BGCNN is 3 as we have trained this model with three classes. As it is a multiclass classification, so the softmax activation function have been used in the FC connected layer of BGCNN which is a more generalized logistic activation. To update the weights of CNN, the loss function is used to calculate the gradients. Sparse categorical cross-entropy has been used as a loss function in this study. As an optimizer, stochastic gradient descent (SGD) has been used. SGD is performed while training [19]. The proposed BGCNN model have run for 80 epochs and have fixed 0.001 as the learning rate. Figure 3 shows the architecture of proposed BGCNN.

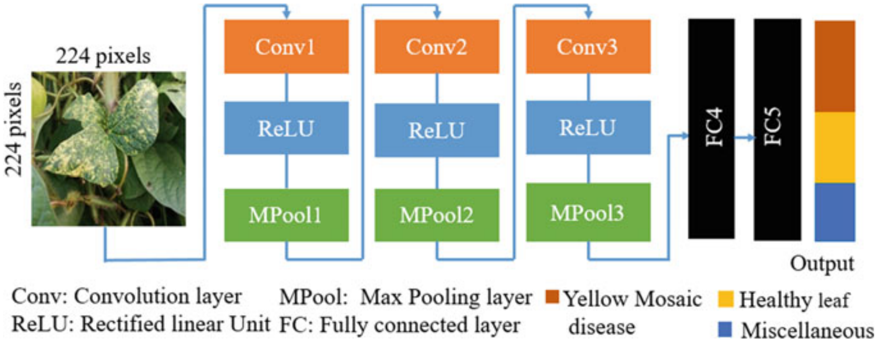


Fig. 3 Schematic representation of BGCNN

3.4 Experiments

In this study, all deep learning models have complied in Google Colab with GPU support. The multiclass classification has been performed in this study as the dataset contains three classes, yellow mosaic disease, healthy leaf, and miscellaneous. Both the original and expanded dataset of black gram leaf has been used in this study and the dataset is divided randomly into training, validation, and test sets. For training and fitting models, the training and validation sets have been used. On the other hand, the test set has been used to examine the recognition performance on images that models did not see before. With the transfer learning approach, existing CNN models have been used in this study. All the layers of these models have been set as trainable. The performance of deep learning models has been measured in this study using different metrics such as Precision (Pre), Recall (Rec), $F1$-Score ($F1$), Specificity (Spe), and Accuracy (Acc). The metrics given between Eqs. 1 and 10 are calculated using indices such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) by considering the values in the confusion matrix obtained in classifications. Here, TP is the number of total images in each class that is correctly classified by models. On the other hand, TN is the total number of correctly classified images in all classes without the relevant class. FP is the total number of misclassified images in all other classes without the relevant class, while FN represents the number of misclassified images of the relevant class. Accuracy is most commonly used to evaluate the performance of a model that is the fraction of predictions our models got right. For multi-class classification using macro-averaging, these metrics and their extended calculations are given in between Eqs. 1 and 10 [20].

For class *i*,

$$\text{Pre}(i) = \frac{\text{TP}(i)}{\text{TP}(i) + \text{FP}(i)} \tag{1}$$

$$\text{Rec}(i) = \frac{\text{TP}(i)}{\text{TP}(i) + \text{FN}(i)} \tag{2}$$

$$F1(i) = \frac{2 \times \text{Precison } \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Spe}(i) = \frac{\text{TN}(i)}{\text{TN}(i) + \text{FP}(i)} \tag{4}$$

$$\text{Acc}(i) = \frac{\text{TP}(i) + \text{TN}(i)}{\text{TP}(i) + \text{TN}(i) + \text{FP}(i) + \text{FN}(i)} \tag{5}$$

$$\text{AveragePre} = \frac{1}{\text{classes}} \sum_{i=1}^{\text{classes}} \text{Pre}(i) \tag{6}$$

$$\text{AverageRe} = \frac{1}{\text{classes}} \sum_{i=1}^{\text{classes}} \text{Rec}(i) \tag{7}$$

$$\text{Average}F1 = \frac{1}{\text{classes}} \sum_{i=1}^{\text{classes}} F1(i) \tag{8}$$

$$\text{AverageSpe} = \frac{1}{\text{classes}} \sum_{i=1}^{\text{classes}} \text{Spe}(i) \tag{9}$$

$$\text{AverageAcc} = \frac{1}{\text{classes}} \sum_{i=1}^{\text{classes}} \text{Acc}(i) \tag{10}$$

4 Result and Discussion

The main aim of this study is to examine the performance of BGCNN architecture in recognizing yellow mosaic disease of black gram and to compare it with the performance of AlexNet, VGG16, and Inception V3. Both original and expanded datasets have been used for all experimental studies. Table 3 summarizes the input size, number of parameters, optimization method, and learning rate used for four models.

Classification means classifying images of the dataset to a specific class. The TP, TN, FP, FN, precision, recall, $f1$-score, and specificity values acquired by four deep learning models for each class in the expanded dataset, are presented in Table 4. Considering the precision value of four deep learning models, BGCNN has shown the best performance ranged from 94.50 to 98.31%, among pre-trained models Inception V3 has shown the best performance 96.38–97.12%. The highest recall value of 97.80% has achieved by BGCNN in the healthy leaf class. On the other hand, BGCNN and Inception V3 have achieved the highest $f1$-score of 0.97 in the healthy leaf and yellow mosaic disease class. Inception V3 has achieved the highest specificity value of 99.17% in the miscellaneous class while the highest specificity of

Table 3 The image resolutions, optimization method, learning rate, and number of parameters for deep learning models

Model name	Input size	Optimization method	Learning rate	Number of total parameters
AlexNet	227 × 227	Adam	0.001	25,723,471
VGG16	224 × 224	SGD	0.01	134,272,835
Inception V3	299 × 299	Adam	0.001	21,808,931
BGCNN	224 × 224	SGD	0.001	3,240,163

Table 4 Class wise classification performance of deep learning models

Model name	Class	TP	TN	FP	FN	Pre (%)	Re (%)	$F1$ (%)	Spe (%)
AlexNet	Y	1373	1864	103	56	93.02	96.08	0.95	94.76
	H	1216	2004	68	108	94.70	91.84	0.93	96.72
	M	596	2713	40	47	93.71	92.69	0.93	98.55
VGG 16	Y	1402	1878	74	42	94.99	97.09	0.96	96.21
	H	1234	2041	50	71	96.11	94.56	0.95	97.61
	M	607	2720	29	40	95.44	93.82	0.95	98.95
Inception V3	Y	1423	1881	53	39	96.41	97.33	**0.97**	97.26
	H	1247	2066	37	46	97.12	96.44	**0.97**	98.24
	M	613	2732	23	28	96.38	95.63	0.96	**99.17**
BGCNN	Y	1451	1867	25	53	**98.31**	96.48	**0.97**	98.68
	H	1246	2084	38	28	97.04	**97.80**	**0.97**	98.21
	M	601	2743	35	17	94.50	97.25	0.96	98.74

(Y) Yellow mosaic disease, (H) healthy leaf, and (M) miscellaneous
Bold value indicates highest value of performance metrics obtained by the model

the yellow mosaic disease class has achieved by BGCNN. For the yellow mosaic disease class, BGCNN has been achieved precision the highest precision, 98.31%, and Inception V3 has the highest precision for both healthy leaf and miscellaneous class, 97.12%, and 96.38%, respectively.

On the expanded dataset, the average precision, recall, $f1$-score, specificity, and accuracy values obtained by four deep learning models are given in Table 5. In the training phase of all deep learning models, 80 epochs have been used. By dividing the total training time of each model by 80, the time per epoch has been calculated and presented in Table 5. Inception V3 has achieved the highest true classification rate of the samples that the model classifies as positive (precision). On average recall, BGCNN has performed better than other models. Inception V3 and BGCNN has achieved the highest average $f1$-score. The training of inception V3 has taken more

Table 5 Average results of deep learning models

Model name	Average-Pre (%)	Average-Re (%)	Average-$F1$	Average-Spe (%)	Average-Acc (%)	Time per epoch (s)
AlexNet	93.81	93.53	0.94	96.67	95.86	351
VGG16	95.51	95.15	0.95	97.59	96.99	1183
Inception V3	**96.63**	96.46	**0.97**	98.22	97.78	1904
BGCNN	96.61	**97.17**	**0.97**	**98.54**	**98.07**	**283**

Bold value indicates highest value of performance metrics obtained by the model time than others, completed in 42 h 19 min. BGCNN has achieved highest average specificity, accuracy with the lowest training time per epoch.

A comparative experiment was introduced in this section to examine the effect of data augmentation on classification accuracy. Figure 4 shows a remarkable change in test accuracy of all deep learning models after using the expanded dataset of black gram. Test accuracy of BGCNN using the original dataset is 82.67% while using the expanded dataset it has achieved 97.11% test accuracy. Inception V3 has performed better than other pre-trained deep learning models using the original dataset. On the other hand, BGCNN has achieved the highest test accuracy with the expanded dataset. Using the original dataset, AlexNet, VGG16, and Inception V3 have achieved recognition accuracy of 71.14%, 76.35%, and 79.24%, respectively. On the other hand, AlexNet, VGG16, and Inception V3 have achieved 93.78%, 95.49%, and 96.67%, respectively, recognition accuracy with the expanded dataset of black gram. The outcome of this experiment demonstrates that models trained with the expanded dataset can learn more suitable features, under various environments it enhances the anti-interference performance.

To illustrate the performance of deep learning models with expanded datasets more perceptibly, the total number of incorrect classifications for each class was given in Table 6. One remarkable performance has been performed by BGCNN for the yellow mosaic disease class. On the other hand, the lowest number of misclassifications has been performed by Inception V3 for the miscellaneous class.

In this study, a new CNN model has developed from scratch which achieved 86.06% training accuracy and 82.67% test accuracy using the original dataset, after

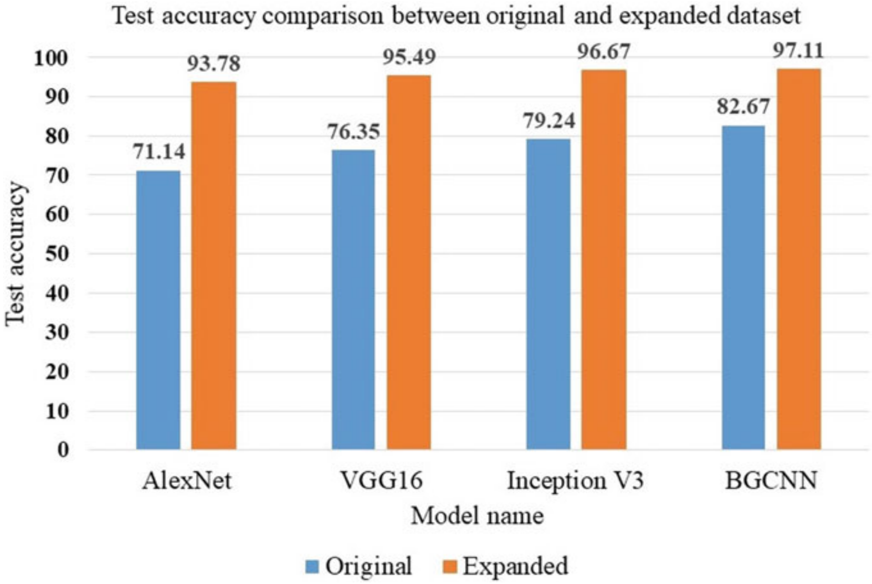


Fig. 4 Test accuracies for deep learning models on both original and expanded datasets

Table 6 Misclassification numbers of deep learning models for each class

Class name	AlexNet	VGG16	Inception V3	BGCNN
Yellow mosaic disease	103	74	53	25
Healthy leaf	68	50	37	38
Miscellaneous	40	29	**23**	35
Total false predictions for expanded dataset	211	153	113	98

Bold value indicates highest value of performance metrics obtained by the model

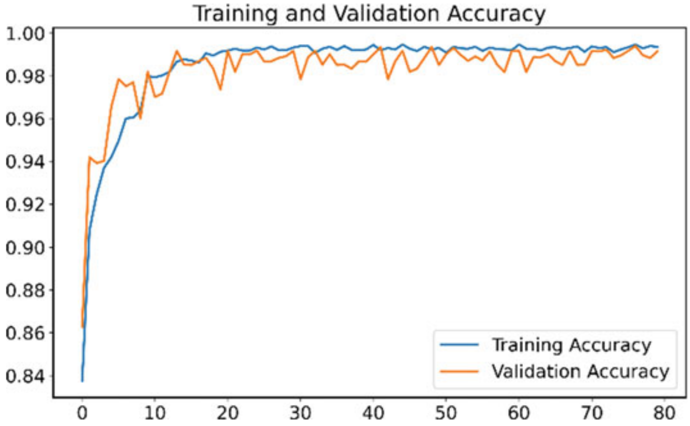


Fig. 5 Training and validation accuracy curve of BGCNN for expanded dataset

using the expanded dataset it has shown superior recognition performance, training accuracy increased to 97.81%, and test accuracy increased to 97.11%. Accuracy and loss curve of both training and validation of BGCNN with expanded dataset are shown in Figs. 5 and 6. The other three pre-trained models have also performed better with the expanded dataset, among these Inception V3 has performed better than others.

State-of-the-art CNN models need more training time and contain a large number of layers and parameters compare to BGCNN. These models have taken more time than BGCNN during the prediction of unknown images.

5 Conclusion

This paper has proposed a deep learning architecture to recognize the yellow mosaic disease of black gram, namely, BGCNN. Using image augmentation 16,980 images have been created based on 2830 images of black gram leaves. The performance of BGCNN has also compared with the state-of-the-art deep learning architectures used in leaf disease recognition in the literature. The success of BGCNN has significantly

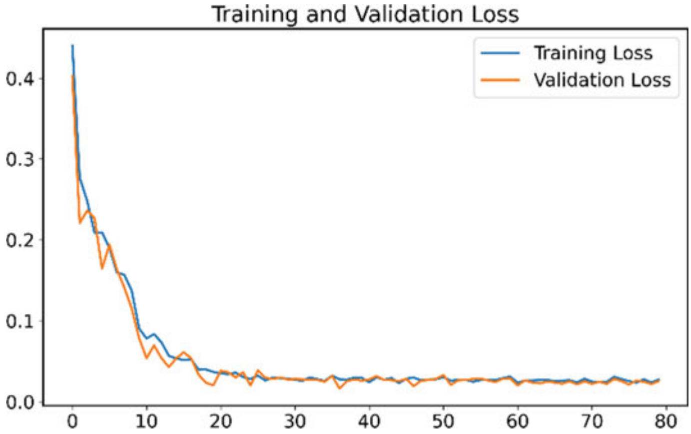


Fig. 6 Training and validation loss curve of BGCNN for expanded dataset

changed after using the expanded dataset. The BGCNN model has achieved 82.67% test accuracy with the original dataset, while BGCNN with the expanded dataset has achieved 97.11%. Besides, this study also illustrates the advantage of using pre-trained models, especially if the training dataset is not large. On the other hand, when the training time of models analyzed, BGCNN had taken the lowest time with 80 epochs in both the original and expanded dataset. This study aimed to develop a recognition approach for rapid and accurate diagnosis of yellow mosaic disease, our model is unable to recognize other diseases of black gram which is a limitation of this work. In feature works, it is planned to expand the black gram leaf disease dataset and the number of classes.

References

1. Vigna mungo. https://en.wikipedia.org/wiki/Vigna_mungo
2. Black Gram: Nutrition, Therapeutic Benefits, Uses for Skin and Hair. https://www.netmeds.com/health-library/post/black-gram-nutrition-therapeutic-benefits-uses-for-skin-and-hair
3. Mohiuddin, M., Akter, N., Khanum, R.: Economics of black gram cultivation and its impact on farmers livelihood in two selected districts of Bangladesh. SAARC J. Agric. **16** (2018)
4. TNAU Agritech Portal: Crop Protection. https://agritech.tnau.ac.in/crop_protection/black_gram_disease/blackgram_d8.html
5. Mia, M.R., Roy, S., Das, S.K.: Mango leaf disease recognition using neural network and support vector machine. Iran J. Comput. Sci. **3**, 185–193 (2020)
6. Sorte, L.X.B., Ferraz, C.T., Fambrini, F., dos Reis Goulart, R., Saito, J.H.: Coffee leaf disease recognition based on deep learning and texture attributes. Procedia Comput. Sci. **159**, 135–144 (2019)
7. Han, K.A.M., Watchareeruetai, U.: Classification of nutrient deficiency in black gram using deep convolutional neural networks. In: 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), Chonburi, Thailand, pp. 277–282 (2019)

8. Liu, B., Ding, Z., Tian, L., He, D., Li, S., Wang, H.: Grape leaf disease identification using improved deep convolutional neural networks. Front. Plant Sci. **11** (2020)
9. Atila, U., Uçar, M., Akyol, K., Uçar, E.: Plant leaf disease classification using EfficientNet deep learning model. Ecol. Inform. **61** (2019)
10. Rao, A., Kulkarni, S.B.: A hybrid approach for plant leaf disease detection and classification using digital image processing methods. Int. J. Electr. Eng. Educ. (2020)
11. Sun, G., Jia, X., Geng, T.: Plant diseases recognition based on image processing technology. J. Electr. Comput. Eng. **2018**, 1–8 (2018)
12. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Front. Plant Sci. **7**, 1419 (2016)
13. Chandy, A.: Pest infestation identification in coconut trees using deep learning. J. Artif. Intell. **1**(01), 10–18 (2019)
14. Shakya, S.: Analysis of artificial intelligence based image classification techniques. J. Innov. Image Process. (JIIP) **2**(01), 44–54 (2020)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: 25th International Conference on Neural Information Processing Systems, vol. 1, pp. 1097–1105 (2012)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. http://arxiv.org/abs/1409.1556 (2014)
17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.B.: Rethinking the inception architecture for computer vision (2016). https://doi.org/10.1109/CVPR.2016.308
18. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: International Conference on Engineering and Technology (ICET), pp. 1–6 (2017)
19. Wijnhoven, R.G.J., de With, P.H.N.: Fast training of object detection using stochastic gradient descent. In: 20th International Conference on Pattern Recognition (ICPR), pp. 424–427 (2010)
20. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45**, 427–437 (2009)

Irrelevant Racist Tweets Identification Using Data Mining Techniques

Jyothirlatha Kodali, Vyshnavi Kandikatla, Princy Nagati, Veena Nerendla, and M. Sreedevi

Abstract In recent times, Twitter is one of the major sources to access information. Its feature of the hashtag is something that grabs more attention from the users. One can write one's mind and heart out on Twitter at any given minute. Due to which there is a rapid increase in the generation of irrelevant content on Twitter. Lately, a new hashtag called "#whitelivesmatter" was used as a counter for another hashtag "#blacklivesmatter". A lot of anti-government protests and various other violent activities were conducted, recorded, and posted on Twitter with this hashtag. A lot of Kpop fans had taken over this hashtag and flooded Twitter with extremely irrelevant content. Due to which the main and important content of the protests was drowned in these irrelevant tweets, which made it extremely hard for the officials to find and reinforce the law and order. This paper aims at building a model that helps in finding the relevance of text content in the tweet and its hashtag #whitelivesmatter in specific. In this paper, supervised data analysis techniques like text classification are used to get the required output.

Keywords Tweets · Racism · Text classification · Naive Bayes (NB) · Support vector machine (SVM)

1 Introduction

Social media has now radically changed the way people share their ideas, viewpoints/opinions, informational knowledge. As the world locked down due to the pandemic of Coronavirus (COVID-19) since the early days of 2020, people began to spend their time on social media like never before. In the year 2020, the number of active users of social media is 3.81 billion, 9.2% increase from 3.48 billion in 2019.

Twitter is one of the leading microblogging services and social media sites for the public to express their views and share information with their followers. With millions of daily users who are networking with each other through tweeting, retweeting,

J. Kodali · V. Kandikatla · P. Nagati · V. Nerendla · M. Sreedevi (✉)
Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur,
Andhra Pradesh, India
e-mail: msreedevi_27@kluniversity.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_15

replying, sharing, etc. Twitter is considered to be a large platform for celebrities to promote and update their fans, government officials updating their people, leaders to share their views and provide reliable information. Twitter has proven to be very helpful in emergencies and disasters, by quick spread and response [1].

Twitter hasn't just been a platform for tweets and information but either for scams, hacks, and spread of false information [2]. The study of Vosoughi explains that the reach of false news and false rumors is six times greater than the accurate stories on Twitter [3]. One among those misleading information is the WhiteLivesMatter movement, in contradiction to BlackLivesMatter movement, that raged with the death of Black-American George Floyd.

From the point of rising of the Black Lives Matter movement in 2013, with the aid of a Twitter hashtag #BlackLivesMatter, the movement has been depicted and documented in various films, songs/music albums in various television programs, in electronic and print media, literature, and many other visual arts. "White Lives Matter" is a chauvinistic phrase for white people that originated in early 2015 as a racist response to the Black Lives Matter movement, which originated as protest against the brutality against the African-Americans by the patrolling officers garnered ample amount of publicity and exposure in 2014 for the protests in Ferguson.

The BlackLivesMatter hashtag gained popularity with celebrities and influencers posting in support of the movement. These racism tweets have the hidden or buried crimes that were made against the black people and protests. In rivalry to the BlackLivesMatter movement, the movements such as WhiteLivesMatter, AllLivesMatter, and BlueLivesMatter, exposing the indiscrimination in their respective races. People used #WhiteLivesMatter to expose the protestors and their relentless behavior towards them. A sudden increase in the hashtag mentioned later has created confusion in the netizens, as there was almost no proper reason for the hashtag to be in trend. A deep dive into the hashtag revealed that the Fans of Korean pop music (K-pop) and TikTok users have taken over the #WhiteLivesMatter hashtag by tweeting with random text and images and videos of their favorite singers, memes, GIFs. Use of this particular hashtag was increased as a retaliation to the #BlackLivesMatter movement, with many posters opposing or criticizing the protests. Later on K-pop fans, especially, flooded the hashtag with photos and videos which were extremely unrelated to the topic, the hashtag started to gain popularity on Twitter platform as kpop than as racism. This caused trouble to Dallas Policemen, who are identifying the illegal activities through the WhiteLivesMatter hashtag. Instead, the tweets have shoved down the essential information against the protestors who are causing the damage and decreasing trust on the platform.

Hashtags are one of the key components of tweets as they are used to categorize the tweets and are useful in detecting tweets with specific topics [4]. With irrelevant hashtags being tagged to the tweets, the search engine in Twitter is providing the users with unimportant tweets. Identifying such misleading tweets and providing the relevant tweets for the tagged hashtags is necessary, especially when a user searches for them specifically.

This paper mainly focuses on the tag Whitelivesmatter which trended on Twitter with million tweets, in which the majority of the tweets were posted with various fancams and media related to Kpop idols and unrelated tweet text. This is achieved through Natural Language Processing (NLP) techniques and Machine Learning (ML) algorithms by training them accordingly.

2 Literature Review

One of the referred articles proposed a framework for supervised sentiment classification, to be used on Twitter data, a popular social media channel. This framework is useful in eliminating the need for labor-intensive manual annotation by using 50 Twitter tags and 15 smileys [5] as sentiment classifiers, allowing different types of tweets posted to be classified and labeled based on their sentiment types. But this paper mainly focuses on whether the tweet posted is related to its hashtag or not but not on their sentiment types.

The reviewed papers mainly concentrate on the information posted on the Twitter platform and classify them into spam or ham tweets. Advertisements related to different blogs, products that are posted on Twitter using the currently popular hashtags are to be considered as spam messages. Here, the advertisements posted can either be related or irrelevant to the hashtag added. In either case, the spammers intend to promote their content using the popularity of that hashtag [6]. Due to this, a lot of relevant and useful information gets drowned among all the spam messages and it becomes challenging for legitimate users to find the content they need.

Pervin [7] analyzed and concluded that if a hashtag is appearing along with some other hashtags which are similar to it in terms of context and area, it’s popularity increases. But in the same scenario if the similar ones are replaced with random irrelevant tags then popularity declines. Vijayakumar and Vinothkanna [8] compared CapsNet algorithm and existing algorithms in classifying the texts based on their font style using tokenized images of the text. Dann [9] suggests that when the process of data collection is based upon searching the suitable resources for answering a specific question then the data which is needed and not needed can be segregated easily. Herzallah et al. [10] utilized features based on graphs, behavioral patterns of the users, and content being posted to classify users either as a spammer or a nonspammer. Inuwa-Dutse et al. [11] has devised an approach to determine the pattern in which spam tweets are posted. Narasamma and Sreedevi [12] has attempted an approach using diverse algorithms related to Support vector Machine integrated with some clustering techniques to recognize spam or repeated tweets. Twitter data analysis book written by Kumar [13] has provided a lot of insights for extraction of different types of data from Twitter. In the referred paper, Sungheetha and Sharma [14] advocated an approach where scalar-output feature detection is replaced with vector-output capsules so that it can be used in preserving extra information like thickness and position of the features.

This present article aims to find whether the content posted and the hashtag given to it is relevant or not. The similarity between the above-referred paper and this paper is that in both cases the aim is to find whether the information published is useful to the public or not. The variation in these cases is that the above-referred paper mainly focuses on whether the posted content is spam (mostly advertisements) or not whereas this paper mainly focuses on determining whether the tweet is related to the attached hashtag or not. In this paper, a specific hashtag "#whitelivesmatter" is considered which has recently been popular on various social media platforms.

3 Gathering Twitter Data

To train and experiment with the model, racism and kpop related tweets which consists of the hashtag "whitelivesmatter" are required. As there is no free-source dataset that is required for the paper, the tweets are extracted through multiple procedures.

Tweets are accessible to researchers, scholars, and practitioners through public Twitter API. It is considered the best legal approach to extract the tweets from Twitter without violating the Twitter Terms of Service [15]. The Twitter API is authenticated using the access tokens provided by Twitter for the registered developer account. The requests are made through Twitter REST API using Tweepy to extract the tweets. The Cursor of Tweepy responded with an object of JavaScript Object Notation (JSON) format [16]. The object is casted into Pandas DataFrame. The limitation for a developer is restricted to retrieve past 7 days' data and limited requests in a specific period of time. The process resulted in retrieval of only 500 tweets. Most of the extracted tweets are of English (en) language.

Furthermore, tweets are extracted through the custom-built module using available packages of Python 3.8. A total of 1800 unique tweets with hashtag #whitelivesmatter between May and August are extracted from Twitter. Additionally, 800 tweets with #kpop are extracted to provide a balancing data to train the model.

4 Preparing Data

The extracted tweets dataset has the following attributes:

1. avatar: the link of the user's profile picture
2. data-conversation-id: the unique id for the original conversation thread
3. data-id: the unique id for the tweet
4. date: date of tweet posted
5. tweet: the text of tweet
6. username: the unique id of user posted the tweet.

In all the retweets, mentions of the users and the duplicated tweets are ignored and omitted. The extracted tweets of K-pop and Racism tweets are merged together as one dataset. The text of the tweet is not perfect in grammar and is mostly not suitable for a machine to understand. The tweets have emojis, acronyms, expressions, special characters, hashtags, mentions, newline characters, links of pictures, and videos. So, preparing the tweet text for the machine to understand better is the key factor for most of the text-based projects. It is also stated that the efficiency or accuracy of the classification model can be affected by the appropriate preprocessing tasks [17]. The hashtags, mentions, external links, and media links are removed and placed in hashtag, mention, external_link, picture_link attributes respectively. As the main focus is on identifying the tweet's content, throughout the paper, only date and tweet attributes are considered.

The data extracted doesn't have the information of the category of the tweet or any label that defines the relevancy of the tweet to the hashtag. Therefore the 1116 tweets are manually labeled with 0 and 1. The relevant tweets are labeled as 0 and the remaining tweets are labeled as 1. The empty tweets are labeled as 1, i.e., irrelevant. From 1116 tweets, 716 tweets are treated as training data, and the remaining 400 tweets as testing data (Fig. 1).

Using the in-built packages, mostly Regular expression, the basic texts are corrected. The new line characters and emojis are replaced with single white space using the Regular expressions. The ISO formatted text of tweets are encoded to UTF format to apply the techniques using FtFy library available for Python 3.8. In order to ensure uniformity and improve the efficiency during vectorization, all the words are changed into lowercase and special characters are also omitted. The omitted special characters include "[‘-=~!@#$%^&*()_+[]{}0123456789 ;\'\\:"|<,./<>?]". The numericals in the tweet don't contribute to classification, rather they affect the model when the numericals are attached to the words. For instance, the word racism is different from racism1 but semantically, they are the same. Therefore, the numericals are removed from the tweets.

Performed the required pre-processing techniques upon the words from texts/tweets using Natural Language Processing (NLP) techniques available through NLTK library in Python:

1. Tokenization

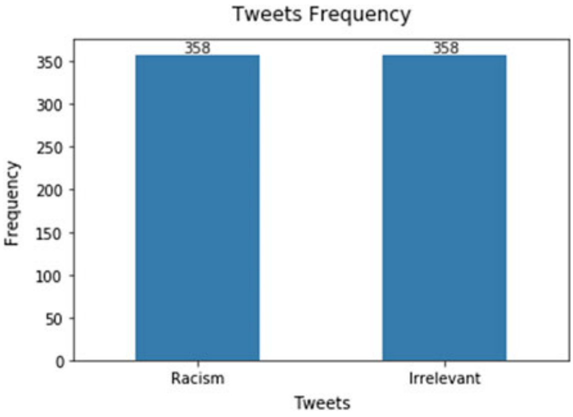


Fig. 1 The training data frequency

2. Stopwords
3. Lemmatization.

Tokenization

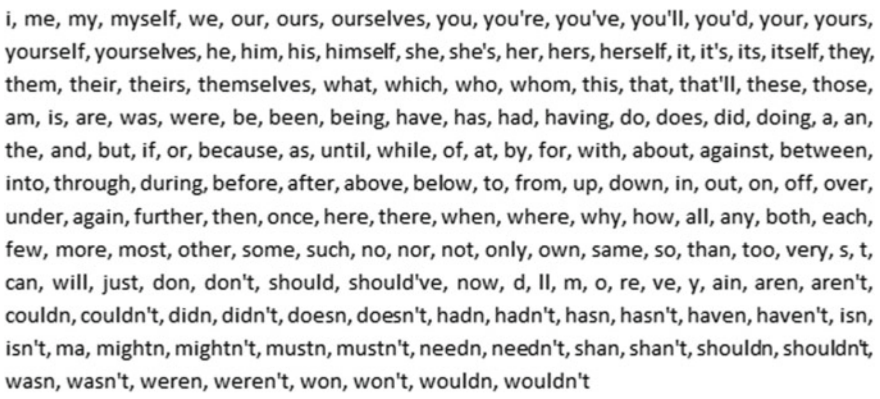
The text of tweets are combinations of multiple sentences and each sentence has multiple words. So, all these words get parted from the sentences into individual lists of singular words, eliminating accentuation, and converts all letters to lowercase. This technique is used to make a sentence into a list of words. After performing the tokenization process, the obtained tokens (words) are then used in further preprocessing steps like building a vocabulary set. Later on, this vocabulary set is sent for further processing where the frequencies of the tokens will be figured out and are also used to determine the significance of the words/tokens in the classification model.

Stopwords

Stopwords are often discarded from the text before training the models since they occur in abundance, hence providing little to no unique information that can be utilized for classification. These stopwords are available in the corpus module. The tokenized words are searched in this corpus and all those matched words are removed from the text (Fig. 2).

Lemmatization

Lemmatization is the process of eliminating the inflectional endings and returning the dictionary form of the words, i.e., lemma [18]. Lemmatization is preferred over stemming for this paper as stemming usually considers the stem part of the word and discards the remaining part. In such instances the words acquired may not have a proper meaning. As opposed to stemming, lemmatization means to acquire the standard (syntactically right) types of words, the supposed lemmas. The reason that lemmatization was opted for this paper is that the resultant words after processing



i, me, my, myself, we, our, ours, ourselves, you, you're, you've, you'll, you'd, your, yours, yourself, yourselves, he, him, his, himself, she, she's, her, hers, herself, it, it's, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, that'll, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, don't, should, should've, now, d, ll, m, o, re, ve, y, ain, aren, aren't, couldn, couldn't, didn, didn't, doesn, doesn't, hadn, hadn't, hasn, hasn't, haven, haven't, isn, isn't, ma, mightn, mightn't, mustn, mustn't, needn, needn't, shan, shan't, shouldn, shouldn't, wasn, wasn't, weren, weren't, won, won't, wouldn, wouldn't

Fig. 2 The list of stopwords

can be used for finding frequencies and significances. This may not be efficiently achieved if the words do not have any semantic or syntactic meaning.

Term Frequency

Another approach to characterize documents consisting of text—other than binary values—is called the term frequency (tf(t, d)). The term frequency can be defined as the number of times a term t (i.e., word or token) appears in document d.

$$\text{normalized term frequency} = \frac{\text{tf}(t, d)}{n_d} \tag{1}$$

where,

- $\text{tf}(t, d)$: Raw term frequency (the count of term t in document d).
- n_d: The total number of terms in document d.

Term Frequency-Inverse Document Frequency (TF-IDF)

The term frequency-inverse document frequency (TF-IDF) is utilized for portraying text documents. The inverse document frequency downscales the words that appear frequently among the documents and determines the importance of the word in classification. The TF-IDF approach imagines that the significance of a word is inversely relative to how frequently and regularly a feature (words) occurs in all the documents. TF-IDF is mostly used to rank documents in various content mining undertakings. TF-IDF values of the words are utilized to train the model efficiently.

5 Methodologies

The data is trained upon the following algorithms of text classifications:

1. Naive Bayes
2. Support Vector Machine (SVM).

Naive Bayes

Naive Bayes Algorithm is used as it classifies the text according to the probability of the occurrence of the tokenized words in that particular class. Naive Bayes is based on Bayes' theorem, in which Naïve says that features in the dataset are mutually independent. The occurrence of one feature does not influence the probability of occurrence of the other feature.

Steps of working:

1. Calculate prior probability for given class labels.
2. Calculate conditional probability with each attribute for each class.
3. Multiply same class conditional probability.

4. Multiply the same with step 3.
5. See which class has a higher probability, and it belongs to the given input set step.

As generally known, Naive Bayesian classification is of three types: Gaussian, Multinomial, and Bernoulli.

In this paper, Multinomial Naive Bayes has been used, as it predicts on the basis of the probability of occurrence of a term and its inverse document frequency.

Advantages:

1. Naive Bayes algorithms work rapidly and can save a lot of time.
2. It is suitable for solving multi-class prediction problems.
3. Naive Bayes suit categorical input variables than numerical variables.

Support Vector Machine (SVM)

"Support Vector Machine" is one of the widely used supervised machine learning algorithms which is used for classification as well as regression techniques. But it is more often used in classification than in regression. In the SVM algorithm, each data item is plotted as an n-dimensional space point with the value of each and every feature as the value of a particular coordinate. Then, classification is performed through a repetitive process of finding the hyper-plane which can efficiently differentiate the data points into required labels. SVM is used as it is good in handling high-dimensional data and also makes sure that the margin between the labels is maximized while decision making.

Advantages:

1. SVM gives very good results if the data points have a distinguished separation when represented in dimensional space.
2. Minor changes in data sat do not greatly affect the final results.
3. Non-Linear can be handled in an efficient way using the kernel functions.

6 Results

Evaluating a ML model is as important as building it. The models created through the data available from the extracted tweets can be used to estimate the irrelevancy upon newly gathered tweets. Hence a versatile and keen evaluation of the model is required to make it robust. The methods to assess a classification model are classification accuracy, confusion matrix, precision and recall score, F1 score.

The accuracy of any classification algorithm shows how many of the predictions are correct. In some scenarios, it represents how good a model is but in certain scenarios just the accuracy is not enough.

An accuracy of 94.5% was obtained by Naive Bayesian (NB) model and an accuracy of 96.0% was obtained by Support Vector Machine (SVM) model which have

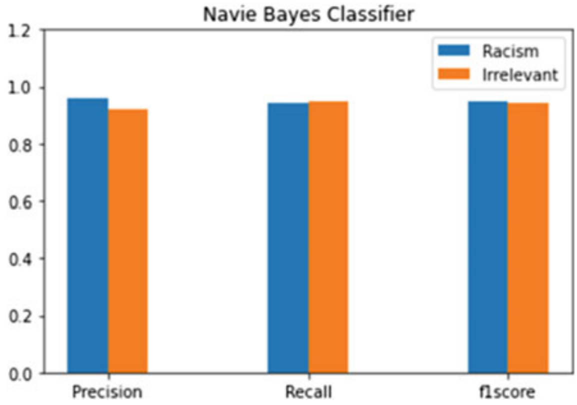


Fig. 3 Evaluation metrics—precision, recall, and f1 score of Navie Bayes Classifier

been trained to classify whether the posted context/text content in the tweets were relevant to the hashtag "#whitelivesmatter" or not.

A confusion matrix is not directly used to evaluate a model, but it provides a key insight towards the predictions. Confusion matrix helps comprehend other classification metrics such as precision and recall.

Precision: The estimation of how good the model is when the prediction of class labels is positive. It indicates how many positive predictions are true.

Recall: The measure of how good the model is in correctly predicting positive classes. It indicates how many positive classes the model can predict correctly.

F1 score can be explained as the Weighted average of the precision and recall scores. It is used when the problem consists of unevenly distributed data as it considers both false positive (FP) and false negative (FN). The ideal value of a F1 score is 1 and the worst is 0.

The Results of Naive Bayes Classifier

After applying the Naive Bayesian algorithm, the evaluation metrics are represented graphically in Fig. 3 and calculated using the confusion matrix shown in Fig. 4. Out of all 223 racism predictions by the model, 215 tweets were accurately predicted thus the precision acquired for the racism class was 0.96 and similarly the precision value for the irrelevant class was 0.92. Out of 229 actual racism tweets the model has predicted 215 correctly thus the recall obtained for the racism class was 0.94 and similarly the recall value for irrelevant class was 0.95. And the F1 score obtained for racism class was 0.95 and for the irrelevant class was 0.94. All the above-mentioned evaluation metrics are shown in Table 1.

The Results of Support Vector Machine (SVM)

After applying the Support Vector Machine (SVM) model, the evaluation metrics are represented graphically in Fig. 5 and calculated using the confusion matrix shown in Fig. 6. Out of all 217 racism predictions by the model, 215 tweets were accurately predicted thus the precision acquired for the racism class was 0.99 and similarly the

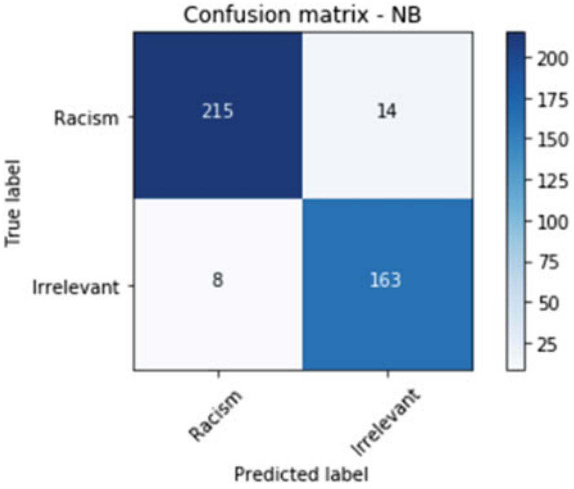


Fig. 4 Evaluation metric—confusion matrix of Navie Bayes Classifier

Table 1 Representation of evaluation metrics of Navie Bayes Classifier

	Precision	Recall	F1 score
Racism	0.96	0.94	0.95
Irrelevant	0.92	0.95	0.94

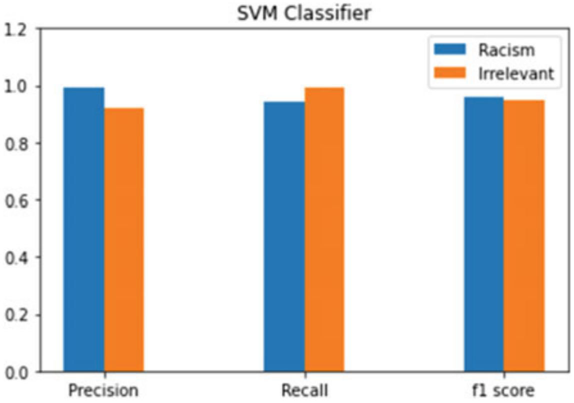


Fig. 5 Evaluation metrics—precision, recall, and f1 score of SVM classifier

precision value for the irrelevant class was 0.92. Out of 229 actual racism tweets the model has predicted 215 correctly thus the recall obtained for the racism class was 0.94 and similarly the recall value for irrelevant class was 0.99. And the F1 score obtained for racism class was 0.96 and for the irrelevant class was 0.95. All the above-mentioned evaluation metrics are shown in Table 2.

Though SVM and Naive Bayes resulted in the same f1 score and had similar values of recall. SVM resulted in much better accuracy and precision of classifying the tweets as either relevant (racist) or irrelevant tweets.

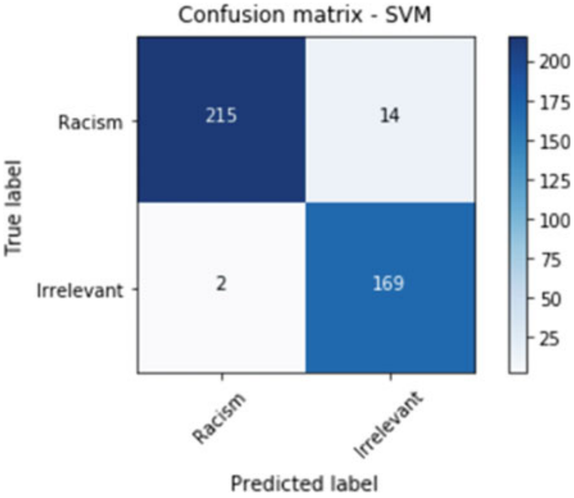


Fig. 6 Evaluation metric—confusion matrix of SVM classifier

Table 2 Representation of evaluation metrics of SVM classifier

	Precision	Recall	F1 score
Racism	0.99	0.94	0.96
Irrelevant	0.92	0.99	0.95

7 Conclusion

This paper inspects whether the tweets were related to their hashtag or not specifically for the hashtag whitelivesmatter. To classify the tweets based on relevancy with their hashtags, we have utilized SVM and Naive Bayes algorithms. They were used so that they can perform well even when the data set is not big enough unlike other deep learning techniques. Thus, by the performance of SVM and Naive Bayes models upon the data extracted from Twitter, it has been deduced that SVM models have performed well in predicting the irrelevancy of the tweets related to racism.

By using this model in the backend of Twitter search with the hashtag-whitelivesmatter can result in the appropriate tweets by avoiding all those that are irrelevant from the context of that hashtag.

The scope of this paper can further be broadened by considering various hashtags and diversified by implementing behind the search engines of various social-media platforms. This can be further improved through extracting more data of the tweets, images/videos posted along with them, and including them while classification. It can also be improved by finding out the regularity of the features and utilizing them in decision-making.

References

1. Cooper, Jr., G.P., Yeager, V., Burkle, Jr., F.M., Subbarao, I.: Twitter as a potential disaster risk reduction tool. Part I: introduction, terminology, research and operational applications. PLoS Curr. **7** (2015)
2. Halawi, B., Mourad, A., Otrok, H., Damiani, E.: Few are as good as many: an ontology-based tweet spam detection approach. IEEE Access **6**, 63890–63904 (2018)
3. Vosoughi, S.: Automatic detection and verification of rumors on Twitter. Mit.edu. Url: https://www.media.mit.edu/cogmac/publications/Soroush_Vosoughi_PHD_thesis.pdf
4. Kolos, S.: Hashtag as a Way of Archiving and Distributing Information on the Internet
5. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using Twitter hashtags and smileys. In: Coling 2010: Posters, Aug 2010, pp. 241–249
6. Sedhai, S., Sun, A.: An analysis of 14 million tweets on hashtag-oriented spamming. J. Assoc. Inf. Sci. Technol. **68**(7), 1638–1651 (2017)
7. Pervin, N., Phan, T.Q., Datta, A., Takeda, H., Toriumi, F.: Hashtag popularity on Twitter: analyzing co-occurrence of multiple hashtags. In: International Conference on Social Computing and Social Media, Aug 2015, pp. 169–182. Springer, Cham (2015)
8. Vijayakumar, T., Vinothkanna, M.R.: Capsule network on font style classification. J. Artif. Intell. **2**(02), 64–76 (2020)
9. Dann, S.: Twitter data acquisition and analysis: methodology and best practice. In: Maximizing Commerce and Marketing Strategies Through Micro-Blogging, pp. 280–296. IGI Global. Twitter, 2020, Terms of Service. Url: https://twitter.com/en/tos
10. Herzallah, W., Faris, H., Adwan, O.: Feature engineering for detecting spammers on Twitter: modelling and analysis. J. Inf. Sci. **44**(2), 230–247 (2018)
11. Inuwa-Dutse, I., Liptrott, M., Korkontzelos, I.: Detection of spam-posting accounts on Twitter. Neurocomputing **315**, 496–511 (2018)
12. Narasamma, V.L., Sreedevi, M.: A Comparative Approach for Classification and Combined Cluster Based Classification Method for Tweets Data Analysis. Url: https://link.springer.com/chapter/10.1007%2F978-981-32-9690-9_33
13. Kumar, S., Morstatter, F., Liu, H.: Twitter Data Analytics, pp. 1041–4347. Springer New York, New York, NY (2014)
14. Sungheetha, A., Sharma, R.: Transcapsule model for sentiment classification. J. Artif. Intell. **2**(03), 163–169 (2020)
15. Twitter, 2020, Terms of Service. Url: https://twitter.com/en/tos
16. Wolny, W.: Knowledge gained from Twitter data. In: 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk, pp. 1133–1136 (2016). https://doi.org/10.15439/2016F149
17. Uysal, A.K., Gunal, S.: The impact of preprocessing on text classification. Inf. Process. Manage. **50**(1), 104–112 (2014)
18. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008). https://doi.org/10.1017/CBO9780511809071

Smart Farming System Using IoT and Cloud

Neha Patil and Vaishali D. Khairnar

Abstract Due to the unprecedented increase in human population, agriculture plays an indispensable role in satisfying their daily needs. Henceforth, improving farm productiveness is indeed a huge challenge in the existing farming industry, which lacks continuous record management to satisfy the constantly emerging food needs. Along with the increasing population, global warming and climate transition also remain as an increasing challenge in the agricultural sector. In this scenario, this research has attempted to develop a smart farm management method, which incorporates cloud as well as the Internet of things (IoT) to take appropriate action. For instance, smart farming helps to provide a variety of important data such as air temperature. The paper has provided a smart device for farm field tracking, which controls dry run, motion detection, soil moisture detection, rainwater detection, humidity, and temperature. Also, this research work implements proper measures for those concepts on receiving the collected information without human input and later the detected quantities are stored for further data analysis within the cloud. Real-time feeds are being supervised upon this webpage as well as on the cell phone messaging. These would encourage farm workers and cultivate agricultural crops in a more modern way.

Keywords IoT · Agriculture · Monitoring · Sensors · Smart farming · Soil · Crop · Temperature · Production

1 Introduction

IoT is a modern communication and computing paradigm in which microcontrollers, sensors and transceivers have been fitted with everyday objects to feel the environmental parameters around them. In addition, the exchange of sensed data with each other or users becomes an essential element of the web system. By IoT, any object included with a unique identifier in our everyday life is linked so that they can transfer data without human interference over the network. Examples include the

N. Patil · V. D. Khairnar (✉)
Terna Engineering College, Mumbai, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_16

home control system, which uses Wi-Fi or Bluetooth to share data between different home devices [1].

IoT is a global network of devices used for intercommunication. Ubiquitous connectivity, ubiquitous computing and ambient intelligence are also combined. “The IoT is a vision, where ‘things’ become readable, identifiable, locatable, addressable and/or controllable through the Internet, particularly everyday items, such as all home appliances, furniture, clothing, cars, roads, smart materials, etc”. IoT refers to the networked interconnection coverage area of everyday tools, devices, artefacts or computers. Regarding place and time, these items may differ and they may be big or small. The idea is to use a related sensor or RFID or electronic technology like GPS to tag any object [2, 3].

In smart farming, the vital job is performed only by IoT systems. Smart agriculture is an evolving phenomenon since IoT sensors are able to provide data about their fields of agriculture. Thanks to their less costly sensors, IoT and wireless sensor networks (WSNs) allow new technologies and development ways for further accurate, feasible farming in the form of smart cultivation [4, 5].

Cloud computing refers to services on the cloud Internet that can be a centralized computer system or a distributed computer system. Parallel or distributed computing refers to the cloud or both. Clouds may be created over a large centralized or distributed data centre with virtualized or physical resources [2].

A new form of automation that is based on computing services, virtualize and service-based architecture is cloud computing. Many information technology (IT) firms, such as Google, Yahoo and Amazon, already provide customers with cloud services. Users might not think about hardware, software or other external tools in cloud technology. Here, they are not aware where the information is located on the cloud storage. Cloud computing offers users with an interface for data sharing [6].

It was instrumental in advancing farm IoT production. To begin with, cloud computing will offer growers low-cost storing information resources for text, image, video and other farm content, lowering expense of storing of farming businesses significantly [7].

Furthermore, using this pure information to draw predictions, terms of technical standard of growers are challenging. Just statistical study allows agriculture sector specialists to make reliable conclusions and recommendations. Thoughtful huge computational applications could only be supported with cloud services [8, 9].

In India, It seems to be the key essential profession of several other families. For the cultivation of crops such as beans, bajra, wheat, apples, tomatoes, bananas, corn, rice, cotton, jowar and cereals, about 60% of total of a land are being ploughed and then used. The IoT model and about the need for automated driving are being taken advantage of by agriculture. Interconnected vehicles can operate the farm of the twenty-first century: an enormous opportunity are being created mostly by the incorporation of various technologies deliver autonomous activities requires low oversight [10, 11].

During these respect, smart agriculture can thus benefit farmers and ensure that the country’s economy is strengthened when viewed on a massive level. A method is named precision agriculture, in which almost all the atmospheric impacts necessary for plant to produce were continuously examined. Tracking itself cannot enhance the well-being of its plants, and it is also necessary, if possible, to manage these aspects. Furthermore, all of the whole data gets preserved which could be utilized to additionally forecast the preferable product to be cultivated within this particular environment. To create an approach that helps the situation, the principles of IoT and cloud might be included [12].

Field automotive detection, farmed animals management, space checking, and various agricultural choices are among the IoT systems available in organic agriculture. Wise Natural Cultivation, that becomes presently common and widespread, could be incorporated on a huge scale, demonstrating that it would not be limited to big estates [13].

Since this activity would significantly reduce environmental inefficiencies of advanced farming, agricultural production should grow and evolve from where it is now. Smart metropolitan areas capture and analyse information using IoT tools like linked sensors, cameras and meters. These details are then used in cities to develop infrastructure utility services among other things [14].

2 Literature Review

In Xu et al., the author concentrated primarily mostly on IoT as well as its present research and its key strategies of encouraging (identification and tracking, communication, service management, networks involved). It sheds light on the industry's major IoT applications, and numerous research developments and challenges are described (standardization, complexity of design, accessibility, integration, etc.). Industry point of view discusses latest IoT studies. It begins by discussing the history of SOA structures of IoT before moving over to the recent applications which could been included in IoT. Following that, it goes over a few as the most industrially important IoT networks. Following that, it includes the methodology issues and IoT's development plans [15].

In Liqiang et al., the author collaborated with IoT for remote field tracking. There is also a picture sensing node that receives the crop images and uses different parameters for the inference, such as temperature and humidity. Low power consumption and reliable operation with an atmosphere of high precision end up making excellent choice for crops management. Two networking protocols are used in the control network. Tree-based collection standard protocol collects network information as well as sends it to an access point. The dissemination operation is a supplement to the gathering process [16].

In Keerthi et al., the author addressed IoT-based greenhouse monitoring schemes. Agribusiness schemes, in various types, have been on the rise in recent years, including in city environments. The farming industry grows rapidly as a result of scientific advances, that is aided by cloud IoT in this case. The Internet of things (IoT) can drastically alter how people manage out daily routine and how humans preserve knowledge regarding ourselves. Using different sensors, including temperature, humidity, and soil moisture, multiple parameters are effectively controlled. The coordinator node gathers the data from the harvest every 30 s and is processed on a cloud. It allows the user to review the information at any moment [17].

In Gutiérrez, an intelligent irrigation sensor was created by the writer. This paper uses the concept and implementation of automatic irrigation in the field of cultivation. In this procedure, digital images are obtained, using a smartphone, of the surrounding soil and the kernel of the crop. This helps to estimate the water content optically. A construction about an automated irrigation system centred on microcontrollers including wireless communication is introduced for an exploratory level to rural locations. An intention of introduction is just to indicate whether automated irrigation is being employed for focus on saving water [18].

In Gondchawar et al., The author addressed an agricultural management system based on IOT. A GPS-based robot is being used in this project. This paper presents smart irrigation including specific benefits and sensible decision and smart warehouse management systems. Several of such activities would be controlled via connecting sensors Wi-Fi or ZigBee units, cameras and actuators using both a microcontroller and perhaps a Raspberry Pi [19].

In Kaur et al., A system of drip irrigation was created by the author. It is a fully automated technique that helps by actually controlling irrigation water using an Android app to minimize intellectual labour. The irrigation is managed to track the environmental conditions set of values of the various sensors such as humidity and temperature applied in the field. A suggested device includes a microcontroller that captures sensors data of soils locations. Three-dimensional diagrams have been used in programming that shows such attributes. The equipment can be easily programmed using an access point GUI. The 3D diagrams created for sensed information around the overall landscape will aid us in visualizing and taking proactive steps in the given circumstance [20].

Krishna et al., to calculate various environmental parameters, and the author suggested a mobile robot fitted with different sensors. To execute the whole operation, it requires Raspberry Pi 2 Model B hardware. Aspects of this new smart cellular robot would perform assignment like detecting dampness in soil, pissing birds, spreading pesticides, going forth and reverse, and changing electric motor ON/OFF. The device is equipped via a camera module, to track the actions at the right life. The suggested wireless devices were tested all over the area and readings were tracked and acceptable results were observed, suggesting that this device is very suitable for smart agricultural practices [21].

Shenoy and Pingle explore potential ways to minimize transport costs for agricultural products and also forecast crop prices on the basis of past knowledge and current market scenarios. It also provides a solution to reduce intermediaries that typically aim to capture percentage of earnings between buyers as well as suppliers. The whole approach creates a balance among farm workers or even purchasers of farm commodities [22].

Khot and Gaikwad have the ability to track the intensity of the light all over the field and store data in a database to even further evaluate and examine it. It is extremely easy to use the field variable database to make an optimum decision within the specified time. As just another variable to be regulated by the agricultural sector, the procedure throughout the paper mainly focused on the brightness [23].

Sheetal et al., by regulating the environmental parameter, i.e. soil ph., attempt to overcome the issue in the grains due to the uneven distribution of rain. As a central controller, an Arduino is used. This governs the mechanism along with the contact process. A standard process would be made based upon this plant needs; unless some climatic weather such as air temp, soil quality, or humid, drops underneath exceeds the given limit, the IoT would detect the change of variables then send the information towards farm workers, which would then make the managing judgment and deliver this into a machine [24].

Brewster et al., have summarized technical limitations as well as obstacles that must be addressed during the implementation of a low-scale pilot project based on IoT in the farming. For all phases of agricultural goods, including crop production, manufacturing, distribution, and the retail industry, IT provides a conceptual idea. An architecture and design structure methodology are introduced, via a focus on the integration perspectives that are crucial mostly to absorption of IoT applications in the agro-based sector. The agriculture research information model framework is developed to recognize reliable communication, and farm information management remedy to ensure information sharing is highlighted [25].

Ayaz et al., has addressed several considerations and emphasized the importance for different innovation mainly Internet of things, for rendering farming wiser and more competitive to relation to meeting demands. Wireless sensors, unmanned aerial vehicles (UAVs), cloud computing, and communication technology are all extensively explored with this reason. In addition, a more in-depth look at latest study attempts is offered. The IoT used with WSN in cultivation implementations is thoroughly examined. Sensors for land readiness plant position, water management, insecticide and pesticide identification and other farming applications are described. Is so how is this new technology assisting farmers at all levels of a seed cycle from seedling to cultivating and packing is addressed in detail [26].

Johnson et al., expected to enable farming, where the author introduces different IoT services as well as sensors. Additionally, it discusses various information analysis approaches used on sensor information presented by IoT devices throughout the field, and their function with data science within agricultural production focuses about its concept of sensors, IoT, even data science in farming for others might want and take advantage including its integration with these technologies [27].

Ragavi et al., possess an adequate on direct seeding, as well as ARM processors and cloud-based IoT farming. This demonstrates basic ability by planting the seeds by coating it with soil, that becomes quite consistent through monitoring structures, like monitor temperature, moisture content, wetness, etc., and livestock motions, all of which would affect crop production [28].

Vineela et al., presented which data being obtained through various sensors but that actual tracking being carried on, but perhaps the priority in this literature review is still on get everything automatic. The writers want to use various technologies to increase agricultural productivity and also show how to use a low-cost WSN to collect data from humidity, soil moisture and temperature sensors. To good food crops, here article proposes an automated framework. The researchers present an approach for intelligent information sensing as well as a smart irrigation system. Various sensors are connected with Raspberry Pi throughout the suggested system resulting in an effective wireless sensor network [29].

Nayyar and Puri offer an appropriate environmental monitoring program which could assist farmers in practicing smart farming, thereby increasing total harvest and customer satisfaction. The farming stick suggested within those projects is built with Arduino technology and a breadboard interface to communicate with different sensors that transmit real-time data to Thingsspeak.com. Major limitation is that, it is costly to design [30].

3 Problem Statement

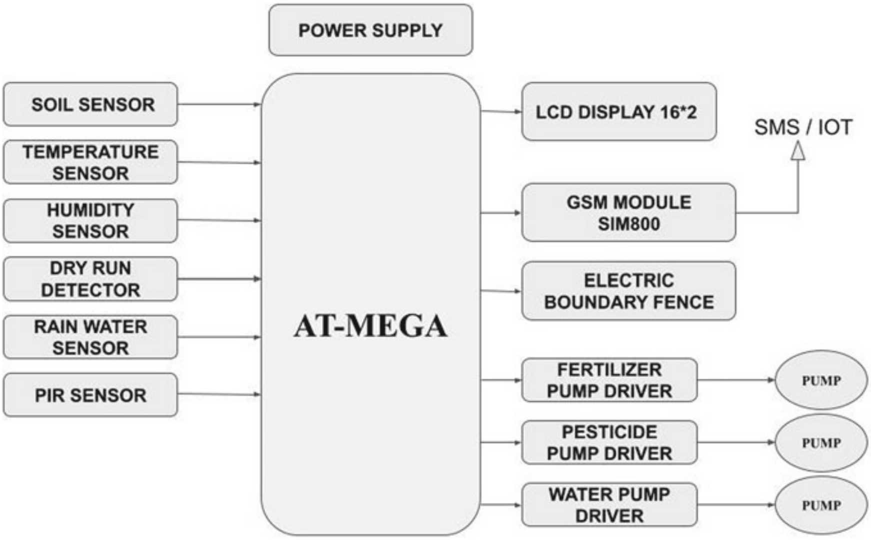
The researcher supposes to implement a "Smart Farming System" using IoT and cloud computing with various sensors, which will help to collect the data and analyse it. The proposed system collects information about different agricultural parameters (temperature and humidity) using an IoT sensor. These values collected are then sent over the mobile via SMS. Farmers can view all the parameters required for a smart farming system through the web page.

Throughout these plans, a GSM module is used that links via a GPRS Internet network. That is also attributable to the fact that its proposed work would have been performed outside, as seen in that midst about an agricultural field, where supplying Wi-Fi can become challenging, but though it sets up a Wi-Fi service outdoors, someone might crack the network. So, for this reason, cellular network is used for Internet connectivity. It will provide better range than Wi-Fi.

4 Proposed Work

The modern agricultural methods allow the use of sophisticated systems to be implemented for the betterment of job quality and increasing productivity. Such systems can provide a helping hand to the farmers and give them a chance to grow. The proposed "Smart Farming System" collects information about different agricultural parameters (temperature, humidity, soil moisture detection, rainwater detection, dry run detection and motion detection) using a wireless sensor network. These values collected are then sent over the mobile via SMS. All the values required by farmers for smart farming are viewed on the web page which is stored in the cloud database. Every system will have a unique ID which will help to identify sensor values for the farmers. The system also facilitates the controlling of pump motors through IoT and SMS. This facility helps farmers to turn the motor in their farm on or off at any time.

5 System Architecture



5.1 Project Working

The system is working around the ATMEGA328 microcontroller. The analogue sensors transform the physical parameters into the equivalent electrical pulses, and the digital sensors transform the physical parameters into the electrical pulses followed by digitization. The central controller samples the signals from all the sensors and stores them into the local memory. The controller also attempts periodic connections with the server to send the values stored in the memory to the server. Also, the values obtained after reading from sensor are processed further to make the decisions about pump operations and fencing supply. The LCD is used to display the status of the system as well as sensor values periodically.

The system operates in two modes, AUTO and MANUAL. In AUTO mode, as soon as the system gets power supply by turning the switch ON depending on the soil moisture data and if the soil needs moisture then, the water pump switches ON, while pesticide and fertilizer pumps will switch ON and OFF according to the time interval set in seconds. The sensors values are sent to the cloud via GPRS through GSM and the mobile message is also sent. Farmers can view the status on the website too. On the other hand, in manual mode, all three pumps, i.e. water, fertilizer and pesticide can be turned ON and OFF from the website.

5.2 Hardware Specification

GSM Module
GSM is functioning as the communication device to communicate over the mobile network. It is used to send SMS on mobile phones and the data over the server. The server communication is done via GPRS. The device connects to the Internet via an HTTP connection. It hits the PHP page on the server with the data in parameters. The link is renewed every time according to the data available. The GSM returns the echoed text to the device which is decoded by the central controller to get the commands from the server. The PHP file with "GET" method is to send data from the server. The PHP file communicates with the database on the server. It also extracts the data from the database and echoes it back. The front page is communicating with the database and updates it according to the user commands. It also fetches the data from the database and displays it on the screen. It also plots the data for graph.

ATmega328
In the AVR family, the ATmega328 is a low-power, low-cost and high-performance microcontroller built by Atmel. The ATmega328 is a single 28-pin chip with serial communication. The analogue input from of the sensor is collected, analysed, and the actuators are activated. Besides that, the details of the sensor would be submitted to a cell phone and a website through the use of the GSM module. The user's computer receives frequent notifications about the field's state for tracking.

LCD display 16 * 2
This is an electronic display unit used to illustrate parameters and their status in the device. The 2 lines and 16 characters per line are represented by the 16 × 2 LCD. All the sensors sense the values, and information is displayed on LCD in the form of status.

Relay
Relay is a switch that controls circuits opening and closing electromechanically. Without any human interference, it turns to switch ON or OFF to make or break contact by using signal. External adapter of 12 V is connected to the relay module.

5.3 Sensors

Soil Sensor
This sensor measures the conductivity of the soil. The water content and the minerals contribute to the soil conductivity. It measures the voltage drop across the resistance offered by soil. A volumetric liquid limit of specimen is determined using a soil moisture sensor. This sounds perfect for experimenting in courses like soil mechanics including environmental sciences, among many others. Soil moisture sensors are

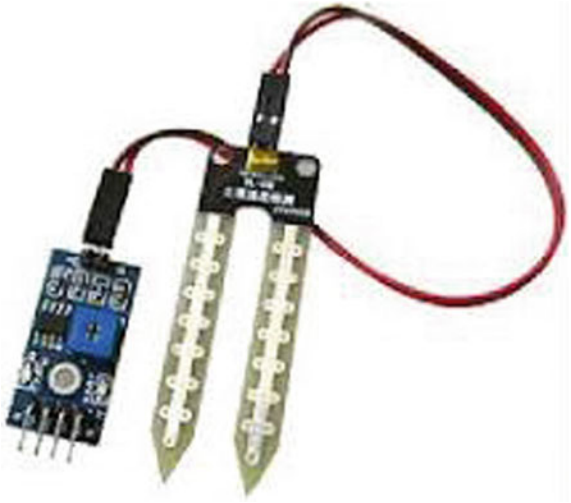


Fig. 1 Soil sensor

required to determine how much water is present in soil. Sensory threshold: 2.5 V (Fig. 1).

DTH11

DTH11 is basically the humidity and temperature sensor.

Temperature sensor: This sensor measures the temperature and its output voltage is proportional to the temperature. Sensor threshold = 350 mV.

Humidity sensor*:* This sensor measures the relative humidity of air and gives the output in digital format.

Sensor threshold = Relative humidity of 70% (Fig. 2).

Rainwater Sensor

This sensor senses, if there is any conductivity between the two plates due to rainwater.

Sensory threshold: Conductivity present or not (Fig. 3).

PIR Sensor

PIR stands for passive infrared sensor. This sensor detects the movement in the field and gives the output in form of a pulse.

Sensory threshold: Pulse present or not (Fig. 4).

Dry Run Detector

This sensor is the current sensor that detects the changes in the motor current due to no-load or only frictional load. Dry run is detection which is simulated via a potentiometer.

Sensory threshold: Lagging current (Fig. 5).

Fig. 2 Temperature and humidity sensor

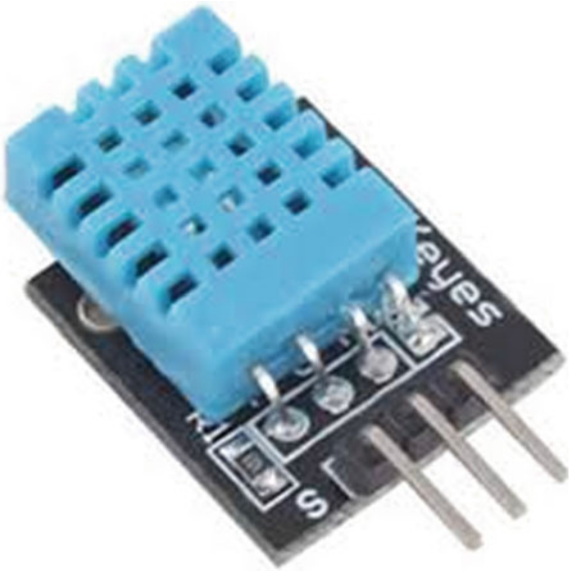
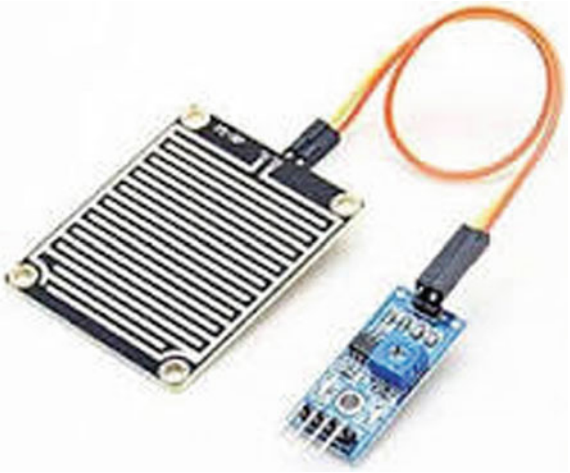


Fig. 3 Rainwater sensor



5.4 Software Specifications

Arduino IDE
The Arduino IDE is a software application that allows the Arduino board to compile and upload the C program. It is designed to work with many kinds of microcontrollers. If the code is compiled and published, the required action is carried out. Arduino IDE software performs microcontroller programming.

Fig. 4 PIR sensor

Fig. 5 Dry run via potentiometer

Application Web

With actual observation as well as measurement of smart farming, graphical interface application was created. It is based on soil moisture, temperature, humidity, motion detection and rainwater detection data. The program enables the user to use any device over the Internet to visualize the data graphically. In addition to graphical presentation, it also offers manual water supply, pesticide, and fertilizer control and all the records are collected on a server. For control as well as coding, the webpage is PHP-encoded. The database implementation is in MySQL.

My SQL

Using MySQL database and the connection to it is made through PHP. It connects to the database in PHP by using the MySQLi connect method with the parameters that host server, database name, username and password. If all the parameters are correct, the method returns an object of connection that is stored into a variable. This object is later used for any query in the database. Two queries namely INSERT and

SELECT are used. Whenever a hardware device hits the link with data, the PHP file on the server fetches the data and stores it into the local variables. It also fetches the date and time at that movement and stores it into the local variables. This data is then inserted into the database by using an INSERT query. On the display page, the information from the database is extracted by using a SELECT query and it is displayed in an HTML context.

AT commands

The GSM works on the AT commands. The central controller sends "AT" commands along with data to the GSM. Each "AT" command with data is counted as 1 step and we need 12 steps to access to the internet and get the information from it.

If expected data is received at the end of the 12th step then the initialization part is not repeated and the step number is resumed from step 10. The parameters in the link are updated as per the status of sensors and retransmitted to the server. When the expected data is not received due to a network issue or server issue, the step number is set to 1 and the whole process will be started from the beginning.

Mobile SMS

Another set of AT command is required to send the SMS. This command initiates a context in which the controller sends the SMS text data which is recorded. The decimal 26 is sent as a character to terminate the context and send SMS to the number.

6 Experiment and Results

Figure 6 shows the experimental set-up of proposed system. The proper and successfully interface of microcontroller, sensors, motor pump, relays and GSM module is done. External adapters are also connected to the circuit. After assembling the system, the sensor readings are checked and tested in different situations.

Figure 7 shows the soil moisture detection, where soil moisture sensor is inserted into soil in order to check the moisture level of the soil.

Figure 8 shows rain water detection, where rain sensor will detect the presence or absence of rain.

Figure 9 shows the status of sensors proposed in the project.

- D denotes dry run; it states Y if dry run and N if no dry run.
- S denotes soil moisture, if moisture is present it shows Y and if there is no moisture in soil then it shows N.
- P denotes person movement, if it detects any motion then Y or else if no motion then N.
- R denotes rain water, if it is detecting rain then Y, if no rain then Y.
- H denotes humidity in percentage.
- T denotes temperature in Celsius.

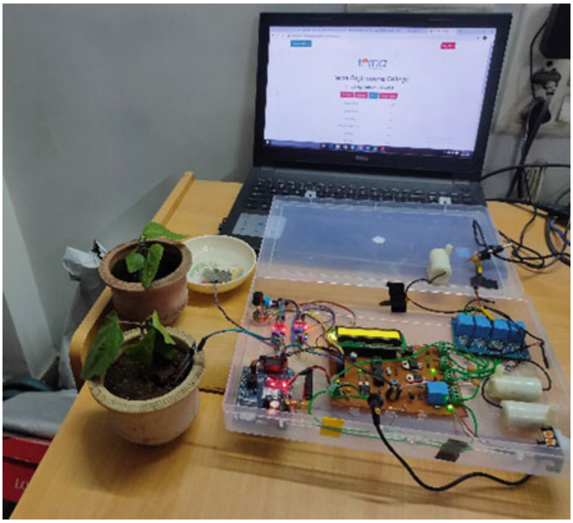


Fig. 6 Experimental set-up (interfacing microcontroller, sensors, relay and motor pump)

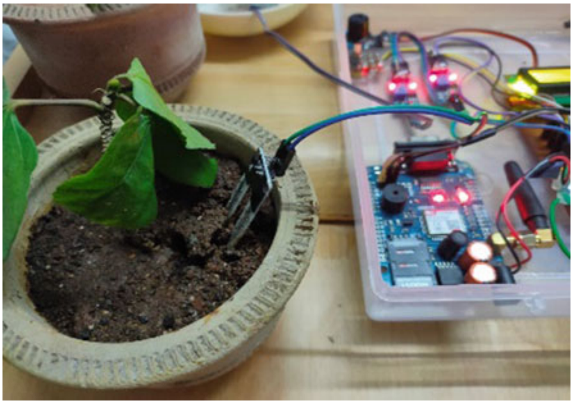


Fig. 7 Soil moisture detection

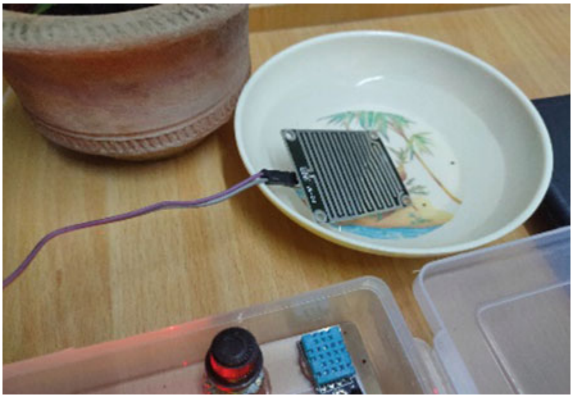


Fig. 8 Rainwater detection

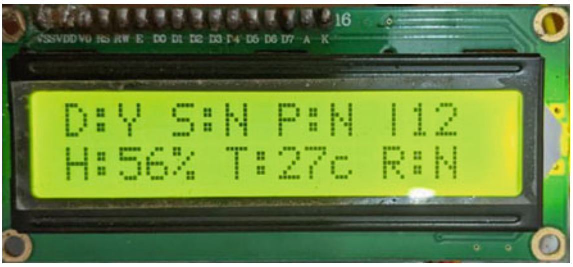


Fig. 9 Output displayed on LCD (sensors values)

Number indicates the steps required to send the data to the server, totally 12 steps are proposed in this system.

Figure 10 shows the details of the sensors via mobile SMS. The status obtained from the sensors are humidity, temperature, soil moisture, dry run and motion and further they are sent through the SMS.

Figure 11 is showing all the sensor details as well as it updates both date and time on the web page. Whereas during manual mode, farmers can manually switch ON/OFF the motor pump as per the water, pesticide and fertilizer requirements.

Figure 12 shows the graphical representation of temperature and humidity on the web page with respect to date and time.

Figure 13 shows the detailed records of various values, which have been stored in the database. Farmers can easily download this excel sheet from the web page directly in order to take any required decisions.

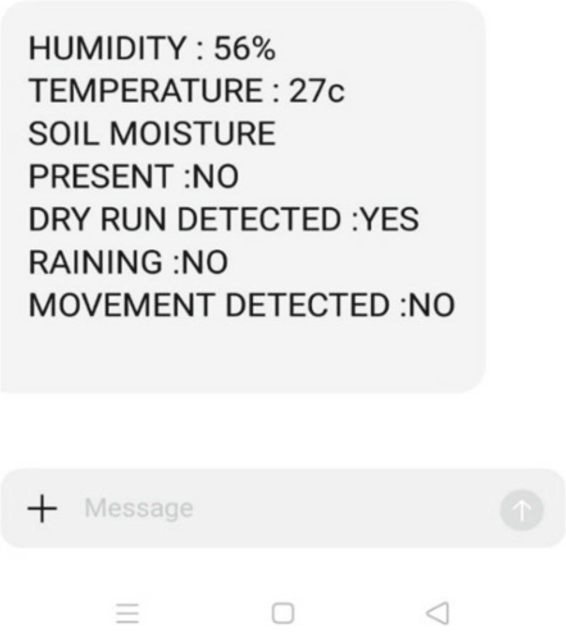
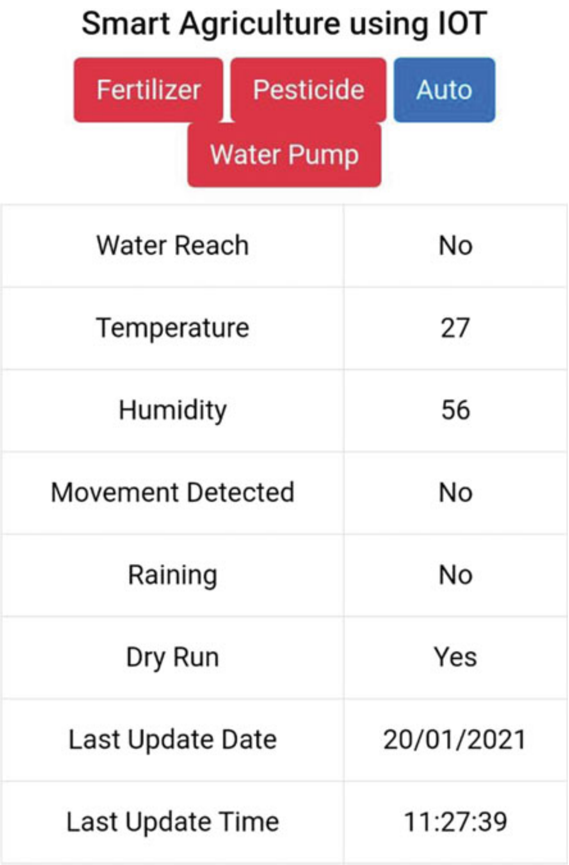

HUMIDITY : 56%
TEMPERATURE : 27c
SOIL MOISTURE
PRESENT :NO
DRY RUN DETECTED :YES
RAINING :NO
MOVEMENT DETECTED :NO

Fig. 10 Output via mobile SMS (sensor values)



Smart Agriculture using IOT

Water Reach	No
Temperature	27
Humidity	56
Movement Detected	No
Raining	No
Dry Run	Yes
Last Update Date	20/01/2021
Last Update Time	11:27:39

Fig. 11 Output displayed on website

7 Conclusion

By incorporating IoT technology using cloud, the proposed smart farming system has efficiently measured the agricultural parameters including temperature, humidity, soil humidity, dry run, motion detection and rainwater detection. The proper use of water, pesticides and fertilizers is also well-managed by using the proposed system. Thus, this system provides a greater precision and effectiveness in retrieving live parameters. This will help farmers in increasing their agricultural yield and ensures an effective agricultural food supply.

The proposed system works properly in normal scenario but it should get good and continuous Internet connectivity for sending and receiving the recorded values of the sensors. The future scope is to include a greater number of sensors in order to obtain a huge amount of data from which the farm worker can get help to take proper and quick decisions in order to increase the farm production.

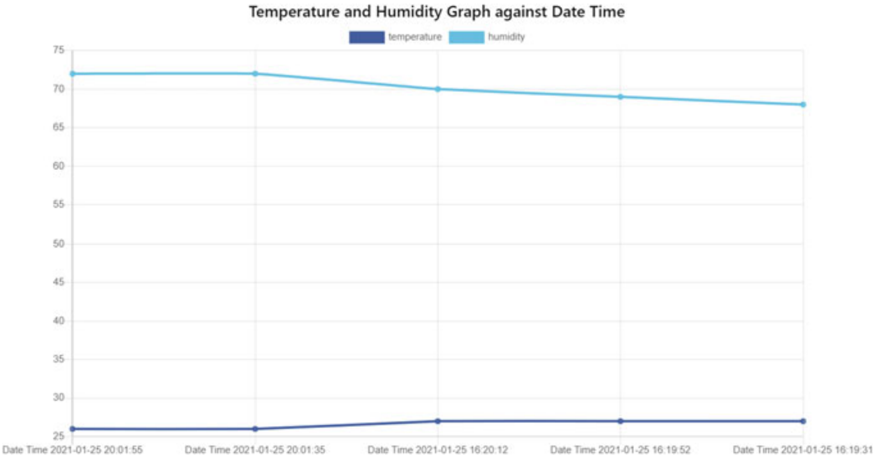
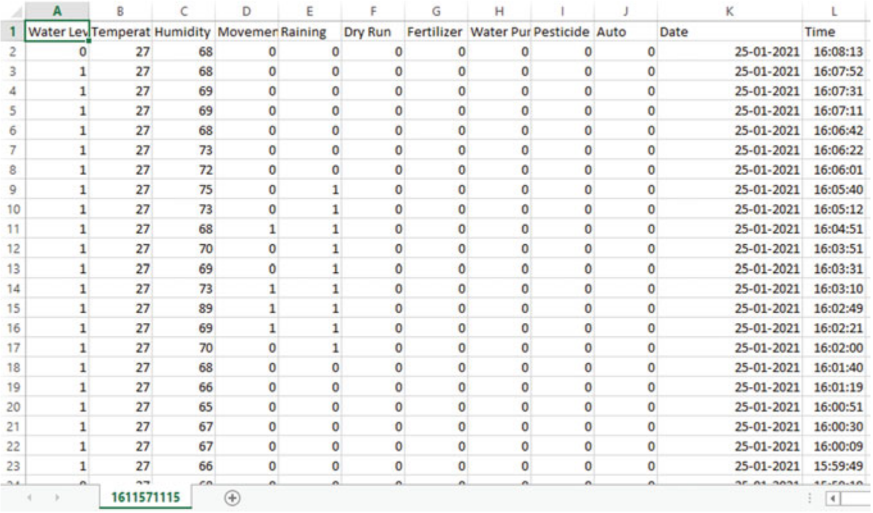


Fig. 12 Monitored data from temperature and humidity sensor



	A	B	C	D	E	F	G	H	I	J	K	L
1	Water Lev	Temperat	Humidity	Movemen	Raining	Dry Run	Fertilizer	Water Pur	Pesticide	Auto	Date	Time
2	0	27	68	0	0	0	0	0	0	0	25-01-2021	16:08:13
3	1	27	68	0	0	0	0	0	0	0	25-01-2021	16:07:52
4	1	27	69	0	0	0	0	0	0	0	25-01-2021	16:07:31
5	1	27	69	0	0	0	0	0	0	0	25-01-2021	16:07:11
6	1	27	68	0	0	0	0	0	0	0	25-01-2021	16:06:42
7	1	27	73	0	0	0	0	0	0	0	25-01-2021	16:06:22
8	1	27	72	0	0	0	0	0	0	0	25-01-2021	16:06:01
9	1	27	75	0	1	0	0	0	0	0	25-01-2021	16:05:40
10	1	27	73	0	1	0	0	0	0	0	25-01-2021	16:05:12
11	1	27	68	1	1	0	0	0	0	0	25-01-2021	16:04:51
12	1	27	70	0	1	0	0	0	0	0	25-01-2021	16:03:51
13	1	27	69	0	1	0	0	0	0	0	25-01-2021	16:03:31
14	1	27	73	1	1	0	0	0	0	0	25-01-2021	16:03:10
15	1	27	89	1	1	0	0	0	0	0	25-01-2021	16:02:49
16	1	27	69	1	1	0	0	0	0	0	25-01-2021	16:02:21
17	1	27	70	0	1	0	0	0	0	0	25-01-2021	16:02:00
18	1	27	68	0	0	0	0	0	0	0	25-01-2021	16:01:40
19	1	27	66	0	0	0	0	0	0	0	25-01-2021	16:01:19
20	1	27	65	0	0	0	0	0	0	0	25-01-2021	16:00:51
21	1	27	67	0	0	0	0	0	0	0	25-01-2021	16:00:30
22	1	27	67	0	0	0	0	0	0	0	25-01-2021	16:00:09
23	1	27	66	0	0	0	0	0	0	0	25-01-2021	15:59:49

Fig. 13 Various sensors' reading monitored from the field

References

1. Agrawal, S., Das, M.L.: Internet of Things—a paradigm shift of future internet applications. In: International Conference on Current Trends in Technology, IEEE, pp. 1–7 (2011)
2. Mekala, G.M.S., Viswanathan, P.: A novel technology for smart agriculture based on IoT with cloud computing. In: International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) SMAC 2017J. Maxwell, C. (ed.) A Treatise on Electricity and Magnetism, 3rd edn., vol. 2. Oxford, Clarendon, pp. 68–73 (1892)

3. Ravindranath, K., Bhargavi, C.S., Reddy, K.S., Chandana, M.S.: Cloud of things for smart agriculture Int. J. Innovative Technol. Explor. Eng. (IJITEE) **8**(6S), ISSN 2278–3075 (2019)
4. Prathibha, S., Hongal, A., Jyothi, M.: IOT based monitoring system in smart agriculture. In: 2017 International Conference on Recent Advances in Electronics and Communication Technology (2017)
5. Bauer, J., Aschenbruck, N.: Design and implementation of an agricultural monitoring system for smart farming. In: 2018 IoT Vertical and Topical Summit on Agriculture—Tuscany (2018)
6. Namasudra, S., Roy, P., Balusamy, B.: Cloud computing: fundamentals and research issues. In: 2017 Second International Conference on Recent Trends and Challenges in Computational (2017)
7. Nativi, S., Mazzetti, P., Santoro, M., Papeschi, F., Craglia, M., Ochiai, O.: Big data challenges in building the global earth observation system of systems. Environ. Model. Softw. **68**, 1–26 (2015)
8. Rupanagudi, S.R., Ranjani, B.S., Nagaraj, P., Bhat, V.G., Thippeswamy, G.: A novel cloud computing based smart farming system for early detection of borer insects in tomatoes. In: Proceedings of the 2015 International Conference on Communication, Information and Computing Technology (ICCICT), Mumbai, India, pp. 1–6 15–17 Jun 2015
9. Ferrández-Pastor, F.J., García-Chamizo, J.M., Nieto-Hidalgo, M., Mora-Pascual, J., Mora-Martínez, J.: Developing ubiquitous sensor network platform using internet of things: application in precision agriculture. Sensors **16**, 1141 (2016)
10. Krishna, K.L., Silver, O., Malende, W.F., Anuradha, K.: Internet of Things application for implementation of smart agriculture system. In: International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC) (2017)
11. Bacco, M., Berton, A., Ferro, E., Gennaro, C., Gotta, A., Matteoli, S., Ruggeri, F.P.M., Virone, G., Zanella, A.: Smart farming: opportunities, challenges and technology enablers. In: 2018 IoT Vertical and Topical Summit on Agriculture—Tuscany (IOT Tuscany) (2018)
12. Nagaraja, G.S., Soppimath, A.B., Soumya, T., Abhinith, A.: IoT based smart agriculture management system. In: 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS) (2019). 978-1-7281-2619-7/19/$31.00 © IEEE
13. Doshi, J., Patel, T., Bharti, kumar, S.: Smart farming using IoT, a solution for optimally monitoring farming conditions. Procedia Comput. Sci. **160**, 746–751 (2019). https://doi.org/10.1016/j.procs.2019.11.016
14. Srivastava, R., Sharma, V., Jaiswal, V., Raj, S.: A research paper on smart agriculture using IoT. Int. Res. J. Eng. Technol. (IRJET) **7**(7) (2020). e-ISSN: 2395-0056. www.irjet.net p-ISSN: 2395-0072
15. Xu, L.D., He, W., Li, S.: Internet of Things in industries: a survey. IEEE Trans. Ind. Inf. **10**(4) (2014)
16. Liqiang Z. et al, Shouyi, Y., Leibo, L., Zhen, Z., Shaojun, W.: A crop monitoring system based on wireless sensor network. Procedia Environ. Sci. **11**, 558–565 (2011)
17. Keerthi, V., Kodandaramaiah, G.N.: Cloud IoT based greenhouse monitoring system. J. Eng. Res. Appl. **5**(10), 35–41 (Part-3) (2015). ISSN: 2248-9622
18. Gutiérrez, J., Villa-Medina, J.F., Nieto-Garibay, A., Porta-Gándara, M.A.: Automated irrigation system using a wireless sensor network and GPRS module. IEEE Trans. Instrum. Meas. **63**(1) (2014)
19. Gondchawar, N., Kawitkar, R.S.: IoT based smart agriculture. Int. J. Adv. Res. Comput. Commun. Eng. **5**(6) (2016)
20. Kaur, B., Inamdar, D., Raut, V., Patil, A., Patil, N.: A survey on smart drip irrigation system. Int. Res. J. Eng. Technol. (IRJET) **3**(2) (2016)
21. Krishna, K. L., Silver, O., Malende, W. F., Anuradha, K.: Internet of Things application for implementation of smart agriculture system. In: 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (2017)
22. Shenoy, J., Pingle, Y.: IoT in Agriculture. In: International Conference on Computing for Sustainable Global Development (INDIA Com) (2016)

23. Khot, S., Gaikwad, M.: Development of cloud-based light intensity monitoring system for greenhouse using Raspberry Pi. In: IEEE International Conference on Computing Communication Control and Automation (ICCUBEA) (2016)
24. Sheetal, V., Bakshi, A., Tanvi, T.: Green house by using IoT and cloud computing. In: IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology (2016)
25. Brewster, C., Roussaki, I., Kalatzis, N., Doolin, K., Ellis, K.: IoT in agriculture: designing a Europe-Wide Large Scale pilot. IEEE Commun. Mag. (2017)
26. Ayaz, M., Ammad-uddin, M., Sharif, Z., Mansour, A., Aggoune, H.M.: Internet-of-things (IoT) based smart agriculture: toward making the field talk, IEEE Access vol. 7 (2019). https://doi.org/10.1109/ACCESS.2019.2932609
27. Johnson, N., Kumar M.B.S., Dhannia T.: A study on the significance of smart IoT sensors and data science in digital agriculture. In: 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA) (2020)
28. Ragavi, B., Pavithra, L., Sandhiyadevi, P., Mohanapriya, G. K., Harikirubha, S.: Smart agriculture with AI sensor by using Agrobot. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) (2020). https://doi.org/10.1109/iccmc48092.2020.iccmc-00078
29. Vineela, T., NagaHarini, J., Kiranma, C., Harshitha, G.,AdiLaksh, B.: IoT based agriculture monitoring and smart irrigation system using Raspberry Pi. Int. Res. J. Eng. Technol. (IRJET) **5**(1) (2018)
30. Nayyar, A., Puri, V.: IoT Based Smart Sensors Agriculture Stick for Live Temperature and Moisture Monitoring. Duy Tan University Publication, 12 Nov 2018

Multipartite Verifiable Secret Sharing Based on CRT

Rolla Subrahmanyam, N. Rukma Rekha, and Y. V. Subba Rao

Abstract In (t, n) threshold secret sharing scheme, the dealer distributes secret among a group of n participants, and any t threshold number of participants can reconstruct the secret. However, $t - 1$ or lesser number of participants can not retrieve the secret. In verifiable secret sharing schemes (VSSS), participants can verify their share after receiving shares from the dealer to ensure that a dealer is not malicious. In multipartite secret sharing based on CRT scheme, a set of participants is divided into disjoint partitions, and whatever action performed at a single level is repeated at all other partitions. However, till date there is no mechanism to verify if dealer is malicious or not in multipartite secret sharing based on CRT. Two schemes are proposed for verification of a dealer, namely multipartite verifiable secret sharing based on CRT by using Iften, and multipartite verifiable secret sharing-based CRT by using kameer Kaya. Both proposed schemes are perfectly secure, and the security of both the schemes dependent on discrete logarithm problem.

Keywords Secret sharing · Multipartite secret sharing · Verifiable secret sharing · CRT (Chinese remainder theorem)

1 Introduction

A secret sharing problem is a very popular problem in the field of cryptography with the following objectivity: Any secret S is to be shared among n participants in such a way that, the secret S can be constructed from collaboration among any t or more

Supported by organization x.

R. Subrahmanyam (✉) · N. Rukma Rekha · Y. V. Subba Rao
University of Hyderabad, Hyderabad, India
e-mail: rukmarekha@uohyd.ac.in

Y. V. Subba Rao
e-mail: yvsrcs@uohyd.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_17

number of those participants/shares. The effort to reconstruct the secret must fail, when the number of collaborating shares is $t - 1$ or less.

Every secret sharing scheme features two distinct phases: secret sharing and secret reconstruction. The share distribution phase dictates the method to be followed to share the secret among n participants. The reconstruction phase defines the method in which the secret can be retrieved from t or more number of participants. Both two phases should be in compliance of the following: Accumulation of $t - 1$ or lesser number of shares should not guarantee, in any manner whatsoever, the revelation of the secret. If the length of both shares and secret is equal, then it is called as ideal secret sharing scheme [1].

Shamir [1] and Blakely et al. [2] indepently proposed (t, n) threshold secret sharing schemes using interpolation and hyperplane geometry, respectively.

Asmuth Bloom proposed a modular way of threshold secret sharing scheme which reduced the number of key recovery operation from $O(t \log^2 t)$ to $O(t)$. One common feature among all the above secret sharing schemes is that they are based on the assumption that dealer is a trusted person.

However, if the dealer is malicious, he may send the wrong shares to participants. With these wrong shares, the participants will never be able to reconstruct the secret.

To overcome this problem, verifiability is introduced in secret sharing. Here every participant can verify their respective share for consistency after receiving a share from a dealer.

Benaloh et al. [3], Feldman [4], Pederson [5], Stinson et al. [6], Patra et al. [7], and Changlu et al. [8] developed verifiable secret sharing schemes based on polynomial interpolation. Many verifiable secret sharing schemes based on polynomial are available in literature survey. But in literature, only few verifiable secret sharing schemes having CRT as basis are available

Iften [9], Qiong et al. [10], and Kaya et al. [11] developed verifiable secret sharing schemes having Chinese remainder theorem (CRT) as basis as extension of Asmuth Bloom secret sharing scheme (SSS). In both Iften and Qiong verifiable secret sharing schemes, participants did not check the range of secret due to this problem, and participants cannot reconstruct the secret. In Kaya et al. [11] verifiable secret sharing scheme, participants can check the range of secret and then reconstruct the secret.

All the access structures such as Asmuth Bloom secret sharing are unipartite access structures. A multipartite access structure was introduced by Shamir [1, 12]. Based on this approach, multipatite secret sharing scheme having CRT as basis was designed by Hsu et al. [12].

In multipartite secret sharing scheme, a set of participants are divided into disjoint partitions, and whatever action performed in a single partition is repeated at all other partitions. In threshold secret sharing, each and every participant is equivalent; but in multipartite secret sharing, a set of participants are divided into disjoint partitions, and each and every participant is equivalent. Multipartite access structure has notable attention in secret sharing scheme. It is natural generalization of threshold secret sharing. Instead of having one threshold condition on number of participants in a given subset, a smaller set of conditions are imposed on the number of participants in the subset from each of the partitions [13].

In multipartite secret sharing having CRT as basis, participants cannot verify their respective shares after receiving shares from the dealer. To overcome this problem, multipartite verifiable secret sharing scheme having CRT as basis is proposed in this paper. Where the participants can verify their respective shares and other shares as well. Two schemes are proposed, the first scheme uses Iftene verifiable secret sharing and second scheme uses Kamer Kaya, et al. verifiable secret sharing scheme.

The organization of paper is as follows. In Sect. 2, related work is briefly explained. In Sect. 3, the proposed methods are discussed along with security analysis. Lastly, Sect. 4, conclusion, and future work are discussed.

2 Related Work

2.1 Asmuth–Bloom SSS [14]

Asmuth–Bloom proposed secret sharing based on CRT in 1983. In this (t, n), threshold secret sharing scheme, at least t number of the participants can reconstruct the secret out of n total number of participants. However, $t - 1$ participants cannot reconstruct the secret. Shamir secret sharing scheme takes $O(t \log^2 t)$ to construct secret, but Asmuth Bloom secret sharing scheme takes only $O(t)$ to build the secret. This scheme is of two phases, share distribution and secret reconstruction.

In Asmuth Bloom secret sharing scheme (SSS), participants have no mechanism to verify their respective share from dealer for consistency. To overcome this aspect, verifiable secret sharing schemes are proposed by various authors like Qiong et al. [2] in 2005, Iftenes [9] in 2007, and many more.

Iftene [9] proposed verifiable secret sharing scheme (VSSS) in 2007 as an extension for Asmuth–Bloom SSS each and every participant can verify their respective share and other shares after receiving shares from the dealer. The security of the scheme depends on discrete logarithm problem, and time complexity of secret reconstruction will be of order of O(t) where t is threshold.

2.2 Iftene CRT-based VSSS [9]

To overcome the verifiability aspect in Asmuth Bloom, Iften proposed VSSS in 2007. In this scheme, each and every participant can verify their respective share and other shares after receiving shares from dealer. This scheme is also made of two phases, share distribution, and secret reconstruction.

Share Distribution

Dealer D computes the following steps

1. A set of integers $\{l_0, l_1, l_2, ..., l_n\}$, such that $l_i \leq l_j$ for $i < j$ is chosen with the following conditions: $\gcd(l_i, l_j) = 1$ where $i \neq j$, and

$\gcd(l_0, l_i) = 1$ for every i

2. Let $M = \prod_{i=1}^{t} l_i$
3. Choose Secret $S \in Z_{l_0}$ among n participants
4. Computes $\mathrm{sh} = S + Al_0 < M$
5. Computes share $\mathrm{sh}_i = \mathrm{sh} \mod l_i$ for $i = 1, 2, ..., n$
6. Choose l_i' s such that each $p_i = 2l_i + 1$ is also a prime.
7. Let $g_i \in Z^*_{p_i}$ of order l_i. The dealer distributes share sh_i to the ith participants secretly and computes $c_i = g_i^{sh} \bmod p_i$ for $1 \leq i \leq n$. Here c_i, p_i and g_i are public.
8. The ith participant checks whether the share is valid or not by

$$c_i = g_i^{\mathrm{sh}_i} \bmod p_i$$

Secret Reconstruction
Suppose that a coalition C of participants want to reconstruct the secret.

1. Other participants in C can verify the ith participant share with

$$c_i = g_i^{\mathrm{sh}_i} \bmod p_i$$

2. The coalition C can reconstruct the secret S if all shares are correct.

Demerit
However, dealer may be dishonest in a scenario where dealer chooses $\mathrm{sh} > M$ then coalition C cannot get correct sh and secret S value.

2.3 *Kamer Kaya CRT-based VSSS [11]*

In both Iften [9] and Qiong [10] verifiable secret sharing scheme, dealer may be dishonest and may send wrong shares to participants. With those shares, authorized participants will never be able to get the actual secret. But in Kaya et al. CRT-based VSSS, dealer sends shares secretly to participants and participants verify their respective share and check the range of $sh < M$ with their share by using Boudot range proof technique [15]. The security of the scheme depends on discrete logarithm problem. This scheme made of two phases, share distribution and secret reconstruction.

2.4 *Multipartite Secret Sharing Scheme [13]*

Multipatrite secret sharing can be perceived as the natural process of the mathematical evolution of the threshold secret sharing scheme. This is because a simple generalization when applied over the access structure of threshold secret sharing leads us to the multipartite secret sharing. If we simply apply the condition of each partition being singular, a r-patrite secret sharing becomes a threshold secret sharing with threshold equal to r. In addition to this logic of organic evolution of mathematical entities, there are strong reasons behind the genesis of this multipartite secret sharing, even purely from practical perspectives.

Essentially, threshold secret sharing scheme treats each participant as same, in the sense that accumulation of any t or more number of participants lead to the reconstruction of the secret, t being the threshold. In this way, the scheme is not selective; not selective about which participants are participating, but definitely selective so far as the number of the participants is concerned. But multipartite secret sharing scheme takes both factors, what number of participants & which participants, into account, as demonstrated by the paper [12]. This situation is a natural occurrence in environments involving networks. Certain users coming from a particular portion of the network might behove identical treatment, where as another set of users coming from another portion might naturally be demanding of treatments in distinction to the earlier group. This situation is best represented, mathematically, by incorporating a multipartite access structure. In this structure, all elements within the same partition are treated as one, but transgression into other partitions does not preserve this equivalency. To reconstruct the secret, we need representatives from each of the partitions.

2.5 Multipartite Access Structure [12]

Let $\mathcal{P}(P)$ stand for power set of P. Further let $\Omega = \{P_1, ..., P_r\}$ be a partition of the set P, this means $\cup_{i=1}^{r}$ and $P_i \cap P_j = \phi$, for any $1 \leq i < j \leq r$. Again let σ be a special kind of permutation on P. Special in the sense that σ must map each member P_i of Ω onto itself nationally speaking $\sigma(P_i) = P_i, \forall P_i \in \Omega$. Let β_Ω collection of all such permutations σ. Let Λ be a collection of subsets on P so indeed $\Lambda \subseteq \mathcal{P}(P)$. Λ is the subject of our testing. As always let $\sigma(\Lambda) = \{\sigma(A) : A \in \Lambda\}$, so indeed $\sigma(\Lambda) \subset \mathcal{P}(P)$. The collection Λ is Ω-Partite if and only if the following condition holds $\sigma \in \beta_\Omega \Rightarrow \sigma(\Lambda) = \Lambda$. Λ is said to r partite for any positive integer r. If it is Ω partite for some partition Ω on P of cardinality r.

Consider the set $J_r = \{1, 2, ..., r\}$. Let Z_+^r denote the set of vectors $u = (u_1, ..., u_r) \in Z^r$ with $u_i \geq 0$ for every $i \in J_r$. For a partition $\Omega = \{P_1, ..., P_r\}$ of P and Subset $A \subseteq P$ and $i \in J_r$. Define $\Omega_i(A)$ is the number of participants in $A \cap P_i$, i.e., $|A \cap P_i|$ then define a map $\Omega : \mathcal{P}(P) \to Z_+^r$ as $\Omega(A) = (\Omega_1(A), \Omega_2(A), ..., \Omega_r(A))$.

2.6 Multipartite SSS Based on CRT [12]

For a partition $\Omega = \{P_1, ..., P_r\}$ of $P = \{p_i : 1 \leq i \leq n\}$, we suppose that an access structure τ is an Ω-partite, where $|P_1| = n_1, ..., |P_r| = n_r$ and $n_1 +, ..., +n_r = n$. Then the partition is a transformation $\Omega : \mathcal{P}(P) \longrightarrow Z^r_+$. Let the corresponding minimal access structure is τ_0, and maximal prohibited access structure is Δ_1, so that $\Omega(\tau_0) \subset Z^r_+$ and $\Omega(\Delta_1) \subset Z^r_+$ can be determined. It consists of two phases, share distribution and secret reconstruction

In share distribution, dealer selects coprime integers

$\{l_0, l_1 <, ..., < l_{n_1}, l_{n_1+1} <, ..., < l_{n_1+n_2}, ..., l_{n-n_r+1} < ..... < l_n\}$ and computes M_3 and M_4[12]. Dealer computes shares and distribute to all participants.

In secret reconstruction, coalition C of τ participants want to reconstruct the secret.

3 Proposed Methods

In multipartite secret sharing based on CRT, participants cannot verify their respective shares after receiving shares from the dealer. To overcome this problem, a novel multipartite secret sharing scheme based on CRT is proposed where the participants can verify their respective shares and other shares as well. The scheme is proposed in two variants. The first variant uses Iftene verifiable secret sharing scheme (VSSS), and second scheme uses Kamer Kaya, et al. VSSS.

These two schemes have two phases, namely share distribution and secret reconstruction. In share distribution phase, dealer computes shares and secretly sends to participants. In secret reconstruction phase, coalition C number of participants can reconstruct the secret.

Notations

1. Let $P = \{p_i : 1 \leq i \leq n\}$ be set of participants, assume Dealer D and S is secret.
2. Let τ be set of subsets of P, $\tau \subseteq 2^P$ Subsets in τ are called authorized subsets.
3. $\Delta = 2^P \backslash \tau$ is called prohibited access structure. Subsets in Δ are called unauthorized subsets.
4. Super set of authorized set is again an authorized subset if it satisfies monotone increasing property. if $B \in \tau$ and $B \subseteq C \subseteq P$, then $C \in \tau$
5. τ_0 is basis of τ
$$\tau = \{C \subseteq P : B \subseteq C, B \in \tau_0\}$$
6. Δ_1 is maximal unauthorized subset
$$\Delta = \{C \subseteq P : C \subseteq B, B \in \Delta_1\}$$

3.1 Proposed Multipartite VSS Based on CRT by Using Iftene

For a partition $\Omega = \{P_1, ..., P_r\}$ of $P = \{p_i : 1 \leq i \leq n\}$, we suppose that an access structure τ is an Ω-partite, where $|P_1| = n_1, ..., |P_r| = n_r$ and $n_1+, ..., +n_r = n$. Then the partition is a transformation $\Omega : \mathcal{P}(P) \longrightarrow Z_+^r$. Let the corresponding minimal access structure is τ_0 and maximal prohibited access structure is Δ_1, so that $\Omega(\tau_0) \subset Z_+^r$ and $\Omega(\Delta_1) \subset Z_+^r$ can be determined. This consists of five phases, share distribution, commitment, verification, and secret reconstruction.

Share Distribution
Dealer D distributes n shares between members of a set $P = \{p_i : 1 \leq i \leq n\}$, known as the participant set P, each p_i being an individual participant.
The dealer does the following:

1. A set of integers
   $\{l_0, l_1 <, ..., < l_{n_1}, l_{n_1+1} <, ..., < l_{n_1+n_2}, ..., l_{n-n_r+1} <, ..., < l_n\}$,
   where $0 \leq S \leq l_0$ is chosen in the following ways
   $\gcd(l_i, l_j) = 1$ where $i \neq j$
   $$M_3 = \min\left(\prod_{j=1}^{r}\prod_{i=1}^{u_j} l_{s_{j-1}+i}, \text{for all}(u_1, u_2, ..., u_r) \in \Omega(\tau_0)\right)$$
   where $s_i = \sum_{j=1}^{i} n_j$ and
   $0 = s_0 < s_1 < s_2 < s_3 < s_4, ..., < s_r = n$
   $$M_4 = \max\left(\prod_{j=1}^{r}\prod_{i=1}^{v_j} l_{s_{j-1}+i-1}, \text{for all}(v_1, v_2, ..., v_r) \in \Omega(\Delta_1)\right)$$
   $$M_3 > l_0 M_4$$
2. Dealer computes $\text{sh} = S + Ap$, where $A \in Z^+$ is generated randomly with $0 \leq sh < M_3$.
3. The n_j shares, $j = 1, 2, ..., r$,
   $$sh_i = \text{sh} \bmod l_i, \quad i = 1, ..., n_j$$
   are distributes to each participants in P_j randomly
   f: $\{\text{sh}_1, ..., \text{sh}_n\} \longrightarrow P$.

   Commitment
4. It is useful for verify the participant share.
5. Let $g_i \in Z_{p_i}^*$ of order l_i. The dealer distributed share sh_i to the ith participant secretly and computes
   $$c_i = g_i^{\text{sh}} \bmod p_i \qquad (1)$$

Here c_i, p_i and g_i are public.

Verification

6. The ith participant checks whether the share is valid or not by

$$c_i \equiv g_i^{\mathrm{sh}_i} \bmod p_i$$

7. Other participants can verify the ith participant share with verification equation

$$c_i \equiv g_i^{\mathrm{sh}_i} \bmod p_i$$

Secret Reconstruction

8. Suppose that a coalition C of τ participants want to reconstruct the secret. Let $M_C = \prod_{f(\mathrm{sh}_i)\in C} l_i$ and $\mathrm{sh} \equiv \mathrm{sh}_i \bmod l_i$, for $f(\mathrm{sh}_i) \in C$. Solve sh in $GF(M_C)$ uniquely using the CRT.
9. Calculate the secret as $S = \mathrm{sh} \bmod l_0$

Proof of Correctness for Verification

Every participants can verify their respective shares as follows

$$c_i \overset{?}{\equiv} g_i^{\mathrm{sh}_i} \bmod p_i$$

$$\begin{aligned}
c_i &\equiv g_i^{\mathrm{sh}} \bmod p_i, \text{(from equation1)}\\
c_i &\equiv g_i^{\mathrm{sh}_i + l_i a} \bmod p_i, \text{since } \mathrm{sh}_i = \mathrm{sh} \bmod l_i\\
c_i &\equiv g_i^{\mathrm{sh}_i} (g_i^{l_i})^a \bmod p_i\\
c_i &\equiv g_i^{\mathrm{sh}_i} \bmod p_i, \text{ since order of} g_i \text{ is } l_i.
\end{aligned}$$

Security Analysis

Lemma 1 *Commitment $c_i = g_i^{sh}$ mod p_i does not leak any information about sh,*

Proof Let G be a cyclic group of order q and g_i generates G. Let e be the identity element of G. Given $c_i = g_i^{\mathrm{sh}} \bmod p_i$, For any $\mathrm{sh} \in Z_p$, choose a random integer $a \in Z_p$ such that

$$c_i = g_i^{qa+\mathrm{sh}} \bmod p_i$$

$$c_i = g_i^{qa} g_i^{\mathrm{sh}} \bmod p_i$$

$$c_i = e^a g_i^{\mathrm{sh}} \bmod p_i$$

$$c_i = g_i^{\mathrm{sh}} \bmod p_i$$

Hence, $c_i = g_i^{\mathrm{sh}} \bmod p_i$ is uniformally shared in G, i.e., the information about sh is secret.

Theorem 1 *The proposed multipartite iftene VSSS realizing multipartite access structures is a perfect SSS [12]*

Proof In our multipartite scheme, we get that sh can be computed uniquely in $GF(M_C)$ using CRT. Also the solution is unique in $GF(M_3)$ as $\mathrm{sh} < M_3 < M_C$. Hence, it satisfies that $H(S|C) = 0$, $\forall C \in \tau$ (authorized participants can able to reconstruct the secret). We consider that a coalition C' unauthorized participants in Δ has assembled. Let sh' denote the unique solution for $\mathrm{sh} \in GF(M'_C)$, hence $\mathrm{sh}' + jM'_C \bmod l_0$. From $M_3 > l_0 M_4$ and $M_4 > M'_C$, we get $\frac{M_3}{M'_C} > m_0$ for $0 \le j < l_0$. For $0 \le j < m_0$ all $\mathrm{sh}' + jM'_C \bmod l_0$ are different since gcd $(M'_C, l_0) = 1$, and there l_0 such values exits. That is $S \in GF(l_0)$ and coalition participants cannot get any information about the secret. Hence, it satisfies that $H(S|C) = H(S), \forall C \in \Delta$ (Any unauthorized participants can not get any information about the secret.).

3.2 Proposed Multipartite VSSS Based on CRT by Using Kamer Kaya, et al.

For a partition $\Omega = \{P_1, ..., P_r\}$ of $P = \{p_i : 1 \le i \le n\}$, we suppose that an access structure τ is an Ω-partite, where $|P_1| = n_1, ..., |P_r| = n_r$ and $n_1+, ..., +n_r = n$. Then the partition is a transformation $\Omega : \mathcal{P}(P) \longrightarrow Z^r_+$. Let the corresponding minimal access structure is τ_0, and maximal prohibited access structure is Δ_1, so that $\Omega(\tau_0) \subset Z^r_+$ and $\Omega(\Delta_1) \subset Z^r_+$ can be determined. It consists of five phases, share distribution, commitment, verification, and secret reconstruction.

Share Distribution

Dealer D distributes n shares between members of a set $P = \{p_i : 1 \le i \le n\}$, known as the participant set P, each p_i being an individual participant.

the dealer does the following:

1. A set of integers
 $\{l_0, l_1 <, ..., < l_{n_1}, l_{n_1+1} <, ..., < l_{n_1+n_2}, ..., l_{n-n_r+1} < ... < l_n\}$,
 where $0 \le S \le l_0$ is chosen in the following ways
 $\gcd(l_i, l_j) = 1$ where $i \ne j$

$$M_3 = \min\left(\prod_{j=1}^{r}\prod_{i=1}^{u_j} l_{s_{j-1}+i}, \text{ for all}(u_1, u_2, ..., u_r) \in \Omega(\tau_0)\right)$$

 where $s_i = \sum_{j=1}^{i} n_j$ and

$0 = s_0 < s_1 < s_2 < s_3 < s_4, ..., < s_r = n$

$$M_4 = \max\left(\prod_{j=1}^{r}\prod_{i=1}^{v_j} l_{s_{j-1}+i-1},\ \text{for all}(v_1, v_2, ..., v_r) \in \Omega(\Delta_1)\right)$$

$$M_3 > l_0 M_4$$

2. Let $g_i \in Z^*_{p_i'}$ of order l_i. Let $P' = \prod_{i=1}^{n} p_i'$
   Here $p_i' = 2l_i + 1$, and l_i' s both are large primes for $1 \leq i \leq n$. Where $P_i'' = (\frac{P'}{p_i'})^{-1} \bmod p_i'$ for all $1 \leq i \leq n$, i.e., $g \in Z_{P'}$ is unique satisfying

$$g \equiv g_i \bmod p_i'$$

3. Dealer computes $\text{sh} = S + Al_0$ where $A \in Z^+$ is generated randomly with $0 \leq sh < M_3$.
4. The n_j shares, $j = 1, 2, ..., r$,

$$\text{sh}_i = \text{sh} \bmod l_i, \quad i = 1, ..., n_j$$

   are distributed to each participant in P_j randomly
   $f\colon \{\text{sh}_1, ..., \text{sh}_n\} \longrightarrow P$.
   Commitment
5. Assume both dealer and participant do not know N prime factorization. Compute

$$E(\text{sh}) = g^{\text{sh}} \bmod P'N \tag{2}$$

   Verification
6. ith participant check whether the share is valid or not by $E(\text{sh}) \equiv g_i^{\text{sh}_i} \bmod p_i'$ to verify $\text{sh}_i = \text{sh} \bmod l_i$, Then participants can verify validity of the range proof by checking $\text{sh} < M_3$.
7. Other participants can verify the ith participant share with verification equation

$$g_i^{\text{sh}_i} \equiv E(\text{sh}) \bmod p_i'$$

   Secret Reconstruction
8. Suppose that a coalition C of τ participants want to reconstruct the secret. Let $M_C = \prod_{f(\text{sh}_i)\epsilon C} l_i$, and $sh \equiv \text{sh}_i \bmod l_i$ for $f(\text{sh}_i) \in C$. Solve sh in $GF(M_C)$ uniquely using the CRT.
9. Compute the secret as $S = \text{sh} \bmod l_0$

Proof of Correctness for Verification
Every participants can verify their respective shares as follows

$$g_i^{\text{sh}_i} \overset{?}{\equiv} E(\text{sh}) \bmod p_i'$$

Correctness

$$\begin{aligned} E(\mathrm{sh}) \bmod p_i' &\equiv g^{\mathrm{sh}} \bmod P'N \bmod p_i{}', \text{(from Eq. 2)}.\\ E(\mathrm{sh}) &\equiv g_i^{\mathrm{sh}} \bmod p_i{}',\ \text{since } g_i \equiv g \bmod p_i{}'\\ E(\mathrm{sh}) &\equiv g_i^{\mathrm{sh}_i+l_i a} \bmod p_i{}',\ \text{since } \mathrm{sh}_i = \mathrm{sh} \bmod l_i\\ E(\mathrm{sh}) &\equiv g_i^{\mathrm{sh}_i} (g_i^{l_i})^a \bmod p_i{}'\\ E(\mathrm{sh}) &\equiv g_i^{\mathrm{sh}_i} \bmod p_i{}',\ \text{since order of } g_i \text{ is } l_i. \end{aligned}$$

Security Analysis

Lemma 2 *Commitment $E(sh) = g^{sh}$ mod $P'N$ does not reveal any information about sh.*

Proof Let G be a cyclic group of order q and g generates G. Let e be the identity element of G. Given $E(\mathrm{sh}) = g^{\mathrm{sh}}$ mod $P'N$, For any sh $\in Z_{p_i'}$, choose a random integer $a \in Z_{p_i'}$ such that

$$E(\mathrm{sh}) = g^{qa+\mathrm{sh}}$$

$$E(\mathrm{sh}) = g^{q^a} g^{\mathrm{sh}}$$

$$E(\mathrm{sh}) = e^a g^{\mathrm{sh}}$$

$$E(\mathrm{sh}) = g^{\mathrm{sh}}$$

Hence, $E(\mathrm{sh}) = g^{\mathrm{sh}}$ mod $P'N$ is uniformally shared in G, i.e., the information about sh is secret.

Theorem 2 *The proposed multipartite Kamer Kaya, et al. VSSS realizing multipartite access structures is a perfect SSS [12]* (Table 1).

4 Conclusion and Future Work

Two schemes multipartite VSSS, namely multipartite VSSS based on CRT by using Iften, and multipartite VSSS based on CRT by using Kamer Kaya, et al., are proposed. In the first scheme, the dealer may be malicious because participants can not verify if $sh < M_3$ after verifying their respective share. But in the second scheme, the dealer not be malicious because participants can verify if $sh < M_3$ after verifying

Table 1 Comparision table

Scheme name	Unipartite	Multipartite	Verifiability	Dealer not malicious
Asmuth Bloom SSS based on CRT	Yes	No	No	No
Iften et al. VSSS based on CRT	Yes	No	Yes	No
Kameer Kaya et al. VSSS based on CRT	Yes	No	Yes	Yes
Multipartite SSS based on CRT	No	Yes	No	Yes
Multipartite VSSS based on CRT by using Iften, et al.	No	Yes	Yes	No
Multipartite VSSS based on CRT by using Kameer Kaya, et al.	No	Yes	Yes	Yes

their respective share. Both schemes are perfectly secure, and security depends on discrete logarithm problem.

In both multipartite VSSS based on CRT by using Iften and multipartite VSSS based on CRT by using Kamer Kaya, et al. schemes, dealer involvement is there. Our future work is to extend this scheme into decentralized VSSS based on CRT.

References

1. Shamir, A.: How to share a secret. Commun. ACM **22**(11), 612–613 (1979)
2. Blakley, G.R.: Safeguarding cryptographic keys. In: International Workshop on Managing Requirements Knowledge. IEEE Computer Society (1979)
3. Benaloh, J.C.: Secret sharing homomorphisms: Keeping shares of a secret secret. In: Conference on the Theory and Application of Cryptographic Techniques. Springer, Berlin, Heidelberg (1986)
4. Feldman, P.: A practical scheme for non-interactive verifiable secret sharing. In: 28th Annual Symposium on Foundations of Computer Science (SFCS), pp. 427–438. IEEE (1987)
5. Pedersen, T.P.: Non-interactive and information-theoretic secure verifiable secret sharing. In: Annual International Cryptology Conference. Springer, Berlin, Heidelberg (1991)
6. Stinson, D.R., Wei, R.: Unconditionally secure proactive secret sharing scheme with combinatorial structures. In: International Workshop on Selected Areas in Cryptography, Springer, Berlin, Heidelberg (1999)

7. Patra, A., Choudhary, A., Pandu Rangan, C.: Efficient statistical asynchronous verifiable secret sharing with optimal resilience. In: International Conference on Information Theoretic Security. Springer, Berlin, Heidelberg (2009)
8. Lin, C., Harn, L.: Unconditionally secure verifiable secret sharing scheme. AISS Adv. Inf. Sci. Service Sci. **4**(17), 514–518 (2012)
9. Iftene, S.: Secret sharing schemes with applications in security protocols. Sci. Ann. Cuza Univ. **16**, 63–96 (2006)
10. Qiong, L., et al.: A non-interactive modular verifiable secret sharing scheme. In: Proceedings. International Conference on Communications, Circuits and Systems, vol. 1. IEEE (2005)
11. Kaya, K., Selçuk, A.A.: A verifiable secret sharing scheme based on the Chinese remainder theorem. International Conference on Cryptology in India. Springer, Berlin, Heidelberg (2008)
12. Hsu, C.-F., Harn, L.: Multipartite secret sharing based on CRT. Wireless Pers. Commun. **78**(1), 271–282 (2014)
13. Farrs, O., Jaume, M., Padr, C.: Ideal multipartite secret sharing schemes. J. Cryptol. **25**(3), 434–463 (2012)
14. Asmuth, C., Bloom, J.: A modular approach to key safeguarding. IEEE Trans. Inf. Theory **29**(2), 208–210 (1983)
15. Boudot, F.: Efficient proofs that a committed number lies in an interval. In: International Conference on the Theory and Applications of Cryptographic Techniques. Springer, Berlin, Heidelberg (2000)

Implementation of Smart Parking Application Using IoT and Machine Learning Algorithms

G. Manjula, G. Govinda Rajulu, R. Anand, and J. T. Thirukrishna

Abstract By considering the ever-increasing traffic in metropolitan areas, vehicle parking has become a great hindrance, especially while finding the available parking space nearby any office space or shopping mall, which is located on the narrow roadways. As the attempt to manually search for a parking slot consumes more time, commercial parking slots are designed to balance the demand and availability of vehicle parking spaces. Since constructing and monitoring a private parking space requires more money and workforce, parking charge has become very expensive. Due to the non-affordability of drivers, they waste more time in looking for empty parking slots. To overcome these challenges, the proposed research work helps to automatically identify the empty parking spaces, so that the car can be parked even in the most comfortable spot via video image processing and neural networks techniques, which develops a parking management software that actually identifies the existence of parking areas. The data from video footage is used to train the Mask R-CNN architecture, where a computer vision image recognition model is used to automatically identify the parking spaces. To label the car parking place mostly on the source images of a whole parking lot, a pre-processed region-based convolutional neural network (Mask R-CNN) is used. All of this could be solved by impelmenting a smart application, which could also send a text information to the customer, whenever a parking slot becomes available. Only at end of the day, it is required to have an appropriate and possible approach for solving all parking issues in and around the neighbourhood.

G. Manjula · J. T. Thirukrishna
Department of Information and Science, Dayananda Sagar Academy of Technology and Management (DSATM), Bengaluru, India
e-mail: manjula-ise@dsatm.edu.in

G. Govinda Rajulu
Department of Computer Science and Engineering, St. Martin's Engineering College, Secunderabad, India
e-mail: drgovindacse@smec.ac.in

R. Anand (✉)
Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, India
e-mail: anand.r@cmrit.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_18

Keywords Object detection · Smart parking · Mask R-CNN · Basic CNN · COCO dataset · Image mask

1 Introduction

Finding a parking spot nowadays increases our anxiety level. Most of the time we get stressed out looking for a free space to park in. After putting in a lot of effort and precious time we get a spot. Thus, this trouble is not worth it. As we are heading towards a mission of Smart cities that make our parking systems smart too. It is claimed that the world population in urban areas is expanding and the forecast is that 68% of the global population will live in cities by 2050. Streamlined parking is the only way to solve our future problems.

There are many parking spots, which are not available most of the time like shopping malls, tourist places, educational institutions, famous religious places, etc. Even if the spots are vacant we are not aware of them and hence either cancel our visit or roam here and there in search of free parking spots.

Thus, to ease up the situation and provide a helping hand, we can use a novel system where one can be notified about a free parking space through a simple text message. We introduced a deep learning-based parking space detection method. The Mask R-CNN algorithm is being used to segments an object in our vacancy detection phase [1]. This would detect the cars in the parking space and determine if the parking area is available or not. The Mask R-CNN algorithm, which is qualified with the COCO dataset [2] has more than 12,000 images of cars along with other objects would help us to determine the vehicles already parked in a spot.

2 Literature Survey

In [3], video capturing is offered by the usage of digital digicam networks. The complete gadget includes a community of floor cameras to seize license plates, a community of pinnacle view cameras to cowl the complete vicinity of the automobile parking space, gadgets like Nvidia Jetson and a Raspberry Pi. The gadget additionally protected a cloud server and a database to save the statistics approximately the occupancy of the parking slots. All this became related the usage of an internet web page which might show all of the important and applicable statistics.

In this paper [4], the proposed for detecting electronic gadgets through clever cameras. A Raspberry Pi module ready with a widespread Raspberry Pi digital digicam. The deep convolution neural community, referred to as AlexNet, was being used. The systems consist of 60 million parameters and 500,000 neurons. Five convolutional layers, observed via means of max pooling layers, and absolutely related layers with a 1000-manner softwax.

The exceptional usage of CNNs' algorithm for identifying the different objects using pre-trained model and datasets [5]. The skilled community was then used to determine approximately the occupancy reputation as acquired via way of means of the video digital digicam. The image of the automobile parking space became captured periodically and used to educate the neural community. The image became filtered via means of masks for parking spaces. The masks became constructed as soon as manually.

In the paper [6], proposed a way the usage of SSD (Single Shot Multibox Detector) for detecting items the usage of an unmarried deep neural community. It is capable of stumbling on items in snapshots without extracting the place proposal, warding off the downside of the historical past segmentation. There are hard and fast default bins with more than one element ratios at every location.

The intelligent traffic monitoring system [7] has vehicle parking integration module for finding the illegally parked vehicles. But, the module has implemented in different road environments and different weather conditions. Hence, this work addresses vehicle detection based on the differences in background and foreground identification models with temporal analysis, vehicle detector, and tracking.

Every vehicle that enters or exits in the parking area has been counted by using counter-based systems [8], which includes automated gate-arm counters and inductive loop detectors at the entry and exit points in this paper. This type of equipment could provide information about the entire number of available parking areas in a closed parking lot, but that's not very useful in guiding the driver to the actual location of the available space. Because of its low cost, it is frequently used in large outdoor or indoor parking systems.

3 Methodology

The proposed method utilizes a system, which does not have much requirement in the hardware section. Our method should work effortlessly with software by not relying much on the hardware components. The secondary aim of this research work is to inform the user about a free parking space. In the first stage, the video footage from the camera is fed into the neural network. The network then detects the valid parking spaces by analyzing various conditions. Upon detecting a valid parking, the algorithm detects whether the parking space is free or not. If it is free, the user will be notified by a message on their phone, and if not the algorithm will again check for a free space.

Figure 1 shows the process of automated parking space detection system by integrating the Internet of Things (IoT) models [9] and machine learning algorithms. Initially, the Raspberry Pi camera captures a video of the parking slot in which the captured video is converted into a sequence of frames. The client streaming program runs in Raspberry Pi and it will be continuously streaming a video to the computing server. The computing server captures sequence of frames and stores it as a video in the server. Even though the server captures an image from streaming video and finds

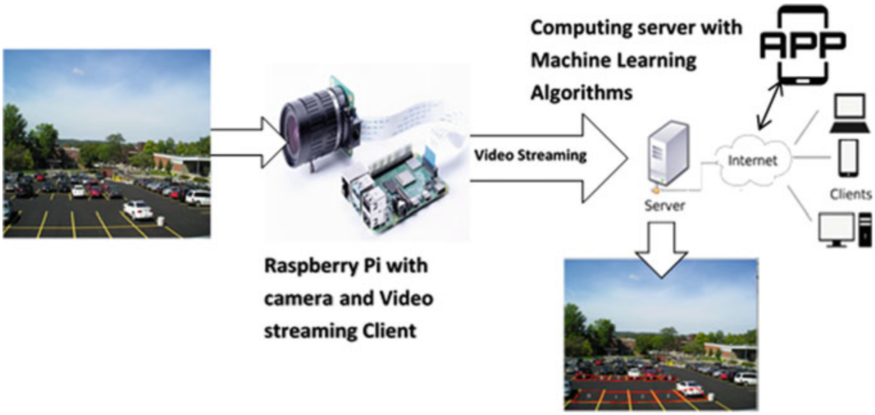


Fig. 1 Parking slot detection architecture

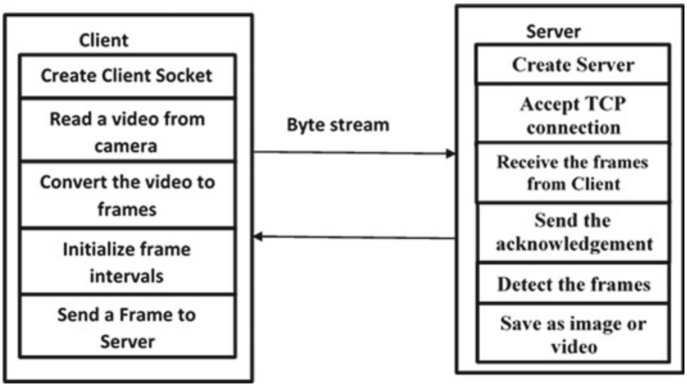


Fig. 2 Python socket IO video streaming

the list of vehicles parked in the respective slots as well as list of available slots in the parking area. The parking slot identification and detection are carried out through a machine learning algorithm. Even the ML algorithms support for improving the accuracy and avoid the false positive information. The application is built in the server, which can be accessed by registered visitors to know the status of the parking slot through a mobile or web application. It frequently updates the status of parking slots either by sending SMS or email notification.

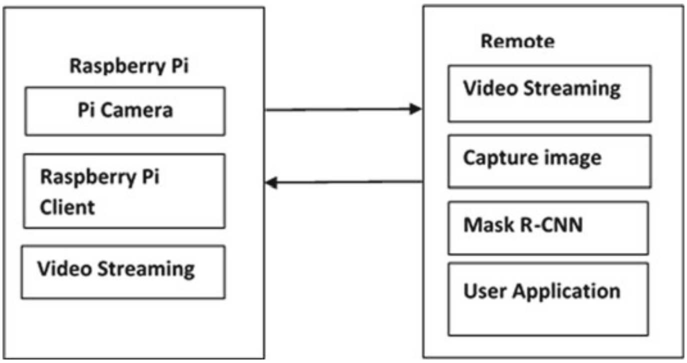


Fig. 3 Application functional blocks

3.1 Socket IO Video Streaming

Figure 2 shows actually the client–server relationship has to be established to process the communication between client and server. The socket IO protocol provides an API that utilizes client and server communication with network functionality [10]. The client is responsible for capturing a video from Raspberry pi camera and converting it into the sequence image frames with specific intervals for sending image frames to the server. The server is ready and running to accept all sequence of bytes for recognizing image frames. The recognized image frames may be converted to image or video and stored in the server. The python socket IO libraries [11] are used to initiate the client–server communication.

3.2 Application of Functional Blocks

Figure 3 shows the set of functional components are integrated into user application for implementing user application. The user application will update the list of parking slots available in parking area. But in backend, the captured image is processed by Mask R-CNN algorithm for identifying and detecting list cars and available parking slots in the parking location [11, 12]. Surely, this application helps users for searching the available parking in the nearby location.

3.3 Dataset

Dataset used for the training purpose is obtained from a popular dataset called COCO (Common Objects in Context). The COCO dataset contains images annotated with

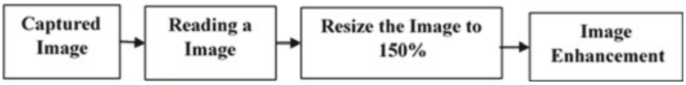


Fig. 4 Pre-processing

object masks [13]. The dataset contains more than 12,000 images of cars, which are already outlined.

3.4 Pre-processing

In the pre-processing stage, the video feed from the camera is fed into the neural network architecture [14]. Before sending the image to respective neural network, the image has to be enhanced through the proper pre-processing functionalities. The pre-processing functionalities such as capturing the image from video and storing the image in local path. The stored image should be processed by using pre-processing techniques, which improves the image enhancement and accuracy. The pre-processing steps are clearly mentioned in Fig. 4.

Reading an image is nothing but analyzing the captured image from stored location and then loading the image into array variables. The captured images by a camera are processed by using the machine learning algorithm of varying size, but it should be launched with fixed size for all of the images to be processed. The image size has also increased 50% more than the original image [15]. The important factor is reduction of noise for smoothening the image [16]. The Gaussian factorization method is used for smoothening the image that takes part in computer vision techniques to enhance the image with Gaussian kernel of (5, 5).

3.5 Deep Learning

It consists of numerous layers of nonlinear nodes by combining a computer file with a group of weights in order to give importance to inputs for the corresponding project, where the set of rules is making an attempt to be instructed in supervised and/or unsupervised behaviour. The sum of the manufactured and weights is surpassed through the activation characteristics of nodes. Every output layer is served simultaneously from the input layer starting from the first layer. Learning is frequently finished in a couple of tiers of representations that correspond to several tiers of abstraction [17].

3.6 Mask R-CNN

Mask RCNN is a deep neural community aimed to resolve example segmentation issues in device mastering or laptop vision. There are different ranges of Mask RCNN. First, it generates proposals approximately the areas wherein there is probably an item primarily based on the entire photo.

Figure 5 shows the architecture of Mask R-CNN and it predicts the elegance of the item, refines the bounding container and generates a mask in pixel stage of the item primarily based totally on the primary degree proposal. Mask R-CNN will no longer require a massive quantity of records for schooling the neural community. Mask R-CNN is a great desire that mixes the accuracy of CNNs with smart layout and performance hints that significantly accelerate the detection process. This will run exceedingly fast (on a GPU). The Mask R-CNN structure is designed in one of these ways, wherein it detects the gadgets throughout the complete photo in a computationally green manner. In different words, it runs pretty quickly [18]. In addition, Mask R-CNN offers more amount of data approximately in every detected item. Most item detection algorithms simply go back to the bounding container of every item. But Mask R-CNN will now no longer offer the simplest delivery in the region of every item, however, it will additionally deliver us an item outline (or masks).

The following steps are used for parking slot detection through Mask_RCNN model

1. Setup configuration parameters for model.
2. Implement a Mask R-CNN model.
3. Load the model weights based existing data set.
4. Capture the input image through a camera.
5. Find cars in the image and classify the objects.
6. Visualize the results and automate the application.

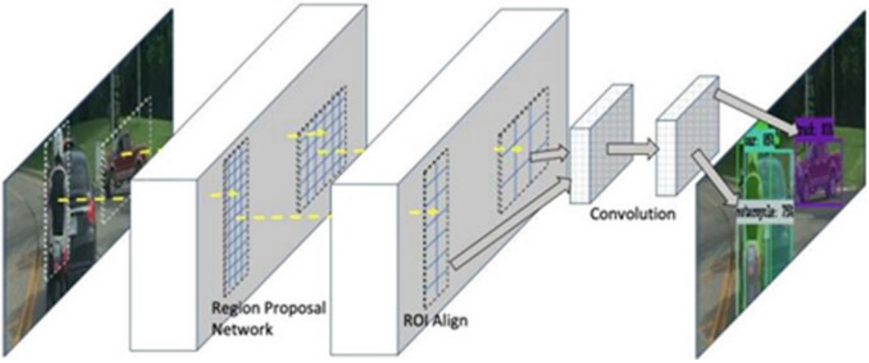
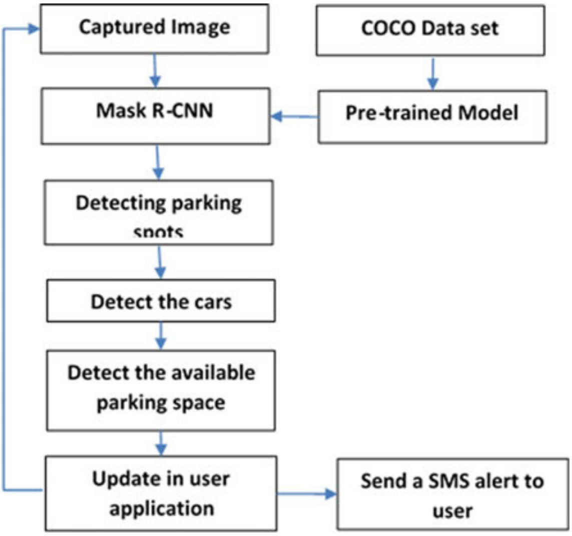


Fig. 5 Architecture of Mask R-CNN

Fig. 6 Flow diagram of application



The model is implemented with an instance segmentation with bounded regions using Mask R-CNN. The pre-trained COCO set library [19] referred for obtaining the labels of image data set.

3.7 *Application Flow Diagram*

Figure 6 shows the flow diagram for working user application, the parking slots are updated in user all application [20]. The camera captures video and sent to the raspberry pi. The raspberry pi sends a video to remote server and stores in it. Every frequent intervals, image will be captured for processing the Mask-RCNN algorithm. It detects cars and available parking slot and updates it in the user application.

4 Results and Snapshot

This framework contains the smart parking system's graphical user interface (GUI) with various functionalities. The proposed system shows the list of pages of the user interface such as live streaming of the parking area, available parking slots and sends the message to registered users with a message notification. It has user login and registration modules with user-friendly applications.

Figure 7 shows the login page for the user to login to the system by entering the respective user data. If the user's credentials satisfy the condition, it redirects to the application page.

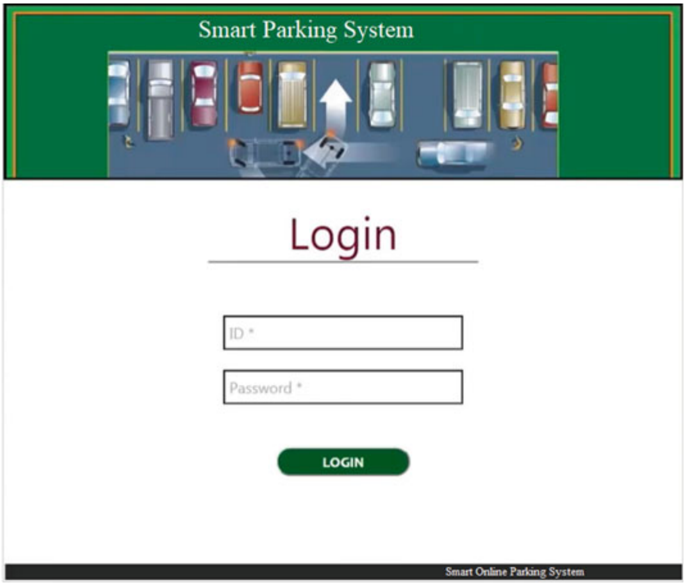
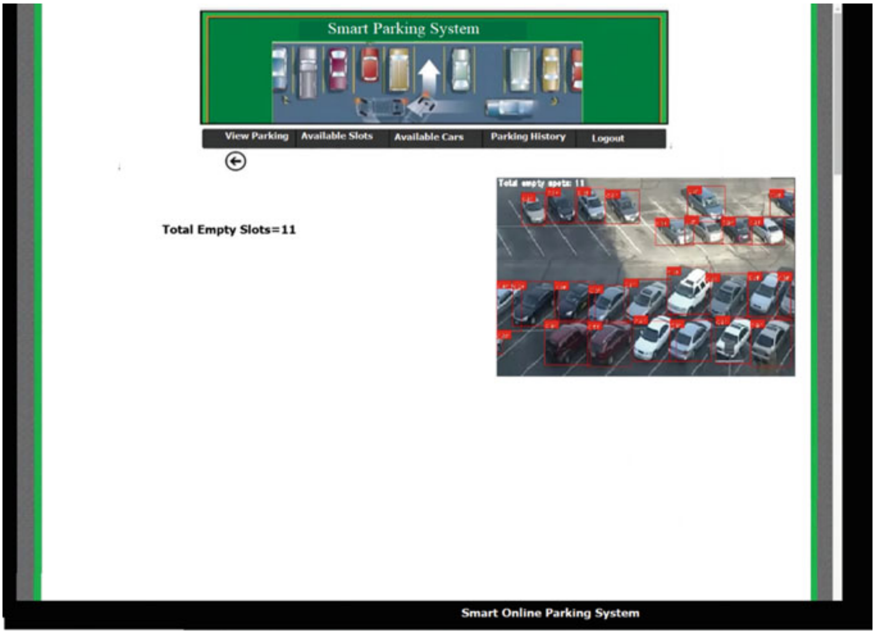


Fig. 7 Login page

The application interface with list of functionalities such as parking location, available slots, parking history, and message notifications as shown in Fig. 8. The user can use the features of the above list for their operations.

5 Conclusion

By analyzing the existing methods, it has been observed that these methods require hardware devices like Raspberry Pi and smart camera. The datasets required for training some of these methods are also manually entered. These methods do not have any implementation for alerting the user of a free space. The proposed approach includes a method, which does not require more hardware equipments. The Mask R-CNN can be used to detect the objects in a fast and efficient way even with high-resolution video feed. The basic idea of the proposed research work is to use the Internet of Things [IoT] and machine learning algorithms to build innovative parking systems, where parking slots can be viewed in a web application.



Total Empty Slots=11

Fig. 8 Application interface

6 Future Enhancements

The potential improvement of the system seems to be where users would be notified about the nearest parking lot in their geographical location. This framework may be extended to determine the type of vehicle and the automated parking system.

References

1. Wang, C., Peng, Z.: Design and implementation of an object detection system using faster R-CNN. In: 2019 International Conference on Robots & Intelligent System (ICRIS), Haikou, China, pp. 204–206 (2019). https://doi.org/10.1109/ICRIS.2019.00060
2. Puri, D.: COCO dataset stuff segmentation challenge. In: 2019 5th International Conference On Computing, Communication, Control and Automation (ICCUBEA), Pune, India, pp. 1–5 (2019). https://doi.org/10.1109/ICCUBEA47591.2019.9129255
3. Bura, H., Lin, N., Kumar, N., Malekar, S., Nagaraj, S., Liu, K.: An edge based smart parking solution using camera networks and deep learning. In: 2018 IEEE International Conference on Cognitive Computing (ICCC), San Francisco, CA, pp. 17–24 (2018). https://doi.org/10.1109/ICCC.2018.00010
4. Chen, L.-C., Sheu, R.-K., Peng, W.-Y., Wu, J.-H., Tseng, C.-H.: Video-based parking occupancy detection for smart control system. Appl. Sci. **10**, 1079 (2020). https://doi.org/10.3390/app10031079

5. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Meghini, C., Vairo, C.: Deep learning for decentralized parking lot occupancy detection. Expert Syst. Appl. (Online) (2016)
6. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: Computer Vision—ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol. 9905. Springer
7. Sarker, M.M.K., Weihua, C., Song, M.K.: Detection and recognition of illegally parked vehicles based on an adaptive gaussian mixture model and a seed fill algorithm. J. Inf. Commun. Converg. Eng. **13**(3), 197–204 (2015)
8. De Almeida, P.R., Oliveira, L.S., Britto, A.S., Silva, E.J., Koerich, A.L.: PKLot–a robust dataset for parking lot classification. Expert Syst. Appl. **42**, 4937–4949 (2015)
9. Patchava, V., Kandala, H.B., Babu, P.R.: A smart home automation technique with Raspberry Pi using IoT. In: 2015 International Conference on Smart Sensors and Systems (IC-SSS), Bangalore, pp. 1–4 (2015). https://doi.org/10.1109/SMARTSENS.2015.7873584
10. Rai, R.: The Socket.IO protocol, Chap. 5. In: Socket.io Real-Time Web Application Development. Packt Publishing. ISBN: 9781782160786
11. Cadenhead, T.: Creating real-time dashboards, Chap. 2. In: Socket.IO Cookbook
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 142–158 (2016). https://doi.org/10.1109/TPAMI.2015.2437384
13. Fleet, D., Pajdla, T., Schiele, B., Tuytelaars T. (eds.): Microsoft COCO: common objects in context. In: Computer Vision—ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol. 8693. Springer, Cham
14. Chapel, M.-N., Bouwmans, T.: Moving objects detection with a moving camera: a comprehensive review. Comput. Sci. Rev. **38**, 100310 (2020)
15. Patankar, J.B.: A method for resizing images by content perception. In: 2017 IEEE International Conference on Image Processing (ICIP), Beijing, pp. 3725–3729 (2017). https://doi.org/10.1109/ICIP.2017.8296978
16. Majeeth, S.S., Babu, C.N.K.: Gaussian noise removal in an image using fast guided filter and its method noise thresholding in medical healthcare application. J. Med. Syst. **43**, 280 (2019). https://doi.org/10.1007/s10916-019-1376-4
17. Herrero-Jaraba, E., Orrite-Uruñuela, C., Senar, J.: Detected motion classification with a double-background and a neighborhood-based difference. Pattern Recogn. Lett. **24**, 2079–2092 (2003)
18. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN, Facebook AI Research (FAIR)
19. Kamel, K., Smys, S., Bashar, A.: Tenancy status identification of parking slots using mobile net binary classifier. J. Artif. Intell. Capsule Netw. **02**(03), 146–154 (2020)
20. Banerjee, S., Choudekar, P., Muju, M.K.: Real time car parking system using image processing. İn: International Conference on Electronics Computer Technology, pp. 99–103 (2011)

Deep Face-Iris Recognition Using Robust Image Segmentation and Hyperparameter Tuning

Dane Brown

Abstract Biometrics are increasingly being used for tasks that involve sensitive or financial data. Hitherto, security on devices such as smartphones has not been a priority. Furthermore, users tend to ignore the security features in favour of more rapid access to the device. A bimodal system is proposed that enhances security by utilizing face and iris biometrics from a single image. The motivation behind this is the ability to acquire both biometrics simultaneously in one shot. The system's biometric components: face, iris(es) and their fusion are evaluated. They are also compared to related studies. The best results were yielded by a proposed lightweight Convolutional Neural Network architecture, outperforming tuned VGG-16, Xception, SVM and the related works. The system shows advancements to 'at-a-distance' biometric recognition for limited and high computational capacity computing devices. All deep learning algorithms are provided with augmented data, included in the tuning process, enabling additional accuracy gains. Highlights include near-perfect fivefold cross-validation accuracy on the IITD-Iris dataset when performing identification. Verification tests were carried out on the challenging CASIA-Iris-Distance dataset and performed well on few training samples. The proposed system is practical for small or large amounts of training data and shows great promise for at-a-distance recognition and biometric fusion.

Keywords At-a-distance · Convolutional neural network · Lightweight · Face · Iris · Multimodal biometrics

Supported by the National Research Foundation (120654). This work was undertaken in the Distributed Multimedia CoE at Rhodes University.

D. Brown (✉)
Rhodes University, 6140 Grahamstown, South Africa
e-mail: d.brown@ru.ac.za
URL: https://www.ru.ac.za/computerscience/people/academicstaff/drdanebrown/

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_19

1 Introduction

Biometrics measure unique behavioural and biological traits to detect or verify a human being's identity, typically for citizen registration, access control, or law enforcement [12, 16]. Government and civilian biometric applications offer advantages over traditional authentication approaches because they are convenient and cannot be forgotten or lost. However, due to their widespread use through networks and sensors, spoofing and other security risks continue to plague biometric systems [8, 12]. Furthermore, dirty sensors, user error and other noise from real-world conditions can cause biometric data degradation. Therefore, developers often bias these systems towards increased security (stricter) or convenience (easier to acquire).

Multimodal biometrics can provide a solution to smartphone authentication and rapid access to applications by combining multiple biometric data sources. However, leveraging the combined data is not appropriate to every application; for instance, when user convenience is compromised. A growing trend towards mobile and at-a-distance biometric acquisition has been observed due to the recent focus on frontal sensor variety and quality [21]. An example is using the face and iris region to unlock the mobile device [27]. Since these two biometrics can be acquired simultaneously at-a-distance, the user's inconvenience[1] is similar to utilizing a single biometric but with improved accuracy and security. On the other hand, this provides opportunities for robustness in capturing the face, iris, or ideally both based on the angle at which the mobile device is held. This is a secure alternative to fingerprints for replacing text passwords to reduce user inconvenience.

Face recognition is a highly visible and user-friendly biometric used for authentication, or as a secondary biometric as a profile picture [11]. Acquiring the face in uncontrolled applications is generally non-trivial because of pose angle variations and occlusions. Using the face as a primary means of authentication may thus not be feasible.

Iris recognition is well established as the United Arab Emirates have used iris recognition for border control since 2001. In this old example, about 2.7 billion iris cross-comparisons were done per day [6]. Its growth is attributed to it being the most accurate external image biometric [14]. Recently, iris sensors have attained improved capturing range at less cost for commercial use [27]. The longer capturing range also increases human error and can reduce the periocular region and other image-based biometrics detail. However, using newer computer vision techniques may prove useful for field advancements.

The impact of this paper includes a tiny, efficient CNN architecture geared towards effective image-based biometric recognition. A deep learning approach to face landmark detection is used to segment the face [7] followed by frontalization. Although based on previous studies [30], iris preprocessing and segmentation is improved and uses lightweight image processing techniques. This can especially be useful for mobile devices as it can provide sufficient security while minimizing inconvenience to the user.

[1] Memorizing and typing of secure passwords.

The application is for the face and iris to be captured simultaneously by requiring a single gaze at the capturing device. Furthermore, face and iris fusion combinations are investigated to obtained higher accuracies and security with the extra information and dual templates. The system is evaluated on an at-a-distance iris dataset that excludes the bottom part (and sometimes the sides) of the face. Limitations include the dataset's use of a DSLR camera that is not typically used in cellphone iris scanners. This is thus not implemented on the mobile phone for real-time recognition.

Document structure: Section 2 reviews literature on the face, iris and their fused systems. Section 3 explains the proposed image preprocessing on the face and iris before classification. This is followed by Sects. 6 and 7 on the proposed methods for traditional and deep learning, respectively. In Sect. 8, the proposed system is compared to related systems and results are analysed. Summary of results and conclusions are drawn Sects. 9 and 10.

2 Related Studies

This section consists of related studies of the face, iris and their combined features.

2.1 Face

Face segmentation relies on feature extraction to prune away background noise, such as an individual's background and hair. Consistently removing these dynamic features during segmentation is a well-researched problem and is essential for accurate face registration. [15] introduced a face segmentation technique that uses gradient boosting for learning an ensemble of regression trees. While the system surpassed real-time performance, there was room for improvement in accurate landmark positioning and robustness to wide pose angles.

Haghighat et al. [10] produced a complete face recognition system. It performed accurately with only one training sample and achieved it without 3D modelling or deep learning. It made use of [15]'s landmark detection algorithm to produce those high recognition results. Furthermore, their approach frontalized all segmented faces using a base mesh. However, a disadvantage is the time-consuming computation of 40 Gabor filters in eight orientations and five scales. The top-performing approaches use deep learning for its fast inference, and accuracy [7, 34].

Yang et al. [34] performs face transformation to achieve translation, scale and rotation invariant registration of each face to reduce the variance of the regression target. A novel deep Convolutional Neural Network (CNN) called Stacked Hourglass Network increases the regression model's capacity on the registered face.

Deng et al. [7] employs feature pyramid level and was inspired by PyramidBox [28]. A shared loss head across different feature maps are used, and scale specific anchors are used on the feature pyramid levels to cater for varied face sizes from 16 × 16 to 406 × 406. This enabled the study to outperform state-of-the-art object detectors, such as R-CNN and variants geared towards face segmentation.

2.2 Iris

While it is well known that the iris is the most accurate external biometric, this is achieved on short-range sensors [14]. The challenge is improving iris recognition when using at-a-distance sensors.

There are two main iris segmentation methods, namely integro-differential operator and Hough transform (HT) [21]. The integro-differential operator applies an exhaustive search for both the iris' centre and radius independently by calculating the maximum in the blurred derivative with respect to the increasing radius of normalized contour integrals along circular trajectories. On the other hand, HT uses binary edge maps to localize pupil and iris boundaries. Votes are accumulated to estimate the parameters of the boundary concentric circles.

Umer et al. [30] used an HT algorithm suitable for iris extraction. The algorithm, known as Restricted Circular Hough transformation (RCHT), searches for circles bounded by the upper and lower eyelids. Texture within two concentric circles is extracted, constituting the feature vectors of the iris. The resulting feature vectors of each eye are combined into a larger vector, classified using linear support vector machines (SVMs). Later [31] improved this system by using an ensemble of patch statistics features, also classified by the linear SVM.

Cho et al. [4] considered concatenation of different feature sets of local feature extractors such as PCA, LBP, PSO and variants. They found that PSO performed best.

Jamaludin et al. [13] recently demonstrated the effectiveness of GPU parallel processing on the CASIA iris dataset. In terms of noise reduction, the iris region's segmentation accuracy was improved similarly as [30], by bounding the upper and lower eyelid regions, thus not using the whole iris. This resulted in a high GAR of 96%. The iris recognition algorithm executed in 719 ms, which is 200 ms faster than the state-of-the-art. This is useful for the training phase, but the model should run inference on a mobile device in close to real-time. Measuring whether complex models can be run on mobile devices is beyond the scope of this paper.

2.3 Face and Iris Fusion

Eskandari and Toygar [8] used an effective fusion scheme that combined the face and iris at the feature level. They emphasized the importance of feature alignment for accurate segmentation. Irises are segmented and rotated for alignment based on the face's pose position. Furthermore, their system is made robust to noise by applying particle swarm optimization (PSO). However, this substantially increases computational complexity.

Umer et al. [31] used Borda count rank-level fusion which may outperform the feature-level fusion method, based on accuracy improvement versus single irises.

In addition to recognition, CNNs have also been used for auxiliary tasks such as ROI extraction instead of relying on image processing techniques. The robustness of CNNs against misalignment and distortion in colour object detection, such as ImageNet [18] is fairly well known. However, it may be particularly interesting to determine whether their feasibility extends to greyscale iris images.

3 Proposed System

The system was coded in C++ using the OpenCV, Dlib image processing libraries and Caffe. Figure 1 provides an overview of the algorithm: face on the left; iris on the right of the diagram.

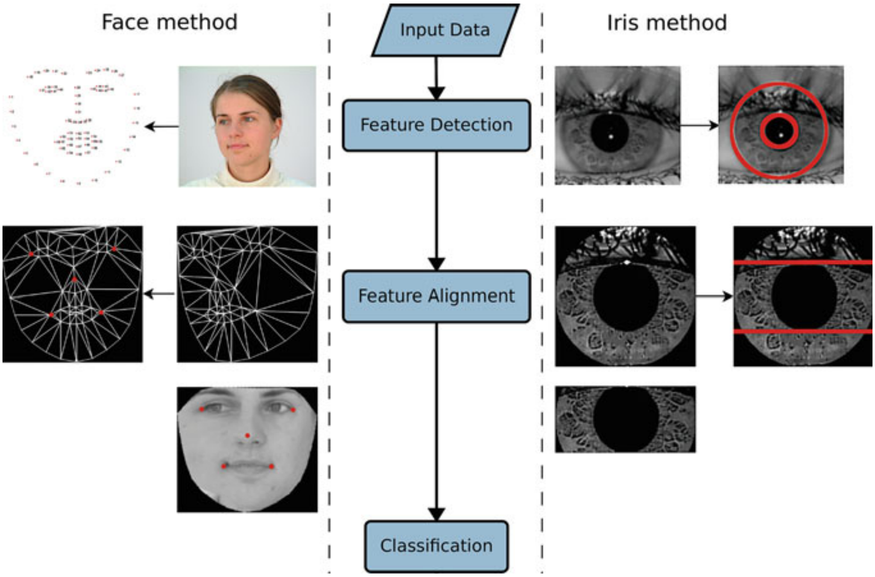


Fig. 1 Proposed face and iris recognition solution

4 Feature Detection

4.1 Face

An initial region of interest (ROI) is determined by detecting the face window. [7]'s object detection was used to localize the face to be frontal with five landmarks, the two eyes, nose edges, and two mouth edges. Figure 2 shows the 68 interpolated landmarks that are subsequently detected within the initial face ROI using [34]'s stacked hourglass CNN model.

4.2 Iris

The Restricted Circular Hough transformation (RCHT) method [30] is applied to the eye region that is obtained from the earlier face localization step [7]. The RCHT method's consistency is improved by applying image processing: Laplacian of Gaussian (LoG) filter with 5 × 5 and 17 × 17 kernels, respectively.

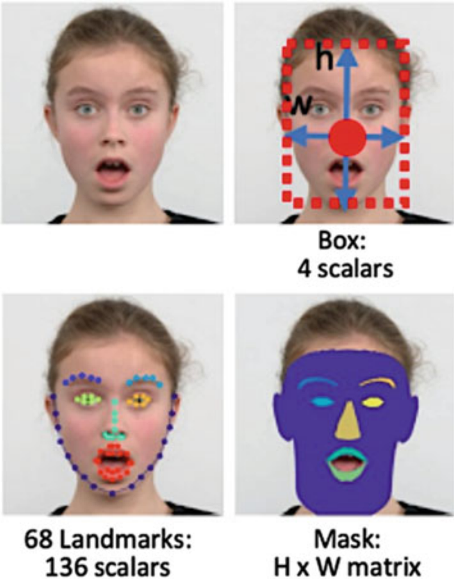


Fig. 2 **Top**: Face detection box. **Bottom**: 68 landmarks to mask the face region for segmentation

5 Feature Alignment for Robust Segmentation

5.1 Face

The detected landmarks are used for face frontalization using Delaunay triangulation and affine transformations. The triangles are created such that no landmark is inside the circumcircle. The example left facing pose shown in Fig. 1, is corrected to a frontal pose before it is ready for classification.

5.2 Iris

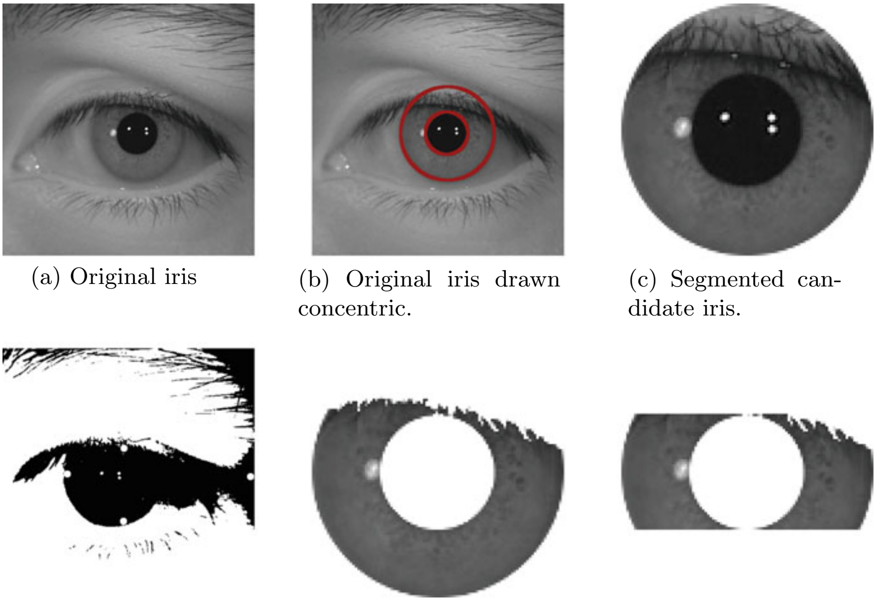
This algorithm is constructed towards accurate iris segmentation to establish an effective iris recognition system. The Hough Transform (HT) method [33] is improved similarly to [30]'s work. The improvement hinges upon finding the centre of the pupil with high precision. The improved process proceeds as follows.

1. Candidate pupil regions are detected by restricting the HT to approximate circles of a pixel-sized radius of $r[15, 45]$, determined empirically for periocular regions. The parameters used for the HT are a Canny edge threshold, $\text{Th} = 100$, and a minimum circle centre inter-distance of 10 pixels.
   In contrast with the original HT method, the pupil (smallest circle) is first detected. When overlapping circles are found, the smaller one with higher contrast is selected as it is assumed the pupil is always smaller and darker than the iris.

2. An approximate concentric circle (candidate iris), with a 3-pixel tolerance for the common centre coordinate, is sought using the same parameters, except with a radius range of $r[30, 80]$, instead of $r[15, 45]$. The candidates that fit within the periocular region are selected. This reduced the number of candidate irises to one in more than 90% of the images when carried out on the right iris for each individual. The candidate iris is highlighted in Fig. 3b.

3. Segmentation of the outer region without considering eyelid boundaries produces the image in Fig. 3c. To counteract occlusions and other inconsistencies, the rest of the process follows a new approach on obtaining the best iris candidate with no known commonalities with the original HT method but a similar final pruning as [30].

4. The eyelid boundaries are determined by contrasting the sclera and the iris region, vertically and horizontally, to the first connected pixel found from the centre of the pupil, producing Fig. 3d. This thresholding removes non-iris pixels better to approximate the amount of occlusion and lower intra-class variations when dealing with non-ideal iris image acquisition such as long eyelashes, squinting



(a) Original iris

(b) Original iris drawn concentric.

(c) Segmented candidate iris.

(d) Otsu binarized.

(e) Bitwise-and on c and d.

(f) Bottom-half iris.

Fig. 3 Iris image segmentation procedure

and off-centre eye gazing. The candidate that falls outside the boundaries by the least number of white pixels is selected as the best candidate, as shown in Fig. 3e. This new approach improves the consistency of the best iris candidate ROI using the bitwise-and operation on the binarized image and the candidate iris.

5. The resulting image has occlusions on the iris's top and bottom extremes due to eyelid positioning. This step further deals with squinting and off-centre eye gazing. The final iris segmentation result is shown in Fig. 3f.

Although one may argue that the last step throws away valuable information, the benefits of cropping a consistent area were found to outweigh that. The pruned concentric circle is the final ROI for the iris. An inverted Gaussian filter with 11 × 11 kernel is used to sharpen the ROI and used as input for classification.

6 Traditional Classification

6.1 Feature Extraction

A combination of LoG and circular local binary patterns (LLBP) was found to be particularly effective at maximizing inter-class variation for image-based biometrics in previous research [2, 3]. The Gaussian and Laplacian kernels were 15 × 15 and 7 × 7, respectively. This was applied to all of the normalized and segmented data. While the LLBP operator usually is not used this way, it reduces lighting differences with less noisy side effects according to [2]. This approach is only applied to the SVM classifier in this paper, as it was not appropriate for CNNs.

6.2 Support Vector Machine

The linear Support Vector Machine (SVM), famous as a general classifier, is used for its scalability over kernel-based versions and since it removes more data points that do not adhere to the maximum margin without requiring substantial parameter tuning [29]. Although the linear SVM might operate better for verification as it is a binary class problem, multiclass problems of identification and probability estimates to reject imposters is made possible through one-versus-rest. Overfitting the model is mitigated by using internal five-fold cross-validation. The Liblinear library allows substantial parallelization when using the above method.

7 Classification Using Deep Learning

The Convolutional Neural Network (CNN) is an effective deep learning algorithm on image analysis [26]. Convolutional layers map inputs to output feature maps using a 2D filter. Each filter's weights are updated during supervised learning to extract relevant discriminant features from the data [35]. There may be some other layers inserted here. In any case, the flattened output is input to a softmax activation layer for multiclass classification. This result is compared to the known labels of validation data, and the validation loss is computed to guide how the weights are updated per epoch.

Three popular CNN architectures were considered: VGG-16 [23], Xception [5] and MobileNet. However, they were mainly designed for problems such as ImageNet [18], which contains a vast number of objects within colour images. The ImageNet weights and top layers were thus discarded and trained from scratch, as they proved ineffective otherwise. Therefore, a custom architecture is also proposed by utilizing Keras-Tuner to determine the optimal number of blocks of layers by iteratively adding layers in a loop.[2] The high-level resulting structure is the first 2/5 blocks of the VGG-16, and one Fully Connected (FC) flattened layer and softmax classification layer, and some augmentation and regularization layers.

7.1 Overview of the Proposed CNN

As stated above, the proposed architecture consists of only two blocks of layers. It is visualized as a block diagram in Fig. 4 with layers highlighted in bold when referring to them in text. **Keras-Tuner** and conditional statements were used to help determine the architectural choices based on achieving the highest validation accuracy. Hyperparameter tuning results are provided in Sect. 8.1 for all classifiers.

1. **Image augmentation**: A typical method to reduce overfitting on image data is to enlarge the training dataset [24] artificially. Furthermore, CNNs often require many training examples so that they can extract more features at each layer. The most effective operations were rotation, shearing, zooming, and horizontal and vertical shifts. Furthermore, nine augmented images per original training image were sufficient, e.g. augmenting three training images results in a total of 30 training images that are used as input.
   The **Resizing** layer involves bilinear resizing the resolution and is tuned for values from 16 × 16 to typical ImageNet size of 224 × 224, in steps of 16 (fixed aspect ratio).
   The **Normalization** layer refers to the application of scaling to unit variance. Finally, convolutions are applied as follows.

2. **Convolution layer**: Feature extraction is in the form of 2D filters. A very small receptive field[3] of 3 × 3 with a 1-pixel stride is used [23].

3. **Non-linear layer with Batch Normalization**: [18] explain that modelling a neuron's output f as a function of its input x is with $f(x) = \tanh(x)$ or $f(x) = \left(1 + e^{-x}\right)^{-1}$. However, training time is substantially slower with those saturating than non-saturating nonlinearity $f(x) = \max(0, x)$. Neurons with the latter nonlinearity are known as Rectified Linear Units (ReLUs).
   The proposed approach first applied **Batch Normalization** followed by ReLU in both blocks as it achieved better validation accuracy. However, in the FC layer, the order was reversed.

4. **Pooling layer**: The pooling layer summarizes semantically similar features in the same kernel map [20]. A **Max pooling 2D** layer replaces $n \times n$ neighbourhoods with their highest activation result.

[2] Random search algorithm of Keras-Tuner was used.

[3] The smallest filter that captures left/right and up/down from a centre pixel's perspective.

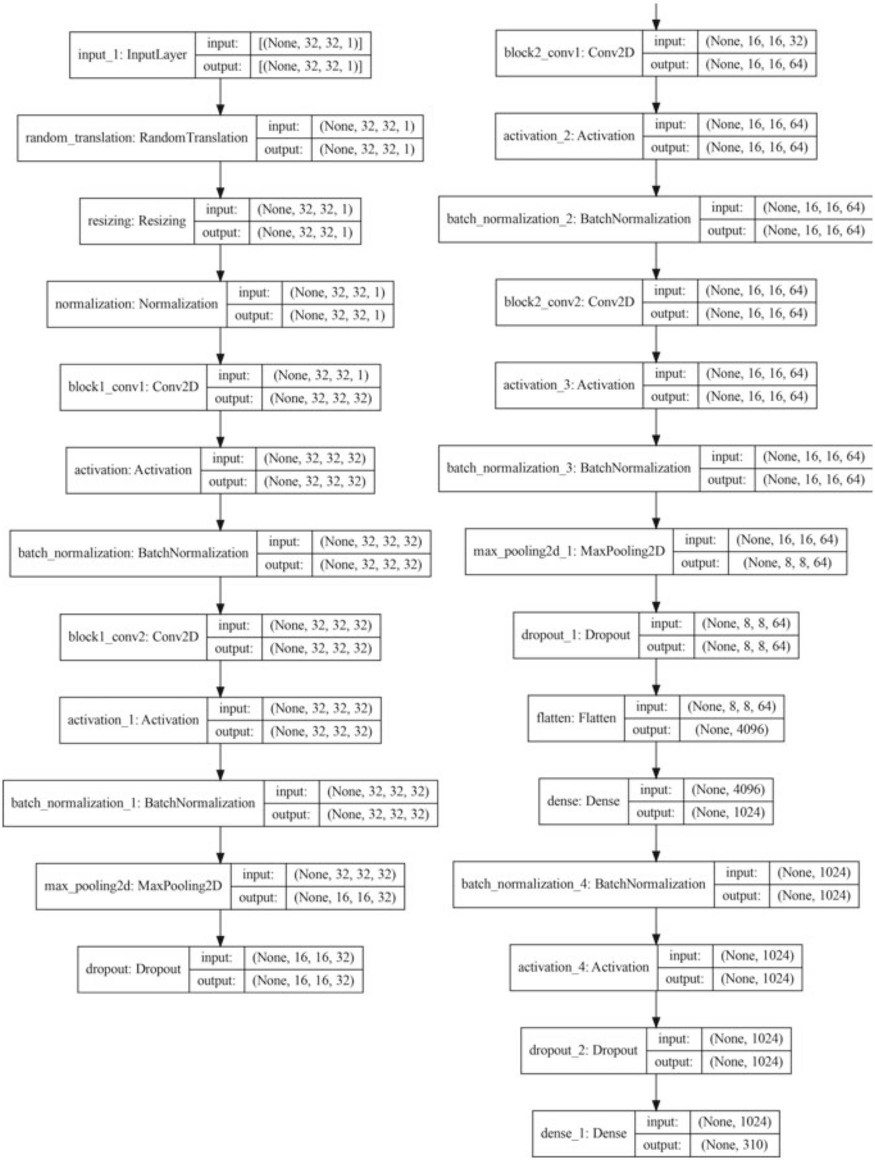

input_1: InputLayer
input: [(None, 32, 32, 1)]
output: [(None, 32, 32, 1)]
random_translation: RandomTranslation
input: (None, 32, 32, 1)
output: (None, 32, 32, 1)
resizing: Resizing
input: (None, 32, 32, 1)
output: (None, 32, 32, 1)
normalization: Normalization
input: (None, 32, 32, 1)
output: (None, 32, 32, 1)
block1_conv1: Conv2D
input: (None, 32, 32, 1)
output: (None, 32, 32, 32)
activation: Activation
input: (None, 32, 32, 32)
output: (None, 32, 32, 32)
batch_normalization: BatchNormalization
input: (None, 32, 32, 32)
output: (None, 32, 32, 32)
block1_conv2: Conv2D
input: (None, 32, 32, 32)
output: (None, 32, 32, 32)
activation_1: Activation
input: (None, 32, 32, 32)
output: (None, 32, 32, 32)
batch_normalization_1: BatchNormalization
input: (None, 32, 32, 32)
output: (None, 32, 32, 32)
max_pooling2d: MaxPooling2D
input: (None, 32, 32, 32)
output: (None, 16, 16, 32)
dropout: Dropout
input: (None, 16, 16, 32)
output: (None, 16, 16, 32)
block2_conv1: Conv2D
input: (None, 16, 16, 32)
output: (None, 16, 16, 64)
activation_2: Activation
input: (None, 16, 16, 64)
output: (None, 16, 16, 64)
batch_normalization_2: BatchNormalization
input: (None, 16, 16, 64)
output: (None, 16, 16, 64)
block2_conv2: Conv2D
input: (None, 16, 16, 64)
output: (None, 16, 16, 64)
activation_3: Activation
input: (None, 16, 16, 64)
output: (None, 16, 16, 64)
batch_normalization_3: BatchNormalization
input: (None, 16, 16, 64)
output: (None, 16, 16, 64)
max_pooling2d_1: MaxPooling2D
input: (None, 16, 16, 64)
output: (None, 8, 8, 64)
dropout_1: Dropout
input: (None, 8, 8, 64)
output: (None, 8, 8, 64)
flatten: Flatten
input: (None, 8, 8, 64)
output: (None, 4096)
dense: Dense
input: (None, 4096)
output: (None, 1024)
batch_normalization_4: BatchNormalization
input: (None, 1024)
output: (None, 1024)
activation_4: Activation
input: (None, 1024)
output: (None, 1024)
dropout_2: Dropout
input: (None, 1024)
output: (None, 1024)
dense_1: Dense
input: (None, 1024)
output: (None, 310)

Fig. 4 Overview of the proposed CNN

In the proposed approach, spatial pooling is carried out by a 2 × 2 max-pooling layer, with a 2-pixel stride, as the second last layer per block. Additional feature extraction is performed in the final layer per block using dropout.

5. **Dropout**: The output of each hidden neuron is randomly set to zero based on a certain probability (normally 0.5) [25]. The neurons which are removed do not contribute to the forward pass or backpropagation.
   The **Dropout** layer at the end of each block was initially set to 0.5, similarly to [23]. However, this was found to reduce validation accuracy. Instead, Keras-Tuner provided different optimal dropout values per block of layers.

6. **Fully connected**: For classification problems involving $K \geq 2$ classes, the softmax function is popular [9]. At this stage, the flattened stack of the FC layer contains 4096 channels before a dense layer with 1024 filters. As seen in Fig. 4, an additional **Dropout** layer is added. This made an insignificant difference during parameter tuning[4] and can thus be discarded. The result is used to backpropagate the parameters for training the network using the ADAM optimizer [17]. The softmax function is used for multiclass classification with a channel per class.

8 Experiments

First, hyperparameter tuning on iris validation data demonstrates how decisions were made. In two separate experiments, verification and identification are evaluated to emulate the literature's sampling strategy for a fair comparison. Fusion was carried out at the feature level using vector concatenation.

8.1 Hyperparameter Tuning Results

Random Search with 100 trials using fivefold cross-validation (CV) was used to evaluate tunable parameters on all of IITD [19] right-iris data.[5] This included a shared translation factor for rotation, shearing, zooming, and horizontal and vertical shifts.

Referring to Fig. 5 and Table 1, the graph shows the epochs required for the best Early Stopping accuracy of the proposed CNN. The proposed architecture peaked at 29 epochs, much earlier than VGG-16 and nine epochs earlier than Xception. The proposed CNN also preferred a significantly lower resolution, as shown in the table. Overfitting problems were as expected. Identification accuracies were all high for deep learning methods and reasonably comparable. The SVM had the lowest variance

[4] Very low probabilities ≤ 0.04 worked best.

[5] 1120 right-iris images from 224 subjects.

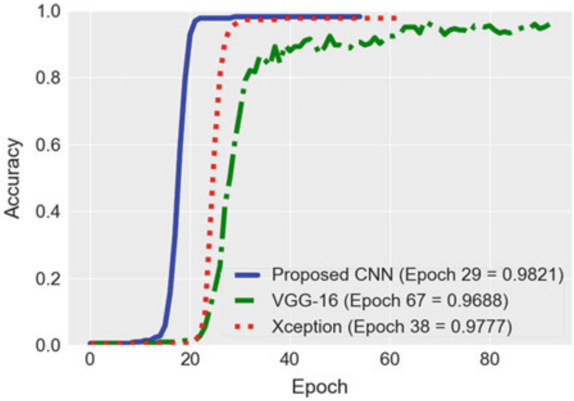


Fig. 5 Epochs required for the best Early Stopping accuracy per model

Table 1 Best CNN parameters fivefold cross validation (right) IITD iris

Classifier	Trans.	Resize	LR	Neurons (B1, B2)	Dropout (per block)	Accuracy (%)
VGG-16	0.32	224 × 224	50^{-4}	32, 64, 128, 256, 256	0.2, 0.2, 0.4, 0.4, 0.2	99.13 ± 0.21
VGG-16	0.32	128 × 128	50^{-4}	32, 64, 128, 128, 128	0.1, 0.1, 0.2, 0.2, 0.1	98.84 ± 0.21
VGG-16	0.16	128 × 128	10^{-3}	64, 128, 128, 256, 256	0.16, 0.16, 0.32, 0.32, 0.1	98.84 ± 0.24
Xception	0.64	96 × 96	10^{-4}	64, 128, 256, 512, 728	0.32	99.14 ± 0.27
Xception	0.32	64 × 64	10^{-3}	32, 64, 128, 256, 512	0.32	99.10 ± 0.21
Xception	0.16	128 × 128	10^{-3}	16, 32, 64, 128, 256	0.24	98.94 ± 0.211
Proposed CNN	0.16	64 × 64	10^{-4}	32, 64	0.1, 0.2, 0	99.74 ± 0.18
Proposed CNN	0.16	32 × 32	50^{-4}	24, 72	0.1, 0.32, 0.04	99.53 ± 0.23
Proposed CNN	0.08	16 × 16	10^{-3}	24, 64	0.1, 0.2, 0.04	99.12 ± 0.19
SVM	–	32 × 32	–	–	$C = 10$	92.15 ± 0.05
SVM	–	16 × 16	–	–	$C = 100$	91.82 ± 0.07
SVM	–	64 × 64	–	–	$C = 1$	91.80 ± 0.05

Table 2 EER of face verification on the CASIA-Iris-Distance dataset using five training and test images for each of the 90 subjects: both related studies perform comparably to VGG-16 and Xception, but the proposed CNN outperforms them

Approach	Face (%)	Fused iris (%)	All fused (%)
Proposed CNN	5.4	6.45	2.55
VGG-16	6.02	7.25	3.52
Xception	5.95	6.97	3.33
SVM	12.12	15.44	6.11
[8]	5.50	10.66	3.78

SVM is significantly worse than the rest, which may indicate bad cases in segmentation or feature extraction. However, it gains substantially from fusion

across splits but yielded a significantly lower accuracy than the rest. Augmentation yielded reasonably similar gains across deep learning architectures—between 4–8% improvement. It tended to provide better gains in lower resolutions. The remaining experiment subsections compare the approaches to related studies.

8.2 CASIA-Iris-Distance Test Results

The CASIA-Iris-Distance v4 database contains images of left and right irises collected from 142 subjects for 2567 samples. The images were captured with the subjects 3 metres away by a 2352 × 1728 NIR camera. Both dual-eye and incomplete face images are in some of the (at-a-distance) captured images, making this a challenging iris dataset.

Furthermore, few studies include the face recognition performance of the CASIS-Iris-Distance dataset, as they are essentially partial faces. Ammour et al. [1]'s verification accuracy of the face component was compared, where 90 subjects with complete were used for training and test sets. This reduced the challenge significantly. Each subject had ten samples selected randomly, while five samples were used for training and testing. Results shown in Table 2 indicate that accuracies are similar between the systems (Tables 3 and 4).

9 Summary of Results

The overall results indicate that a smaller designed CNN can achieve impressive results on challenging iris data. The proposed CNN required fewer epochs than even Xception during validation testing. It outperformed the other variants and related studies in all tests. The SVM did, however, appear to scale better on fused components.

Table 3 Iris verification twofold Cross Validation on the CASIA-Iris-Distance dataset

Approach	EER (%)
Proposed CNN	4.5
VGG-16	5.1
Xception	5.0
SVM	10.1
[4]	10.02
[22]	8.64
[32]	4.91

Worse accuracies are expected with fewer training samples than before. VGG-16 and Xception are outperformed by [32]'s recent system. However, the proposed CNN performs better. The SVM and [4]'s achieved an almost equivalent verification rate

Table 4 Identification accuracy on components of the CASIA-Iris-Distance dataset

Method	**Left (%)**	**Right (%)**	**Fused Irises (%)**
Proposed CNN	93.66	94.32	96.55
VGG-16	91.72	91.98	94.35
Xception	92.77	93.35	95.57
SVM	88.15	89.23	93.22
[30]	91.85	92.21	94.47

Again SVM has subpar accuracy, but its fused irises result is impressive. VGG-16 performed similarly to [30]'s work. The proposed CNN outperforms them significantly

10 Concluding Remarks

This paper utilized image preprocessing techniques on traditional machine learning algorithm—SVM. Popular CNN architectures and an empirically determined one were compared to related studies. The proposed CNN's basic structure was largely based on VGG-16, and 'good' dropout rates were about 0.1 for the first block and about double for the second block.[6] This is lower than the original VGG-16 and Xception studies, and this may indicate that the proposed architecture propagates less noise. However, this may simply be limited to the CASIA-Iris-Distance dataset. Overfitting was also combatted via augmentation. Excluding augmentation resulted in an accuracy drop of 4–8%. This significant change is especially due to the limited number of samples available for training and is a positive result.

The traditional classifier—SVM was generally not as effective as the other classifiers. This trend requires further investigation. On the other hand, the proposed CNN proved to be extremely effective and outperformed the related studies as well as all the other proposed approaches. The CNNs did not benefit from fusion as much as SVM. Trends among classifiers were reasonably consistent in general.

[6] The second block preferring double the dropout rate was a consistent trend.

The following warrants future investigation. Although the proposed iris segmentation method was successful in previous work on traditional classifiers, CNNs may prefer all the information available, including the bottom region of the iris. Additional fusion schemes on other data may also be prudent for further investigation, especially since the SVM gained significantly more during fusion than deep learning methods.

References

1. Ammour, B., Bouden, T., Boubchir, L.: Face–iris multi-modal biometric system using multi-resolution Log-Gabor filter with spectral regression kernel discriminant analysis. IET Biometrics **7**(5), 482–489 (2018), ISSN 2047-4938, 2047-4946, 10.1049/iet-bmt.2017.025
2. Brown, D., Bradshaw, K.: An investigation of face and fingerprint feature-fusion guidelines. In: Kozielski, S., Mrozek, D., Kasprowski, P., Małysiak-Mrozek, B., Kostrzewa, D. (eds.) Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery: 12th International Conference, Ustroń, Poland, May 31–June 3, 2016, Proceedings, pp. 585–599, Springer International Publishing (2016)
3. Brown, D., Bradshaw, K.: Improved palmprint segmentation for robust identification and verification. In: 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 1–7, IEEE (2019)
4. Cho, S.R., Nam, G.P., Shin, K.Y., Nguyen, D.T., Pham, T.D., Lee, E.C., Park, K.R.: Periocular-based biometrics robust to eye rotation based on polar coordinates. Multimedia Tools and Applications 76(9), 11177–11197 (2017)
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
6. Daugman, J.: Iris recognition border-crossing system in the UAE. Int. Airport Rev. **8**(2) (2004)
7. Deng, J., Guo, J., Yuxiang, Z., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. In: arxiv (2019)
8. Eskandari, M., Toygar, Ö.: Selection of optimized features and weights on face-iris fusion using distance images. Comput. Vis. Image Underst. 137, 63–75 (2015)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. arXiv:1706.04599 (2017)
10. Haghighat, M., Abdel-Mottaleb, M., Alhalabi, W.: Fully automatic face normalization and single sample face recognition in unconstrained environments. Expert Syst. Appl. 47, 23–34 (2016)
11. Jain, A., Hong, L., Pankanti, S.: Biometric identification. Communications of the ACM 43(2), 90–98 (2000a)
12. Jain, A.K., Prabhakar, S., Hong, L., Pankanti, S.: Filterbank-based fingerprint matching. IEEE Transactions on Image Processing 9(5), 846–859 (2000b)
13. Jamaludin, S., Zainal, N., Zaki, W.M.D.W.: Sub-iris technique for non-ideal iris recognition. Arabian J. Sci. Eng. 43(12), 7219–7228 (2018)
14. Kaur, G., Verma, C.: Comparative analysis of biometric modalities. Int. J. Adv. Res. Comput. Sci. Softw. Eng. 4(4), 603–613 (2014)
15. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)
16. Kindt, E.J.: An introduction into the use of biometric technology. In: Privacy and Data Protection Issues of Biometric Applications: A Comparative Legal Analysis, pp. 15–85. Springer Netherlands (2013)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)

18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM 60(6), 84–90 (2017)
19. Kumar, A., Passi, A.: Comparison and combination of iris matchers for reliable personal authentication. Pattern Recogn. 43(3), 1016–1026 (2010)
20. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015)
21. Rathgeb, C., Uhl, A., Wild, P.: Iris Biometrics: From Segmentation to Template Security, vol. 59. Springer Science & Business Media (2012)
22. Shekar, B., Bhat, S.S.: Iris recognition using partial sum of second order taylor series expansion. In: Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1–8 (2016)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
24. Song, J.M., Kim, W., Park, K.R.: Finger-vein recognition based on deep densenet using composite image. IEEE Access 7, 66845–66863 (2019)
25. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15(1), 1929–1958 (2014)
26. Sundararajan, K., Woodard, D.L.: Deep Learning for Biometrics: A survey. ACM Comput. Surveys (CSUR) 51(3), 1–34 (2018)
27. Tan, T.: CASIA Iris Image Database. http://biometrics.idealtest.org/ (2016), [Online; Accessed 14-March-2016]
28. Tang, X., Du, D.K., He, Z., Liu, J.: Pyramidbox: A context-assisted single shot face detector. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 797–813 (2018)
29. Tang, Y.: Deep learning using support vector machines. Clin. Orthopaedics Related Res. (CoRR) **2** (2013)
30. Umer, S., Dhara, B.C., Chanda, B.: Iris recognition using multiscale morphologic features. Pattern Recogn. Lett. 65, 67–74 (2015)
31. Umer, S., Dhara, B.C., Chanda, B.: Nir and vw iris image recognition using ensemble of patch statistics features. Visual Comput. 35(9), 1327–1344 (2019)
32. Wang, K., Kumar, A.: Toward more accurate iris recognition using dilated residual features. IEEE Trans. Inf. Forensics Secur. 14(12), 3233–3245 (2019)
33. Wild, P., Hofbauer, H., Ferryman, J., Uhl, A.: Segmentation-level fusion for iris recognition. In: Proceedings of the International Conference on Biometrics Special Interest Group (BIOSIG), pp. 1–6. IEEE (2015)
34. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 79–87 (2017)
35. Zhang, L., Zhang, L., Du, B.: Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4(2), 22–40 (2016)

Text-Based Sentiment Analysis with Classification Techniques—A State-of-Art Study

M. S. Kalaivani and S. Jayalakshmi

Abstract Social media acts as a bridge between people to widely share any data and communication. In recent years, the textual data content is increasing rapidly, where the text can contain any kind of information about people, product or service. Manually reading each text from online is not possible, and also, it is a challenging task to decide whether the user has positive stance or negative stance on the topic. To solve this problem, text processing techniques and algorithms are required. Sentiment analysis is the technology that processes any online text and classifies it into positive, negative and neutral. To analyze online content, new models are proposed by incorporating the machine learning concept. The unstructured information from online documents is analyzed and classified as results, which has been described as user sentiment analysis. The outcome of sentiment analysis can be used for business development, understand the customer expectations and also to know the public opinion toward a particular product or service. This paper focuses on various sentiment analysis processes and also the most used classification techniques from machine learning concepts.

Keywords Sentiment analysis · Machine learning · Maximum entropy · Support vector machine · Naïve Bayes · Feature extraction · Classification

1 Introduction

In the recent decade, websites and social media make a path for user opinions and feedback. People share information and communicate with each other through social media such as Instagram, Facebook, YouTube, WhatsApp, Twitter and others. Twitter is one of the famous micro-blogging system in social media, which is used to send

M. S. Kalaivani (✉)
Tagore College of Arts and Science, Chennai, India
e-mail: kalaims2007@gmail.com

S. Jayalakshmi
Vels Institute of Science, Technology and Advanced Studies, Chennai, India
e-mail: jai.scs@velsuniv.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_20

and receive short messages with character length 140. People can liberally express their views and thoughts, so each second, huge amount of data are generated as online content. Sentiment analysis, a subfield of natural language processing, which process with user opinions and feedback. It is also known as opinion mining, a study about user reviews, feedback, stance, thoughts, estimation or judgment toward a product, people, movie, hotel or anything as online text. To analyze online text, a sequence of tasks such as data preprocessing, feature extraction, classification and result analysis is involved in sentiment analysis. Text classification problems are solved by using NLP methods. Analyzing the meaning of text and context behind the text can be done by using natural language understanding. Sentiment analysis can be categorized into three levels.

Document Level. A document with single topic is considered for the process. An online document can have various opinions or reviews for a single theme. The whole document will be reduced to a single sentiment score. The result classifies positive, negative or neutral opinions from the document. Comparative learning is not considered in document level.

Sentence Level. The sentences are analyzed and sentiments identified as positive, negative and neutral view. The sentence without any opinions considered as neutral. In subjective statements, polarity of the sentiments termed as good or bad.

Aspect Level. This category provides more detail on data analysis. It finds the aspect of given text. For example, a product—television review by a customer, "The clarity is good, but the screen size is very small." In this feedback, the aspects are clarity and screen size. The opinion about screen clarity is positive, and screen size is negative. Based on the analysis, an unstructured information is transformed into structured text, which can be more useful in qualitative and also in quantitative analysis.

2 Approaches in Sentiment Analysis

There are three types of approaches in sentiment analysis. Statistical/machine learning-based approaches, knowledge/lexicon-based and hybrid approaches [1].

Statistical/Machine Learning Approach. It uses various machine learning algorithms applied in linguistic features. It is very useful to study the data and recognizes data designs for regression analysis. Machine learning methods are divided into supervised, unsupervised and semi-supervised algorithms. Bag of words, support vector machines, latent semantic analysis and semantic orientation are some methods used in this approach for text processing.

Lexicon-Based Approach. This method fully depends on lexicons, which is a special dictionary used for text analysis. The words in the dictionary are associated with scores, used to compute sentiments. The input text is separated as single tokens, and every token is compared with lexicons in the dictionary. If there is a positive match, and the score of the input text will be incremented; otherwise, the score will be reduced.

Hybrid Approach. This method is combination of lexicon-based and machine learning approaches. It applies the speed of machine learning approach and accuracy of lexicon-based approaches.

3 Process in Sentiment Analysis

Sentiment analysis is used to identify the user reviews and classify the opinions, in order to determine that customer expectations toward a product, to find service are fulfilled or not. It helps to find the user opinions and classify text as positive, negative and neutral. Input data collection, data preprocessing, feature extraction, classification and result analysis are basic steps executed in sentiment analysis process. In classification part, machine learning algorithms are used for training and testing input dataset. Stopword removal, stemming, tokenization, and negation removal are done in data preprocessing stage (Fig. 1).

3.1 Data Preprocessing

It is the process of removing noisy data, irrelevant or unwanted data from input text. Removing the URLs, special characters, punctuations, stop words are the task involved in preprocessing stage. Stemming and tokenization are also included in preprocessing. Articles, prepositions and pronouns are called stop words. Stemming is a method which is used to find the base form of the word by removing affixes.

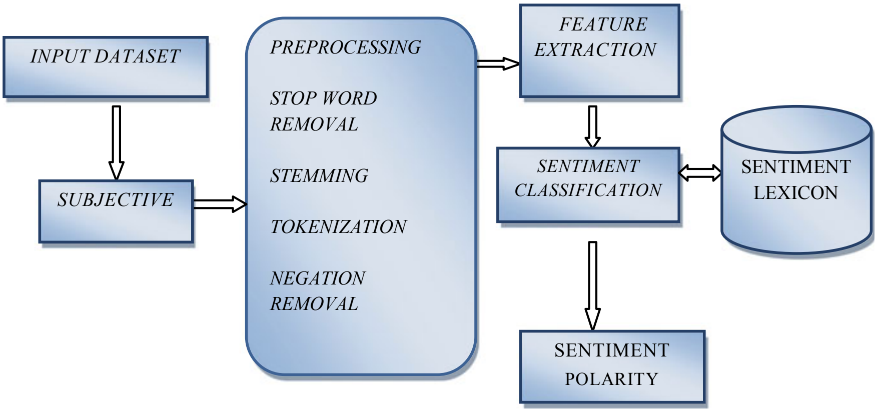


Fig. 1 Sentiment analysis process

Tokenization is the process of separating input text into small units known as tokens. In this phase, the input data is normalized by removing noisy text. After preprocessing, the data will be taken to the next step feature extraction where the required features are separated for training phase.

3.2 Feature Extraction

In this process, the special aspects similar to adjectives, nouns and verbs are extracted and classified as positive and negative words, which are used to find polarity at sentence level [2]. The special features are used to train the data with machine learning algorithms.

Term frequency and presence: It refers identifying individual and distinct word from the input text. It is used to count the occurrence of the words. Negative phrases: The words such as "not" and "never" have sentiment reversing outcome for the input text. So it is important to take negative words in account, while preprocessing the text [3].

Parts-of-speech tagging: This is fundamental task of NLP, which labels each input text as noun, verb, adjective and adverb. The labeled text will be classified by using various methods.

3.3 Sentiment Classification

Lexicon-based approaches and machine learning approaches are mainly used in sentiment classification. The lexicons are collection of words, and every word is allocated with particular score which indicates positive, negative and neutral type of the text. The scores were summed up separately, and the highest score indicates overall polarity of input text [4].

- **Corpus-Based Approach**. It refers a collection of writings about a particular topic. The seed list can be prepared and analyzed by using corpus text [5]. This is categorized into two types, statistical approach and semantic approach. Statistical approach is used in finding the occurrence words. The polarity value is assigned depend on positive occurrence or negative occurrence of a word. In semantic approach, sentiment values are calculated by using principle of similarity among words.
- **Machine Learning Approach**. In this method, classification is done by using the features extracted from the input text. Machine learning can be categorized into supervised learning and unsupervised learning (Fig. 2).
- **Supervised**. The machine is trained with the help of labeled data. After training, the system is tested with new data; by using training knowledge, the system has to classify the new text. The most used supervised machine learning algorithms

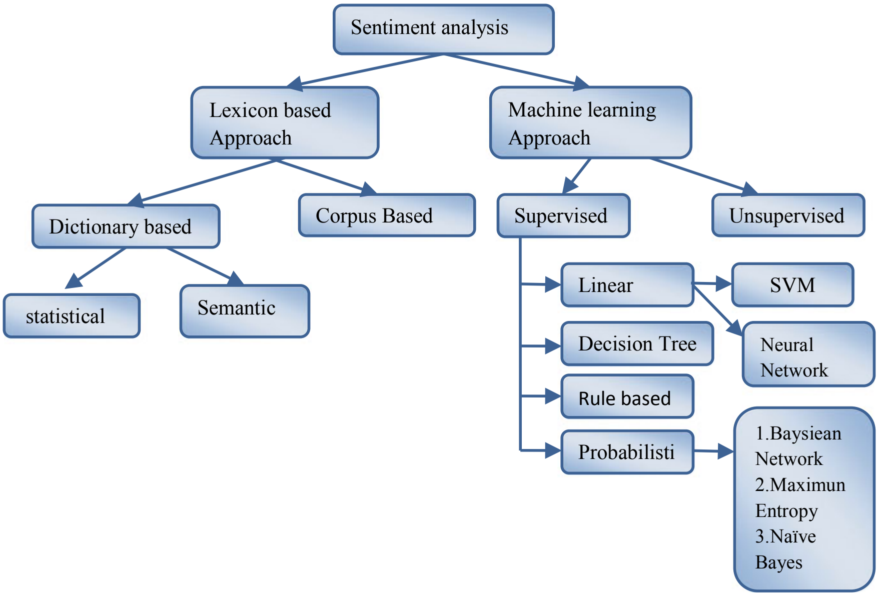


Fig. 2 Sentiment classification techniques

are Naïve Bayes, decision trees, support vector machine, maximum entropy and random forest. Based on the requirement, these algorithms can be applied for sentiment analysis.

- **Unsupervised**. This method will operate on its own and find the patterns for the dataset, and the user intervention is not required in this model.

Naïve Bayes Classifier. This is a simple classifier from probability model and most used for sentiment classification problems. It uses Bayes theorem concept and finds the maximum probability of any word which is fit enough for given class.

$$P(X_i|c) = \frac{\text{Count of } X_i \text{ in document of class } c}{\text{Total number of words in document of class } c} \tag{1}$$

P—probability, X_i—given Term, C—Predefined class model [6]. Hash tables are used to store count of the word occurance.

$$C^* = \arg\max P(c|d) \tag{2}$$

As per the probability definition, document d is further categorized by using the above equation [7].

Bayesian Network. This network is used to identify the relationship amid large number of text. It consists of a set of variables and a set of local probability distribution for each variable. To define the conditional probability distribution, a table—Conditional Probability Table (CPT) is formed, and probabilities are assigned to the variables [8]. The classification can be done by learning the structure of the network and relevant CPTs.

Maximum Entropy. In this classifier, the labeled features are converted to vectors by using encoding methods [9]. This vectors are used to calculate weight of the features and predict label for those features. This is most used method for text classification, in natural language processing. It refers the process of distributing the unknown data uniformly. The basic concept of maximum entropy is the information about the data which is unknown, but distribution is particularly uniform. So, the possibility of non-uniform distribution can be eliminated.

$$P(c|d) = \frac{1}{z(d)\{\exp(\Sigma \ell_i f_i(c, d))\}} \tag{3}$$

where $f_i(c, d)$—feature, λ_i—parameter to be predicted $Z(d)$—normalization function.

Support Vector Machine. It is a supervised learning method used to classify the input text and find regression values. Classification is used to predict the labels and regression which is used in predicting the continuous values. It is a non-probabilistic classifier which is used to categorize the input data into separate classes (Fig. 3).

In the diagram, three hyperplanes *A*, *B* and *C* are computed to classify circle and triangle shapes. The separator *C* is performed well, because the data from both side are at maximum distance from the separator (Table 1).

Fig. 3 SVM classification

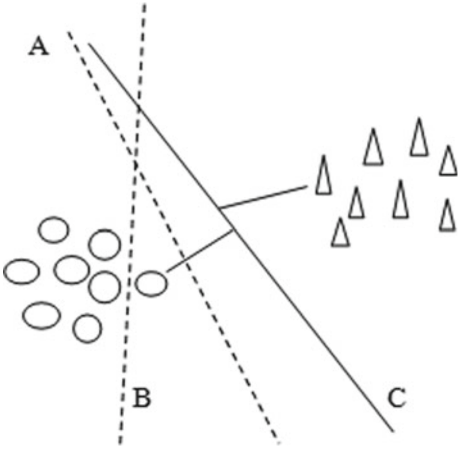


Table 1 Machine learning algorithms used in sentiment analysis

Reference	Year	Task	Dataset	Algorithm
[10]	2015	Sentiment analysis	Twitter dataset—election review, movie review	Naïve Bayes, maximum entropy, support vector machine
[11]	2016	Sentiment analysis	Book review	SVM, multinomial Naive Bayes, stochastic gradient descent, Naive Bayes and decision trees
[12]	2014	Sentiment analysis	Product review	Naïve Bayes, maximum entropy
[13]	2016	Sentiment analysis	Stock market movements	Random forest algorithms
[14]	2016	Sentiment analysis	Movie review and hotel review	Naïve Bayes, K-nearest neighbor
[15]	2017	Sentiment classification	Movie review	Naïve Bayes, decision trees with fuzzy-based approach
[16]	2018	Sentiment analysis	Political reviews	SVM, Naive Bayes
[17]	2017	Sentiment analysis	Product reviews, movie reviews	Naïve Bayes, J48, BFTree and OneR
[18]	2016	Sentiment analysis	Twitter data	Naïve Bayes, decision trees
[19]	2019	Sentiment analysis	Movie review	Bernoulli Naïve Bayes, decision tree, SVM, maximum entropy, multinomial Naïve Bayes
[20]	2018	Sentiment analysis	Electronic product review from Amazon	Naïve Bayes, SVM
[21]	2016	Sentiment classification	Twitter data	Naïve Bayes, SVM, random forest, logistic regression

4 Conclusion and Future Scope

In this paper, sentiment analysis process and classification techniques are discussed. Sentiment analysis has some challenges also, while reading the online text unwanted contents should be removed. These are known as noisy text, which are irrelevant to the topic and cannot be categorized. There is a lot of open challenges in this field, for upcoming researches. The input dataset can be taken from IMDB, Amazon, Flipkart or any online document. Every second, huge amount of data is uploaded online, so efficient techniques are required to analyze these data. Various algorithms can be implemented for same input data set to compare effective outcomes.

In future work, machine learning algorithms SVM, Naïve Bayes, maximum entropy and Bayesian networks can be implemented in text processing. Depending on the input dataset and the project requirement, machine learning methods can be applied. The complexity of the algorithm decides the cost for deploying machine learning projects, and by training the machine with many input dataset, accurate results can be achieved.

References

1. Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. 15–21 (2013)
2. Mehta, M.A., Desai, M.: Techniques for sentiment analysis of Twitter data: a comprehensive survey. In: International Conference on Computing, Communication and Automation (ICCCA2016) (2016)
3. Kaur, J., Sidhu, B.K.: Sentiment analysis based on deep learning approaches. In: Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018) (2018)
4. Kang, H., Yoo, S.J., Han, D.: Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst. Appl. **39**, 6000–6010 (2012)
5. Fazel, K., Diana, I.: A bootstrapping method for extracting paraphrases of emotion expressions from texts. Comput. Intell. (2012)
6. Kaur, H., Mangat, V., Nidhi: A survey of sentiment analysis techniques. In: International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud). IEEE (2017)
7. Desai, M., Mehta, M.A.: Techniques for sentiment analysis of Twitter data: a comprehensive survey. In: International Conference on Computing, Communication and Automation (ICCCA2016) (2016)
8. Grosan, C., Abraham, A.: Intelligent Systems, vol. 17. Springer (2011)
9. Manoharan, S.: Geospatial and social media analytics for emotion analysis of theme park visitors using text mining and GIS. J. Inf. Technol. Digit. World **02** (2020)
10. Kanakaraj, M., Guddeti, R.M.R.: NLP based sentiment analysis on Twitter data using ensemble classifiers. In: 3rd International Conference on Signal Processing, Communication and Networking (ICSCN-2015) (2015)
11. Aliane, A.A., Aliane, H., Ziane, M., Bensaou, N.: A genetic algorithm feature selection based approach for Arabic sentiment classification. IEEE (2016)
12. Gautam, G., Yadav, D.: Sentiment analysis of Twitter data using machine learning approaches and semantic analysis. IJCSI Int. J. Comput. Sci. Issues **9**(4), No 3 (2012)
13. Pagolu, V.S., Challa, K.N.R., Panda, G., Majhi, B.: Sentiment analysis of Twitter data for predicting stock market movements. In: International Conference on Signal Processing, Communication, Power and Embedded System (2016)
14. Dey, L., Chakraborty, S., Biswas, A., Bose, B.: Sentiment analysis of review datasets using Naïve Bayes and K-NN classifier. Int. J. Inf. Eng. Electron. Bus. **4**, 54–62 (2016)
15. Jefferson, C., Liu, H., Cocea, M.: Fuzzy Approach for Sentiment Analysis. IEEE (2017)
16. Hasan, A., Moin, S., Karim, A., Shamshirband, S.: Machine learning-based sentiment analysis for Twitter accounts. Math. Comput. Appl. **23** (2018)
17. Singh, J., Singh, G., Singh, R.: Optimization of sentiment analysis using machine learning classifiers. In: Human Centric Computation and Information Sciences, 7. Springer (2017)
18. Jain, A.P., Dandannavar, P.: Application of machine learning techniques to sentiment analysis. In: 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). IEEE (2016)
19. Rahman, A., Hossen, M.S.: Sentiment analysis on movie review data using machine learning approach. In: International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE (2019)

20. Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N.: Sentiment analysis on product reviews using machine learning techniques. In: Proceeding of CISC (2017)
21. Jianqiang, Z., Xiaolin, G.: Comparison research on text pre-processing methods on Twitter sentiment analysis. IEEE Access (2016)

Face Mask Detection Using MobileNetV2 and Implementation Using Different Face Detectors

Kenneth Toppo, Neeraj Kumar, Preet Kumar, and Lavi Tanwar

Abstract Face recognition and object detection have been around the artificial intelligence field for a couple of years now and are constantly evolving and being pushed in many devices which we might not be even aware of. Many of such face and object detection techniques use Convolution Neural Network (CNN) architecture at the core for understanding and classifying any image passed on to the system. The neural networks identify many characteristics and distinguishing features present in the image and then provide us with a prediction. In this paper we discuss the implementation of a face mask detection technique using Mobile NetV2, observe the accuracy of our model and compare the performance of the trained model by incorporating three different face detector models. The result achieved from the trained model brings forth the opportunity for implementing such techniques on low computational powerful devices thereby making mask detection algorithm integration much easier than other techniques.

Keywords Convolution neural networks · MobileNetV2 · Machine learning · Haar cascade · MTCNN

1 Introduction

The Covid-19 virus and the pandemic declared by the WHO (World Health Organization) had brought challenges in everyone's life. This resulted in people following new

K. Toppo (✉) · N. Kumar · P. Kumar · L. Tanwar
Department of Electronics and Communication Engineering, Delhi Technological University,
New Delhi, India
e-mail: kennethtoppo_2k17ec82@dtu.ac.in

N. Kumar
e-mail: neerajkumar_2k17ec111@dtu.ac.in

P. Kumar
e-mail: preetkumar_2k17ec130@dtu.ac.in

L. Tanwar
e-mail: lavi.tanwar@dtu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_21

lifestyles and habits that were earlier missing from people's life. One major requirement at all places was the need of wearing a mask. Mask has proven to be effective for containing any viral infection as the Covid-19 virus spreads mainly through respiratory droplets. Every office, shopping complex, and many public places have made it mandatory to wear a mask. Constant monitoring of people is required to check whether they are wearing a mask or not. We have hence proposed a technique to easily detect faces that will be wearing masks or not wearing masks. This technique is efficient as it can provide good accuracy with fewer data parameters into consideration and save time. It can be installed at places to automate granting access to people to places based on whether they are wearing a mask or not. This ability to detect faces is based on the concept of object detection in deep learning. Deep learning includes algorithms and techniques that mimic how a brain would approach any image or any piece of information, process it, and then devise some results for the same. A comparative study on most of the techniques and algorithms used in AI such as fuzzy logic, neural networks, and evolutionary algorithms was presented in [1] and many techniques for intelligent video surveillance that used deep learning in real-time detection of objects are discussed in [2–4]. Few face recognition techniques and face mask detectors used include, Face mask detection using multi-task cascaded neural networks [5], a RetinaFaceMask detector which uses high-level semantic information with feature maps [6], and face mask detection using Inceptionv3 and transfer learning [7]. Few important concepts used in our model are that of Convolutional Neural Networks and MobileNetV2 which are discussed in the below sections.

1.1 Convolutional Neural Network

Convolution Neural Networks are the most common neural networks used in Deep Learning. One of the main reasons convolution networks are used is because as technology and camera sensors are improving the image quality and size are increasing too, and computing on such large input data would make our system and outcomes slow. Such a situation would be undesirable as we try to achieve processing speeds similar to or faster than the human brain. Using CNN helps in saving space for processing while not skipping important features in the input data. Space or size reduction facilitates the easy computation of the data and thus giving us timely outputs.

Convolution Neural Networks can achieve this efficiency by making use of kernels or filters. These kernels/filters are two-dimensional assigned weights that are multiplied with a small portion of the input (for example a two-dimensional image) using dot product element multiplication after which it is summed up and the value is stored. Often it can be possible that a repeating 2-D pattern may be present in the input image at different locations, CNN provides the ability to detect these similar overlapping areas (feature sharing). For RGB input images, the input 2-D array is divided into ***three channels*** representing their respective RGB colors, however, the input matrix size remains the same. After the filter is convolved with the input and a

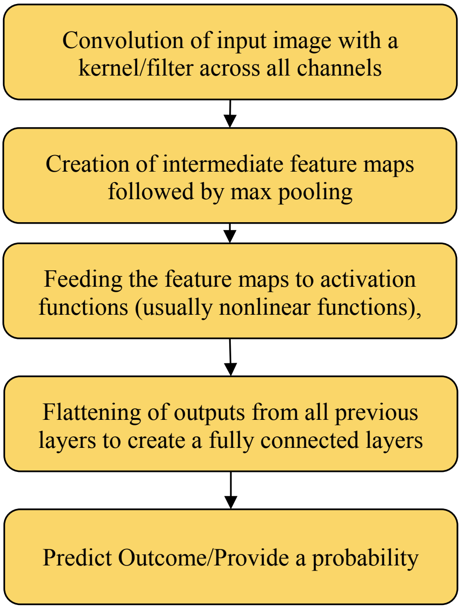

Convolution of input image with a kernel/filter across all channels
Creation of intermediate feature maps followed by max pooling
Feeding the feature maps to activation functions (usually nonlinear functions),
Flattening of outputs from all previous layers to create a fully connected layers
Predict Outcome/Provide a probability

Fig. 1 Flow diagram for processing of input image using traditional convolutional neural networks

new layer is formed, ***pooling*** is conducted on these layers. Pooling is another technique to reduce the spatial dimension of these formed layers. ***Activation functions*** are used to apply this logic and simplify our feature map values, these functions are also responsible for firing up neurons in the next layer. We used the ***ReLu activation function*** (or Rectified Linear function) as our activation function, it sets the values to 0 if it's a negative value, else if positive returns the same value (Fig. 1).

1.2 MobileNetV2

MobileNetV2 CNN [8] is a type of CNN architecture that is used for implementing deep learning models. It is quite an efficient technique for implementing convolutional neural networks. One of the features that makes it unique is the use of ***depthwise separable convolution***. In a normal convolution process in CNN, we have a lot of parameters and dimensions such as the dimension of input, the dimension of the output layer, and dimensions of the filter/kernel. Let the dimensions of the kernel be $G_k \times G_k$, dimensions of feature map size $G_f \times G_f$ with 'H' input channels, and 'N' output channels then the total size would be

$$G_k \times G_k \times H \times G_f \times G_f \times N$$

Now in a depthwise separable convolution, the same filter is first multiplied over different channels which would total to a size of,

$$H \times G_f \times G_f \times G_k \times G_k$$

And pointwise convolution uses a 1 × 1 kernel/filter across all channels over the feature map to give us a total size of the output

$$H \times G_f \times G_f \times N$$

Advantages of MobileNetV2. [8] MobileNetV2 consists of two different blocks - a residual block of stride 1 and another block of stride 2. The first layer is convolution with ReLU, the second a depthwise convolution and the third layer is again convolution without nonlinearity. When used for face mask detection and object detection technique results in fewer parameters as we will be only modifying the input image once when compared with normal convolution and it provides better processing times and better integrability with low-end devices. Now if we consider the total computation size it will be a sum of depthwise and pointwise convolution which comes out to be,

$$H \times G_f \times G_f\ (G_k \times G_k + N)$$

and if we compare this size with the total size of a regular CNN it becomes $(1/N + 1/D^2{}_k)$ times less than a CNN. 1).

1.3 Face Detectors

Face detector models are used to implement the face detection process after the model has been trained. Quite many face detectors are available in the field of computer vision. Haar Cascade face detector given by Viola and Jones [9], is one of the first face detector models developed using open-cv and is the most traditional one, Caffe modeled DNN (Deep Neural Network) face detector [10], multi-task cascaded neural network (MTCNN) based detector [11, 12], and Dlib frontal face detectors. Every face detector has its advantages and disadvantages, so it comes down to a choice based on the hardware used and how efficient is the program that will run on the system. An efficient face detection model will save a lot of computing power and provide better frames per second.

2 Methodology

In our paper, we construct a deep learning model based on the MobileNetV2 architecture for better efficiency as our aim in mind and not too computationally heavy enough to be not implemented on low-end devices. After our model is prepared/trained we

Table 1 MobileNet body architecture [8]

Type/stride	Filter shape	Input size
Conv/s2	3 × 3 × 3 × 32	224 × 224 × 3
Conv dw/s1	3 × 3 × 32 dw	112 × 112 × 32
Conv/s1	1 × 1 × 32 × 64	112 × 112 × 32
Conv dw/s2	3 × 3 × 64 dw	112 × 112 × 64
Conv/s1	1 × 1 × 1 × 64 × 128	56 × 56 × 64
Conv dw/s1	3 × 3 × 128 dw	56 × 56 × 128
Conv/s1	1 × 1 × 128 × 128	56 × 56 × 128
Conv dw/s2	3 × 3 × 128 dw	56 × 56 × 128
Conv/s1	1 × 1 × 128 × 256	28 × 28 × 128
Conv dw/s1	3 × 3 × 256 dw	28 × 28 × 256
Conv/s1	1 × 1 × 256 × 256	28 × 28 × 256
Conv dw/s2	3 × 3 × 256 dw	28 × 28 × 256
Conv/s1	1 × 1 × 256 × 512	14 × 14 × 256
5 × Conv dw/s1 Conv/s1	3 × 3 × 512 dw 1 × 1 × 512 × 512	14 × 14 × 512 14 × 14 × 512
Conv dw/s2	3 × 3 × 512 dw	14 × 14 × 512
Conv/s1	1 × 1 × 512 × 1024	7 × 7 × 512
Conv dw/s2	3 × 3 × 1024 dw	7 × 7 × 1024
Conv/s1	1 × 1 × 1024 × 1024	7 × 7 × 1024
Avg pool/s1	Pool 7 × 7	7 × 7 × 1024
FC/s1	1024 × 1000	1 × 1 × 1024
Softmax/s1	Classifier	1 × 1 × 1000

will apply different face detector models which are—DNN Caffe detector (based Caffe model or Single Shot detector), Haar cascade frontal face detector, and lastly MTCNN (multi-task cascaded rural network) face detector and then study the results emerging from these face detector models. We will use the same trained model made using MobileNet architecture and apply the three face detectors to analyze the difference. The further steps in methodology are divided into two sections—model training and detection of faces.

2.1 Model Training

For model building, we first used a dataset that contains around 1900 images each in a folder labeled with_mask and without_mask. The dataset used has been taken from Kaggle titled 'Face Mask Detection Dataset' [13]. We took around half the number of images that is 1900 images from both the labels present in our dataset to train and test our model on (Figs. 2 and 3).

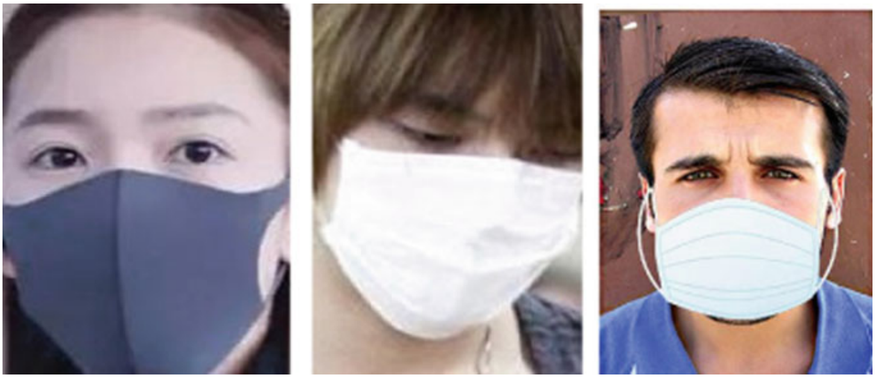


Fig. 2 Some images from the dataset under the with_mask label [13]

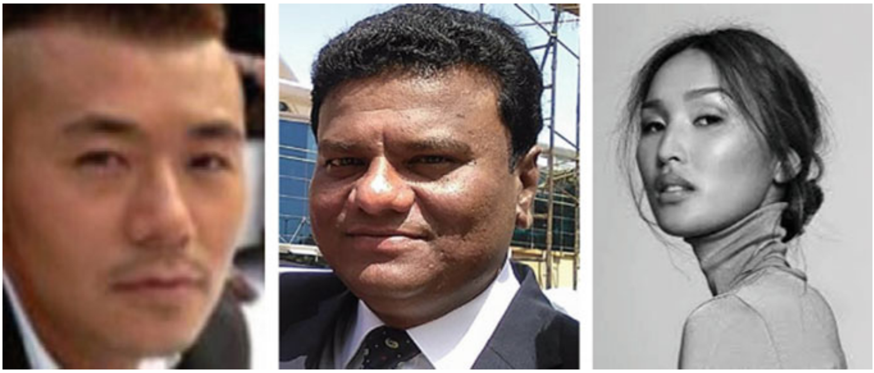


Fig. 3 Some images from the dataset under the without_mask label [13]

We then preprocess all the images in the folders and set the dimension of height and width as 224 and 224 (224 × 224) respectively to make our data more uniform and also it is the dimension accepted by MobileNet, and finally save the images in an array format from Keras.preprocessing.image module which is required when using MobileNet models, performed One-hot encoding using LabelBinarizer for the variables (labels) "with mask" and "without a mask".

ImageDataGenerator is used for data augmentation which creates more datasets for our model from the same dataset by slight alterations made to the same image. This is required as our initial dataset has only 1900 images for training for both categories, i.e. with_mask and without_mask, and because deep learning models perform better with a large number of the dataset.

Base Model and head models are then created which use the MobileNet V2 architecture for implementing CNN and use some pre-trained models and send the images via 3 channels (BGR). MobileNetV2 uses a convolution filter of dimensions 3 × 3 × 3 (sent across 3 channels) and has predefined and pretrained weights that are applied

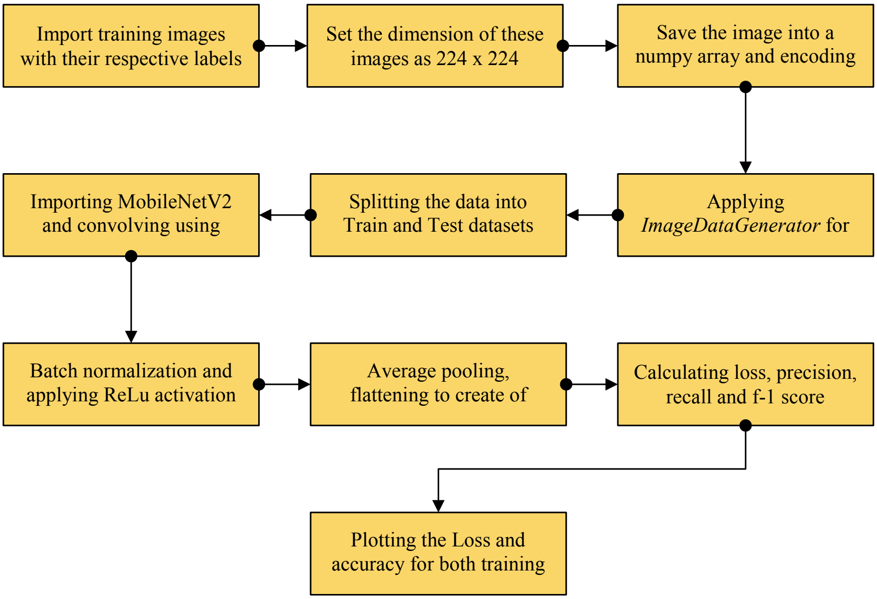

Import training images with their respective labels
Set the dimension of these images as 224 x 224
Save the image into a numpy array and encoding
Applying *ImageDataGenerator* for
Splitting the data into Train and Test datasets
Importing MobileNetV2 and convolving using
Batch normalization and applying ReLu activation
Average pooling, flattening to create of
Calculating loss, precision, recall and f-1 score
Plotting the Loss and accuracy for both training

Fig. 4 Flow diagram for the training of the face detection model

to the input at the various layers. The data from the input layer is convolved with the kernel and is standardized using ***batch normalization*** and pointwise convolved with 1×1 kernel.

Relu activation function was used as the activation function. Finally, average pooling is done before flattening and creating fully connected layers, ***softmax functions*** are used in the end to provide us a probability at the output. The loss is calculated using the binary_crossentropy loss function which is a logarithmic loss function. We then compile our model with parameters set as follows: Initial learning rate—0.0001, batch size = 32, epochs = 17, and save it with .h5 format. Fitted the model on our training set and then predicted on our test set and plotted the following curve on the same graph—training_loss, training_accuracy, validation_loss, and validation_accuracy and a table for precision and f-1 score. We tried three different values of epochs which are 17, 20, and 25 to find the best max number of epochs which will give us low loss and high accuracy for the model (Fig. 4).

2.2 Detection of Faces

For our face detection purpose as discussed above, we included three face detectors namely—DNN Caffe model face detector, the Haar cascade frontal face detector, and the MTCNN face detector. Open-cv library was used to load the face detector model and capture real-time video from the system.

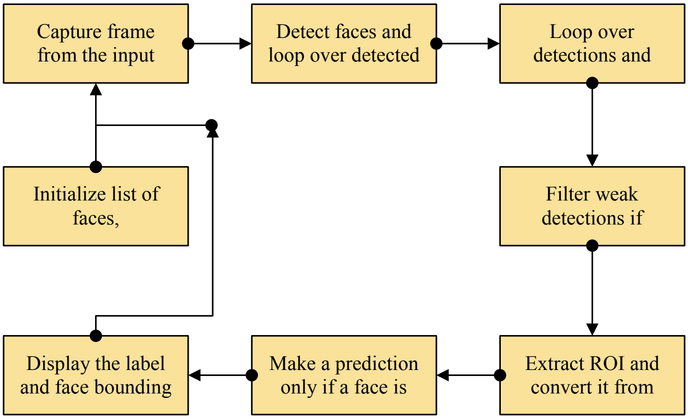

Capture frame from the input
Detect faces and loop over detected
Loop over detections and
Initialize list of faces,
Filter weak detections if
Display the label and face bounding
Make a prediction only if a face is
Extract ROI and convert it from

Fig. 5 Flowchart algorithm for detection of masked and unmasked faces

Region of interest (ROI) was extracted from the image frame and probabilities for each prediction made were included. A confidence score variable was included to compute how strong the predictions generated by the model were. Weak predictions were filtered out and only those predictions were made whenever a face was detected by including a confidence score. Box-boundary for face detection was color-coded that is green if a person is wearing a mask (with_mask) and red if a person is without a mask (without_mask). Frames per second (FPS) was also included in the frame to get a better idea about the efficiency of different models. Comparison and performance for each face detector were then noted and compared (Fig. 5).

3 Observation and Results

3.1 Training Time and Confusion Matrix Output

The model took around one hour and thirty minutes for training completion. Training time was calculated by aggregating the time taken for the training of the model for different numbers of epochs. Training time includes the time taken for calculating precision, recall, f-1 scores, and plotting the graphs regarding accuracy and loss.

System Hardware Details. Processor: Intel® Core™ i5-7200U (2.5 GHz base frequency, up to 3.1 GHz with Intel® Turbo Boost Technology, 3 MB cache, 2 cores), Video Graphics: AMD Radeon 520 (2 GB), Memory: 8 GB DDR4-2133 SDRAM (1 × 8 GB).

Confusion Matrix Output. Precision and Recall values help us to get information regarding our model's performance based on the closeness of predicted values with

Table 2 Precision, recall, and F-1 score for 17 epochs

	Precision	Recall	F-1 score
with_masks	0.99	0.79	0.88
without_masks	0.82	0.99	0.90

Table 3 Precision, recall, and F-1 score for 20 epochs

	Precision	Recall	F-1 score
with_masks	0.99	0.83	0.90
without_masks	0.85	0.99	0.92

Table 4 Precision, recall, and F-1 score for 25 epochs

	Precision	Recall	F-1 score
with_masks	0.99	0.83	0.89
without_masks	0.84	0.99	0.91

the actual values. The tables for Precision, Recall, and F-1 score value generated by the system for our model for various epochs.

Precision is defined as the total number of true positives produced divided by the sum of true positives and false positives.

Recall is defined as the total number of true positives divided by the sum of true positive and false negative predictions.

Tables 2, 3, and 4 show us the precision, recall, and f-1 score values for all the epoch test values that are 17, 20, and 25 respectively and we find that by setting epochs = 20 we get the highest and close values of precision and recall and subsequently high f-1 score.

3.2 *Loss and Accuracy Plots*

Loss and accuracy are used to obtain a graphical perspective of the overall effectiveness of the model concerning the number of epochs carried out. For each different max epoch, we fit the model on the training dataset and then make predictions over the testing/validation dataset. The difference between the values of those predicted and the ones which are actual values helps us to plot the graph. Loss and accuracy plots on both training and validation datasets are given for different numbers of epochs (Figs. 6, 7 and 8).

We found that for 20 epochs we achieve the maximum accuracy of the model. Occlusion tests were carried on different face detector models and the following was the result.

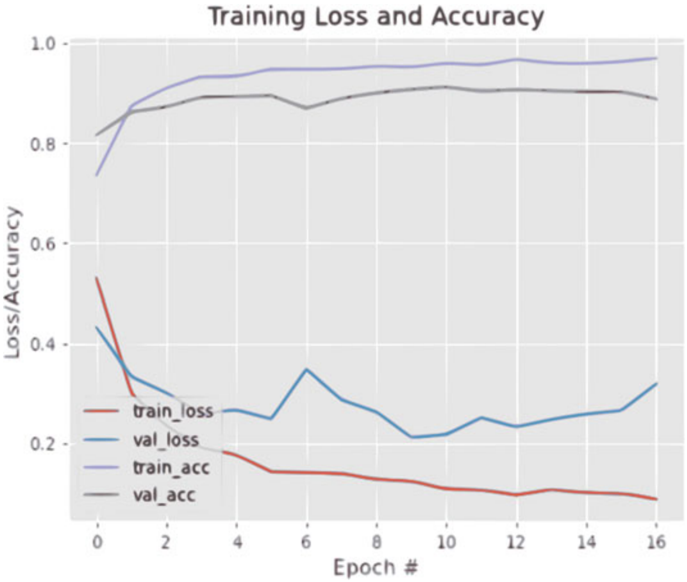


Fig. 6 Training loss, training accuracy, validation accuracy, and validation loss for 17 epochs

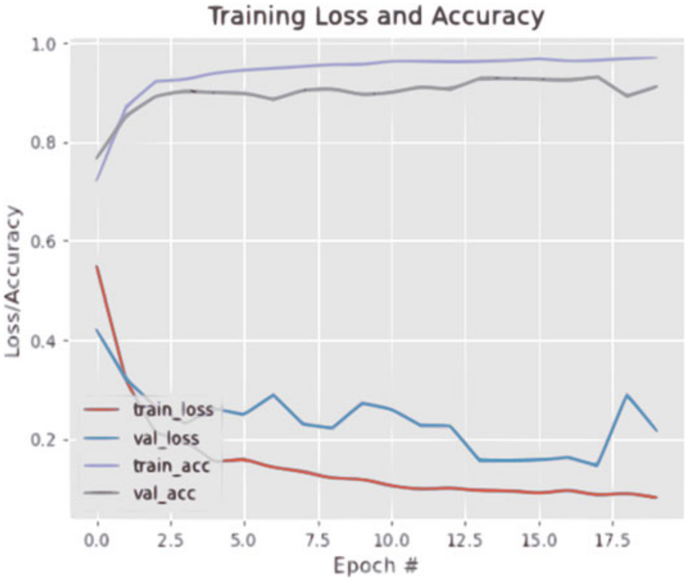


Fig. 7 Training loss, training accuracy, validation accuracy, and validation loss for 20 epochs

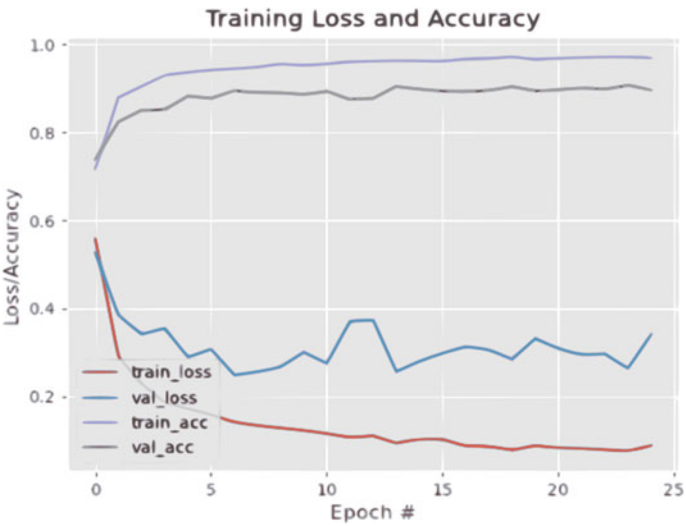


Fig. 8 Training loss, training accuracy, validation accuracy, and validation loss for 25 epochs

3.3 Outputs Using Different Face Detectors

Haar Cascade Outputs. Frontal faces are detected properly and quickly (Fig. 9).

Multi-Task Cascaded Neural Network Outputs. Frontal faces are detected accurately and easily (Fig. 10).

Caffe Model DNN Face Detector. Frontal faces are detected easily and have similar performance to MTCNN (Fig. 11).

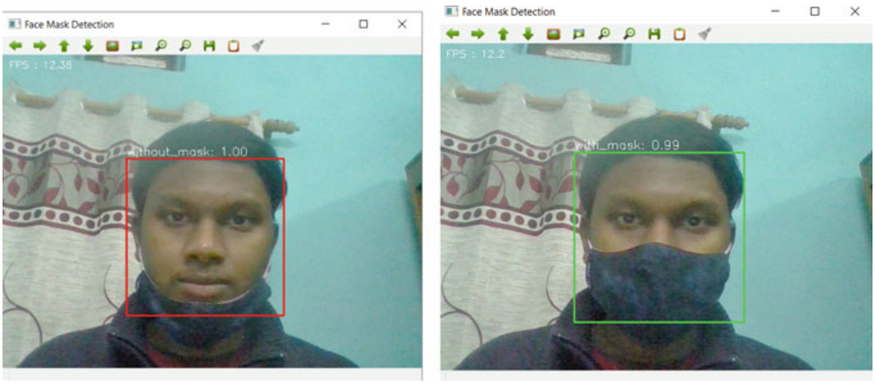


Fig. 9 Images of real-time face detection using haar cascade classifier without_masks and with_masks

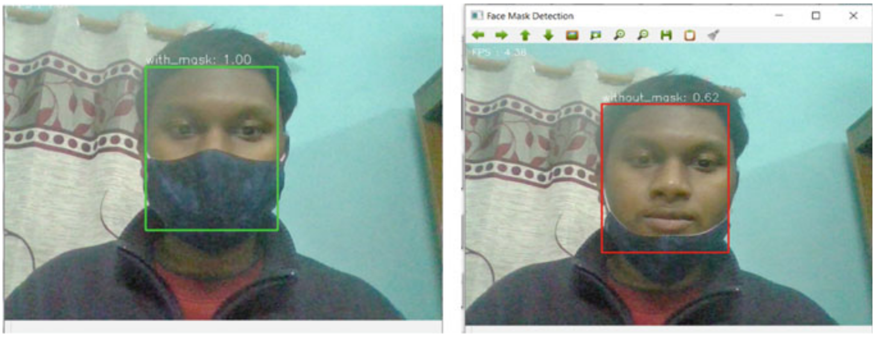


Fig. 10 Images of real-time face detection using MTCNN face detector without_masks and with_masks

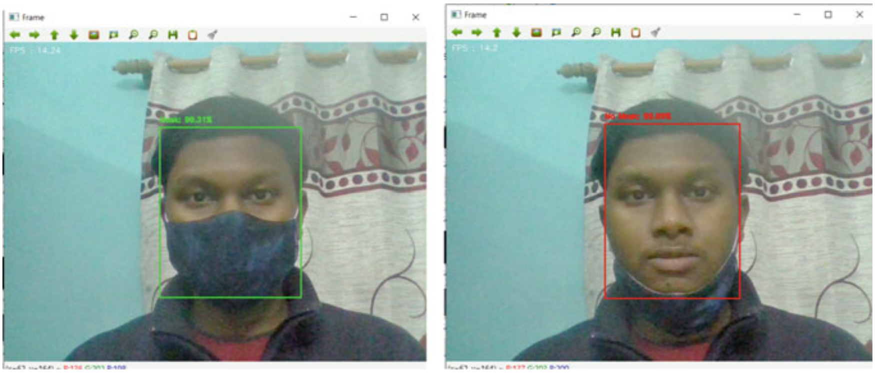


Fig. 11 Images of real-time face detection using Caffe model DNN face detector without_masks and with_masks

3.4 Frames Per Second (FPS)

FPS for each model was calculated and averaged after 30 min of usage for each model. We note the range of FPS for each face detector after we get stabilized FPS (Table 5).

Table 5 Average FPS range for each face detector using our model

Haar cascade frontal face detector	DNN (Caffe model) face detector	Multi-task cascaded neural network detector (MTCNN)
15–17	8.5–10	4.5–5.5

3.5 Side Face Views and Vertical Head Movements

Side face view detections were checked by carrying out head movements from left to right and vertical head movements were detected by moving the head up and down within the camera's range.

Haar Cascade Face Detector. Using this detector neither the side views nor the vertical head movements were tracked (Fig. 12).

Caffe Model Face Detector. The face detector was able to detect masked faces in side-views easily. Upward vertical face detection was not consistent however downward vertical face detection was detected exactly (Fig. 13).

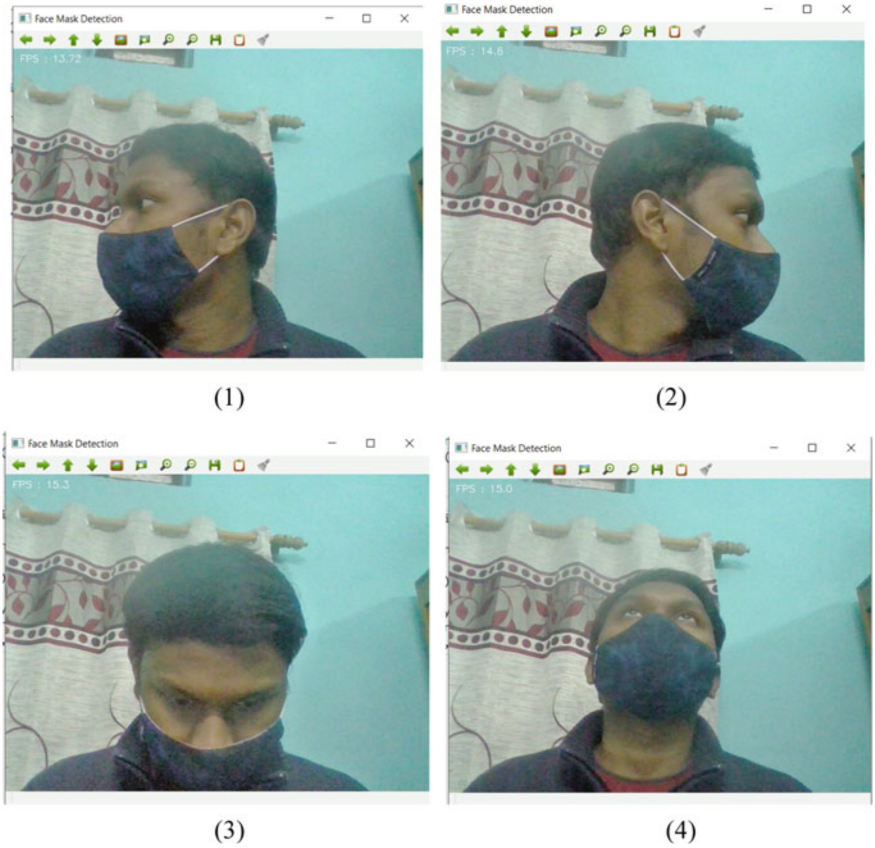


Fig. 12 Side face view mask detection (1), (2) and vertical head movement (3), (4) tracking using haar cascade detector

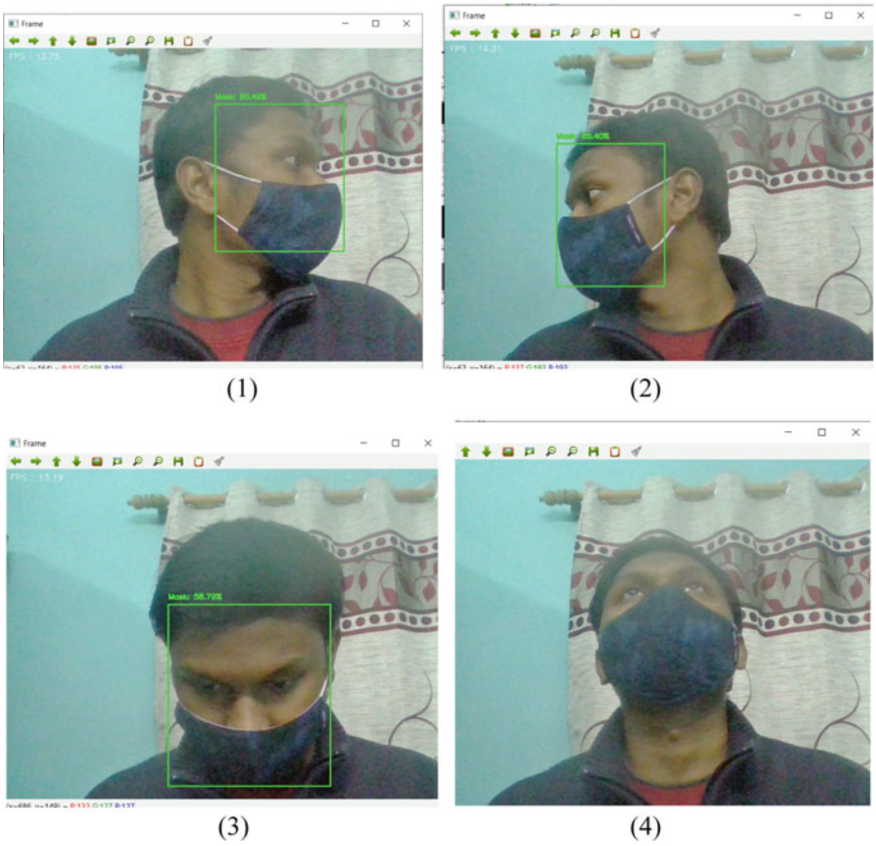
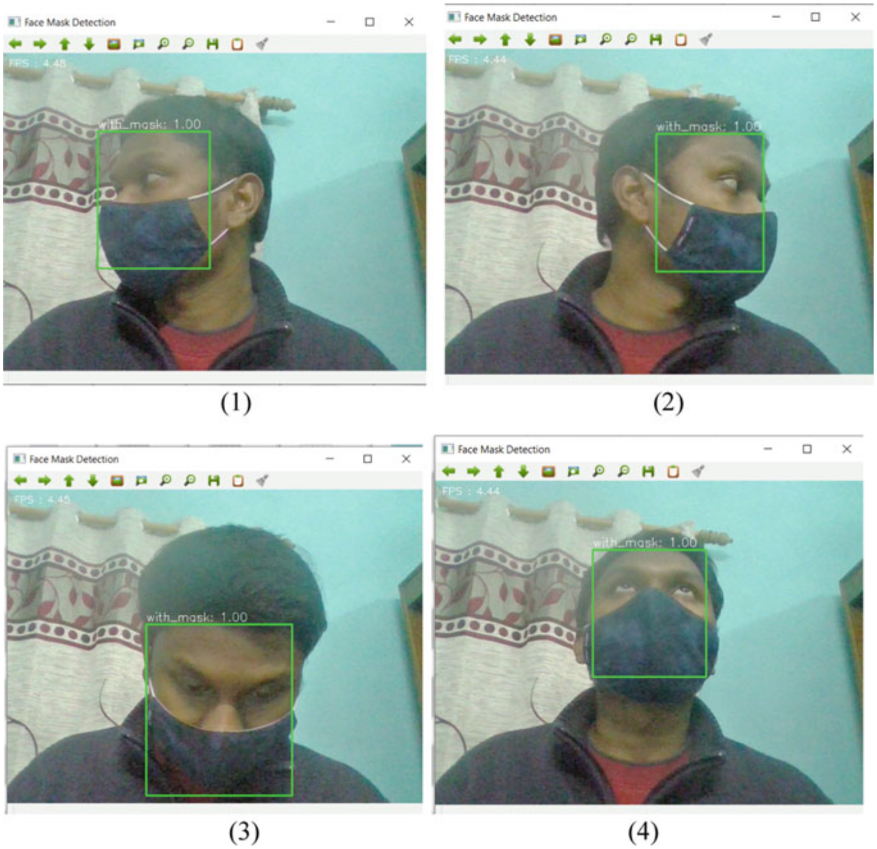


Fig. 13 Side face view mask detection (1), (2), and vertical head movement (3), (4) tracking using Caffe model face detector

MTCNN Face Detector. Using the MTCNN face detector we were able to get both side movement tracking and vertical head movement tracking much more accurately (Fig. 14).

3.6 *Comparing the Results Among All Three Face Detectors*

Comparing our mask detection model using different face detectors we conclude the following for model's performance using different face detectors—**Haar cascade frontal face detector**. This detector gives us the highest achievable FPS out of the three. Front face views were detected correctly however required good lighting conditions and little to no head movements for correct detection. For side face views



(1)

(2)

(3)

(4)

Fig. 14 Side face view mask detection (1), (2), and vertical head movement (3), (4) tracking using MTCNN face detector

and vertical head movements, no detections were made and those which were slightly detected gave false positives and false negatives.

This was likely to happen as haar cascades are the face detectors that are known to provide better speeds and track only a certain feature at any instant. They are least accurate and are much useful for purposes where objects appear on the frame at a certain orientation and have minimum occlusions.

MTCNN Face Detector. The MTCNN detector provides us with the most accurate detection. It works well in most lighting conditions and slight head movements don't affect the predictions. Side face views and vertical head movements were detected properly. However, the only drawback that our model faced was the number of frames that the system was able to process was the lowest among the three.

Caffe Model DNN Face Detector. The Caffe model face detector was accurate for most of the head movements however upward vertical movements were not detected properly. Side movements were detected by the detector and gave correct predictions for the same. FPS for this detector was between that of the MTCNN face detector and the haar cascade frontal face detector. Slight occlusions don't affect the performance of the mask detector and work well in most lighting conditions.

4 Conclusion

Our face detection model created using MobileNetV2 is a lightweight model as it has to process fewer parameters in comparison to the traditional convolutional network architectures. MobileNetV2 architectures are known to have better efficiency than other convolutional networks for training a machine learning model with only a small expense in performance. The face mask detector model built using MobileNetV2 is fast and does not require very high computational power for providing outcomes as it uses depthwise and pointwise convolution. Hence it can be implemented on low-end devices which don't possess good hardware power. Various public places, offices, institutions, etc. can use this technique for face mask detection. In addition to the above, we have compared the trained model's real-time performance in some common situations where face detection might become tricky and presented in this paper. Table 6 presents a comparison of how a particular model will behave when different face detectors are used for the detection of masks and faces. The effect of using each face detector on the trained model results in its own pros and cons hence we presented a detailed comparison of using each face detector, which will surely

Table 6 Tabular comparison of the performance of all the face detectors on our mask detection model

	Haar cascade face detector	MTCNN face detector	Caffe model face detector
FPS	15–17	4.5–5.5	8.5–10
Horizontal and vertical head movement detections	Neither vertical nor horizontal head movements were detected	All head movements were detected easily	Most head movements were detected easily however few face angles were not detected
Lighting conditions required	Requires well-lit surroundings	Requires medium lit to well-lit surroundings	Requires medium lit to well-lit surroundings
Occlusions effect	Affects the face mask detection most	Least affected by occlusions	Mostly unaffected by occlusions however few conditions might affect it
Accuracy	Least accurate face detector among the three	Most accurate detector among the three	Accuracy is less than MTCNN but much more than haar cascade

help in making choices for which face detectors to include while designing a specific project.

In addition to this, we intend to further increase the scope of mask detection by adding more facial tracking modules to this face mask detector. This will not only increase the accuracy of the actual face mask detection in real-time but also will make it faster.

References

1. Raj, S.J.: A comprehensive study on the computational intelligence techniques and its applications. J. ISMAC **1** (2019)
2. Dhaya, R.: CCTV surveillance for unprecedented violence and traffic monitoring. J. Innov. Image Process. (2020)
3. Hoque, M.A., Islam, T., Ahmed, T., Amin, A.: Autonomous face detection system from real-time video streaming for ensuring the intelligence security system. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, pp. 261–265 (2020)
4. Sreenu, G., Saleem, D.: Intelligent video surveillance: a review through deep learning techniques for crowd analysis (2019)
5. Ejaz, M.S., Islam, M.R.: Masked face recognition using convolutional neural network. In: 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, pp. 1–6 (2019)
6. Jiang, M., Yan, H., Fan, X.: Retina face mask: a face mask detector (2020)
7. Chowdhary, G.J., Punn, N.S., Sonbhadra, S.K., Agarwal, S.: Face mask detection using transfer learning of InceptionV3 (2020)
8. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. C.: MobilenetV2: Inverted Residuals and Linear Bottlenecks.4510–4520. https://doi.org/10.1109/CVPR.2018.00474. (2018)
9. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. IEEE Conf. Comput. Vis. Pattern Recognit. **1**, 1–511. https://doi.org/10.1109/CVPR.2001.990517. (2001)
10. Ghenescu, V., Mihaescu, R.E., Carata, S., Ghenescu, M.T., Barnoviciu, E., Chindea, M.: Face detection and recognition based on general purpose DNN object detector. In: 2018 International Symposium on Electronics and Telecommunications (ISETC), Timisoara, pp. 1–4 (2018)
11. Zhang, K., Zhang, Z., Li, Z.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10) (2016)
12. Zhang, N., Luo, J., Gao, W.: Research on face detection technology based on MTCNN. In: International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, pp. 154–158 (2020)
13. Face Mask Detection Dataset. The dataset is available at https://www.kaggle.com/omkargurav/face-mask-dataset

Image Encryption Using Diffusion and Confusion Properties of Chaotic Algorithm

J. N. Swaminathan, S. Umamaheshwari, O. Vignesh, P. Raziya Sulthana, A. Hima Bindu, M. Prasanna, and M. Sravani

Abstract Based on Chaos theory, cryptographic techniques show several new and successful ways to build reliable image encryption schemes. We present an image encryption in this paper using a logistic map of 1D. The proposed model framework is based on a main stream generator for the mechanism of uncertainty. A secret key of 256 bits, which is itself created by a logistic map, initiates the confusion process. In order to make the cipher more dynamic against any attack, after encrypting each block of the image, the secret key is changed. The experimental results show that the proposed approach offers an effective and safe way to encrypt and transfer images in real-time.

Keywords Image encryption · Diffusion · Confusion · Cryptographic · Logistic map · Secret key

1 Introduction

Society development leads to the significance of the data used, so huge amounts of digital visual data are stored every day on various media and exchanged over different types of networks [1]. In our lives, it has been used in digital form and is becoming increasingly necessary due to greater machine speed efficiency, media storage and network bandwidth [2]. Therefore, the weakness of this type of knowledge to be compared to the paper-based images, attacks such as alteration and manufacturing are greater [3]. In recent decades, information security has become a major issue in which new encryption algorithms based on algebraic methods or chaotic dimensions have been proposed [4]. The encryption technique involves two primary operations in this project, permutation at the pixel level and masking and permutation at the bit

J. N. Swaminathan (✉) · P. Raziya Sulthana · A. Hima Bindu · M. Prasanna · M. Sravani
QIS College of Engineering and Technology, Ongole, Andhra Pradesh 523272, India

S. Umamaheshwari
Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India

O. Vignesh
Easwari Engineering College, Chennai, Tamil Nadu, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_22

level [5]. Simulations show that the encryption technique proposed is efficient and has a high level of security [6].

For the Chaotic Digital Image Encryption implementation, MATLAB was used. For images of various sizes [7], this encryption technique was tested. But in the system, there is indeed a downside that it is only working with images contain the same number of pixels horizontally and vertically [8]. We recommend a chaotic framework of picture encryption and a key generator. The developed key can be used for the 1D logistics map as an initial condition [9]. The plain image is divided into block of specified number and each block is encrypted through the 1D logistic maps with different key obtained. The block size is known to be 8 bits. For different pictures, the experimental findings using the image database indicate the efficiency and intensity of the proposed chaotic image encryption. Finally, security analysis reveals that the proposed device is capable of generating useful for understanding that is statistically random [10]. The complexity of the problem, prior work, intention, and contribution of the paper should be described in the introduction of the paper [11]. It's indeed possible to include the contents of each section to easily comprehend the material.

1.1 Related Previous Works

The chaos-based image cryptosystem consists mainly of two phases. At its input, the plain representation is obtained [12]. In the chaos-based picture cryptosystem, there are two steps. The stage of complexity is the transposition of the pixels, in which location of the pixels is scrambled across the whole picture without affecting the pixel value and the image becomes unrecognizable [13]. A chaotic system performs the pixel permutation. The chaotic behavior, derived from the 16-character key, is guided by the initial conditions and specific resources [14].

The pixel values are changed in the diffusion process. The sequence is generated sequentially from one of the three chaotic processes of external key choices [15]. The whole confusion-spreading circular repetitions for a number of times to achieve a sufficient protection standard. Randomness of the property associated with chaotic maps makes it more fitting for Encryption of images.

2 Proposed Method

This Specific chaotic processes are used for uncertainty and diffusion Stages. Complex chaotic maps are often chosen rather than the simple maps. In order to increase overall complexity of the algorithm and therefore improve security. The cryptography input is a simple image that has to be encrypted. The Two stages consist of the cryptosystem (Fig. 1).

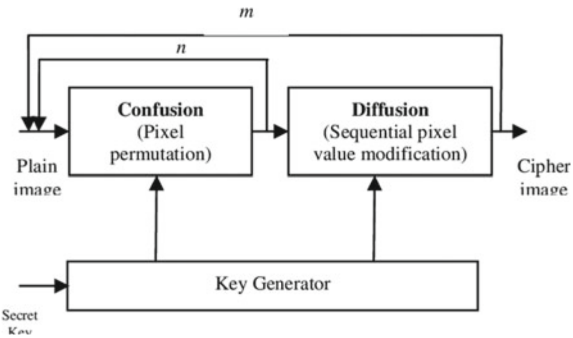


Fig. 1 Proposed architecture of chaotic-based system using confusion and diffusion property

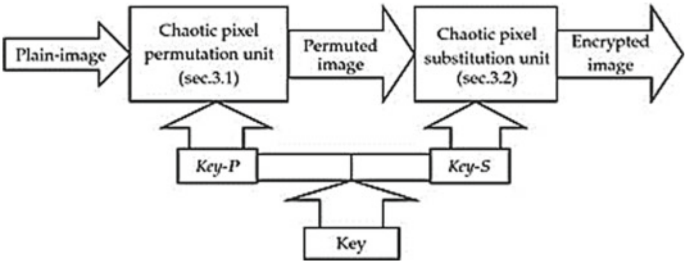

Plain-image
Chaotic pixel permutation unit (sec.3.1)
Permuted image
Chaotic pixel substitution unit (sec.3.2)
Encrypted image
Key-P
Key-S
Key

Fig. 2 Block diagram of the chaotic-based encryption

In addition, the plain image is masked with randomly generated mixing series in the recommended encryption technique and then the recursively performed diffusion method with the minimum required number of rounds (Rounds = 3) to achieve the maximum degree of security (Fig. 2).

During the permutation process, the mixed image will be transformed to binary sequence of the size $N \times M \times 8$. And then, given to the possible combination and distribution block. Integer values are generated to use them in between 0 and 255 levels. Only with generated random integer sequence, the basic input images are then masked. The chaotic algorithm is used in this system so that the created integer sequence is also chaotic. A simple and highly secure encryption is being attempted to encrypt digital images. Most recent studies have already shown that security level of uncertainty module is relatively poor. As the histogram of the shuffled image is completely unchanged, it is weak against several types of attacks, especially statistical attacks. The security of the cryptosystem is thus mainly based on the process of diffusion.

Algorithm Steps

Step 1: Integer stream mixing is provided using (1).

Step 2: Mix the plain image pixels with the produced MI series.

Step 3: The mixed image will be transformed to a binary image of 0s and 1s and the mixed image bits are shuffled using the Permutation Order.

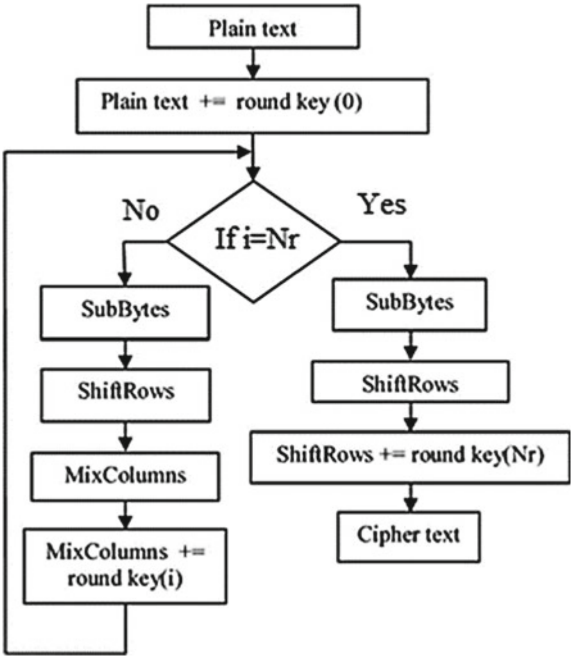

Plain text
Plain text += round key (0)
If i=Nr
No
Yes
SubBytes
ShiftRows
MixColumns
MixColumns += round key(i)
SubBytes
ShiftRows
ShiftRows += round key(Nr)
Cipher text

Fig. 3 Flow chart representation of chaotic algorithm

Step 4: With the random binary sequence (DFB), permuted picture pixel bits are diffused.
Step 5: Step 3 and Step 4 is repeated for (Rounds = 3) to attain the maximum degree of security (Fig. 3).

3 Result and Implementation

The current technique of image encryption uses the chaotic pixel location permutation method and one of the same chaotic mechanisms for pixel value changes Key space analysis, statistical analysis and sensitivity analysis was implemented in order to demonstrate the satisfactory security of the new system. The image histogram indicates that the pixels in an image are allocated by counting the number of pixels in each intensity color. It is recognized that the histogram of the original image and the process of misunderstanding are the same as, therefore, was carried out for that diffusion purpose. The final encrypted image histogram is relatively uniform and this varies greatly from that of the original image (Figs. 4, 5, 6 and 7).

In addition to the analysis of histograms, the relation is in between two pixels that are parallel vertical, two horizontal Pixels adjacent and two pixels adjacent diagonally in simple the image/cipher image is evaluated accordingly. In this, analysis of the correlation coefficient indicates the relation in the encrypted images, between pixels (Table 1).

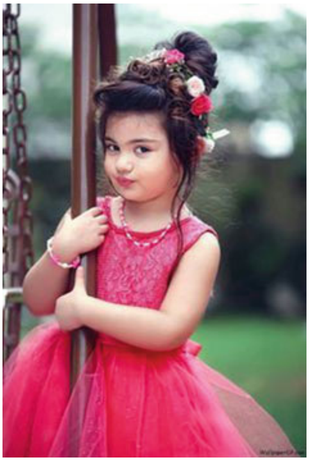


Fig. 4 Input image

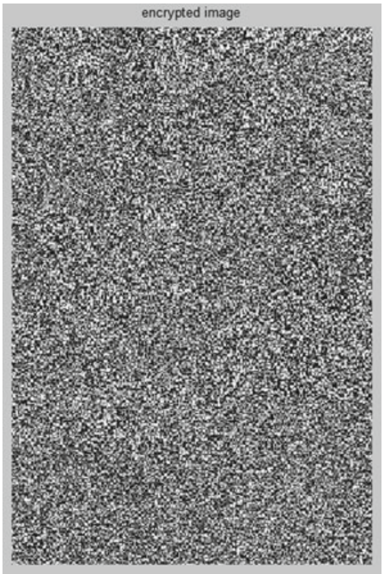


Fig. 5 Chaotic encrypted image

4 Conclusion

In encryption of image which based on a chaotic algorithm, a new method is proposed. The image pixels were replaced and the gray level values are used continuously are modified. The experimental experiments demonstrated that the proposed method based on chaotic shuffling and gray value adjustment is resistive through various attacks, such as cryptanalytic attacks, brute-force and mathematical attacks. The key

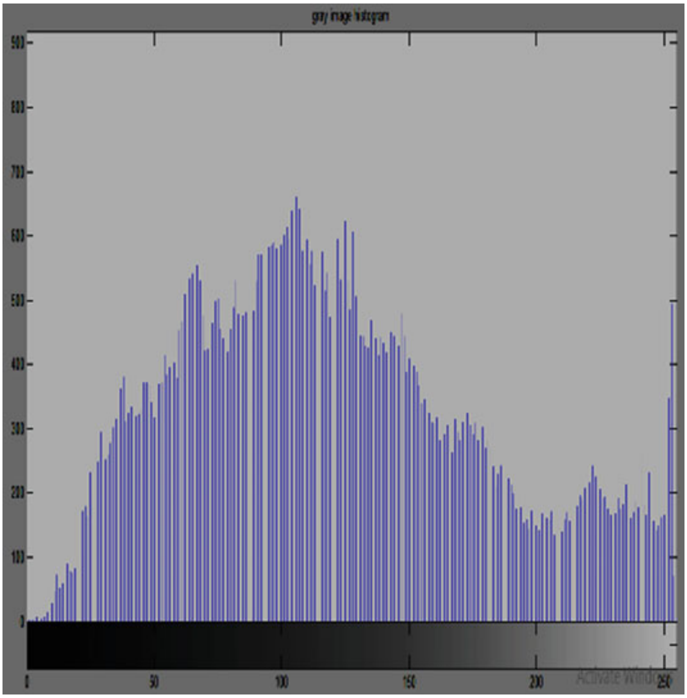


Fig. 6 Histogram-input image

space is increased by choosing a high-dimensional chaotic model. By choosing appropriate chaotic behaviors, complex non-linearity is preserved. Repeated permutations are avoided, but the diffusion method changes pixel values. The proposed cryptosystem avoids all the cryptographic weaknesses of earlier chaos-based encryption schemes by combining these all features.

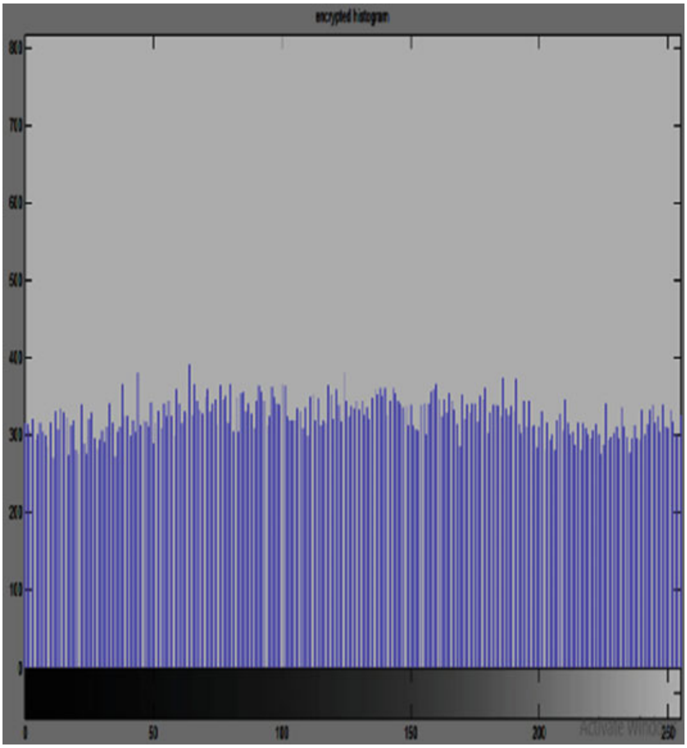


Fig. 7 Histogram-Chaotic encrypted image

Table 1 Entropy comparison

Methods	Entropy values
Proposed	8.9978
Wrong *k* [4]	7.9999

Acknowledgements The authors want to acknowledge Dr. N. S. Kalyan Chakravarthy, Chairman and Correspondent, QIS Group of Institutions, Ongole, Andhra Pradesh.

References

1. Kadir, A., Aili, M., Sattar, M.: Color image encryption scheme with the multiple impulse injection using a combined hyper chaotic method. Optik **129**, 231–238 (2017)
2. Wu, X., Zhu, B., Hu, Y.: Using rectangular transform-enhanced chaotic container maps, a novel color image encryption scheme. IEEE Access **5**, 6429–6436 (2017)
3. Murugan, C.A., Karthigaikumar, P.: Image encryption survey, bio-cryptography and the effective data encryption algorithms. Mob. Netw. Appl. **24**, 1–6 (2018)

4. Norouzi, B., Mirzakuchaki, S.: Breaking a new image encryption scheme based on improper chaotic fractional system model. Multimed. Tools Appl. **76**, 1817–1826 (2017)
5. Li, C., Lin, D., Feng, B.: Cryptanalysis of a chaotic entropy-based image encryption algorithm. IEEE Access **6**, 75834–75842 (2018)
6. Ye, G., Pan, C., Huang, X.: A chaotic entropy-based image encryption algorithm. Int. J. Build. Confus. **28**, 9 (2018)
7. Li, C., Feng, B., Li, S.: Dynamic analysis through the state-mapping channels of digital chaotic systems. IEEE Trans. Circuits Syst. I **66**, 2322–2335 (2019)
8. Rau, C.: Floating point library with the half-precision. http://half.sourceforge.net/index.html (2017). Accessed 16 May 2019
9. Joshua, C.D., Kamachi, M.G., Jain, L., et al.: Encryption technique based on the DWT for medical images. In: 2016 13th Wavelet Active Media Technology and Information Processing International Computer Conference (ICCWAMTIP), pp. 252–255 (2016)
10. Essaid, M., Akharraz, I., Saaidi, A.: A new controversy based image encryption scheme which uses the improved skewed tent plot. Procedia Comput. Sci. **127**, 539–548 (2018)
11. Shivkumar, S., Kavitha, A., Swaminathan, J., Navaneethakrishnan, R.: General self-organizing tree-based energy balance routing protocol with clustering for wireless sensor network. Asian J. Inform. Technol. **15**(24), 5067–5074 (2016)
12. Umamaheshwari, S., Swaminathan, J.N.: Man-in-middle attack/for a free scale topology. In: IEEE ICCCI, pp. 1–4 (2018)
13. Swaminathan, J.N., Kumar, P.: Design of stego-linearizer in HPA linearization. In: Proceedings of 6th IEEE International Conference on Advanced Computing (IACC'16), pp. 1–3 (2016)
14. Umamaheswari, S.: Capsule network-based data pruning in wireless sensor networks. Int. J. Commun. Syst. **33**, e4145 (2020)
15. Umamaheswari, S.: Performance analysis of wireless sensor networks assisted by on-demand-based cloud infrastructure. Int. J. Commun. Syst. **33**, e4272 (2020)

A Sentiment Analysis of a Boycott Movement on Twitter

Sooraj Bhooshan, R. Praveen Pai, and R. Nandakumar

Abstract Sentiment analysis refers to determining emotional content from a textual input. The Internet world is now run by the term “Web 2.0,” and one of the major platforms that made this web to web 2.0 is Twitter, a microblogging social network where one can post short messages known as tweets. The general public uses Twitter social networking platform very often to post opinions on various topics ranging from reviews to current affairs. And one of the trending topics in India in the year 2020 was the boycotting online shopping services such as Amazon and Flipkart due to issues related to political clash between India and China and also issues related to nepotism in Bollywood. The border issue between India and China was a big talking point as it involved clashes between the Chinese and Indian soldiers which resulted in few martyrs. This made a major public anger toward China and its products which then resulted into the boycott movement. The other talking point was the nepotism in Bollywood. The recent suicide of Sushant Singh Rajput has brought out the public sadness and blamed Bollywood nepotism as the result of his suicide. This resulted in public boycotting products that were endorsed by actors/actresses that were having a favoritism which was granted by their relatives in the Bollywood industry. Hence, we decided that it was a great opportunity to extract tweets related to this boycott movement and do sentiment analysis on it to determine various emotions expressed by the general public.

Keywords Online boycott movement · Sentimental analysis · Tweepy · Twitter · Data mining

S. Bhooshan (✉) · R. Praveen Pai · R. Nandakumar
Department of Computer Science and IT, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

R. Nandakumar
e-mail: nandakumar@asas.kh.amrita.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_23

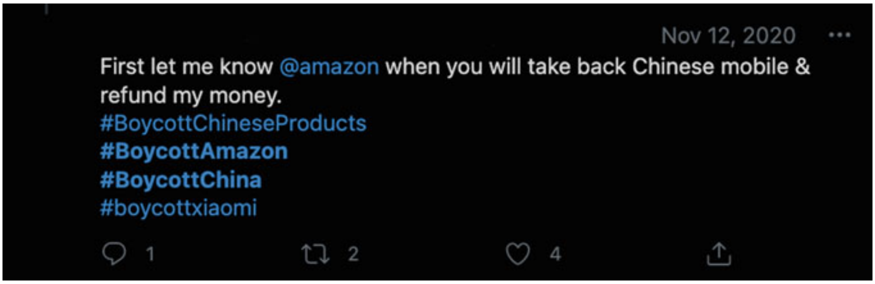
1 Introduction

Sentiment analysis is a text analysis method that uses natural language processing (NLP) to broadly classify a textual input to different emotions (mainly positive, negative and neutral). It is a popular research area where researchers try to extract sentiments from generally sensitive topics, surveys, reviews and even about the research area of sentiment analysis itself [1]. Nowadays, the general public plays an important role in deciding how to run online services. Businesses no longer have a free hand in the market; it is the customer's opinion that is now increasingly important. Opinions posted on social media act as a valuable source of information to business. These opinions can be used to learn how to plan business strategies and engagement activity so that customers are fully satisfied. And one of the most popular social media platforms to express these opinions is Twitter.

Twitter is a microblogging platform or in simple terms, a social networking platform where people post their opinions in the form of short messages known as tweets. Tweets have many unique characteristics such as message length, writing technique, emoticons and special characters such as hashtags (#) and mention (@) [2]. The tweets are mostly opinions on different subjects ranging from reviews, surveys, or even niche and specific topic areas such as apparels [3]. Since these are opinions expressed by the public, it acts as a valuable source of data to extract and do sentiment analysis.

One of the trending topics in the year 2020 on Twitter was the boycott of online e-commerce services such as Amazon and Flipkart due to the China–India border clash and also due to the issue related to nepotism in Bollywood that was brought by the recent suicide of Sushant Singh Rajput. One of the examples of the boycott movement tweet is shown in Fig. 1.

The main objective of this study is to make use of the tweets pertaining to this topic by extracting it and doing sentiment analysis to determine the emotions expressed by the general public. The sentimental analysis can be done in different ways, either by lexicon-based approach or by machine learning approach or you can use both together as a hybrid approach [4]. In this paper, we will be using a lexicon-based approach using various Python packages to do the sentiment analysis.



Nov 12, 2020

First let me know @amazon when you will take back Chinese mobile & refund my money.
#BoycottChineseProducts
#BoycottAmazon
#BoycottChina
#boycottxiaomi

Fig. 1 Example of a tweet that mentions the boycott movement

The paper will have the following sections: Sect. 2 will cover the previous research works related to this topic. Section 3 will cover the proposed system where we explain how the tweets are extracted, cleaned and analyzed, and in Sect. 4, we will examine the final results, and the conclusions derived from it are mentioned in Sect. 5.

2 Literature Review

For many years, sentiment analysis has been widely used as a research area—the data is easily available and is a source of valuable information that can be put to various strategic uses and can also be used for various applications such as business intelligence, smart homes and website reviews [5]. Below are the research works which we went through to learn more about sentiment analysis and its working in different domains.

In Suzanne C. Makarem and Haeran Jae's research on consumer boycott behavior using Twitter data [6], they performed content analysis on Twitter feeds and used human sentiment analysis to examine the relationship between the motives behind boycotts and the emotional intensity of boycott messages [6]. Their main objective was to identify different themes that arise from Twitter and the motivation behind it.

Devang Jhaveri, Aunsh Chaudhari, Lakshmi Kurup performed sentiment analysis on major e-commerce websites in India (Flipkart and Snapdeal) through data collected from Twitter. The paper discusses various lexicon-based approaches that are applied to the Twitter data and the accuracy derived from it. The main drawback in their research paper was the involvement of advertisement tweets on their data which resulted in more positive effect hence failing to elicit the accurate picture [7].

Sudarshan Sirsat, Sujata Rao, Bharti Wukkadada did sentiment analysis on Twitter data for product evaluation [8]. The paper discusses how Twitter data can be harnessed to review product performance and its success. In this work, the authors created a sentiment classifier with the help of Tweepy and TextBlob to classify tweets into positive and negative.

Hamid Bagheri and Md Johirul Islam [9] used the similar approach mentioned in Sirsat et al. research work [8]. They used a wealth of already existing libraries to do the sentimental analysis. The authors worked on different queries including movies, politics, fashion and fake news and finally showed the polarity of each as the results. Major drawback mentioned in the paper was that the neutral score was very high which concludes that there is a need of improvement in the sentiment analysis.

Another related work is the Twitter sentiment analysis by Faizan [10]. In this paper, the author used a machine learning approach and created a model for the analysis of feelings using the KNN algorithm with unigram, bigram and n-gram features [10]. The model was used on tweet dataset that included #USairline. This work makes use of the TextBlob package to automatically set a target for each tweet. Another similar and related study approves that the unigram model helps to achieve a baseline target for sentiment analysis [11].

Dr. U Ravi Babu researched on sentimental analysis of reviews for E-Shopping websites [12]. In this work, the author does not make use of the Twitter data but instead uses the comments published on the respective websites. The author used different preprocessing techniques such as stemming, tokenization and URL and used POS tagging to do sentiment analysis. This work mainly focused on finding out which website was the best by taking the comments into consideration.

A Twitter sentiment analysis done by Sarlan et al. involves a deep analysis on Twitter data [13]. Their work mentions two approaches for extracting sentiment automatically: a lexicon-based approach and machine learning-based approach. The main intention of this paper was to study sentiment analysis of tweets and also develop a python program to collect, analyze and show sentiment results in a pie chart. The drawback seen in their work is that the pie chart represented more neutral sentiment. This may be due to inaccurate analysis of the tweets. They also failed to create the program due to technical limitations.

3 Proposed System

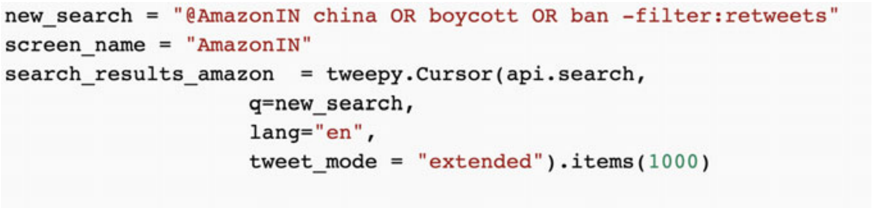
Broadly, the three steps of sentiment analysis are as follows:

1. Data extraction.
2. Data preprocessing.
3. Sentiment analysis.

3.1 Data Extraction

The first procedure is the data extraction. Data extraction is the process of retrieving data from data sources with the intention of further processing or storage [14]. There are various ways to gather data from Twitter. You can scrape Twitter for data which is an unconventional way, or you can make use of APIs available for fetching tweets. With Twitter API, you can either fetch real-time data by giving a hashtag or username [15] or you can fetch tweets that were posted long ago by giving required inputs. Here, we have used Tweepy, a user-friendly Python library for accessing Twitter API [16] for extracting data that we need.

By using Tweepy, we were able to fetch data related to the boycott movement by feeding search terms such as “@AmazonIN china OR boycott OR ban -filter:retweets” and “@Flipkart china OR boycott OR ban -filter:retweets”. The -filter:retweets will avoid all the retweets or replies. Retweets are filtered out because it may contain replies that are advertisement or bot messages and will result in inaccurate data. Using Tweepy, we collected around 1000 tweets for our dataset. The code (Fig. 2) shows how we extracted tweets related to the boycott movement.



```
new_search = "@AmazonIN china OR boycott OR ban -filter:retweets"
screen_name = "AmazonIN"
search_results_amazon  = tweepy.Cursor(api.search,
                   q=new_search,
                   lang="en",
                   tweet_mode = "extended").items(1000)
```

Fig. 2 Code that is used to extract tweets related to the boycott movement

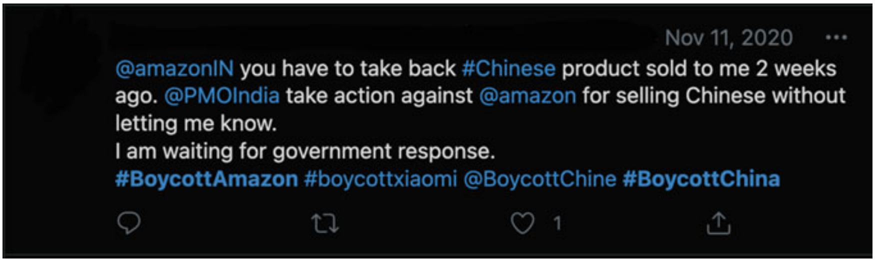
3.2 *Data Preprocessing*

The next step after data extraction is data preprocessing. Tweets are usually very noisy, and it contains a lot of junk data. An example tweet is given in Fig. 3.

The example shows that a tweet can contain hashtags (#), links, mentions (@) and other special characters. Therefore, it is necessary to clean the tweets before sending it to analysis and that is where data preprocessing is used. For data preprocessing, we used the following procedures:

- Created a function that uses Python regex library to remove mentions, hashtags and other special characters.
- Tokenization: It is the process of extracting bags of cleaner terms from raw comments by deleting the stop words. We used the nltk.tokenize package from NLTK library [17] to remove stop words from the tweets. The stop words are imported from nltk.corpus package.
- Lemmatization: Lemmatization usually means proper application of a vocabulary and the morphological analysis of words, usually with a view to removing only inflectional endings and to retrieve the base or dictionary form of a word, which is called the lemma [18]. We used the WordNetLemmatizer module from nltk.stem package to lemmatize our tweets and further clean our data and bring more meaning to it.

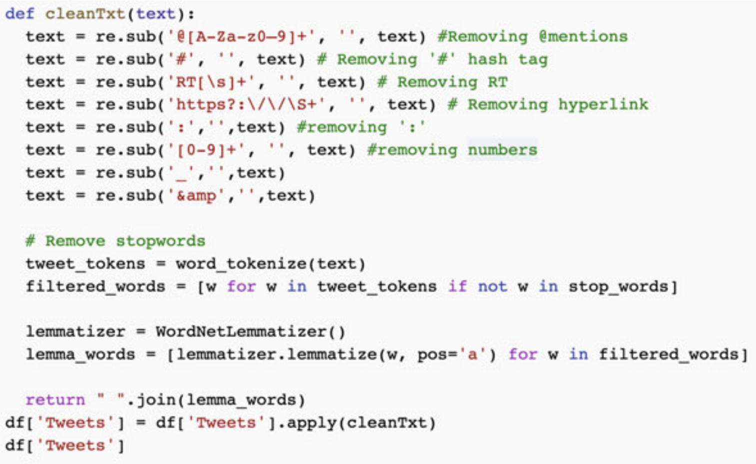
Figure 4 shows the function that is used for data preprocessing.



Nov 11, 2020 ···

@amazonIN you have to take back #Chinese product sold to me 2 weeks ago. @PMOIndia take action against @amazon for selling Chinese without letting me know.
I am waiting for government response.
#BoycottAmazon #boycottxiaomi @BoycottChine #BoycottChina

Fig. 3 Example of a noisy tweet

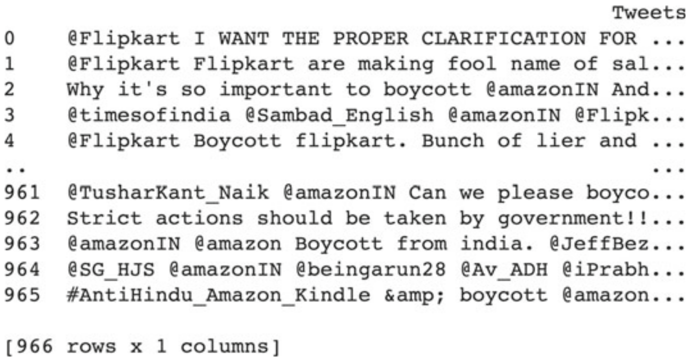


```
def cleanTxt(text):
  text = re.sub('@[A-Za-z0–9]+', '', text) #Removing @mentions
  text = re.sub('#', '', text) # Removing '#' hash tag
  text = re.sub('RT[\s]+', '', text) # Removing RT
  text = re.sub('https?:\/\/\S+', '', text) # Removing hyperlink
  text = re.sub(':','',text) #removing ':'
  text = re.sub('[0-9]+', '', text) #removing numbers
  text = re.sub('_','',text)
  text = re.sub('&amp','',text)

  # Remove stopwords
  tweet_tokens = word_tokenize(text)
  filtered_words = [w for w in tweet_tokens if not w in stop_words]

  lemmatizer = WordNetLemmatizer()
  lemma_words = [lemmatizer.lemmatize(w, pos='a') for w in filtered_words]

  return " ".join(lemma_words)
df['Tweets'] = df['Tweets'].apply(cleanTxt)
df['Tweets']
```

Fig. 4 Data preprocessing function

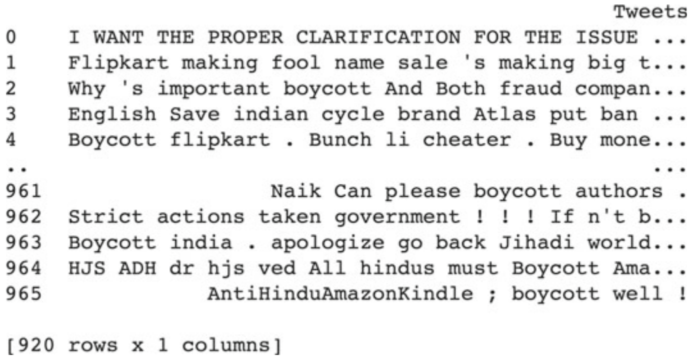
The above three steps have helped us clean the data to an extent but there were duplicate tweets inside our dataset. We used the drop_duplicates() function of the Python data frame to delete all the duplicate tweets. Now, the data is fully cleaned and ready for the analysis stage. Figure 5 shows the tweet dataset before the cleaning process, and Fig. 6 shows the dataset after cleaning.

By comparing Figs. 5 and 6, we can notice that the tweets are now cleaned, and most of the special characters are removed, and the dataset is now reduced from 966 to 920 rows as the duplicate tweets were removed by using drop_duplicates() function. Now, the data is clean and is now ready for sentiment analysis.



```
                                                Tweets
0     @Flipkart I WANT THE PROPER CLARIFICATION FOR ...
1     @Flipkart Flipkart are making fool name of sal...
2     Why it's so important to boycott @amazonIN And...
3     @timesofindia @Sambad_English @amazonIN @Flipk...
4     @Flipkart Boycott flipkart. Bunch of lier and ...
..                                                 ...
961   @TusharKant_Naik @amazonIN Can we please boyco...
962   Strict actions should be taken by government!!...
963   @amazonIN @amazon Boycott from india. @JeffBez...
964   @SG_HJS @amazonIN @beingarun28 @Av_ADH @iPrabh...
965   #AntiHindu_Amazon_Kindle &amp; boycott @amazon...

[966 rows x 1 columns]
```

Fig. 5 Tweets before cleaning



```
                                                  Tweets
0      I WANT THE PROPER CLARIFICATION FOR THE ISSUE ...
1      Flipkart making fool name sale 's making big t...
2      Why 's important boycott And Both fraud compan...
3      English Save indian cycle brand Atlas put ban ...
4      Boycott flipkart . Bunch li cheater . Buy mone...
..                                                   ...
961                     Naik Can please boycott authors .
962    Strict actions taken government ! ! ! If n't b...
963    Boycott india . apologize go back Jihadi world...
964    HJS ADH dr hjs ved All hindus must Boycott Ama...
965                AntiHinduAmazonKindle ; boycott well !

[920 rows x 1 columns]
```

Fig. 6 Tweets after cleaning

Table 1 Table showing the sentiment score of a cleaned tweet

Cleaned tweet	Subjectivity	Polarity
They make hype first customer buys products flash sale cancel orders. Worst company. Ban fake companies	0.7777777778	−0.4166666667

3.3 Sentiment Analysis

The next step is to take the preprocessed data for sentiment analysis. There are various tools that are available to do sentiment analysis like SentiStrength, Semantria and social mentions [19]. For our research, we took advantage of the Python library called TextBlob [20]. TextBlob is a Python library that is widely used for performing basic NLP tasks. These tasks include POS tagging, noun phrase extraction, translation and mainly sentiment analysis. We ran our dataset through TextBlob's sentiment property to return the polarity and subjectivity. An example of this based on the real Twitter data is given in Table 1.

After this, we created a simple function that takes the polarity of the tweet to check whether the tweet is having positive, negative, or neutral sentiment. The logic applied is as follows:

```
if polarity < 0:
    return 'Negative'.
elif polarity == 0:
    return 'Neutral'.
else:
    return 'Positive'.
```

Then, we counted the total number of positive, negative and neutral tweets and also plotted this score in a graph. This is shown in Table 2 and Fig. 7, respectively. Table 1 also shows that one of the cleaned tweets returned a negative polarity score which means that the tweet is expressing a negative emotion.

Table 2 Total counts of positive, negative and neutral tweets

Sentiment	Count
Neutral	444
Positive	270
Negative	206

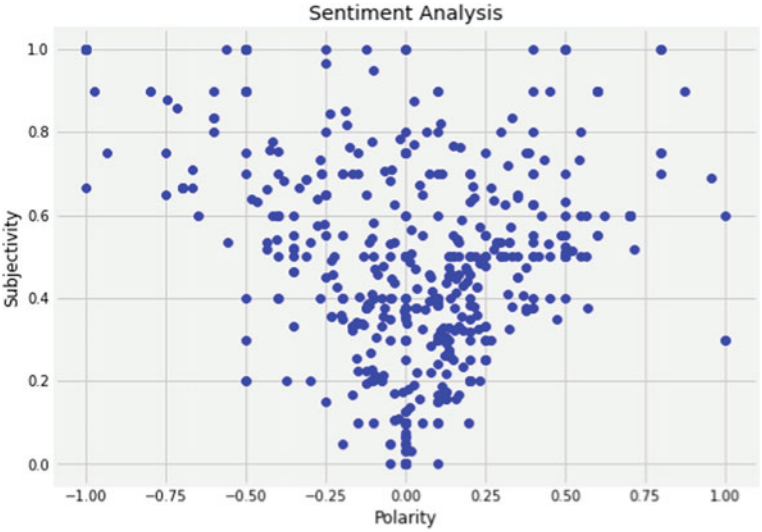


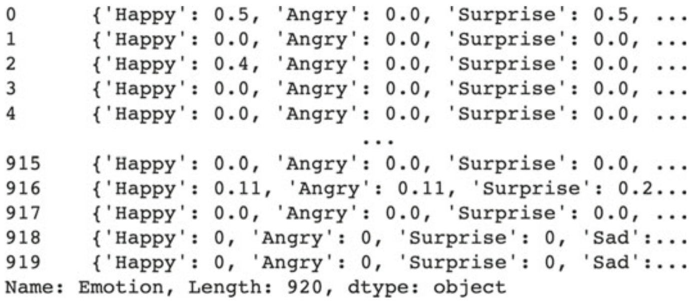
Fig. 7 Plotting the score of sentiment analysis using TextBlob

The sentiment analysis using the TextBlob package shows that there are a lot of neutral and positive tweets. This was a common disadvantage/challenge faced by authors of the research works mentioned in the literature review part. So, to tackle this challenge, we decided to check if we can derive more meaning from these neutral and positive tweets and for that we used a python package called text2emotion [21].

Text2emotion is a Python package which will help to process any textual message, then recognizes the emotion embedded in it and then gives out the output in the form of a dictionary. It is compatible with five different emotion categories: happy, angry, sad, surprise and fear [21].

We created a Python function that uses text2emotion package to return the emotions embedded in each tweet and got the list of dictionaries. Below is the code implemented to return the emotion dictionary, and Fig. 8 shows the dictionary returned by it:

```
def getEmotion(text):
  return te.get_emotion(text)
df['Emotion'] = df['Tweets'].apply(getEmotion)
print(df['Emotion'])
```



```
0      {'Happy': 0.5, 'Angry': 0.0, 'Surprise': 0.5, ...
1      {'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.0, ...
2      {'Happy': 0.4, 'Angry': 0.0, 'Surprise': 0.0, ...
3      {'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.0, ...
4      {'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.0, ...
                                   ...
915    {'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.0, ...
916    {'Happy': 0.11, 'Angry': 0.11, 'Surprise': 0.2...
917    {'Happy': 0.0, 'Angry': 0.0, 'Surprise': 0.0, ...
918    {'Happy': 0, 'Angry': 0, 'Surprise': 0, 'Sad':...
919    {'Happy': 0, 'Angry': 0, 'Surprise': 0, 'Sad':...
Name: Emotion, Length: 920, dtype: object
```

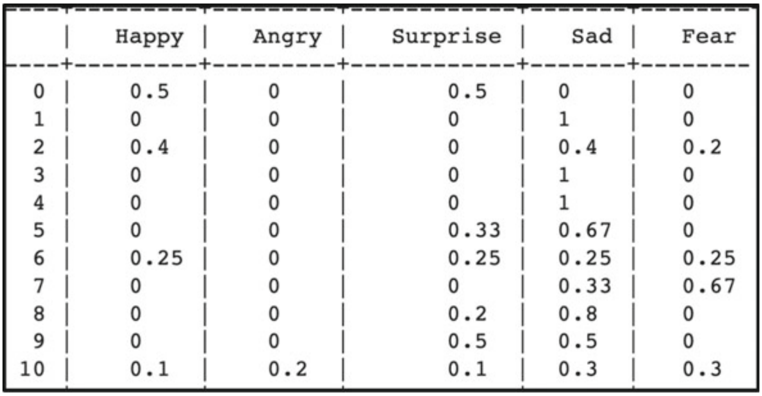
Fig. 8 List of dictionaries returned by text2emotion

Since the output returned by the text2emotion package was a list of dictionaries, we had to convert it into a tabular format for easy analysis. The below code is used to convert the list of dictionaries into a tabular format, and Fig. 9 shows the output returned by it. The figure provided only shows the first 10 rows due to space concerns. The original tabular format returned 920 rows.

```
data = []
for i in df["Emotion"]:
  data.append(i)
dfItem = pd.DataFrame.from_records(data)
print(tabulate(dfItem, headers='keys', tablefmt='psql'))
```

After creating the tabular format, we summed up the emotion score of each tweet and generated a table that shows each emotion, and the total number of tweets that are associated with that emotion. It is shown in Table 3.

The graphical representation of Table 3 is given in Fig. 10.



```
----+---------+---------+------------+-------+--------
    |  Happy  |  Angry  |   Surprise |  Sad  |   Fear
----+---------+---------+------------+-------+--------
  0 |  0.5    |  0      |       0.5  | 0     | 0
  1 |  0      |  0      |       0    | 1     | 0
  2 |  0.4    |  0      |       0    | 0.4   | 0.2
  3 |  0      |  0      |       0    | 1     | 0
  4 |  0      |  0      |       0    | 1     | 0
  5 |  0      |  0      |       0.33 | 0.67  | 0
  6 |  0.25   |  0      |       0.25 | 0.25  | 0.25
  7 |  0      |  0      |       0    | 0.33  | 0.67
  8 |  0      |  0      |       0.2  | 0.8   | 0
  9 |  0      |  0      |       0.5  | 0.5   | 0
 10 |  0.1    |  0.2    |       0.1  | 0.3   | 0.3
```

Fig. 9 Tabular format representing the emotion score of each tweet (920 tweets)

Table 3 Table showing total number of tweets associated with each emotion

Emotion	Tweets
Happy	87
Angry	31
Sad	305
Surprise	149
Fear	208
No emotion	140

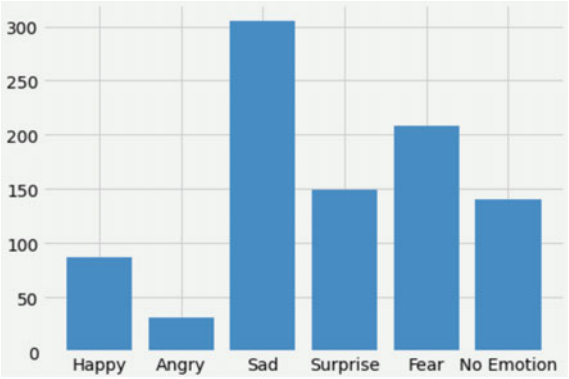


Fig. 10 Graphical representation of Table 3

4 Result Analysis

In Table 3, we can see that the count of neutral tweets has been immensely reduced to 144 from the initial 444 tweets. With the text2emotion package, we were able to successfully derive more meaning from the neutral and positive tweets. We can see that there are 87 tweets that express happiness, 31 tweets that express anger, 305 that express sadness, 149 tweets that express surprise emotion, 208 tweets that express fear and 140 tweets with no emotion. By analyzing Table 3, we can clearly state that the general public is expressing sadness through their tweets.

The cleaned Twitter data included a lot of words that expressed sadness and fear. Words like “cheating,” “liar,” “bad” and “worst” were extensively used in the tweets. The count of anger emotion surprisingly remained very low owing to the presence of sad words in text2emotion package. This means that even though the user was expressing anger in the tweets, they were actually using sad words rather than words that portray anger.

By looking at the results, we can state that the tweets made by the general public regarding the boycott movement have a negative sentiment overall as the tweets express sadness and fear emotions more than the positive emotions.

5 Conclusion

In this paper, we discussed Twitter sentiment analysis on the topic "Online Shopping Boycott Movement" that was brought out due to issues such as the India–China border clash and the Nepotism issue in Bollywood. We used the help of Tweepy, NLTK, TextBlob and text2emotion packages to collect, preprocess and analyze the data. We found out different emotions that the users expressed through their tweets and the overall sentiment of the general public regarding this topic. We conclude that the analysis shows a negative sentiment overall on the tweets that were extracted. This study also helped us to tackle the challenge faced by many authors regarding the issue of having a lot of neutral tweets after their analysis by using the text2emotion Python package.

The Twitter data we extracted contained few tweets that were expressed in a mix of two languages, i.e., Hindi words written out in English. This resulted in some tweets returning no emotions during the sentiment analysis. Another limitation we noticed is that the text2emotion package does not include a large bag-of-words that resulted in some words to be ignored during the emotion analysis. For future work, researchers can try to create models to handle a mix of two languages and find a larger bag-of-words to do the analysis and derive a better meaning out of this.

References

1. Mäntylä, M.V., Graziotin, D., Kuutila, M.: The Evolution of Sentiment Analysis—A Review of Research Topics, Venues, and Top Cited Papers. M3S, ITEE, University of Oulu
2. Kumar, A., Sebastian, T.M.: Sentiment analysis on Twitter. IJCSI Int. J. Comput. Sci. Issues **9**(4), No 3 (2012). ISSN (Online): 1694-0814
3. Rasool, A., Tao, R., Marjan, K., Naveed, T.: Twitter sentiment analysis: a case study for apparel brands. IOP Conf. Ser. J. Phys. Conf. Ser. **1176**, 022015 (2019)
4. Sentiment analysis using machine learning approaches (lexicon based on movie review dataset). J. Ubiquitous Comput. Commun. Technol. (UCCT) **02**(03), 145–152 (2020)
5. Kharde, V.A., Sonawane, S.S.: Sentiment analysis of Twitter data: a survey of techniques. Int. J. Comput. Appl. **139**(11) (2016). 0975-8887
6. Makarem, S.C., Jae, H.: Consumer boycott behavior: an exploratory analysis of Twitter feeds
7. Jhaveri, D., Chaudhari, A., Kurup, L.: Twitter sentiment analysis on E-commerce websites in India. Int. J. Comput. Appl. **127**(18) (2015). 0975-8887
8. Sirsat, S., Rao, S., Wukkadada, B.: Sentiment analysis on Twitter, data for product evaluation. IOSR J. Eng. (IOSRJEN) 22–25. ISSN (e): 2250-3021, ISSN (p): 2278-8719
9. Bagheri, H., Islam, M.J.: Sentiment Analysis of Twitter Data. Computer Science Department, Iowa State University
10. Faizan: Twitter sentiment analysis. Int. J. Innov. Sci. Res. Technol. **4**(2) (2019). ISSN No: 2456-2165
11. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment, Analysis of Twitter Data. Department of Computer Science, Columbia University, New York, NY
12. Ravi Babu, U.: Sentiment analysis of reviews for E-shopping websites. Int. J. Eng. Comput. Sci. **6**(1), 19966 (2017). ISSN: 2319-7242, Page 19965–19968 Index Copernicus Value (2015): 58.10. https://doi.org/10.18535/ijecs/v6i1.20

13. Sarlan, A., Nadam, C., Basri, S.: Twitter sentiment analysis. In: 2014 International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia, 18–20 Nov 2014
14. https://en.wikipedia.org/wiki/Data_extraction
15. Brahmananda Reddy, A., Vasundhara, D.N., Subhash, P.: Sentiment research on Twitter data. Int. J. Recent Technol. Eng. (IJRTE) **8**(2S11) (2019). ISSN: 2277-3878
16. http://docs.tweepy.org/en/latest/
17. https://www.nltk.org/index.html
18. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
19. Harikantra, S.S., Fernandes, R.: Opinion mining on Twitter data. IJIRSET International Journal of Innovative Research in Science, Engineering and Technology **5**(9) (May 2016)
20. https://textblob.readthedocs.io/en/dev/
21. https://pypi.org/project/text2emotion

Implementing the Comparative Analysis of AES and DES Crypt Algorithm in Cloud Computing

R. S. Reshma, P. P. Anjusha, and G. S. Anisha

Abstract Cloud computing helps users to access information over the internet. Efficiency, scalability and resource consumption are all optimized in cloud computing. Data security is an important concern in cloud computing that can be addressed with cryptography. Cryptography can be defined as a measure to protect confidential information that must be protected from others who are not intended to be viewed in the same way. The use of cryptography is widespread from small schools and colleges to the social media platform that we addicted to. AES and DES cryptographic algorithms are the main concern of this paper and also implementing which algorithm is best in the basis of time and also comparing the differences between these two algorithms.

Keywords AES · DES · Cryptography · Encryption · Decryption · Cloud

1 Introduction

Cryptography is mainly used to hide information from others which help to keep the information secured. It is a concept of converting a plain text into some encrypted text. This is done so that only the intended person would be able to understand and decrypt the text. Advanced encryption standard (AES) and data encryption standard (DES) are the main solicitudes of this paper. We implemented the comparative analysis of these two crypt algorithms in cloud computing. AES is a replacement for DES because of this faster speed and larger key size. AES is said to use an iterative method which involves substituting the input signals with the bit stream. Whereas on the other hand, DES is based on a Feistel cipher that takes the operations of encryption and decryption and uses a reverse scheme. We implemented this in cloud computing, as we can say that the use of cloud computing is increasing day by day, all the IT firms and other industries using technology are widely using cloud computing;

R. S. Reshma (✉) · P. P. Anjusha · G. S. Anisha
Department of Computer Science and IT, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_24

mainly because it offers many important services which are useful for them in their industries. The services include storage, networking, etc.

There are four types of cloud development models. They are:

- **Public cloud**
  In public cloud the resource is shared between many users.
- **Private cloud**
  Private cloud takes individual/organization not shared with any other individuals.
- **Hybrid cloud**
  It is a combination of public and private cloud, most of the organizations are using hybrid cloud.
- **Community cloud**
  It is shared by the users of same industry.

1.1 Implemented Algorithms

Advanced Encryption Standard (AES)
AES is a symmetric encryption algorithm. It is a block cipher with a block length of 128 bits. As shown in the Fig. 1, there are different Key length of specific bits 128, 192, or 256 bits. Each iteration of encryption consists of these key lengths having a specific number of rounds where these rounds are calculated from the unique AES key which was used earlier (Fig. 2).

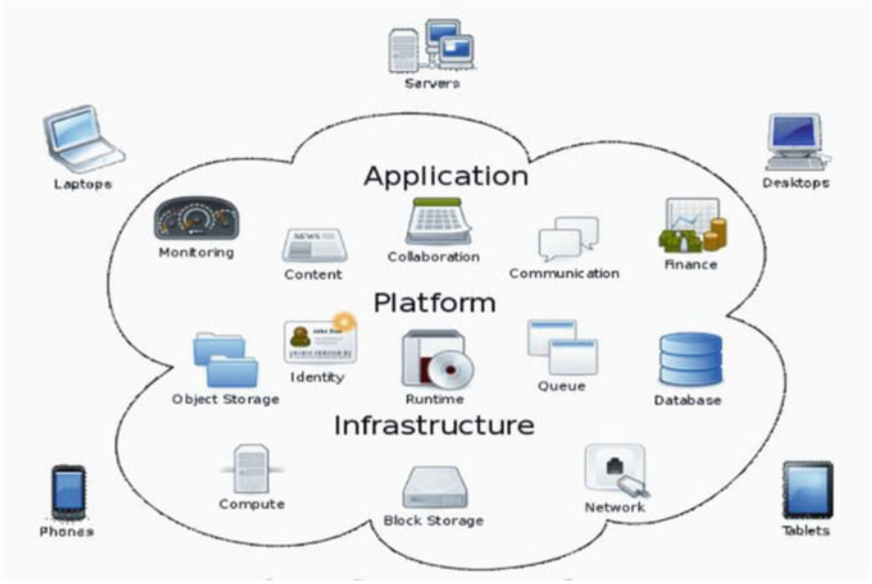


Fig. 1 Cloud computing diagram

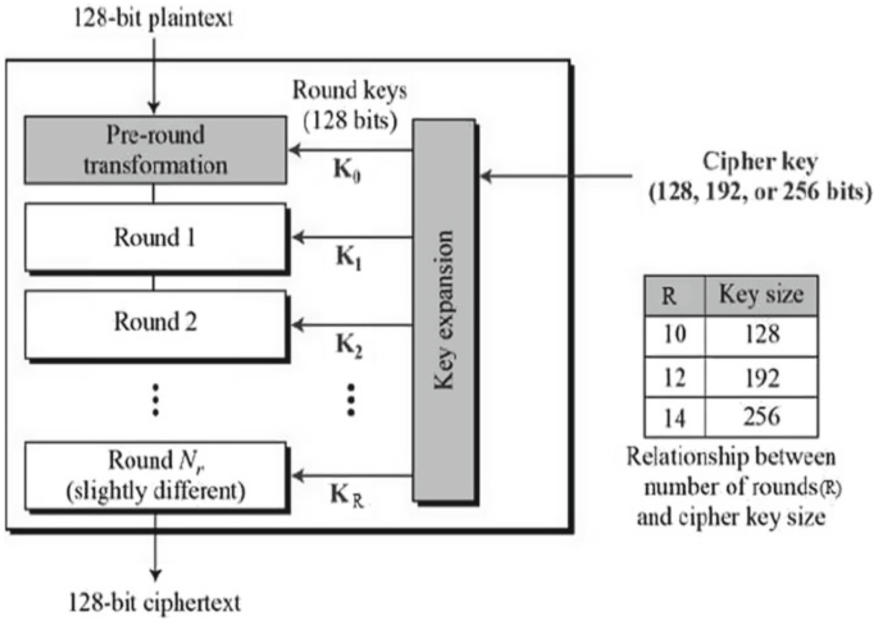


Fig. 2 AES structure

Data Encryption Standard (DES)

DES is a cryptographic algorithm that converts the bit in the input into an encoded output which can only be processed with a DES key. This output is a product of well-defined complex activities. However, DES is not considered to be secured, but as its decryption uses a reverse key, it is used in the implementation of cryptography.

Comparison between AES and DES algorithm

The difference between AES and DES is stated in Table 1. Through which we can conclude the strength and weakness of these encryption methods.

Table 1 Comparison table

AES	DES
AES stands for advanced encryption standard	DES stands for data encryption standard
AES is extra stronger than DES	DES is poor in security
Encrypt 128 bits of plain text	64 bits of plain text
It has excellent security	Security is not enough
AES is comparatively faster than DES	DES is slower

2 Related Work

This paper describes the comparison between DES, AES and RSA cryptographic algorithms and takes a simple literature survey on it [1]. They analyze different security issues and different cryptographic algorithms for the better security of cloud [2]. They also specify the different types of Cloud and their characteristics. They have also introduced five well-known and extensively encrypted methods BLOWFISH, AES, RSA and DES, and then, the comparison of their success based on the study of their encryption, and decryption period in the local system for different file sizes is done [3]. Parsi Kalpana proposed a method by implementing RSA algorithm for ensuring the security of cloud user’s data [4]. However, Sandeep K Sood discussed about the different techniques to secure the data from the intruders [5]. Moreover, this paper helps in the enhancement of data protection which reduces the probability of any attacks possible [6]. They discuss about some of the techniques that were implemented to protect data in cloud [7]. They discussed about various security issues and importance of security in cloud computing [8]. Calculated the performance measures of AES and DES algorithms which is conducted on the basis of the CPU usage, time taken, etc. [9]. AES and DES algorithms have been implemented using MATLAB software [10]. After implementing these techniques, they compared the differences. Based on the simulation time of encryption and decryption process, they tried to implement the AES, DES and RSA algorithms for the encryption process and have eventually compared their performance basis on above-mentioned stimulated time [11]. Here, they have discussed about the various security issues that we face in cloud computing and also about the challenges that comes with the cloud computing mechanism that we eventually face during the cloud engineering [12]. Furthermore, they have also stated the various security algorithms.

3 Proposed Methodologies

In this paper, we are comparing two cryptographic algorithms AES and DES to identify which one is better in cloud in the basic mean of time. Encryption speed of each algorithm for different file size is taken. Both the implementations done exact to make sure that the results will be relatively fair and accurate.

Implementation

We are using Python language for checking the encryption time taken by both the algorithms.

Localhost

localhost is a hostname that pertains to the device that is actually attached to it. It is used to link to the host’s network services via the loopback network interface. If you use the loopback interface, every local network interface hardware is bypassed. We implemented this experiment on a laptop by making it as localhost.

Google Cloud

Google Cloud is basically a platform that provides cloud services like infrastructure as a service, platform as a service, and serverless computing environments [13].

Here in this paper, we have deployed our application into Google Cloud, to check which algorithm takes less time to encrypt among AES and DES.

- The application asks to upload any file that is present in the system.
- Then, we encrypt the same file using both the AES and DES algorithms. Two variables have been used in the entire process; the start time $T1$ (the moment it starts encrypting) and the end time $T2$ (the time when the encryption process ends) to calculate the total time taken by both the encryption algorithms.

$$\text{Total time taken} = T2 - T1 \tag{1}$$

Total time taken decides that which among AES and DES algorithm takes the least amount of time for processing encryption on files.

4 Experiment and Result Analysis

The results of the evaluation of AES and DES cryptographic algorithms are shown in Table 1 for input sizes: 1 KB, 10 KB, 204 KB, 3 MB and 7 MB. These are some of the input files that were taken to implement it. In cloud computing, we have introduced this analogy. Table 2 displays the product of the AES and DES encryption time (in thousand seconds). It thus showed that in execution time, AES is faster than DES. All the relevant values of observations and graphs are taken for analysis of processes of algorithms. Figure 3 shows the graph of the encryption time for different input sizes (Fig. 4).

The encryption time takes by both the methods is given below:

Table 2 AES and DES encryption times for various data sizes

Input size	AES encryption time (ms)	DES encryption time (ms)
1 KB	1.483	7.131
10 KB	1.801	7.398
204 KB	7.578	9.554
3 MB	75.966	78.335
7 MB	161.930	162.64

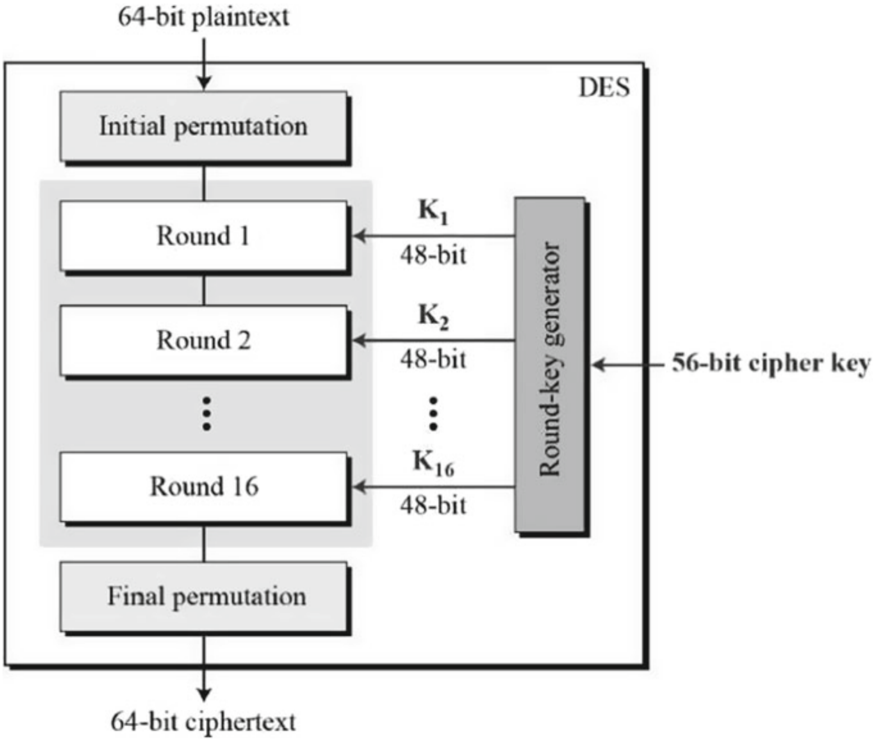


Fig. 3 DES structure

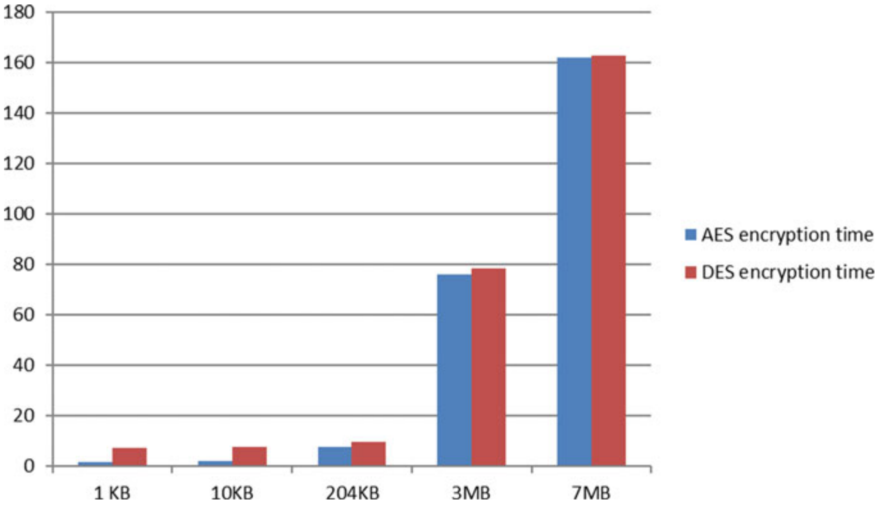


Fig. 4 Graph depicting the encryption time of various input sizes

Graph (Fig 3) shows the encryption time of AES and DES of various input sizes. Encryption time shows in milliseconds. This graph clearly shows that AES takes less encryption time than DES.

5 Conclusion and Future Work

In communication protection, encryption algorithms are important. Based on the encryption time, we conducted a performance comparison of the cryptographic algorithms AES and DES in cloud to check which among these algorithms works better when it comes to time efficiency. Our study estimates that AES is much better than DES as it takes less time for encryption. Compared to DES, AES is relatively quicker and better when the input size is not huge. The results of the simulation suggest that for input sizes such as 1 KB, 2 MB or 7 MB, AES is working so much better and faster than DES. As of now, this paper only suggests AES is better than DES in terms of time efficiency. Further research can also be done on these algorithms to check their capabilities in terms of security provided by these two and better storage size; that way, it would become easier to select one algorithm among these two in terms of reliability.

References

1. Kannan, M., Priya, C., VaishnaviSree, S.: A comparative analysis of DES, AES and RSA crypt algorithms for network security in cloud computing. J. Emerg. Technol. Innovative Res. (JETIR) **6**(3) (2019)
2. Khan, S.K., Tuteja, R.R.: Security in cloud computing using cryptographic algorithms. Int. J. Innovative Res. Compute Commun. Eng. **3**(1) (2015)
3. Hossain, M.A., Hossain, M.B., Uddin, M.S., Imtiaz, S.M.: Performance analysis of different cryptography algorithms. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **6**(3) (2016)
4. Kalpana, P., Singaraju, S.: Data security in cloud computing using RSA algorithm. Int. J. Res. Comput. Commun. Technol. (IJRCCT) **1**(4). ISSN 2278–5841 (2012)
5. Sood, S.K.: A combined approach to ensure data security in cloud computing. J. Netw. Comput. Appl. **35**(6) (2012)
6. Albgumi, A., Alassafi, M.O., Walters, R., Wills, G.: Data security in cloud computing. In: Fifth International Conference on Future Generation Communication Technologies (2016)
7. Sugumran, M., Murugan, B.B., kamalraj, D.: An architecture for data security in cloud computing. In: 2014 World Congress on Computing and Communication Technologies, pp. 252–255 (2014)
8. Kaufman, L.M.: Data security in the world of cloud computing. IEEE Secur. Privacy **7**(4), 61–64 (2009)
9. Rihan, S.D., Khalid, A., Osman, S.E.F.: A performance comparison of encryption algorithms AES and DES. Int. J. Eng. Res. Technol. (IJERT) **4**(12), 151–154 (2015)
10. Bhat, B., Ali, A.W., Gupta, A.: DES and AES performance evaluation. In: International Conference on Computing, Communication and Automation (2015)
11. Mandal, A.K., Prakash, C., Tiwari, A.: Performance evaluation of cryptographic algorithms: DES and AES. In: 2012 IEEE Students' Conference on Electrical Electronics and Computer Science (2012)

12. Arora, R.,Parashar, A.: Secure user data in cloud computing using encryption al gorithms. Int. J. Eng. Res. Appl. (2013)
13. Mahajan, P., Sachdeva, A.: A study of encryption algorithms AES, DES and RSA for security. Global J Comput. Sci. Technol. (2013)

A Model for Predictive and Prescriptive Analysis for Internet of Things Edge Devices with Artificial Intelligence

Dinkar R. Patnaik Patnaikuni and S. N. Chamatagoudar

Abstract In recent years, abundant amounts of data have been accumulated from a huge network of Internet of things (IoT) devices spread around the globe. The collected data is only useful if it creates an action. To forge data actionable, it needs to be broadened with context and creativity. Traditional methods of evaluating structured data and creating action do not contribute to efficiently process the massive amounts of real-time data that stream from IoT devices. The study has shown that most of the IoT gadgets offering cloud storage along with analytics either trade the data or are lost dumped with no use. For instance, consider the trillions of log files that contain metadata, timestamps of a smart bulb which seems useless if used by nobody. But, it is always important to correlate the data with similar data patterns in a different application that helps in forecasting an insight into possible outcomes. Hence, there is a huge scope for improvement in this realm which motivated us to perform experiments and prove the concept with rigid conclusions. This is where AI-based analysis and response become crucial for extracting optimal value from that data. Also the research involved contains sensible prescriptive analysis offering hindsight when one talks about the edge or node devices in the IoT scenario but certainly, it lacks the rigid structure for offering insight and foresight. An in-depth insight at the edge level can be conceived by the existing artificial intelligence building models offered by many IT giants such as AWS Greengrass. Thus, there is an immense need to process the edge device data with enough intelligence and use existing analytics tools to greatly enhance the performance of the cloud and improve overall IoT application in hand by making the cloud requirements less CPU intensive and more economic. In this paper, a model for predictive and prescriptive analysis to improve production capabilities, gain efficiencies, and reduce operating costs by delving into edge computing to produce actionable insight and foresight is demonstrated with the help of a practical experiment.

D. R. P. Patnaikuni (✉) · S. N. Chamatagoudar
Electronics and Telecommunication Engineering, Walchand Institute of Technology, Solapur, India
e-mail: pdrpatnaik@witsolapur.org

S. N. Chamatagoudar
e-mail: snchamatagoudar@witsolapur.org

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_25

Keywords Edge computing · Internet of things · AWS Greengrass · Cloud computing · Prescriptive analysis · Internet of things (IoT) · Artificial intelligence · Raspberry Pi · MPU6050

1 Introduction

IoT-enabled gadgets are currently developing from the ability to handle standard applications and repetitive tasks to a much better system that can change with dynamic tasks, but still, there is a vast scope for development. The prime gaps identified are pointing the researchers to make an improvement in the delays involved in real-time processing of the data collected and processed in an IoT application. Here we need to discuss two basic terms that most often come into the picture namely "Cloud computing" and "Edge computing". Cloud computing refers to the usage of services deployed for a specific purpose describing a suitable business model. Cloud computing is basically classified into three main services (Fig. 1);

(a) Platform as a service
(b) Software as a service
(c) Infrastructure as a service

The above terminologies are very familiar to the majority of the people in the market, and hence, it has proven their advantageous features to many businesses. Yet there is a substantial gap in terms of many aspects which are recently discovered with the tremendously increasing demand of businesses and organizations for more advanced cloud computing services. With this demand, there is a necessity to find novel ways of computing the tasks with lesser delays, improved performance, and

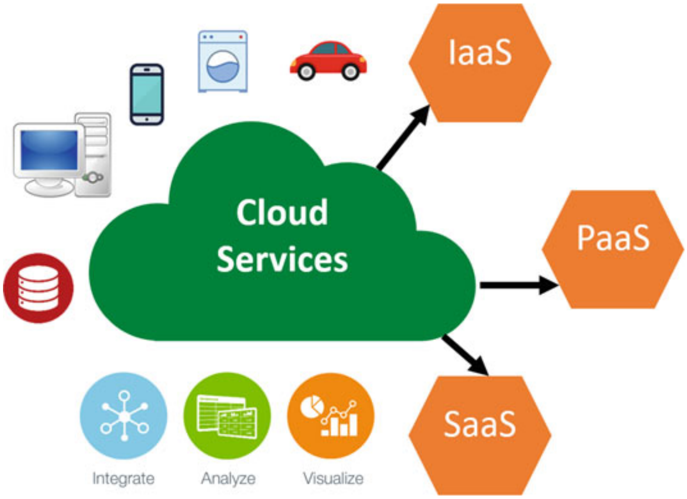


Fig. 1 Cloud services and applications

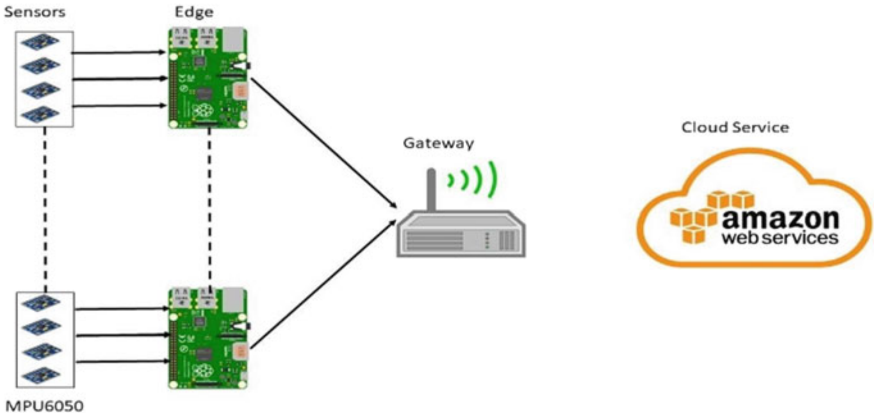


Fig. 2 Block diagram representing edge and cloud service

reduced operational costs at any stage of the IoT architectural model followed by any given application. Figure 2 depicts an iconic representation of today's smart IoT devices like TV, washing machine, computers, and cellphones that fall under the IoT gadget category. These smart devices are capable of connecting to the internet and send useful data to the cloud which is logged and processed for a variety of purposes. On the other hand, some gadgets merely have cloud connectivity to facilitate remote control and surveillance.

The combination of AI and IoT can empower the traditional way of responding to various tasks to a much better adaptive response level. The novelty of this new approach lies in the ability of the system to process some critical data without having to rely on the cloud services at the edge level of an IoT system. This approach greatly enhances the processing time by quickly performing trivial tasks within the edge layer of the IoT architecture. This also has a great benefit in terms of the average response time and a significant reduction in bandwidth usage.

1.1 Real-Time and Post-Event Processing

Post-event processing—This involves identification of data set patterns and extracting the predictive analytics out of it, e.g., correlation between traffic and traffic signal, melting of glaciers with an increase in climatic temperature.

Real-time processing—This involves a quick response to conditions and build-up of knowledge based on decisions taken for an event, e.g., availability of parking space in a parking lot through a video surveillance system.

Adaptive/continuous analytics—This involves the decisions and appropriate actions that need to be taken for efficient system adaptation to all the changes taking place.

Artificial intelligence plays a key role in elevating edge computing that delivers local-level computing at the edge of an IoT system. This further offers computing, processing capabilities, and deployment of standard systems thereby reducing cloud communication to some extent. In circumstances where there is an involvement of private data that cannot be always authenticated via cloud computing, thus, confidentiality can be maintained by processing and further analyzing critical information at the edge level itself thereby reducing round trip time and overall latency for cloud communication.

To save long latencies for communicating with the cloud and relying on roundtrip times and delays in processing and connectivity issues involved therein, simply the data aggregated from sensors can be generalized and processed at the edge level of IoT systems [1–3]. This further saves the resources involved by avoiding unnecessary communication with the cloud and hence leads to reduced bandwidth usage, thus, making the overall system more economic.

Edge computing contributes basically in making faster decisions, aggregation of data and from sensors, and smoothing [4, 5].

Cloud providers such as Google Cloud Platform, IBM, Azure, and AWS offer GPU as a service that offers the ability to use pretrained machine learning models and also provides easy access to the tools via APIs [6]. These cloud services do play an outstanding role in deploying edge computing on a large scale by complementing it with their cloud services.

Often, comparative analysis shows that enterprise cloud infrastructure costs abundant money and complexity. This interlinked complexity thereby introduces additional demand to lower node deployment which merely collects the data and pumps it to the cloud irrespective of the need [7]. This does an indirect harm to the overall computing infrastructure by lending additional burden in terms of computing lag. Recent chip development manufacturers have assisted the integration of artificial intelligence into the chips that are built-in which significantly boosts the processing ability of a node in the IoT scenario. The intelligence in these chips evolved these days to such an extent that they can easily take up the tasks like collecting, sampling, smoothing, and shaping the data along with the substantial implementation of artificial intelligence over the sampled data to accelerate the process of action to be taken instantaneously and elude the need to pump every bit of data to the cloud for analysis [8]. This greatly enhances the overall performance in terms of time and space complexity of the system further leveraging the benefits of the cloud.

Computing at the Edge

IoT devices that tend to offer instantaneous prediction based on integrated AI-enabled intelligence with an associated stack written into the hardware that can detect any minor changes and offer a real-time decision without relying on the cloud. This usually involves a set of AI stack integrated into the hardware in the form of firmware that exhibits the ability to perform a few complex processing tasks and complex computations assisted by the CPU at the edge level itself [9, 10]. Here the data transfer between the edge and central cloud or the gateway and the central cloud is eliminated to achieve truly real-time decision making.

Reliability
The real-time decisions to be made are classified first based on the risk involved in case of a false trigger [11, 12]. For instance, consider how medical equipment senses the heart rate of a patient in an ICU, where a false positive could lead to a considerable wrong decision; thereby crashing the entire system's efficiency and accuracy. Thus there is a need to classify the application before implementing these systems at firsthand where the parameters that identify the risk involved with a false positive need to be categorized into a pool of applications with high risk and the rest into a pool with the risk that doesn't lead to judging the system's efficacy in any means for a given false positive.

Security
Offering computing at the edge not only makes the system affordable and efficient but also adds additional security by keeping the possibilities of cyber attacks over the data on the cloud. This is achieved as sensitive data is kept at local servers and processed immediately to maintain real-time feasibility and the data which is processed only needs to be saved for future reference by eliminating the need to store entire raw data. AI-enabled solutions if implemented can further contribute to prescriptive and predictive analysis of majorly at edge level only [13].

Cost With Edge Computing
Since there is an aggregation and analysis implementation entirely at Edge or local level, the system tends to bring in a lot of savings by eliminating costly bandwidth, connectivity, and maintenance issues [14, 15].

2 Proposed Model Using Raspberry-Pi

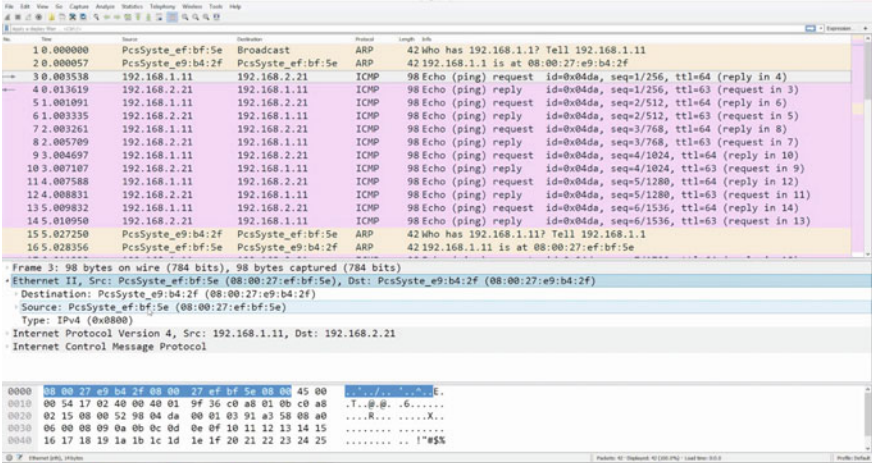
Our proposed solution constitutes an array of accelerometer cum gyroscope connected to separate microcontrollers. Since edge–cloud transfer bandwidth is the key factor affecting performance [16–19], the accelerometer cum gyroscope utilized for this experimental setup was MPU6050 as it needs a bit more complex Kalman filter processing that best suits the goal that this experiment tends to prove. Further, these sensors were interfaced to an Atmega-based board (ATtiny85) which supports SPI protocol. This board was specifically chosen as it supports a very few computing resources with 8 K Flash and 512B RAM clocking at a frequency of 8 MHz. Though these computing resources seem less in comparison with other high-end resources like the Raspberry-Pi or Beaglebone, yet there is enough power with ATtiny85 boards to extract the sensor data and aggregate it so as to further pass it onto the edge computing device (Raspberry-Pi), as shown in Fig. 2. Here, the edge device is powered by Amazon Greengrass which supports local computing, data management, and database-oriented activities being an open-source edge runtime and IoT service. The novelty in this approach is the use of AWS AI pretrained services that greatly enhances the overall bandwidth usage and a reduction in the average response time by bringing the computation of certain percentage of tasks near to the edge layer of the IoT architecture. Here, the edge computing principle used brings in a comprehensive processing and management mechanism by pushing the computational infrastructure closer to the data source where the pre trained AI model takes care of identifying the edge node data and the time it takes to process locally and through central cloud, and then, it classifies the nodes based on time taken by the pool of sensors to be processed at edge level and the rest through central cloud. The aggregated data is effectively analyzed and connected seamlessly to AWS cloud services where the power of the cloud comes into the picture. The experimental setup was composed of four edge devices connected with six MPU6050 sensors to each edge thereby summing up to twenty-four Kalman filter calculations. The data was collected and sent to AWS cloud without taking the computing capabilities at the edge level in Phase-I for comparison. Whereas in Phase-II, the data was collected and processed at the edge level, and then the results were published to the cloud. There was a substantial improvement in the processing time and the ease of access to the data by using Phase-II.

3 Results

The said experiment was conducted with six Raspberry Pis and three of which were set to perform edge computing of the Kalman filter calculations for the accelerometers connected to the edge. The remaining three Raspberry Pis were subjected to Kalman filter calculations performed by relying on an AWS instance-based computing unit. Here, we used several iterations of sniffing the packets to analyze the RTT of the communication involved using Wireshark tool and a screenshot of the same is shown in Fig. 3. We used Linux terminal to iterate the ICMP protocol-based ping requests to check and validate the time intervals needed to ensure a packet sending a receiving cycle for the edge category and cloud category of the Raspberry Pi. Further, we repeated the same experiment of categorizing the Raspberry Pis and validated the improvement in RTT time in the edge category using Wireshark Network Sniffing tool.

Several iterations of the experiment were performed and an average is calculated and is presented as shown in Table 1. Processing time is the time calculated using the python script interpreter for the calculations of the Kalman filter. This shows a clear indication of the fact that under circumstances where there is a possibility of processing partial computing tasks at the edge layer of an IoT architecture, this process of implementing edge computing can greatly improve the overall efficiency of the entire system.

Table 1 shows a very clear indication that phase-I is costly in terms of processing time, bandwidth monitored via the Wireshark tool, and of course the latency.


File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help

```
 1 0.000000   PcsSyste_ef:bf:5e   Broadcast           ARP    42 Who has 192.168.1.1? Tell 192.168.1.11
 2 0.000057   PcsSyste_e9:b4:2f   PcsSyste_ef:bf:5e   ARP    42 192.168.1.1 is at 08:00:27:e9:b4:2f
 3 0.003538   192.168.1.11        192.168.2.21        ICMP   98 Echo (ping) request  id=0x04da, seq=1/256, ttl=64 (reply in 4)
 4 0.013619   192.168.2.21        192.168.1.11        ICMP   98 Echo (ping) reply    id=0x04da, seq=1/256, ttl=63 (request in 3)
 5 1.001091   192.168.1.11        192.168.2.21        ICMP   98 Echo (ping) request  id=0x04da, seq=2/512, ttl=64 (reply in 6)
 6 1.003335   192.168.2.21        192.168.1.11        ICMP   98 Echo (ping) reply    id=0x04da, seq=2/512, ttl=63 (request in 5)
 7 2.003261   192.168.1.11        192.168.2.21        ICMP   98 Echo (ping) request  id=0x04da, seq=3/768, ttl=64 (reply in 8)
 8 2.005709   192.168.2.21        192.168.1.11        ICMP   98 Echo (ping) reply    id=0x04da, seq=3/768, ttl=63 (request in 7)
 9 3.004697   192.168.1.11        192.168.2.21        ICMP   98 Echo (ping) request  id=0x04da, seq=4/1024, ttl=64 (reply in 10)
10 3.007107   192.168.2.21        192.168.1.11        ICMP   98 Echo (ping) reply    id=0x04da, seq=4/1024, ttl=63 (request in 9)
11 4.007588   192.168.1.11        192.168.2.21        ICMP   98 Echo (ping) request  id=0x04da, seq=5/1280, ttl=64 (reply in 12)
12 4.008831   192.168.2.21        192.168.1.11        ICMP   98 Echo (ping) reply    id=0x04da, seq=5/1280, ttl=63 (request in 11)
13 5.009832   192.168.1.11        192.168.2.21        ICMP   98 Echo (ping) request  id=0x04da, seq=6/1536, ttl=64 (reply in 14)
14 5.010950   192.168.2.21        192.168.1.11        ICMP   98 Echo (ping) reply    id=0x04da, seq=6/1536, ttl=63 (request in 13)
15 5.027250   PcsSyste_e9:b4:2f   PcsSyste_ef:bf:5e   ARP    42 Who has 192.168.1.11? Tell 192.168.1.1
16 5.028356   PcsSyste_ef:bf:5e   PcsSyste_e9:b4:2f   ARP    42 192.168.1.11 is at 08:00:27:ef:bf:5e

Frame 3: 98 bytes on wire (784 bits), 98 bytes captured (784 bits)
Ethernet II, Src: PcsSyste_ef:bf:5e (08:00:27:ef:bf:5e), Dst: PcsSyste_e9:b4:2f (08:00:27:e9:b4:2f)
  Destination: PcsSyste_e9:b4:2f (08:00:27:e9:b4:2f)
  Source: PcsSyste_ef:bf:5e (08:00:27:ef:bf:5e)
  Type: IPv4 (0x0800)
Internet Protocol Version 4, Src: 192.168.1.11, Dst: 192.168.2.21
Internet Control Message Protocol

0000  08 00 27 e9 b4 2f 08 00  27 ef bf 5e 08 00 45 00   ..'../.. '..^..E.
0010  00 54 17 02 40 00 40 01  9f 36 c0 a8 01 0b c0 a8   .T..@.@. .6......
0020  02 15 08 00 52 98 04 da  00 01 03 91 a3 58 08 a0   ....R... .....X..
0030  06 00 08 09 0a 0b 0c 0d  0e 0f 10 11 12 13 14 15   ........ ........
0040  16 17 18 19 1a 1b 1c 1d  1e 1f 20 21 22 23 24 25   ........ .. !"#$%
```

Fig. 3 Ping packet capture for RTT calculations

Table 1 Comparative analysis

	Phase I Computing at cloud	Phase II Computing at edge
Processing time	20.67 ms	1.57 ms
Bandwidth	Low	0
Latency	Average	Low

4 Conclusion

The experimental results show that for quick processing of data involved with sensors may have additional latency involved while relying on cloud services. With the deployment of edge-based computing augmented with the smartness of AI which helps to auto classify the available edge devices based on processing time it takes to process at edge of the IoT architecture and through the central cloud pathway, a pool of edge computing and central cloud service-dependent tasks are performed. And in a system involving the majority of sensor data and its processing, this concept of processing at the edge or near edge can bring improved bandwidth usage, reduction in latency, and lesser dependencies on the Internet for trivial operations. Rather, edge computing comes in to be more handy and quick under circumstances where there is a possibility of directly applying the intelligence over the procured data at edge levels saving more time, space, complexity, and cost of the overall system. This scheme can significantly boost productivity for any organization delivering quicker analytics with predictive and prescriptive analysis. However, care must be taken while selecting the underlying edge nodes and other computing units like the ATtiny85 to suffice the needs of the desired application under consideration. In case of IoT systems that rely on private data, proper decision making in selection of the architecture is important as keeping user data local can promote enhanced security by making the data processing easier and isolated. The limitation of this system, however, involves some other factors such as delay in identification of a caused bug at a local level where data is processed and updated at discrete intervals of the day. This can be a scope of future work which involves remote monitoring of the data processing logs at the edge of an IoT system without having to deal with the core data but only the logs established on a real-time basis.

References

1. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple Classifier Systems, Cagliari, Italy, 21–23 June. LNCS, vol. 1857, pp. 1–15, Dec 2000
2. https://digital-library.theiet.org/content/journals/https://doi.org/10.1049/iet-net.2018.5182
3. Calo, S.B., Touna, M., Verma, D.C., Cullen, A.: Edge computing architecture for applying AI to IoT. In: 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA. pp. 3012–3016 (2017). https://doi.org/10.1109/BigData.2017.8258272
4. https://ieeexplore.ieee.org/abstract/document/83270
5. Backes, J., et al.: Reachability analysis for AWS-based networks. In: Dillig, I., Tasiran, S. (eds.) Computer Aided Verification. CAV 2019. Lecture Notes in Computer Science, vol. 11562. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-25543-5_14
6. Pelle, J., Czentye, J.D., Sonkoly, B.: Towards latency sensitive cloud native applications: a performance study on AWS. In: 2019 IEEE 12th International Conference on Cloud Computing (CLOUD), Milan, Italy. pp. 272–280 (2019). https://doi.org/10.1109/CLOUD.2019.00054
7. Cook, B., Khazem, K., Kroening, D., et al.: Model checking boot code from AWS data centers. Form Methods Syst. Des. (2020). https://doi.org/10.1007/s10703-020-00344-2
8. Giménez-Alventosa, V., Moltó, G., Caballer, M.: A framework and a performance assessment for serverless MapReduce on AWS Lambda. Future Gener. Comput. Syst. **97**, 259–274. ISSN 0167-739X (2019). https://doi.org/10.1016/j.future.2019.02.057
9. Song, Z., Cheng, J., Chauhan, A., Tilevich, E.: Pushing participatory sensing further to the edge. In: 2019 IEEE International Conference on Edge Computing (EDGE), Milan, Italy. pp. 24–26 (2019). https://doi.org/10.1109/EDGE.2019.00019
10. Caprolu, M., Pietro, R.D., Lombardi, F., Raponi, S.: Edge computing perspectives: architectures, technologies, and open security issues. In: 2019 IEEE International Conference on Edge Computing (EDGE), Milan, Italy. pp. 116–123 (2019). https://doi.org/10.1109/EDGE.2019.00035
11. Alrowaily, M., Lu, Z.: Secure edge computing in IoT systems: review and case studies. In: 2018 IEEE/ACM Symposium on Edge Computing (SEC), Seattle, WA, USA. pp. 440–444 (2018). https://doi.org/10.1109/SEC.2018.00060
12. Dolui, K., Datta, S.K.: Comparison of edge computing implementations: fog computing, cloudlet and mobile edge computing. In: 2017 Global Internet of Things Summit (GIoTS), Geneva, Switzerland. pp. 1–6 (2017). https://doi.org/10.1109/GIOTS.2017.8016213
13. Ren, P., Qiao, X., Chen, J., Dustdar, S.: Mobile edge computing—a booster for the practical provisioning approach of web-based augmented reality. In: 2018 IEEE/ACM Symposium on Edge Computing (SEC), Seattle, WA, USA. pp. 349–350 (2018). https://doi.org/10.1109/SEC.2018.00041
14. Wei, X., et al.: MVR: an architecture for computation offloading in mobile edge computing. In: 2017 IEEE International Conference on Edge Computing (EDGE), Honolulu, HI, USA. pp. 232–235 (2017). https://doi.org/10.1109/IEEE.EDGE.2017.42
15. Ozcan, M.O., Odaci, F., Ari, I.: Remote debugging for containerized applications in edge computing environments. In: 2019 IEEE International Conference on Edge Computing (EDGE), Milan, Italy. pp. 30–32 (2019). https://doi.org/10.1109/EDGE.2019.00021

16. Personè, V.D.N., Grassi, V.: Architectural issues for self-adaptive service migration management in mobile edge computing scenarios. In: 2019 IEEE International Conference on Edge Computing (EDGE), Milan, Italy. pp. 27–29 (2019). https://doi.org/10.1109/EDGE.2019.00020
17. Y. Li, Wang, S.: An energy-aware edge server placement algorithm in mobile edge computing. In: 2018 IEEE International Conference on Edge Computing (EDGE), San Francisco, CA, USA. pp. 66–73 (2018). https://doi.org/10.1109/EDGE.2018.00016
18. Giang, N., Lea, R., Blackstock, M., Leung, V.C.M.: Fog at the edge: experiences building an edge computing platform. In: 2018 IEEE International Conference on Edge Computing (EDGE), San Francisco, CA, USA. pp. 9–16 (2018). https://doi.org/10.1109/EDGE.2018.00009
19. Loghin, D., Ramapantulu, L., Teo, Y.M.: Towards analyzing the performance of hybrid edge-cloud processing. In: 2019 IEEE International Conference on Edge Computing (EDGE), Milan, Italy. pp. 87–94 (2019). https://doi.org/10.1109/EDGE.2019.00029

Comparative Analysis of SIM-Based Hybrid Modulation Schemes Over Log-Normal Channel Model

Siddhi Gangwar, Kavita, Subhash Burdak, and Yashna Sharma

Abstract To enhance the performance of free space optical systems, the hybrid modulation scheme has been proposed. For this analysis, the performance of different hybrid schemes has been studied in the terms of the calculated bit error rates. In this paper, SIM modulation of hybrid schemes over log-normal model has been studied. Analysis of PPM-FSK-SIM, PPM-BPSK-SIM and PPM-GMSK-SIM-based hybrid modulation schemes over the log-normal channel model has been done using MATLAB software. The performance of PPM-GMSK-SIM is proved to be better than the performance of the other two schemes in the obtained results. Although other modulation schemes may have their own advantages, PPM-GMSK-SIM scheme proposed in this paper is specifically suited to free space optical systems. The variation of BER of PPM-GMSK-SIM with the various parameters such as atmospheric turbulence and link distance has also been analyzed in this paper. In addition, the detailed comparative analysis of different schemes presented in this paper can help in choosing the appropriate modulation technique in accordance with the desired application.

Keywords Subcarrier intensity modulation (SIM) · Free space optical communication (FSO) · Probability density function (PDF) · Pulse position modulation (PPM) · Bit error rate (BER)

1 Introduction

Free space optical communication (FSO) is a communication technique wherein free space acts as medium for light to travel between transmitters and receivers. It transmits data wirelessly by propagating light through free space for communication or networking [1]. Free space can refer to air, space or vacuum. FSO doesn't require any material medium for data transmission. This is in contrast with the use of a solid medium of transmission like glass fiber cable, which is commonly used for

S. Gangwar · Kavita · S. Burdak · Y. Sharma (✉)
Dept. of Electronics and Communication Engineering, Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India
e-mail: yashnasharma@dtu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_26

optical communication purposes [2]. Free Space Optics is a Line of Sight (LOS) communication technology, which means that FSO needs a direct path between transmitter and receiver antennas for propagation of light waves and transmission of data bits [3]. There must not be any obstacles in between this path of propagation and the sender and receiver must be in line of sight of each other [4]. It may be considered as a substitute to the conventional radio relay link communication systems. Free space optical communication has numerous advantages such as its high bandwidth, low power requirement and low cost of installation [5].

FSO is highly susceptible to the various atmospheric conditions and turbulences like fog, clouds, haze, smoke, rainfall, physical obstruction along with scintillation and scattering effects, which affects the performance and quality of the optical system [6]. The quality of the communication system can be analyzed by the BER of transmission exhibited by the system. Performance improvement in the extreme weather conditions is one of the many challenges faced in the design and implementation of an OWC system [7]. Pulse position modulation (PPM) is a pulse modulation technique that varies the position of carrier pulses in accordance with the instantaneous values of the message signal. A number of message bits are encoded into a single pulse by varying its position with respect to time. It is a low power consuming, orthogonal technique with less generation of noise, but it necessarily needs synchronization between receiver and transmitter along with a large bandwidth. Frequency Shift Keying (FSK), is a digital modulation technique, that changes the frequency of analog carrier waves as per the digital message signal which is to be transmitted. M-ary frequency shift keying (MFSK) is similar to the PPM technique in a way, as it is also an orthogonal modulation scheme. It has been highly commended by academics. It exhibits a good power efficiency, increased data transmission rate and also incorporates an easy basic design, but it's inefficient in terms of spectrum. Subcarrier intensity modulation (SIM) is a technique in which pre-modulated subcarriers are used for modulating the signal intensity of the wave. It reduces error for supporting higher orders of modulation and thereby raises the throughput by using more than one subcarrier for transmission of data. SIM-based technique is used to overcome the drawback of FSK. But SIM is less power efficient. Thus, by combining PPM, FSK and SIM hybrid scheme, PPM-FSK-SIM is proposed which provides the power efficiency and large throughput [8].

Phase modulations with subcarrier intensities (SIM) due to their high sensitivity at receiving end, constant amplitude nature and good capability of background noise rejection (by increasing receiver complexity) have emerged as a challenging modulation [9]. In phase shift keying (PSK), the carrier phase is altered according to the modulating waveform which is a digital signal [2]. In BPSK, the sinusoidal wave is transmitted with amplitude which is fixed in nature. Phase difference is a straight angle for the data at one level to the data at the other level (180° phase shift) [10]. BPSK-SIM technique is widely used but its power efficiency is not quite good. So, it has been combined with the PPM and the hybrid scheme has been proposed to overcome this issue [11]. The performance of hybrid scheme PPM-BPSK-SIM is found to be better than the performance of BPSK-SIM in terms of BER [2].

Minimum Shift Keying is a continuous phase binary digital frequency modulation with coherent detection capability [12]. In order to increase the spectral efficiency and reduce the inter-symbol interference of MSK, GMSK was introduced in which pre-modulation of signal is done using an LPF [13]. The Gaussian Minimum Shift Keying (GMSK) modulation is a modified version of the MSK (Minimum Shift Keying) modulation, wherein, the phase is filtered through a Gaussian filter in order to ensure a smooth shift from any point in the constellation to the next [8]. It is commonly used for voice communication in RF cellular backhaul network because of its power efficiency as it has minimum side lobe power. The superiority of GMSK over MSK in terms of error performance has been proved and experimentally verified by researchers [13]. Recently, it was highlighted that GMSK with MIMO (multiple-input multiple-output) has an advantage in terms of spectral efficiency and power efficiency [14] over OFDM usage with MIMO. GMSK has better power and spectral efficiency but its drawback is inter-symbol interference effect. For improving this limitation, the hybrid scheme PPM-GMSK-SIM is proposed. This hybrid scheme increases the BER performance and has good spectral efficiency [8].

The comparison between the various hybrid SIM schemes is necessary to analyze which hybrid scheme is better for the FSO system. Different hybrid schemes have been proposed by researchers but a comparative analysis of such hybrid schemes has not been presented in terms of their error performance. This analysis will make it easier for researchers to find which hybrid scheme is needed for their application.

2 Methodology

2.1 Hybrid SIM Modulation Scheme System Design

The schematic diagram of the transmitter and receiver sections of hybrid modulation scheme is shown in Figs. 1 and 2.

The PPM encoder converts the input data bits into L slot symbol, where $L = 2^m$, m is the input bits. Each slot symbol duration is, $t_s = bt_b/L$, where t_b is the duration of input bits. Before being applied to the modulator (BPSK, FSK or GMSK) the output of the PPM encoder is converted into the serial form and the DC bias is added to the output of the modulator. For laser diode to work, the minimum voltage level must

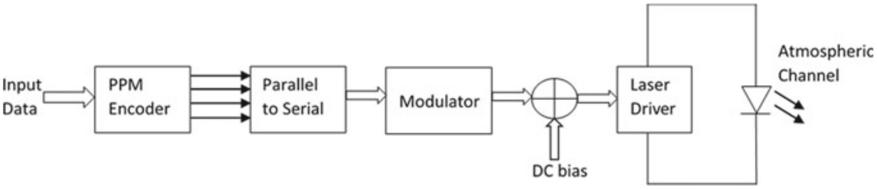


Fig. 1 Schematic diagram of transmitter section of SIM hybrid modulation OWC system

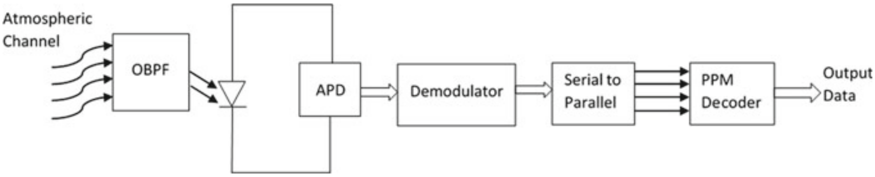


Fig. 2 Schematic diagram of receiver section of SIM-based hybrid modulation OWC system

be higher than the threshold value. Hence, a DC bias is added to the signal in order to overcome the threshold voltage [8].

At the receiver, the signal is filtered by using optical bandpass filter. Then, the APD converts the filtered signal into the electrical signal. These electrical signals are demodulated using the respective demodulator (BPSK, FSK or GMSK) and the output of the demodulator is converted back into parallel form before applying to the PPM decoder. The desired output is derived from the PPM decoder [9].

2.2 Log-Normal Method

When an optical signal is transmitted through the free space, it undergoes amplitude and phase fluctuations [3]. Various channel models like gamma-gamma, negative exponential and log-normal distribution models are available to replicate atmospheric turbulances. The log-normal model is used for weak to moderate turbulance intensities, whereas the gamma-gamma model is applied for strong turbulence conditions [15]. All these models can be used for analyzing the performance of FSO communication system but here we are using the log-normal model. This model is easy to implement mathematically and it gives accurate results in the FSO communication system [16].

In log-normal model, turbulence intensity is weak; therefore, the probability density function (pdf) is based on Rytov approximation.

Probability density function [8] used in this paper is

$$p_I(z) = \frac{1}{z\sqrt{2\pi\sigma_I^2}} \exp\left[-\frac{(\ln(z) + \sigma_I^2/2)^2}{2\sigma_I^2}\right], \quad z > 0 \tag{1}$$

where σ_I^2 is the scintillation index, channel state $z = I/\langle I\rangle$, $\langle I\rangle$ is the average value of instantaneous intensity and I is the instantaneous intensity at the receiver.

Scintillation index [5] is given by, $\sigma_I^2 = 1.23{C_n}^2\, L_d^{\frac{11}{6}}\, k^{\frac{7}{6}}$.

If there is M-symbol data transmission then, all random variable sum in the given channel model with new mean (μ_S) and variance (σ_S^2) is,

$$S = \frac{\sum_{i=1}^{M} I_i}{I_o} \tag{2}$$

where I_o is the received optical irradiance at zero scintillation.

PDF of S is given by [17],

$$p(S) = \frac{1}{S\sqrt{2\pi\sigma_S^2}} \exp\left[-\frac{(\ln(S) + \mu_S)^2}{2\sigma_S^2}\right] \tag{3}$$

$$\mu_S = \ln(L) - \frac{1}{2}\ln\left[1 + \frac{\exp(\sigma_I^2) - 1}{L}\right] \tag{4}$$

$$\sigma_S^2 = \ln\left[1 + \frac{\exp(\sigma_I^2) - 1}{L}\right] \tag{5}$$

Due to different types of weather conditions like rain, fog, pollution, etc. the traveled signal experience some loss and the attenuation in the transmitted signal will occur.

Attenuation constant [2] is given by,

$$\varphi = \frac{A_p}{\pi\left(\frac{\emptyset_A L_d}{2}\right)^2} \exp(-\beta_v L_d) \tag{6}$$

where L_d is the link distance, A_p is the aperture area, $\emptyset_A$ is the divergence angle of the receiver and β_v is the atmospheric extinction coefficient.

2.3 BER of PPM-FSK Hybrid Subcarrier Intensity Modulation Scheme

The performance of error using coherent detection [8] is given by,

$$P_{\text{ec}} = \frac{M}{4}\text{erfc}\left(\sqrt{\frac{mE_b}{2N_o}}\right) \tag{7}$$

where N_o is the Gaussian noise double-sided PSD and E_b is the energy per bit.

At zero scintillation, AWGN channel, the conditional probability error [8] of this hybrid scheme is,

$$P_{\mathrm{ec}} = \frac{M}{4}\mathrm{erfc}\left(\frac{RG\varphi I\sqrt{ml}}{4\sqrt{\sigma_n^2}}\right) \tag{8}$$

where m is the $\log_2 M$, M is the M-symbol FSK, l is the $\log_2 L$ bit and L is for L-symbol PPM. The unconditional probability error over weak turbulence model is given as follows;

$$P_e = \int\limits_0^\infty P_{\mathrm{ec}}\, p(S)\mathrm{d}S \tag{9}$$

This equation is typical. Therefore, we are going to use *gauss-Hermite quadrature integration approximation.* After, calculation the simplified equation [8] is,

$$P_e = \frac{1}{(L-1)\sqrt{\pi}}\sum_{i=1}^{N} w_i\left[1-\left(1-\frac{M}{4}\mathrm{erfc}\left(\frac{RG\varphi I_o}{L\sqrt{2L\sigma_n^2}}\exp(\sqrt{2}x_i\sigma_S+\mu_S\right)\right)^{L-1}\right] \tag{10}$$

where ω_j and x_i are the weighting factor and zeroes of Hermite polynomial and the noise variance [17] is given by,

$$\sigma_n^2 = 2qG^2F_AR\varphi I_o\exp\left(\sqrt{2}x_i\sigma_S+\mu_S\right)\Delta f + \frac{4k_BTF_n\Delta f}{R_L} \tag{11}$$

where G represents APD gain, q is the charge, R is the responsivity, Δf is the effective noise bandwidth. Here, $\Delta f = R_b/2$, R_b is the bit rate, k_B is the Boltzman coefficient, F_n is the noise figure of the amplifier, T is the temperature, R_L represent load resistance and F_A is calculated as $k_A\ G + (1-k_A)(2-(1/G))$.

2.4 BER of PPM-BPSK Hybrid Subcarrier Intensity Modulation Scheme

The conditional probability error of this hybrid scheme [2] is given by,

$$P_{\mathrm{ec}} = Q\left[\frac{m_oR\varphi I_oS}{2\sqrt{2\sigma_n^2}}\right] \tag{12}$$

where m_o is the optical modulation index.

Using Eqs. (9) and (12), the unconditional probability error over weak turbulence model [2] is expressed as,

$$P_e = \frac{1}{\sqrt{\pi}} \sum_{i=1}^{N} w_i \left[Q\left(\frac{m_o R \varphi I_o}{2\sqrt{2\sigma_n^2}} \exp\left(\sqrt{2} x_i \sigma_S + \mu_S\right) \right) \right] \tag{13}$$

where the noise variance [2] is given by,

$$\sigma_n^2 = 2qBR\left(I_{\text{sun}} + I_{\text{sky}}\right) + \frac{4k_B T B}{R_L} \tag{14}$$

where I_{sun} is the irradiance due to the sun and I_{sky} is the irradiance due to the sky and B is the bandwidth.

2.5 BER of PPM-GMSK Hybrid Subcarrier Intensity Modulation Scheme

The conditional probability error of this hybrid scheme [17] is calculated as follows,

$$P_{\text{ec}} = \frac{1}{2} \text{erfc}\left(\frac{\text{GRS}\varphi I}{2} \sqrt{\frac{\gamma}{\sigma_{2-\text{PPM}}^2}} \right) \tag{15}$$

where γ is the degradation factor and $\sigma_{2-\text{PPM}}^2$ is given by, $\sigma_{2-\text{PPM}}^2 = \frac{N_o R_b L}{2l}$.

Using Eqs. (9) and (15), the unconditional probability error over weak turbulence model [17] is expressed as,

$$P_{\text{e}} = \frac{1}{(L-1)\sqrt{\pi}} \sum_{i=1}^{N} w_i \left[1 - \left(1 - \frac{1}{2} \text{erfc}\left(\frac{RG\varphi I_o}{L\sqrt{\frac{L}{l}\sigma_n^2}} \exp\left(\sqrt{2} x_i \sigma_S + \mu_S\right) \sqrt{\gamma} \right) \right)^{L-1} \right] \tag{16}$$

where noise variance is given by Eq. (11).

3 Results and Discussion

The above discussed hybrid schemes have been designed to overcome the issue faced by the different modulation schemes individually. In this paper, analysis of the different hybrid schemes is discussed in terms of BER. PPM-GMSK-SIM is found to be better than the other two schemes.

Here, the different SIM-based hybrid schemes have been studied over the log-normal channel model. The performance of the hybrid schemes is studied in terms of BER. MATLAB software has been used for calculation and analysis. The values of parameters used in the analysis have been shown in Table 5.

In Fig. 3 BER versus irradiance has been plotted for all the above discussed hybrid schemes with wavelength 1550 nm, atmospheric turbulence $C_n{}^2 = 8.5 \times 10^{-15}$ and $L = 2$. At irradiance $I_o = 4$ dBm, PPM-GMSK-SIM is found to have a BER of 2.95×10^{-11} whereas, PPM-FSK-SIM has a BER of 2.5×10^{-8} and PPM-BPSK-SIM has a BER of 1.6×10^{-2}. Thus, it can be concluded that the performance of BER of PPM-GMSK-SIM-based hybrid scheme is better than the other two schemes that have been studied. A summary of the resulting BER at $I_o = 2$, 4 and 3 dBm for different hybrid scheme is presented in Table 1. For irradiance $I_o < -2$ dBm, PPM-GMSK-SIM and

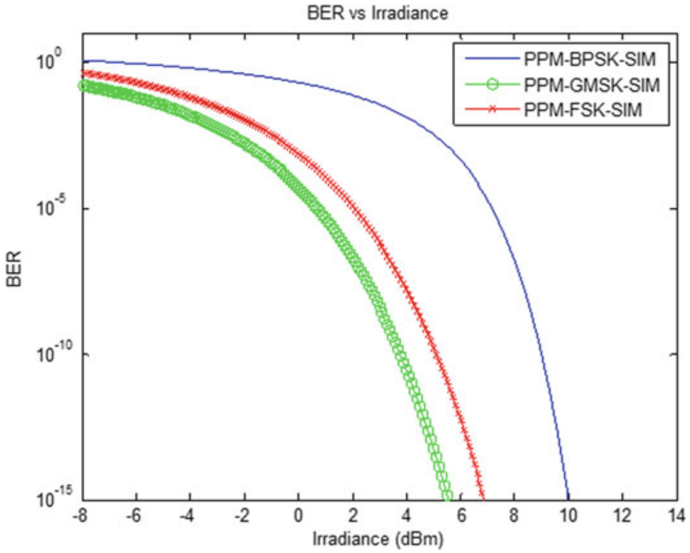


Fig. 3 BER versus irradiance for PPM-GMSK-SIM, PPM-FSK-SIM and PPM-BPSK-SIM with atmospheric turbulence $C_n{}^2 = 8.5 \times 10^{-15}$

Table 1 Comparison of different hybrid schemes in terms of BER for different values of irradiance

Irradiance dBm	Bit error rate		
	PPM-BPSK-SIM	PPM-GMSK-SIM	PPM-FSK-SIM
2	7.401×10^{-2}	2.028×10^{-7}	1.184×10^{-5}
3	3.552×10^{-2}	4.083×10^{-9}	6.461×10^{-7}
4	1.318×10^{-2}	2.958×10^{-11}	1.669×10^{-8}

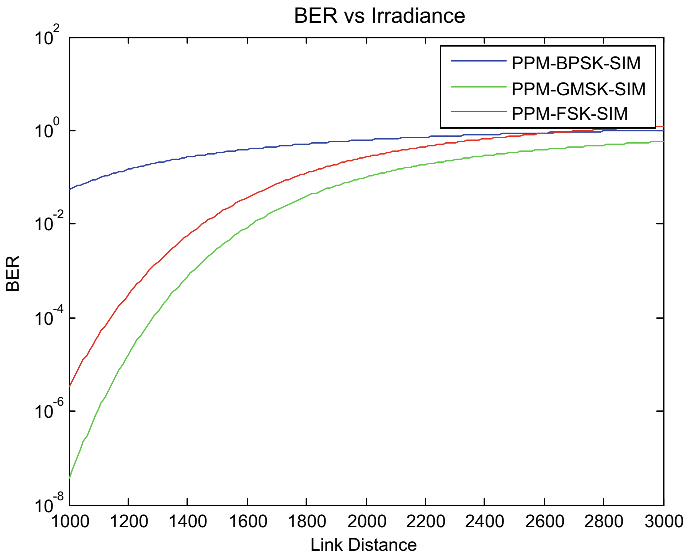


Fig. 4 BER versus link distance for PPM-GMSK-SIM, PPM-FSK-SIM and PPM-BPSK-SIM atmospheric turbulence $C_n{}^2 = 8.5 \times 10^{-15}$

PPM-FSK-SIM have a very poor performance and for PPM-BPSK-SIM when the irradiance $I_o < 7$ dBm BER value is very high.

The variation of bit error rate along the link distance for the three schemes been analyzed, that is, PPM-BPSK-SIM, PPM-GMSK-SIM and PPM-MFSK-SIM, has been shown in Fig. 4. The graph has been plotted for the hybrid schemes for wavelength 1550 nm with the value of atmospheric turbulence $C_n{}^2 = 8.5 \times 10^{-15}$ and $L = 2$. It can again be observed that the BER values for PPM-GMSK-SIM are lowest among the three, thus giving a better performance throughout the range of link distances studied. The BER for PPM-BPSK-SIM is found to be the highest among the studied modulation schemes, making it less suitable for applications requiring low bit error rates. A summary of the resulting BER for link distances equal to 1000, 1400 and 2200 m for different hybrid scheme is presented in Table 2.

Table 2 Comparison of different hybrid schemes in terms of BER for different values of link distances

Link distance (m)	Bit error rate		
	PPM-BPSK-SIM	PPM-GMSK-SIM	PPM-FSK-SIM
1000	5.412×10^{-2}	3.743×10^{-8}	3.359×10^{-5}
1400	2.675×10^{-1}	7.635×10^{-9}	5.739×10^{-7}
2200	7.210×10^{-1}	1.887×10^{-11}	4.501×10^{-8}

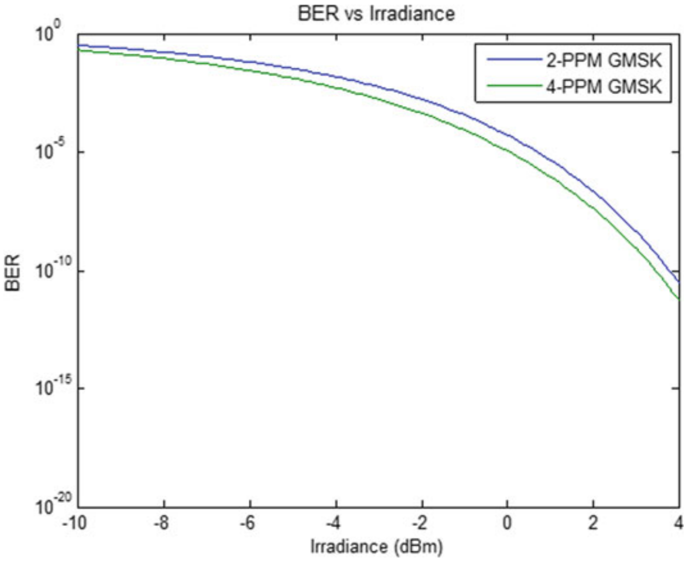


Fig. 5 BER versus irradiance for 2-PPM-GMSK-SIM and 4-PPM-GMSK-SIM with atmospheric turbulence $C_n{}^2 = 8.5 \times 10^{-15}$

In Fig. 5, BER versus irradiance graph has been plotted for different values of L, that is, for 2-PPM-GMSK-SIM and 4-PPM-GMSK-SIM. At irradiance $I_o = 2$ dBm, 4-PPM-GMSK-SIM has a BER of 3.78×10^{-8} and 2-PPM-GMSK-SIM has a BER of 2.02×10^{-7}. This also shows that the bit error rate value of 4-PPM-GMSK-SIM is less than that for 2-PPM-GMSK-SIM. A summary of the resulting BER at $I_o = 1$ and 3 dBm for $L = 2$ and 4 PPM-GMSK-SIM hybrid scheme is presented in Table 3. Thus, it can be concluded that the value of BER decreases with an increase in the Link distance.

In Fig. 6, BER versus irradiance plot has been drawn for different values of atmospheric turbulence. At irradiance $I_o = -2$ dBm PPM-GMSK-SIM has a BER of 1.42×10^{-8} for atmospheric turbulence $C_n{}^2 = 2.5 \times 10^{-15}$, a BER of 1.66×10^{-3} for atmospheric turbulence $C_n{}^2 = 8.5 \times 10^{-15}$, a BER of 1.33×10^{-2} for atmospheric turbulence $C_n{}^2 = 1.2 \times 10^{-14}$ and a BER of 2.2×10^{-1} for atmospheric turbulence $C_n{}^2 = 2.5 \times 10^{-14}$. Thus, it can be noted that for higher value of atmospheric turbulence

Table 3 Comparison of 2-PPM-GMSK with 4-PPM-GMSK in terms of BER for different values irradiance

Link distance (m)	Bit error rate	
	2-PPM-GMSK-SIM	4-PPM-GMSK-SIM
1200	1.661×10^{-5}	7.944×10^{-7}
1600	8.687×10^{-3}	3.641×10^{-4}
2000	1.016×10^{-1}	1.098×10^{-2}

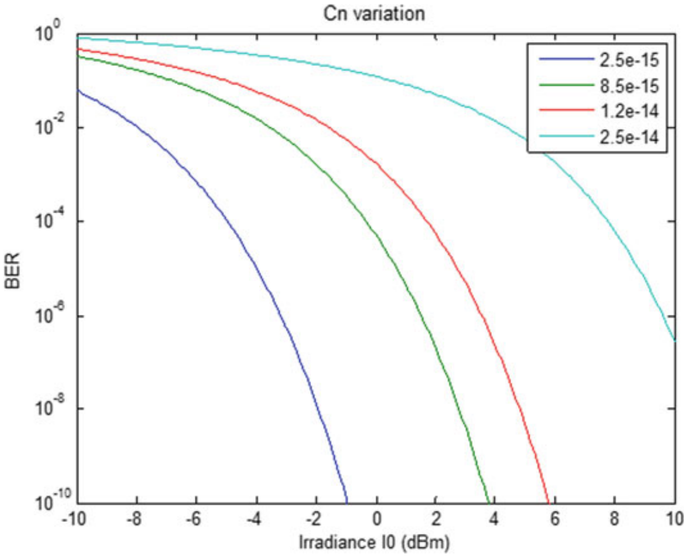


Fig. 6 BER versus irradiance for PPM-GMSK-SIM with atmospheric turbulence values ${C_n}^2 = 2.5 \times 10^{-15}$, ${C_n}^2 = 8.5 \times 10^{-15}$, ${C_n}^2 = 1.2 \times 10^{-14}$ and ${C_\text{n}}^2 = 2.5 \times 10^{-14}$

the bit error rate performance of PPM-GMSK-SIM is degraded, implying that during tornadoes, high wind speed, rain, etc. the system performance will be degraded.

In Fig. 7, BER versus link distance plot has been shown for 2-PPM-GMSK-SIM and 4-PPM-GMSK-SIM. For 4-PPM-GMSK-SIM, at link distance of 1400 m, the BER value is 2.648×10^{-5}, while at link distance of 1600 m, the BER value is found to be 7.773×10^{-4}. With increase in the link distance, the performance of BER is degraded. BER degrades faster in case 4-PPM-GMSK-SIM as compared to 2-PPM-GMSK-SIM. At link distances greater than 2600 m the performance of BER of both (2-PPM-GMSK-SIM and 4-PPM-GMSK-SIM) is almost similar and undesirable. A summary of the resulting BER at link distance = 1200, 1600 and 2000 m for L = 2 and 4 PPM-GMSK-SIM hybrid scheme is presented in Table 4. The various constants used in this study are listed in Table 5.

4 Conclusion

In this paper, SIM-based different hybrid modulation schemes have been analyzed. The bit error rate performance of PPM-GMSK-SIM hybrid scheme is found to be better than that of PPM-FSK-SIM and PPM-BPSK-SIM. At 2 dBm irradiance 2.028×10^{-6} BER is achieved in PPM-GMSK-SIM. The error performance of the PPM-GMSK-SIM enhances if we increase the value of L (i.e., 4-PPM-GMSK-SIM is better than 2-PPM-GMSK-SIM). The BER performance with changes in the atmospheric

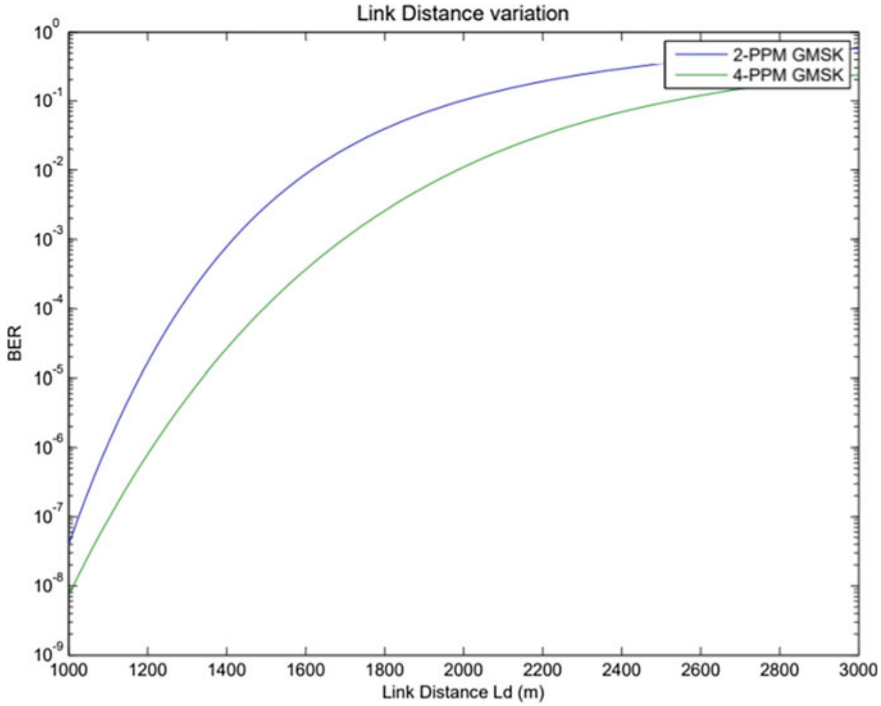


Fig. 7 BER versus link distance (*m*) for 2-PPM-GMSK-SIM and 4-PPM-GMSK-SIM with atmospheric turbulence $C_n{}^2 = 8.5 \times 10^{-15}$

Table 4 Comparison of 2-PPM-GMSK with 4-PPM-GMSK in terms of BER for different values link distance

Irradiance dBm	Bit error rate	
	2-PPM-GMSK-SIM	4-PPM-GMSK-SIM
1	4.466×10^{-6}	8.682×10^{-7}
2	4.083×10^{-9}	7.629×10^{-10}

turbulence is studied and it is observed that the performance of the studied hybrid scheme degrades as the atmospheric turbulence increases. The performance of BER degrades with the increase of the link distance and the BER degrades faster in case 4-PPM-GMSK-SIM as compared to the 2-PPM-GMSK-SIM. In comparison with the normal modulation schemes, hybrid schemes are better for the purpose of FSO communication and among the above discussed hybrid schemes, PPM-GMSK-SIM is found to be the best for increasing the BER performance of the FSO system. However, it must be noted that the other schemes may have their own advantages and for each application, a different scheme may be suitable. We have discussed the merits and demerits of the various available SIM-based hybrid modulation schemes in this paper, using which the researchers can select an appropriate modulation scheme as per their specific application requirements.

Table 5 Values of constants used for BER calculations

Name	Symbol	Value (unit)
APD gain	G	10
Link distance	L_d	2 km
Wavelength	Λ	1550 nm
Receiver noise temperature	T	300 K
Aperture diameter	A_p	0.04 m
Ionization factor	k_A	0.85
Divergence angle	Ω	0.001 radians
Boltzmann's constant	k_B	1.38×10^{-23} WHz/K
Amplifier noise figure	F_n	2
Atmospheric extinction coefficient	B_μ	0.01 dB/Km
Electron charge	Q	1.6×10^{-19} C
Bit rate	R_b	2×10^6 bps
APD load resistance	R_L	1 kΩ
Responsivity	R	1

References

1. Zhu, X., Kahn, J.M.: Free-space optical communication through atmospheric turbulence channels. IEEE Trans. Commun. **50**(8), 1293–1300 (2002)
2. Giri, R.K., Patnaik, B.: BER analysis and capacity evaluation of FSO system using hybrid subcarrier intensity modulation with receiver spatial diversity over log-normal and gamma–gamma channel model. Opt. Quantum Electron. **50**(6) (2018). https://doi.org/10.1007/s11082-018-1499-8
3. Jagadeesh, V., Palliyembil, V., Muthuchidambaranathan, P., Bui, F.M.: Free space optical communication using subcarrier intensity modulation through generalized turbulence channel with pointing error. Microw. Opt. Technol. Lett. **57**(8), 1958–1961 (2015)
4. Pradeep, R., Ravikumar, K., Umesh, S.B.: Comparative analysis of different modulation technique for free-space optical communication. Int. Res. J. Eng. Technol. (IRJET) **5**(3) (2018). e-ISSN: 2395-0056, p-ISSN: 2395-0072
5. Popoola, W.O., Ghassemlooy, Z.: BPSK subcarrier intensity modulated free-space optical communications in atmospheric turbulence. J. Lightwave Technol. **27**(8), 967–973 (2009)
6. Chan, V.W.S.: Free-space optical communication. J. Lightwave Technol. **24**(12), 4750–4762 (2006)
7. Chaleshtory, Z.N., Gholami, A., Ghassemlooy, Z., et al.: Experimental investigation of environment effects on the FSO link with turbulence. IEEE Photonics Technol. Lett. **29**(17), 1435–1438 (2017)
8. Dubey, D., Prajapati, Y.K., Tripathi, R.: Error performance analysis of PPM-and FSK-based hybrid modulation scheme for FSO satellite downlink. Opt. Quantum Electron. **52**(6) (2020)
9. Sharma, K., Grewal, S.K.: Performance assessment of hybrid PPM–BPSK–SIM based FSO communication system using time and wavelength diversity under variant atmospheric turbulence. Opt. Quantum Electron. **52**(10), 1–25 (2020). https://doi.org/10.1007/s11082-020-02547-7

10. Choyon, A.K.M.S.J., Chowdhury, R., Chowdhury, S.M.R.: Optimum link distance and BER performance investigation for BPSK RF sub-carrier coherent FSO communication system under strong turbulence. Int. J. Sci. Technol. Res. **9**(9), 282–287 (2020)
11. Faridzadeh, M., Gholami, A., Faridzadeh, M., Gholami , A.: BPSK-SIM-PPM modulation for free space optical communications. In: 7th International Symposium on Telecom (2014)
12. Hongzhan, L.R, Zhongchao, L., Zhiyun, W., Yaojun Qiao, H.: BER analysis of a hybrid modulation scheme based on PPM and MSK subcarrier intensity modulation. IEEE Photonics J. **7**(4) (2015)
13. Murota, K., Hirade, K.: GMSK modulation for digital mobile radio telephony. IEEE Trans. Commun. **29**(7), 1044–1050 (1981)
14. Jiang, T., et al.: Performance improvement for mixed RF-FSO communication system by adopting hybrid subcarrier intensity modulation. Appl. Sci. **9**(18), 1–14 (2019)
15. Leitgeb, E., Ghassemlooy, Z., Popoola, W.O.: Free-space optical communication using subcarrier modulation in gamma-gamma atmospheric turbulence. In: 9th International Conference on Transparent Optical Networks (ICTON '07) vol. 3, pp. 156–160 (2007)
16. Sood, A., Bala, N., Kumar, M.: BER analysis of hybrid-fso communication system over log-normal atmospheric turbulence. J. Emerg. Technol. innovative Res. **6**(6), 157–161 (2019)
17. Sahoo, P.K., Prajapati, Y.K., Tripathi, R.: PPM-and GMSK-based hybrid modulation technique for optical wireless communication cellular backhaul channel. IET Commun. **12**(17), 2158–2163 (2018). https://doi.org/10.1049/iet-com.2018.5365

Lesion Preprocessing Techniques in Automatic Melanoma Detection System—A Comparative Study

Shakti Kumar and Anuj Kumar

Abstract An automatic melanoma detection system is an image processing-based technique used to detect melanoma. From the infected skin area image, the automatic melanoma detection system produces classification results as benign or melanoma. The automatic melanoma detection system contains four steps. Preprocessing step removes the noise from the infected image. The segmentation step finds the region of interest. Feature extraction is used to obtain lesion features and the classification step predicts lesion image as benign or melanoma. The decision of melanoma detection from such an automatic system depends upon the quality of the input image. Therefore, preprocessing of lesion images is an essential step in the automatic melanoma detection system. It becomes a challenging task due to the presence of various outliers like glare, dust, and hairs on skin lesions. Preprocessing techniques are applied for noise and artifact removal from the lesion. A lot of preprocessing techniques are available in the literature. The selection of appropriate preprocessing techniques may improve the accuracy of the automatic melanoma detection system. Therefore, in this work, we have studied and compared different preprocessing techniques so that the researchers may select appropriate techniques for them. This paper highlights and compares the image enhancement preprocessing techniques based on SNR and PSNR using pepper, salt, and Gaussian noise.

Keywords Lesion preprocessing · Peak signal-to-noise ratio · Image enhancement techniques · Artifacts removal · Automatic melanoma detection system

1 Introduction

An automatic melanoma detection system (AMDS) is an image processing-based system that is used for melanoma detection. Melanoma is a type of skin cancer

S. Kumar (✉) · A. Kumar
DCSA, Panjab University, Chandigarh, India
e-mail: shaktibajpai@pu.ac.in

A. Kumar
e-mail: anuj_gupta@pu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_27

with the highest mortality rate. Melanoma occurs due to high exposure to sunlight. The most inner layer of the epidermis just above the dermis layer contains melanocytes which are responsible for skin pigmentation. Uncontrollable growth of these melanocytes due to UV radiation leads to melanoma, skin cancer. This cancer is common in both youngsters and adults. Early detection of melanoma leads to less death rate. Computer-based diagnosis of melanoma using AMDS is a noninvasive way to detect melanoma by using infected skin area images as input. In an automatic melanoma detection system, the output of one stage is the input of another stage as shown in Fig. 1.

For the computer-based diagnosis of melanoma, preprocessing is the first step to be carried out. As the images are captured in a real-time environment, images may have noise, uncontrollable illumination, and low contrast. The preprocessing step helps to remove outliers like noise and hairs present in the infected skin area of the image. One popular technique to remove the hairs is DullRazor which detects hair edges and repairs them using morphological operations and linear interpolation, respectively [1]. Similarly, the quality of the image can be improved by applying image enhancement techniques. In this way, the preprocessing step performs well either for artifact removal or for image enhancement. Then, segmentation step is used to find out the lesion as a region of interest from the complete skin image. The infected area images when captured for the lesion also consider the surrounding skin area. The removal of unnecessary parts abutting the lesion is required. Segmented lesion boundary will be used for extracting relevant features only instead of considering the complete infected image. In the feature extraction step, different features are extracted from the segmented lesion. The most common method used by the dermatologist in the analysis of skin lesions is the ABCD rule. In this rule, A stands for asymmetry, B stands for border, C stands for color, and D stands for diameter. The feature extraction step helps to identify different lesion features including shape, symmetry, color, and diameter. These values are used to compute the total dermoscopy score (TDS) whose calculation helps the dermatologist as well as automatic systems to study and analyze the lesion. Then, classification step generates the result as benign or melanoma. The classification step identifies the lesion as benign or melanoma using the extracted features. Color correlogram-based classification uses a Bayesian classifier. SFTA-based features are helpful for texture analysis-based classification of the lesion using abnormal skin color. Similarly, KNN-based linear classifier compares the threshold against fixed sensitivity for the classification of the lesion as benign or melanoma. In the case of TDS value, the system with TDS < 4.75 interprets the lesion as benign.

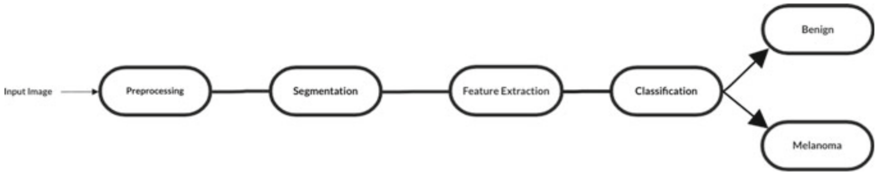


Fig. 1 Stages of AMDS

For TDS ranging between 4.8 and 5.45, the lesion is interpreted as suspicious, and for the TDS > 5.45 lesion is highly indicative of melanoma.

The selection of a good preprocessing technique will affect the accuracy of overall AMDS. The lesion image whose diagnosis is to be carried out must be noise-free. A lesion image with noise and artifacts like hairs may lead to misdiagnosis as most of the publicly available datasets need to be preprocessed first due to the availability of artifacts and unequal size of images, and this may be treated as the novelty of the research. The noise is removed using preprocessing which is the first stage of computer-aided cancer diagnostics. Noise has serious effects on the misleading results for the image-based diagnostic system [2]. Very few papers in the literature are focused on preprocessing techniques even when the overall success rate of automatic melanoma detection systems is dependent on preprocessing [3]. The input image is enhanced and preprocessed so that it is easy to recognize the lesion from the captured area of the skin. The problems associated with skin lesion images of different datasets and their corresponding solution using preprocessing techniques for AMDS are listed in Table 1.

Different datasets are available for the analysis of melanoma [4]. The images of these datasets when preprocessed will lead to more accuracy for automatic image processing-based melanoma detection systems. The ensuing Sect. 2 presents a discussion of preprocessing based on substages. Section 3 presents the methodology used for the analysis of publically available datasets like PH2 and HAM10000, MED-NODE, and DermIS. Section 4 covers results and discussion. The conclusion with the future scope is discussed in Sect. 5.

Table 1 Preprocessing techniques for problems associated with lesion images

Problems associated with lesion ımage in existing datasets	Preprocessing techniques	Expected outcomes
Lesion images of different size and orientation	Image scaling	All preprocessed images will have the same size
Lesion images having a very small difference between background skin and foreground lesion contrast level	Contrast enhancement	Enhanced contrast will provide a sufficient difference between the image background and lesion. After this step, the lesion border detection will become easy
Lesion images having hairs	Morphological methods and ınpainting techniques	The lesion images from which hairs are removed to avoid misleading results

2 Substages of Preprocessing

According to Hoshyar et al., the preprocessing of lesion images can be sectioned into image restoration, image enhancement, and artifact removal [5]. Each substage of preprocessing stage uses different techniques for noise removal and enhancement of lesion images. The selection of preprocessing techniques is dependent on the automatic system used for the diagnosis and detection of melanoma. Filters dealing with mean, median, Gaussian and speckle noise are most popular for preprocessing [6–8]. The further details of the preprocessing stage are depicted in Fig. 2.

2.1 Image Enhancement

To improve the visual appearance of the lesion image, different enhancement techniques are used. The techniques applicable for image enhancement are presented in the ensuing subsection.

Image Scaling. This technique is applicable to make all the lesion images in a standard size acceptable by the automatic melanoma detection system.

Images are resized with a constraint that they will have fixed width pixels but variable size height [9]. The scaling helps to generate a homogeneous size of lesion images for an automatic melanoma detection system as shown in Fig. 3.

Transformation of Color Space. Most of the automatic melanoma detection system considers color parameter as an extracted feature to predict melanoma. RGB color space uses the RED, GREEN, and BLUE spectral wavelength, while to imitate human visual perception of color, HSI and HSV color models are used. HSI and HSV models represent color in terms of hue, saturation, and intensity value. The calculation for these parameters is done based on average color wavelength, amount of white color, and wavelength, respectively. CIE-LAB color space provides uniformity to color, and CIE-XYZ color space is capable of generating positive tristimulus values for each color. A lesion image in a different color space is shown in Fig. 4.

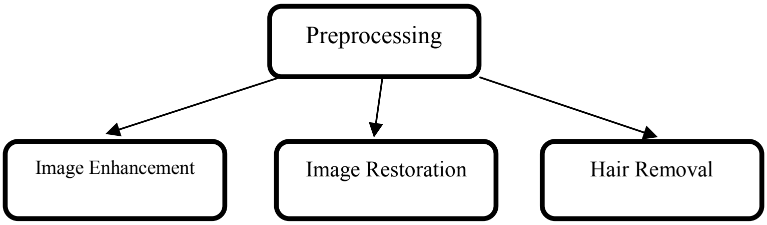


Fig. 2 Preprocessing substages for lesion images

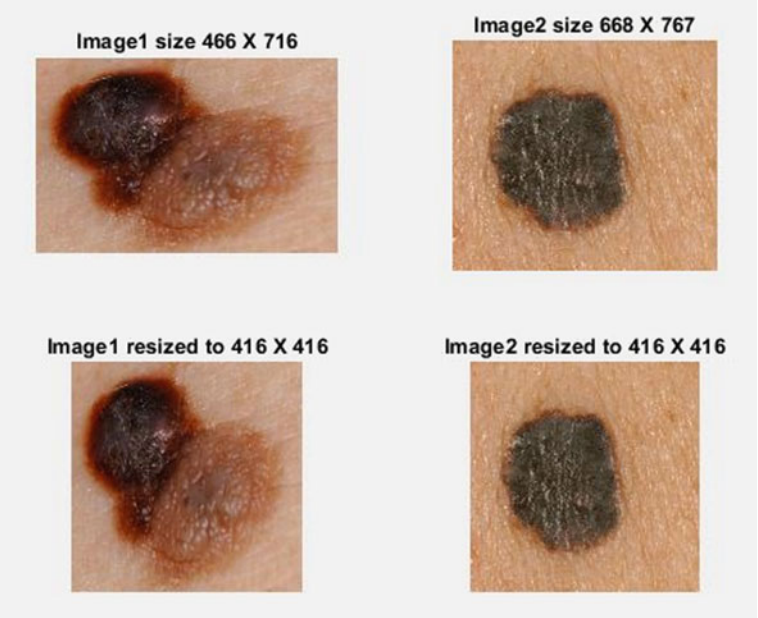


Fig. 3 Scaling operation performed on the lesion images for image normalization. *Image Source* MED-NODE dataset [10]

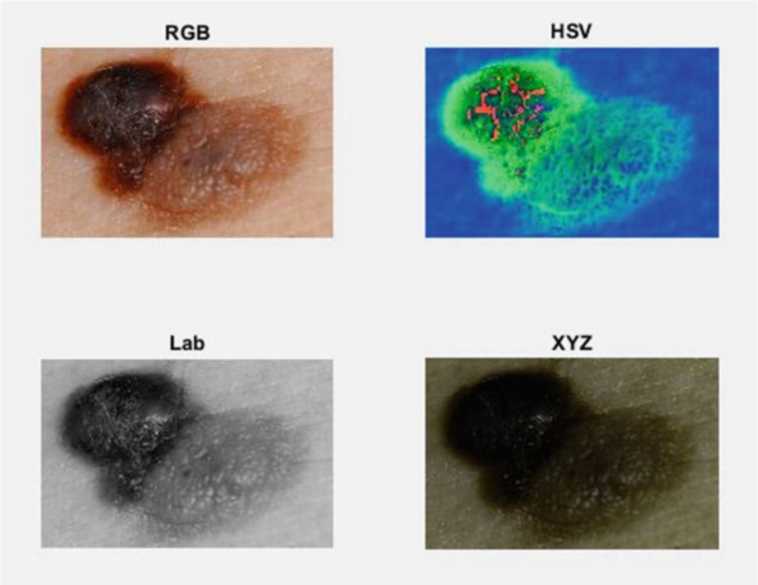


Fig. 4 Lesion image in RGB, HSV, lab, and XYZ color space. *Image Source* MED-NODE dataset [10]

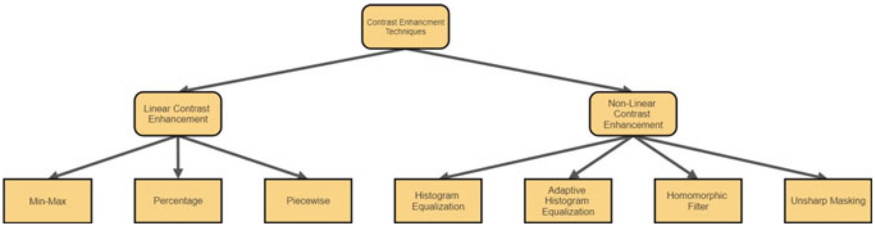

Contrast Enhancment Techniques
Linear Contrast Enhancement
Non-Linear Contrast Enhancement
Min-Max
Percentage
Piecewise
Histogram Equalization
Adaptive Histogram Equalization
Homomorphic Filter
Unsharp Masking

Fig. 5 Contrast enhancement techniques

Contrast Enhancement. The level of intensity in an image is termed as contrast. To sharpen the lesion border to improve the accuracy of the system, the image enhancement techniques are used. According to authors [11, 12] they play a vital role by creating a brightness difference between the foreground and background of the image. Contrast enhancement techniques are linear and nonlinear as shown in Fig. 5.

Contrast enhancement techniques working on the principle of contrast stretching are known as linear contrast enhancements. They remap the grayscale value of the histogram to be spread over the full range. While in nonlinear contrast enhancement, an input image is mapped to many output images at the cost of losing the correct brightness level for the lesion image. Min–max, percentage, and piecewise are the different types of linear contrast enhancement techniques. Histogram equalization, adaptive histogram equalization, homomorphic Filter, and unsharp masking are the part of nonlinear contrast enhancement technique. The impact of the enhancement technique on the lesion is shown in Fig. 6.

There are many types of contrast enhancement techniques. Here, we have studied and discussed contrast adjustment, histogram equalization, and adaptive histogram equalization only with the constraint that multiple enhancement schemes when applied will lead to more accuracy at the cost of more processing time. Also, as discussed in Fig. 6, the models reviewed for the lesion image enhancement are adequate as they provide a sufficient difference between the foreground and background of the lesion for the easy retrieval of the lesion border for further analysis.

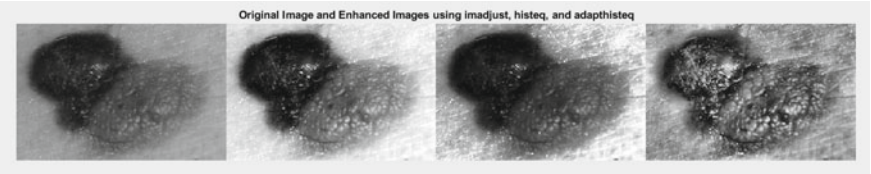

Original Image and Enhanced Images using imadjust, histeq, and adapthisteq

Fig. 6 Lesion image having the effect of contrast adjustment, histogram equalization, and adaptive histogram equalization. *Image Source* MED-NODE dataset [10]

2.2 Image Restoration

The process of recovering the original image from the degraded or blurred images known as image restoration. Reasons for image degradation include bad focus, motion, blur, and noise. Degraded lesion images will always lead to misleading results for an automatic melanoma detection system. A de-noising algorithm will help to recover the original image from the degraded lesion image. The restoration process is performed to recover the original image from a degraded image having noise or blur. A de-noising algorithm is considered good if it suppresses the noise with edge preservation [13]. Different types of filters along with their impact are visualized and compared [14– 17].

Noise-Based Restoration is applicable when the lesion image is having Gaussian and salt pepper noise. By reducing the sharpness of edges, noise reduction is achieved using a mean filter. To remove random noise, an adaptive filter is used. For sharpness preserving noise removal technique, order statistic filter is used. The different filter and their processing capabilities along with expected outcomes are listed in Table 2.

Blur-Based Restoration. Image formation process with imperfection leads to blurring images. For de-blurring of the lesion image, different techniques like Lucy Richard algorithm, inverse filter, and Wiener filter can be applied.

2.3 Hair Removal

The captured images of lesions contain thick hair. These thick hairs are responsible for misleading segmentation processes. Thick hairs present on the skin lead to wrong segmentation and even make the segmentation process difficult. So a prior removal of hairs from the lesion skin is required. To remove thick hairs, different techniques like morphological methods [18], curvilinear structure [19], inpainting methods [20], DullRazor [1], and a combination of bicubic interpolation with top hat transformation are applicable [21]. The hair removal using the DullRazor technique is shown below in Fig. 7 along with the corresponding hair mask.

3 Methodology

Lesion analysis can be done effectively if it undergoes preprocessing techniques such as image enhancement and filtering. Lesion image for the input to the automatic melanoma detection system if preprocessed properly leads to more accurate results. Datasets that are publically available for various researchers are the ISIC

Table 2 Filters and their processing capabilities

Filter	Description and processing capabilities	Pros	Cons
Arithmetic mean	It is a linear filter. It uses neighborhood pixel values to denoise the current pixel. It removes Gaussian noise	A larger filter size produces a less noisy image	Not perform well as reduces the high-frequency details
Geometric mean	It is a linear filter. A window is considered around the noisy pixel which covers the surrounding pixels as well based on which, mean is computed to denoise the given pixel. It removes Gaussian noise	Performs better than arithmetic mean	Strong filter leads to blurring
Harmonic mean	It is a linear filter. A sliding window is used to define the area around the noisy pixels. Removes Gaussian and salt noise	Removes positive outliers efficiently	Do not work with pepper noise and also blurs the image
Contra-harmonic mean	It is a linear filter, and its behavior can be controlled with the order of the filter. With the positive value of the order, it performs well with pepper noise, and with the negative value of the order, it performs well with salt noise. For zero-order value, it behaves as the arithmetic mean; removes Gaussian noise	Preserves edges	Cannot perform simultaneous elimination of salt and pepper noise
Median filter	It is a nonlinear filter. Uses mask to denoise the pixel. Removes salt and pepper noise	It is good for smoothing and preserves sharp details	Along with noise, sometimes it removes fine details as well
Max–Min filter	It is a nonlinear filter. Finds out the brightest or darkest point in the image	Remove pepper noise with max values and finds out the brightest pixel, removes salt noise with min values, and finds out the darkest pixel	Cannot perform simultaneous elimination of salt and pepper noise

(continued)

Table 2 (continued)

Filter	Description and processing capabilities	Pros	Cons
Mid-point filter	It is a nonlinear filter. Removes speckle noise, short-tailed noise, gaussian white noise, or uniform noise	Helps to compute the intensity value between max and min	Produces blurry images

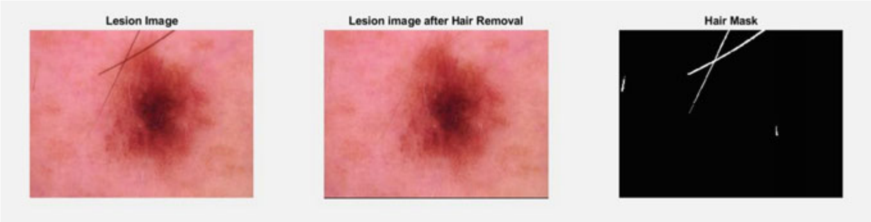


Fig. 7 DullRazor technique to remove hairs of the lesion. *Image Source* HAM10000 [22]

Challenge datasets. The International Skin Imaging Collaboration (ISIC) is an international repository of dermoscopic images to improve the diagnosis of melanoma. This challenge provides 25,331 images under eight different categories [22].

DermIS dataset is available for the educational purpose [11]. PH^2 dermoscopic dataset is provided by Hospital Pedro Hispano and contains 200 images available for educational use [4]. Similarly, the MED-NODE dataset contains 70 melanoma and 100 nevus images from the digital image archive of the Department of Dermatology of the University Medical Center Groningen (UMCG) [10]. These datasets contain images having nevus, melanoma, basal cell carcinoma, and squamous cell carcinoma. Images from these datasets are studied for signal-to-noise ratio (SNR) and peak signal-to-noise ratio (PSNR), and a comparative study is provided by adding salt, pepper, and Gaussian noise to different image enhancement methods. The higher values of PSNR and SNR show that the noise is having less impact on the images studied from the dataset. The impact of noise is calculated using SNR and PSNR on the images preprocessed using grayscale intensity mapping, histogram equalization, and adaptive histogram equalization.

4 Results and Discussion

A comparative study of different preprocessing techniques having image enhancement is provided and is compared based on SNR and PSNR. The study is carried out on the images taken from different publically available datasets including HAM10000, DermIS, PH^2, and MED-NODE. Image enhancement techniques such

Table 3 Impact of noise on ımage enhancement techniques

Image type	Noise	SNR	PSNR
Original RGB	Salt and pepper	16.2974	22.1867
Grayscale ıntensity mapped ımage	Salt and pepper	8.1439	14.0369
Image with histogram equalization	Salt and pepper	10.9840	16.8769
Image with adaptive histogram equalization	Salt and pepper	11.1137	17.0066
Original RGB	Gaussian	14.3037	20.1931
Grayscale ıntensity mapped image	Gaussian	8.0633	13.9562
Image with histogram equalization	Gaussian	10.6825	16.5754
Image with adaptive histogram equalization	Gaussian	10.2644	16.1573

as contrast stretching, equalization of histogram, CLASH, unshapen mask, and filtering techniques such as median, adaptive median, Gaussian, and wiener are studied.

The RGB image is converted into grayscale. After its conversion to grayscale, image enhancement techniques such as grayscale intensity mapping, histogram equalization, and adaptive histogram equalization are applied. For each preprocessed image, we added salt and pepper noise with a density factor of 0.02 which will affect the 2% of the available pixels. After each enhancement technique, salt and pepper and Gaussian noise are added, and PSNR and SNR values are computed as shown in Table 3.

From the above experiment, we observed that after the preprocessing techniques, the salt and pepper noise least affect the image processed with adaptive histogram equalization showing the highest value for SNR and PSNR, and the grayscale intensity mapped images are most affected by salt and pepper noise. Similarly, the Gaussian noise has the least impact on images preprocessed with histogram equalization and more effect on grayscale intensity mapped images.

5 Conclusion and Future Scope

AMDS is a noninvasive way to detect melanoma. The accuracy of overall AMDS is dependent on the preprocessing step; therefore, a meticulous review of preprocessing step is required. The objective of this paper is to study and discuss the stages of preprocessing steps in association with the impact of salt, pepper, and Gaussian noise on different image enhancement techniques. The images are compared using SNR and PSNR values. The SNR (signal-to-noise ratio) defines the ratio of average signal power to average noise power from an image. The PSNR (peak signal-to-noise ratio) is the ratio of peak signal power to average noise power. It also helps to identify the difference of pixel values from the ground truth. The objective also covers the study of datasets including artifacts like hairs and unequal-sized lesion images. Different publically available datasets contain different artifacts. The HAM10000 dataset-based images contain hairs on the skin which were processed using the DullRazor method. Datasets like MED-NODE contain images which are of different size, so such images must be scaled or normalized. For producing color-related features associated with lesions, the infected image can be represented in different color spaces like HSV, lab, and XYZ. The comparisons are done based on SNR and PSNR which is a well-known index for comparing the original image and demised image. Original RGB color-spaced image when preprocessed using adaptive histogram equalization has a high value of SNR and PSNR which is the lowest in the case of grayscale intensity mapped images in case of salt and pepper noise. Also, original RGB color-spaced images when preprocessed using histogram equalization have a high value of SNR and PSNR which is the lowest in the case of grayscale intensity mapped images in case of Gaussian noise. This paper gives ideas to researchers to select the best technique for the preprocessing of infected skin images to provide desirable and accurate results. Evaluation of efficiency in the further stages of the cancer detection system after proper selection of preprocessing technique can be considered as future scope of the paper. The future scope also covers the study of how the selection of a proper preprocessing technique and color space models for a specific type of noise affects the system's accuracy. The combination of filters on a noisy lesion image along with the preprocessing time taken for each fusion can also be considered as the future scope for the researchers.

References

1. Lee, T., Ng, V., Gallagher, R., Coldman, A., McLean, D.: DullRazor: a software approach to hair removal from images. Comput. Biol. Med. **27**, 533–543 (1997)
2. Mesquita, J., Viana, C.: Classification of Skin Tumours Through the Analysis of Unconstrained Images. De Montfort University Leicester, UK (2008)
3. Chucherd, S., Makhanov, S.S.: Sparse phase portrait analysis for preprocessing and segmentation of ultrasound ımages of breast cancer. IAENG Int. J. Comput. Sci. **38**, 2 (2011)
4. PH^2 Database: PH^2 Database (2013) [Online]. Available http://www.fc.up.pt
5. Hoshyar, A.N., Al-Jumaily, A., Hoshyar, A.N.: The beneficial techniques in preprocessing step of skin cancer detection system comparing. Procedia Comput. Sci. **42**, 25–31 (2014)
6. Cheng, H.D., Shan, J., Ju, W., Guo, Y., Zhang, L.: Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recogn. **43**(1), 299–317 (2010)
7. Michailovich, O., Tannenbaum, A.: Despeckling of medical ultrasound images. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **53**, 64–78 (2006)
8. Loizou, C.P., Pattichis, C.S., Christodoulou, C.I., Istepanian, R.S.H., Pantziaris, M., Nicolaides, A.: Comparative evaluation of despeckle filtering in ultrasound imaging of the carotid artery. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **52**(10), 1653–1669 (2005)
9. Tak Lau, H., Al-Jumaily, A.: Automatically early detection of skin cancer: study based on neural network classification. In: International Conference of Soft Computing and Pattern Recognition (2009)
10. Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M.F., Petkov, N.: MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst. Appl. **42**, 6578–6585 (2015)

11. DermIS: DermIS (2016) [Online]. Available https://www.dermis.net/dermisroot/en/17697/image.htm
12. Kong, F.W., Horsham, C., Ngoo, A., Soyer, H.P., Janda, M.: Review of smartphone mobile applications for skin cancer detection: what are the changes in availability, functionality, and costs to users over time? Int. J. Dermatol. (2020)
13. Radhika, N., Antony, T.: Image denoising techniques preserving edge. ACEEE Int. J. Inf. Technol. **01**(02) (2011)
14. Patida, P., Gupta, M., Srivastava, S., Nagawa, A.K.: Image de-noising by various filters for different noise. Int. J. Comput. Appl. **9**(4), 0975–8887 (2010)
15. Shinde, B., Mhaske, D., Patare, M., Dani, A.R.: Apply different filtering techniques to remove the speckle noise using medical ımages. Int. J. Eng. Res. Appl. (IJERA) **2**(1) (2012)
16. Hargaš, L., Hrianka, M., Duga, A.: Noise Image Restoration by Spatial Filters. Department of Electronics and Electrotechnology, University of Žilina, Slovaki (2007). http://www.urel.feec.vutbr.cz/ra2007/archive/ra2003/papers/376.pdf
17. Myler, H.R., Weeks, A.R.: The Pocket of Handbook of İmage Processing Algorithms in C. Department of Electrical and Computer Engineering, University of Central Florida, Orlando, Florida (1993)
18. Schmid, P.: Segmentation of digitized dermatoscopic images by two-dimensional color clustering. IEEE Trans. Med. Imaging (1999)
19. Fleming, M.G., Steger, C., Zhang, J., Gao, J., Cognetta, A.B., Pollak, I., Dyer, C.R.: Techniques for a structural analysis of dermatoscopic imagery. Comput. Med. Imaging Graph. Official J. Comput. Med. Imaging Soc. **22**(5), 375–389 (1998)
20. Zhou, H., Chen, M., Gass, R., Rehg, J.M., Ferris, L., Ho, J., Drogowski, L.: Feature preserving artifact removal from dermoscopy ımages. In: Proceedings of SPIE Medical Imaging: Image Processing, vol. 6914, p. 46 (2008)
21. Alina, S., Mihai Ciuc, C., Radulescu, T., Wanyu, L., Petrache, D.: Preliminary work on dermatoscopic lesion segmentation. In: 20th European Signal Processing Conference EUSIPCO (2012)
22. ISIC Database: ISIC Database [Online]. Available https://www.isic-archive.com/#!/topWithHeader/tightContentTop/challenges
23. Jeong, C.B., Kim, K.G., Kim, T.S., Kim, S.K.: Comparison of ımage enhancement methods for the effective diagnosis in successive whole-body bone scans. J. Digit. Imaging **24**(3), 424–436 (2011)

A Comparatıve Analysis on Three Consensus Algorithms

Proof of Burn, Proof of Elapsed Time, Proof of Authority

Aswathi A. Menon, T. Saranya, Sheetal Sureshbabu, and A. S. Mahesh

Abstract Blockchain technology has attracted immense attention in recent years from research, business, and governments all over the world. It is regarded as a mechanical advancement that is supposed to disrupt a few application areas that come into contact with all aspects of our lives. Nonetheless, a significant number of these blockchain implementations suffer from genuine flaws in their presentation and security, which must be addressed before any wide-scale acceptance can be achieved. The consensus calculation, which determines the exhibition and security of any blockchain system, is a key element. An algorithm for consensus is a tool that allows clients or machines to coordinate themselves in a decentralized environment. It must ensure that all nodes in the network will agree on a single source of truth even if a few nodes fail. They can usually make improvements, but there is not a complex administration structure in place to reach an agreement among multiple administrators. In a decentralized system, it is an entire other story. Assume we are dealing with a distributed database—how can we decide which records to add? Conquering this test in an atmosphere where nodes do not trust each other was perhaps the most demanding breakthrough in blockchain preparation. In this article, we will investigate how consensus calculations that are fundamental to the working of cryptographic forms of money and distributed ledgers. Therefore, a few current new agreement calculations have been presented to resolve the impediments of different blockchain structures. A methodical investigation of these equations can help us see how and why the way it operates is carried out by a particular blockchain. Consequently, in order to address the impediments of the various blockchain frameworks, some of the existing calculations of the new consensus have also been presented. A methodical investigation of these calculations will help to see how and why a specific blockchain plays out the way it works. In this article, we are talking about three consensus algorithms using an exhaustive empirical classification of properties and looking in depth at the implications of the various issues still prevalent in the agreement calculations.

A. A. Menon · T. Saranya · S. Sureshbabu (✉) · A. S. Mahesh
Department of Computer Science and IT, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_28

Keywords Blockchain · Consensus · Proof of authority · Proof of work · Proof of burn · Transaction

1 Introduction

Blockchain is a distributed, immutable ledger that allows the recording of transactions and the monitoring of resources in a business network. A blockchain is a type of computer database that stores data. Or, to put it another way, it is an open, distributed database. The data is distributed (i.e. duplicated) through several machines, and the blockchain as a whole is completely decentralized. A resource may be plentiful or scarce. This means that no one individual or agency has power over the blockchain; this is a drastic departure from the centralized databases that corporations and other organizations control and administer. In the broadest sense, the process is set up of data blocks, each of which is linked to the previous block, forming a chain. As a result, the term "blockchain" was coined. Each block, in addition to the data itself, includes a log of when the block was generated or modified, making it extremely useful for preserving a consistent system of record that cannot be manipulated or destroyed. Essentially, anything of significant value can be monitored and shared on a blockchain network, lowering risk and costs for all involved parties. A blockchain practically functions as a circulated and secure database containing transaction logs.

The innovation of the blockchain is an interface to data management and conversion. A few guidelines must be considered for this invention to be available. This assumes that if the blockchain works, the information would be accessible and functional even after years. The blockchain, as well as the data stored on it, must be unalterable. When data is applied to the blockchain, it should be difficult to remove or modify it. Finally, it circulates the architecture, which contains all of the data. The key aim is to create client trust in the network, which allows businesses and people to grow business. The oriented engineering keeps knowledge in a solitary location and distributes it to everybody. Everyone is associated with others in the decentralized architecture. Thus, regardless of whether we break a link with one node, the large variety of multiple hubs remains linked. Blockchain may be the perfect solution if you need to maintain a long term, open record of properties (e.g. to monitor assets or property ownership). Smart contracts, in general, are excellent for easing virtual connections and transactions. When the partners in a deal accept that their conditions have been fulfilled, automatic payments may be issued using a smart contract. The use of blockchain technology could eliminate the need for a middleman. TUI, for instance, is so persuaded by the potential of blockchain that it is reinventing ways to engage hoteliers and customers directly, so that they can transact through blockchain in a fast, secure, and consistent manner, rather than through a central booking platform. Blockchain is a smarter, secure way to monitor activity and keep data up to date while keeping track of its past. The data cannot be compromised or removed by someone, and you profit from both a past trace of data and an immediately up-to-date database. Given the vastly different demands of businesses and individual users, a

Table 1 Overview of different types of blockchain networks [1]

	Public/permissionless	Private/permissioned
Entry access	Open for everyone	For authorized ones only
Visibility of activity	Everyone	Selected nodes only
Nodes for consensus mechanism	All block generators	Selected nodes only
Level of trust	Not required	Required
Privacy	Low	High

single, standardized blockchain network cannot practically support all markets. This has led to the development of multiple blockchain networks, each with a completely different set of protocols, while the foundations remain the same. There are two significant kinds of blockchain public (permissionless) and private (permissioned).

Permissionless blockchain is blockchain that needs no permit to enter and connect with. They are otherwise called public blockchain. The first absolute form of permissionless blockchain is Bitcoin that uses proof-of-work consensus system (PoW). Ethereum (ETH) is another common public network that uses proof-of-stake consensus system (PoS). It incorporates smart contracts as well.

Permissioned blockchain requires permission to enter and connect with, not every user will join this blockchain. It needs the company manager or the owner to access the network with unusual permission. Ripple is one of the better examples of the permitted blockchain (XRP) (Table 1).

1.1 Fundamentals of Blockchain

Data is the lifeblood of business. The lifeblood of any firm is data. The quicker it can be accomplished and the more accurate it can be, the better. Blockchain is ideal for conveying the data because it offers instant, shared, and fully straightforward data stored on an immutable record that network participants can access. A blockchain network can keep track of orders, transfers, record generation, and much more. Furthermore, since individuals have a solitary perspective on reality, you can see all stages of a transaction from beginning to end, giving you more impactful confirmation.

However, regardless of the type of blockchain, each one needs a consensus calculation. Blockchain is so well-known today on the grounds that since the production of the Bitcoin, no one figured out how to create Bitcoin for itself. The hash function is known as this apparatus (SHA256).

A blockchain is a type of Merkle tree, a change-sensitive, hash-based information system in which hash values are gradually accumulated (hashes are hashed), and a large amount of information can be easily fingerprinted. Transactions are put inside a block as the leaves of the Merkle tree, and the root hash is placed in the block as a

Table 2 Certain vital features of blockchain [3]

Immutable	Once the data enters the chain, it cannot be manipulated or removed, enhancing data integrity
Decentralization	There exists neither prime authority nor prime focus of failure
Transparency	Network node has access to view the data entered in the blockchain
Pseudonymity	Blockchain supports anonymity
Chronology	Transactions are time stamped and traceable
Tamper sensitivity	Hash-based construction that permits effective discovery of even the littlest change

means of fingerprinting the whole trade assortment. Hash values are commonly used to link successive blocks. It is feasible to prove productively, by chaining progressive hash values, that the information stored in one copy of the blockchain is also indistinguishable from the information in another [2, 3]. Finally, the "nonce" field may be any number (arbitrary) that generates a hash value when hashed connected to the block, which is a satisfactory answer to the problem [4] (Table 2).

The following are the steps in the blockchain process:

1. **Transaction initiation**: At first, the user uses a hash function to hash the transaction data for later data integrity checking. Furthermore, the hashed data is encrypted using the user's private key to ensure user confidentiality, and the encrypted result is called the transaction's digital signature. Data and signatures from transactions are sent to the network.
2. **Validation of transactions and blocks**: Each node in the network executes two tasks: (a) decodes the digital signature with proposer's public key to authenticate the user and (b) characterizes the transaction data and compares the decrypted data with the signature recognition data. Confirm the transaction. They are then submitted to the network's block miners. Miners who are chosen view legal transactions (by consensus) and group them according to their block size. The miner sends the block to the network. The authentication node verifies as follows: (1) the block's hash value, (2) if the block timestamp is greater than that of the previous block timestamp, (3) the block's height and scale, (4) the expiry date of the previous block, and (5) all transactions in stack. Each validating node can add to its copy of the book a valid block [1].

A transaction in the blockchain is not considered legal until the network members achieve consensus using a consensus algorithm. In terms of scalability, complexity, cost effectiveness, and energy consumption, there is a difference between the existing consensus algorithms and the implementation criteria. Further enhancement and optimization are necessary for the performance, accuracy, scalability, and efficiency of the blockchain consensus mechanism. Bitcoin, Ethereum and hyper ledger consensus and code deployment methods are studied, debated, and suggested. In terms of computational complexity, fault tolerance, scalability, efficiency, and effectiveness, they vary.

2 Literature Review

A comparative study of three consensus protocols: proof of burn (PoB), proof of authority (PoA), and proof of elapsed time (PoET) bringing a retrospective analysis of the protocols. In accordance with a consensus protocol, blockchain is a distributed public registry that utilizes a tamper-sensitive, append-only data structure to allow mutually untrusting parties to establish a global set of states [5].

2.1 Brief Note on Proof of Work and Proof of Stake

There are several popular consensus algorithms in popular blockchain technology, such as proof of work (PoW) and proof of stake (PoS). Their computational complexity, fault tolerance, scalability, efficiency, and effectiveness differ [6]. The well-known consensus algorithm Bitcoin was the first to use **proof of work**. The convention sets out conditions for what makes a block substantial. It may say, for example, just a block whose hash starts with 00 will be legitimate. The solitary path for the miner to make one that coordinates with that combination is to brute force inputs. They can change a parameter in their data to create an alternate result for each guess until they get the correct hash. The PoW consensus is planned to be utilized in a trustless network, and it is fascinating to test its performance by adding flawed nodes into the system and running transactions over an unreliable network [7]. A PoW approach requires expensive mining equipment gadgets, and this strategy is set back by high power utilization. Since the POW technique is so resource escalated, it is not exceptionally proficient.

Proof of stake is an alternate option for PoW where massive energy for computation, and extensive hardware equipment was not required, but miners need to have a minimal amount of funds (Bitcoin) for staking while mining the block. You and the other validators will usually settle on which transactions will be used in the next block. In a way, you are competing on which block will be chosen, and the protocol will pick one for you. Proof of stake has mostly been used in the creation of smaller digital currencies. As a result, it is unknown if it will be a feasible alternative to PoW. Although it seems to be sound in principle, it can be very different in reality, as there still exist a loophole for retrieving the staked coins. When PoS is implemented on a network with a vast amount of revenue, the network becomes a game theory and financial incentives playing field [8].

2.2 Proof of Burn (PoB)

Proof of burn is the concept of transferring value through the destruction of value [9]. In 2014, Ian Stewart suggested proof of burn (PoB) to combat the high energy consumption of PoW and the problem of recoverable staked coins in PoS. The idea is that miners should be able to prove that they burnt any coins by sending them to an unrecoverable address. There is also a public key in this address that has no private key associated with it, so you cannot collect coins from that account. The coin will be removed from the network when sent to this address and can no longer be used. Since a miner must spend coins to mine a block, this deters malicious miners from mining invalid blocks. PoB is similar to PoW in that miners invest in mining computing resources to increase their chances of mining the next block; however, in PoB, miners burn more coins, which is equivalent to buying digital mining machines. To overcome early user domination, the value of burnt coins depreciates exponentially over time. Coin transactions sent to the unrecoverable address must be recorded separately from all other online transactions. After recording the transaction, SHA-256 will be used to calculate the burn hash value of each transaction, and the miner with the lowest burn hash value is eligible for mining. The burn hash is calculated using the following Eq. 1 [10],

$$\text{Burn hash} = (\text{Internal hash}) \times \text{Multiplier} \tag{1}$$

The internal hash is calculated by adding the hash value of the burnt transaction, and the time elapsed after the coins were burned and the current block number. The multiplier is inversely proportional to the amount of coins burnt, increasing the chance that miners would burn more coins. The multiplier value is exponentially raised to allow miners to continue to compete, thereby reducing the chance of miners winning over time. The multiplier value is determined using the following Eq. 2 [11], where T_b is the time that has passed since the coins were burnt, and T_d is the time that the coin will decay after that time.

$$\text{Multipliers} = \frac{e^{\frac{T_\mathrm{b}}{T_\mathrm{d}}}}{\text{Burned coins}} \tag{2}$$

2.3 Proof of Authority (PoA)

Proof of authority is yet another permissioned consensus algorithm that provides superior fault tolerance performance. The right to build new blocks in PoA is given to nodes that have proven their authority to do so. A node should pass a simple check to obtain this power and the ability to generate new blocks. PoA, which was proposed in 2015, is a reputation-based consensus protocol. Rather than coins, the protocol destroys the integrity of miners. In PoA, the role of miners is played by validators. The verifier (called the authorities) of the algorithm is officially approved by the node, which is publicly obtained through a series of identity cross-validation in the approved notarization system. To become a verifier, an network must maintain a good reputation and defend against malicious behaviour. Each validator creates a block in the process. They will earn a bad remark if reviewers behave maliciously and offer misleading injunctions. PoA is used in PoA trading platform and Vechain network. There is no incentive for PoA, but the authority can be encouraged by increasing reputation. A variant of PoA is proof of reputation (PoR), which uses well-known networks as validators instead of official identifiers. It is marked as an approved node on the network after the entity is authenticated, and the consensus is identical to PoA. A company's credibility is determined by its market appeal, the significance of its name, and whether the network is public or private. Gochain and Menlo One Stage Trade are currently using PoR. The estimation of PoA and PoR reduces the decentralization of the blockchain network since validators perform the mining. Additionally, they were not attempted now for its performance and assurance against security threats.

High-performance equipment is not needed. Contrasted with PoW agreement, the PoA agreement does not expect nodes to invest computing capital to solve complex numerical tasks. The time period of the production of new blocks is predictable. Of course, in the blockchain, the PoA aims to give greater weight to scalability, significantly raising network efficiency while lowering the degree of decentralization, making it particularly appealing for systems that need a greater "control" over the chain's protection through identity verification [12]. This timespan varies with PoW and PoS agreements. Fast transaction rate, blocks are generated by accepted network nodes in a series at the named time period. This speeds up validation for the transaction. Apla (a blockchain platform) executes a node boycott process and methods for violating the privileges of block creation. New blocks can only be created in Apla by the selected nodes called validating nodes. The blockchain network and the public ledger are operated by those nodes. The list of the validating nodes is kept in the blockchain registry. The order of nodes in this list determines the sequence in which nodes generate new blocks.

2.3.1 Steps in Block Creation

The validating node creates the new block as follows: (1) Gathers from its transaction queue all new transactions. (2) Executes one on one transaction. Transactions are refused if they are invalid or cannot be executed. (3) Checks block generation list compliance. (4) Creates a block of legitimate transactions and signatures it with the private key of the node (*ECDSA algorithm*). (5) Sends this block to other nodes that validate.

2.3.2 Steps for New Block Validation

Other validating nodes:

1. Receive and validate the new block: In the event that block approval is successful, add the new block to the node's blockchain. The leader node of the current span does not create another node. The block is created and effectively signed.
2. Execute transactions from the block individually. Check the transactions are proficiently executed as per the rules of block's generation.
3. Add or reject the block, depending on the previous step: If the block validation is successful, add the new block to the node's blockchain.

If the block validation failed, the block would be denied, and a bad block transaction would be submitted. If the validating node which generated this invalid block continues to produce such blocks, it may be prohibited or removed from the list of validating nodes.

2.4 Proof of Elapsed Time (PoET)

Proof of elapsed time is an alternative to proof of work (PoW). On account of PoW, to make an candidate block and spread the block to different nodes in the network, an intense computation is required. It requires special mining hardware utilized by the unique mining equipment (explicitly intended to measure the hash value) which is costly to mine the following block in the blockchain. In the kind of Bitcoin, the node that can discover the hash value initially turns into the new leader and get a reward [13]. Numerous other consensus frameworks have been researched and introduced in different forms since the advent of Bitcoin and the adoption of its proof-of-work (PoW) paradigm as the basis for the validity of a blockchain entity as a transmitted record. Certain new consensus model depends on Byzantine Fault Tolerance and principally centre around lessening the energy failures related with verification of work's mining escalated process. In the Hyperledger Sawtooth (Hyperledger is an open-source project that was created to propel cross-industry blockchain advances. It is a vital platform for money, banking, the Internet of Things, supply chains, and other technology [14]) project, PoET was first introduced. PoET is very similar to PoW, but it substitutes for pointless computing [5]. Essentially, each member of the network is given an abnormal clock object and the novice to lapse "awakens" the node, which transforms into the leader and creates the new block. PoET executes the waiting up in a trusted hardware module, the Intel Software Guard Extensions (SGX) accessible in numerous Intel CPUs. Each node basically calls an enclave (A protected area in an application's address space which provides confidentiality and integrity even in the presence of privileged malware [15]) inside SGX for creating an arbitrary delay and afterwards the one to complete its wait time announcing itself to be the leader in consensus. The platform makes a validation that can be utilized by any node to confirm that the leader effectively waited for the legitimate arbitrary time. Expecting the hardware module cannot be sabotaged, and this makes a similar sort of non-final agreement likewise with mining in PoW [16]. This PoW variants are called "lottery-style" arrangement calculations on the basis that there is a haphazard aspect that manages who governs the allocation of blocks and the extension of the chain in this way.

2.4.1 Process

(1) A node downloads the PoET code and uses SGX to create a credential (key) for the code. (2) When the node asks to access the network, it forwards this key. This key is checked by nodes that are already part of the network. (3) The new node now has its own timer object, which has a random value initialized. The code security provided by SGX guarantees this randomness. (4) At a random time, all nodes are initialized; the first one to expire becomes the winner. This means that a new block is formed, added to the current blockchain, and the reward is received. Then, the nodes are re-initialized.

Each node has to register two things with the system.

1. It is a public/private key pair that will not change in the future.
2. The other is the randomized timer that needs to be updated.

Each node first generates a number as its temporary time limit using a formula. These timeouts can be used to generate multiple blocks while an update is required. Specifically, it decides at random whether the next block will also be generated using this waiting time. By running the code in the TEE using SGX, random time wait is delivered, which produces a marked endorsement verifying the code execution in a trusted environment. By using the code running in TEE, each PoET hub offers a marked arbitrary wait up time and then rests during that duration. Estimation of the random wait up time will be made utilizing the formula determined in Eq. 3 [17] where minimum wait is a fixed framework attribute, local average wait is determined by utilizing the genuine number obtained from the hash calculation of the previous node authentication which is the sum of active blocks in the row, and $r \in [0, 1]$. The larger the number of active blocks in the network, the greater the node waiting season to hold away from the effects of the collision.

$$\text{Time} = \text{Minimum Wait} - \text{Local Average Wait} \times \log(r) \tag{3}$$

At the point when a node makes a block, it is checked by different nodes before it is acknowledged by the system. Anticipated attacks techniques can be prevented from basic verification. Since impermanent waiting time must be utilized a limit of 25 times, if a node utilizes it 26 times or more, the block created by this node ought to be discarded. In any case, once breached by the SGX, a refined attacker may decide to create blocks at an adequately quicker rate while as yet seeming to comply to the scheme. Measurable tests are used in the present situation to identify such an assault. The basic idea is to utilize the z-test to check whether the node is producing blocks excessively fast. The test expects that every node has a similar scoring probability p and that the quantity of chances of winning follows a binomial dissemination $X\ B(m; p)$, where m is the number of block in the network. At the point when m is adequately large, it is approximated by the normal distribution of $X\ B(m; p)$. A z-test is utilized to decide whether a win is predictable. A z-test is used to verify that a node wins steadily. The test assumes that in the event that the chain has m nodes, at that point every node has a similar winning probability p, with the number of wins following a typical dissemination, $N(mp, \sqrt{}(mp(1-p))$. The z-score for a hub is determined by utilizing below Eq. 4 [17],

$$z = \frac{\text{WinNum} - m}{\sqrt{mp(1-p)}} \tag{4}$$

Here, WinNum is the quantity of block effectively made by those nodes. If z exceeds the pre-defined parameter $z_{\max}$, the new block is rejected. POET assigns a variety of candidate $\mathrm{z}_{\max}$ values, including 1.645, 2.325, 2.575, and 3.075. This check is repeated several times from the last to first block on the chain. Transaction cost is received by the leader node making another block. The concepts underlying the Sawtooth Lake Scheme are relatively simple to grasp: (i) increase the waiting time when there are more active nodes to minimize the risk of collisions; (ii) must employ statistical tests to recognize a potentially corrupted node that produces blocks at a higher rate than honest nodes; and (iii) reduce both of them by using a random waiting time several times. PoET consumes less resources than PoW and does not benefit the wealthy. PoET employs an exclusive SGX facilities in any situation. In addition, since the arrangement depends on Intel's SGX facilities, it allows Intel the control role that activates a less decentralized blockchain. What is more, PoET is defenceless against malicious assaults. However, it is unclear how robust a PoET-based blockchain architecture can be used, since this often depends on the security of the underlying computing system. If a node is compromised, it is not needed to follow the pre-defined protocol, which can be used to weaken the entire system. Intel's Sawtooth Lake is only one example of a one-of-a-kind PoET implementation. In general, there is a lack of understanding of PoET at the protocol and theoretical analysis levels as opposed to other systems such as proof of work. The HyperLedger Sawtooth blockchain is currently using PoET. Hyperledger is an umbrella project, with twelve independent blockchain-related systems and instruments in its portfolio. To render Hyperledger the fastest-growing open-source project, the Linux Foundation has collaborated with business leaders such as IBM, Intel, and over 250 additional member organizations with extensive expertise managing open-source projects [14]. The Hyperledger systems are designed to be use case customizable and implemented in business networks, in comparison with the huge, public, single-instance Bitcoin, and Ethereum networks. The six blockchain systems in the Hyperledger umbrella are named Fabric, Sawtooth, Indy, Iroha, Grid, and Burrow. The remaining six projects are supplementary instruments: Calliper, Cello, Composer, Explorer, Quilt, and Ursa [5].

3 Experiment and Result Analysis

Blockchains are intended to give security to a wide range of information so nobody can mess with them. Yet additionally, simultaneously, they are discovered to be burning-through colossal measure of energy. In this way, different associations and organizations are attempting to discover the answer for tackling the energy issues of blockchains, and simultaneously, they are attempting to change to inexhaustible wellsprings of energy for manageable turn of events and for diminishing any ecological peril that is related with the utilization of blockchains. Blockchains are helpful whenever executed; however, to utilize it in each field of the economy, the labourers and the representatives should be given legitimate information on these blockchains. In this way, schooling is the primary factor for the usage of the blockchains.

A blockchain based on proof-of-authority networks appears to be in low demand for the user's computing resources and therefore does not pose a severe impact on the energy required to keep the network going. The transaction time of PoA networks is usually bigger than that of PoW-based networks. PoA networks are extremely scalable, particularly when compared to PoW blockchain, and are ideal for use as a platform. In certain ventures, this model will most likely become vital in the near future for its efficiency, audit, and coordination even in a circumstance where trust between participants is hard to build up. In a greater picture, the more robust and decentralized components, principally PoW, actually appear to be hard to replace. PoA will remain a valuable apparatus to consider until public blockchains improve their exhibition and versatility. Proof-of-authority networks have a severe lack of decentralization. In contrast to PoW blockchain, PoA blockchain can only have a restricted number of validators who are not democratically elected. PoA does not allow blockchain to be built in a way that is as immune to censorship and blacklisting as other consensus mechanisms doA limited number of host validators can openly combine to filter specific types of transactions based on the user's identity or the transaction's purpose. A person's fear of losing his or her reputation does not stop him or her from engaging in malicious conduct. The extent of the benefits that can be gained from a reputation-destroying event could be more advantageous than community reputation. This question also exposes the network to third-party intervention, leaving the possibility of covering the costs of the act's damage [18]. The consensus algorithm can be useful in special cases where protection and dignity must not be compromised. The Energy Web blockchain, for example, has a confirmation time of 3–4 s and can scale to several thousand transactions per second [19].

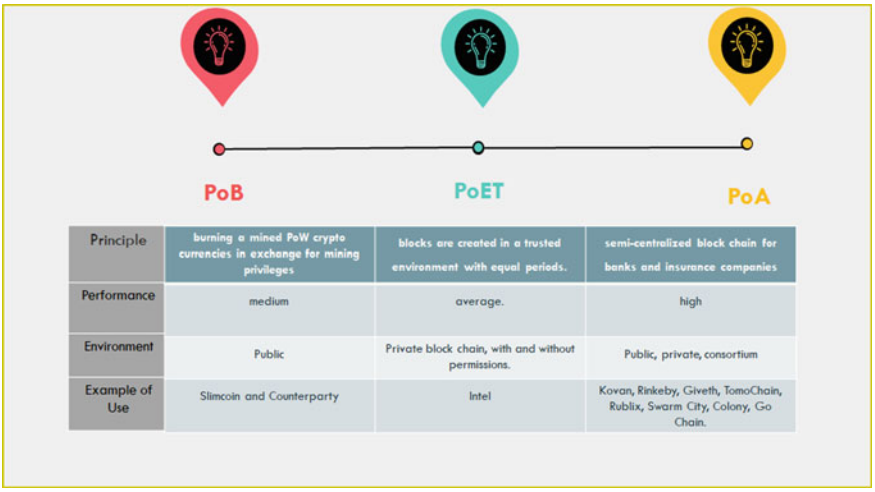
PoB hypothesis is that it does not generally consider building a sustainable long-term model to embrace. No measure of Bitcoin can ensure success. It additionally has the negative impact of eliminating Bitcoin from the complete stockpile. This will be likely to drive up the cost of Bitcoin, which might be useful for financial specialists temporarily. The ultimate objective, in any case, is for individuals to utilize something like Bitcoin as a mechanism of trade and not simply a store of significant worth. There is still time obviously for evidence of consume to substantiate itself, yet so far it appears to be that utilizing coin consume as a component might be more advantageous than similarly as a way to agreement. One point put forward for PoB is that it promotes a long-term commitment and time scale for a project. This potentially provides higher price stability for the cryptocurrency as long-term holders are less likely to sell or spend their coins. PoB is often said to be better than proof of work to make sure that assets are distributed in an equal, decentralized manner. PoB expects to be an alternative of PoW cost for verification by charging validator nodes, who pay in coins to earn the advantage of validating squares. Validator node submits coins that are "burnt" and cannot be recovered to expand the opportunity of being chosen by the arbitrary selection process. Approval relies upon the ability to waste money; subsequently, PoB brings about superfluous wastage of asset. Then again, centralization chances do not rely upon hardware equipment [15]. Although proof-of-burn assert that they do not use energy, researchers argued that proof of burn does cause resource waste as much as the resources used to produce the burned coins are wasted. There seems to be a strong notion to that found in the proof-of-stake consensus, where many more coins are already acquired by those who have a lot of coins. It is the rich who get the richer issue [20].

Furthermore, the advances encompassing SGX would have a significant impact on PoET as a practical consensus model. For the time being, it is an excellent tool for Hyperledger Saw tooth and plays an important role in experimentation and advancements within measured blockchain structures. In line with its SGX Network Guard Extension technology, proof of elapsed time (PoET) is a new innovative consensus algorithm created by Intel and deployed in the open-source Hyperledger Sawtooth blockchain platform. PoET is a form of PoW consensus that aims to reduce the amount of energy used by the latter algorithm [5]. PoET seems to be a major increase in the performance of proof-of-work systems. Simultaneously, it provides a better solution to the "Random Leader Selection Problem" but without the resource-intensive or complex mechanics and incentive mechanisms that are needed with proof-of-stake consensus. The obvious and unavoidable drawback here is the evident and unavoidable reliance on the security of specialized hardware. Not only that, but SGX is solely produced by Intel, expanding the consensus model's reliance to Intel as a business, a third party. The idea of such dependence contradicts the modern paradigm that cryptocurrency is attempting to achieve with blockchain, which is to eliminate trust throughout the system [21]. Yet network efficiency essentially declines as the scale of the network increases, with multiple components staying stable. The key critique levelled at this approach is the necessity of the Intel environment, which means that security must always be placed in a single authority [15].

Relatively, the previous block rate rises and is suggested as the explanatory clarification of reduced throughput. PoET is known to have good scalability but low consistency. At any point, a stale block occurs where a valid block is distributed by more than one node within a comparatively brief period. On the off chance that a huge replication block was as of late circulated however has not yet arrived at a portion of the network, making an network fork, a node that did not get this block will distribute and convey its own successor block.

Table 3 Performance metrics of the algorithms

	Speed	Energy consumption	Level of centralization	Security
Proof of burn	Normal	Very high	Very high	Average
Proof of elapsed time	Normal	Low	Very low	Low
Proof of authority	High	Very low	High	Average



	PoB	PoET	PoA
Principle	burning a mined PoW crypto currencies in exchange for mining privileges	blocks are created in a trusted environment with equal periods.	semi-centralized block chain for banks and insurance companies
Performance	medium	average.	high
Environment	Public	Private block chain, with and without permissions.	Public, private, consortium
Example of Use	Slimcoin and Counterparty	Intel	Kovan, Rinkeby, Giveth, TomoChain, Rublix, Swarm City, Colony, Go Chain.

Fig. 1 Characteristics of three consensus algorithms

4 Conclusion

We prepared a report of blockchain technology highlighting blockchain and its three consensus algorithms. The blockchain innovation was once acquainted longer than 10 years back with operating shared exchanges of computerized economic forms between a gathering of untrusted network members besides the requirement for an outsider. Over the long haul, blockchain was superior to creating decentralized applications past financial exchanges in quite a number of fields. Therefore, specific blockchain designs and consensus conventions multiplied. Recently, there has been an increased demand for a free, adaptable, scalable, and energy-efficient blockchain. This is due to the ever-increasing application requirements for higher administrations in a vast range neighbourhood biological environment. Thusly, blockchain engineering and consensus conventions have skewed with the objectives for a green community computerized environment. The required and contradictory action of cutting-edge models and policies, on the other hand, is the limit that provides the ultimate objective of the emerging computing culture. Thus, you want to alter it to accommodate the thought of dynamic use in addition to adaptive transformation as mentioned in the utility requirements. It can be helpful for the business sector to have a roadmap that encourages them to explore the field of blockchain technology. As a result, we proposed a framework for analysing blockchain consensus mechanisms based on different attributes.

Future works may include the hybrid of blockchain consensus (PoB–PoA) and discuss the efficiency and security connected with blockchain to provide a qualitative overview of various security mechanisms used to improve blockchain security and privacy concentrating on blockchain applications such as cloud computing and IoT. The motive of this article is to take a deeper glance at blockchain from the perspective of a fair urban collective environment. In short, the growth of a range of blockchain designs and practices is reviewed, and an overview of their properties, tasks, and barriers is provided. Further research is expected to build a usable blockchain gadget with specific collaborative efforts, reconfiguration, adaptability, flexibility, and energy use capabilities to bridge the gap between present buildings and agreements and reap the closing purpose of environmentally pleasant and realistic handling, expected for better customer management. This becomes one of the most innovative fields of study right now. Hopefully, this work may provide insights for future research in this field and that we may be able to participate to this rapidly expanding network. Our analysis uses secondary data, and we rely more on it than on primary data. They have been collected from a number of sources, such as blogs, journals, and academic papers.

Acknowledgements We would like to express our gratitude to all who have helped us directly or indirectly in our research. To start with, we thank the almighty for all his blessings showered on us during the tenure of our project. We are obliged to Prof. Dr. U. Krishnakumar, Director, Amrita School of Arts and Sciences, Kochi for giving this opportunity to complete the research. We express our sincere thanks to Dr. Vimina E. R., Head of the Department, Department of Computer Sciences and IT, for the support and encouragement. We extend our heartiest gratitude to our internal guide Mahesh A. S. for his valuable guidance. We are also grateful to all other members of faculty for their valuable guidance. Finally, we wish to express our sincere thanks to our family, friends and all the references and secondary data that helped us in preparing our project.

References

1. Ismail, L., Materwala, H.: A review of blockchain architecture and consensus protocols: use cases, challenges, and solutions. Symmetry **11**(10), 1198 (2019)
2. Lucas, B., Páez, R.V.: Consensus algorithm for a private blockchain. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). IEEE (2019)
3. Faber, B., et al.: BPDIMS: a blockchain-based personal data and identity management system. In: Proceedings of the 52nd Hawaii International Conference on System Sciences (2019)
4. Antonopoulos, A.M.: Mastering Bitcoin: Programming the Open Blockchain. O'Reilly Media, Inc. (2017)
5. Corso, A.: Performance analysis of proof-of-elapsed-time (poet) consensus in the sawtooth blockchain framework (2019)
6. Zhu, X.: Research on blockchain consensus mechanism and implementation. IOP Conf. Ser. Mater. Sci. Eng. **569**(4) (2019)

7. Porat, A., et al.: Blockchain consensus: an analysis of proof-of-work and its applications (2017)
8. https://academy.binance.com/en/articles/what-is-a-blockchain-consensus-algorithm
9. https://blog.iqoption.com/en/this-is-what-you-need-to-know-about-proof-of-burn/
10. Proof of Burn—Bitcoin Wiki. https://en.bitcoin.it/wiki/Proof_of_burn
11. Slimcoin Whitepaper. http://www.doc.ic.ac.uk/~ids/realdotdot/crypto_papers_etc_worth_reading/proof_of_burn/slimcoin_whitepaper.pdf
12. https://affidaty.io/blog/en/2019/08/blockchain-proof-of-authority-poa/
13. https://www.educative.io/edpresso/proof-of-elapsed-time-consensus-algorithm
14. The Linux Foundation (2018) [Online]. Available: https://www.linuxfoundation.org/press-release/2018/07/hyperledger-passes-250-members-with-addition-of-9-organizations/
15. https://sawtooth.hyperledger.org/docs/core/releases/1.0/architecture/poet.html
16. Cachin, C., Vukolić, M.: Blockchain consensus protocols in the wild (2017). arXiv preprint arXiv:1707.01873
17. Chen, L., et al.: On security analysis of proof-of-elapsed-time (poet). In: International Symposium on Stabilization, Safety, and Security of Distributed Systems. Springer, Cham (2017)
18. https://en.bitcoinwiki.org/wiki/Proof-of-Authority
19. Andoni, M., et al.: Blockchain technology in the energy sector: a systematic review of challenges and opportunities. Renew. Sustain. Energy Rev. **100**, 143–174 (2019)
20. https://www.coinbureau.com/education/proof-of-burn-explained/
21. https://blockonomi.com/proof-of-elapsed-time-consensus/

Issues and Challenges in the Implementation of 5G Technology

Mithila Bihari Sah, Abhay Bindle, and Tarun Gulati

Abstract The next-generation mobile communication network (5G) is a heterogeneous network, and it has the added advantage in the wireless communication field. Users will feel uninterrupted communication over the 5G network. It required a higher bandwidth in order to achieve a higher data rate. 5G is classified into three categories such as ultra-reliable low latency communication (URLLC), massive machine-type communication (mMTC), and enhance mobile broadband (eMBB) by the International Telecommunication Union (ITU). 5G will provide the higher data rate (Gbps), low latency, enhance quality of service (QoS), low energy consumption at a low cost per transmission, better spectral efficiency (SE), energy efficiency (EE), quality of service (QoS), improved throughput and better user experience. There will be so many challenges to achieve the above-mentioned factors. The main challenges are to reduce Interference, latency, power consumption, and enhance data rate. The paper highlights the different issues and challenges of 5G and compare different existing methodologies for mitigating these challenges.

Keywords 5G · Heterogeneous network · Millimeter wave · Coverage area · Interference · Latency · Spectral efficiency · Energy efficiency · Low power consumption

1 Introduction to 5G

5G is a heterogeneous network and also called next-generation network which consists of different small cells or nodes to provide a better quality of communication. 5G is a key enabling technology in wireless communication that completely changes the way of communication and supports automation in a different field. For the implementation of 5G network, a high frequency band is required, and 3GPP

M. B. Sah (✉) · A. Bindle · T. Gulati
Department of Electronics and Communication Engineering, MM Engineering College,
Maharishi Markandeshwar Deemed to be University, Mullana, Ambala, Haryana, India

T. Gulati
e-mail: gulati_tarun@mmumullana.org

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_29

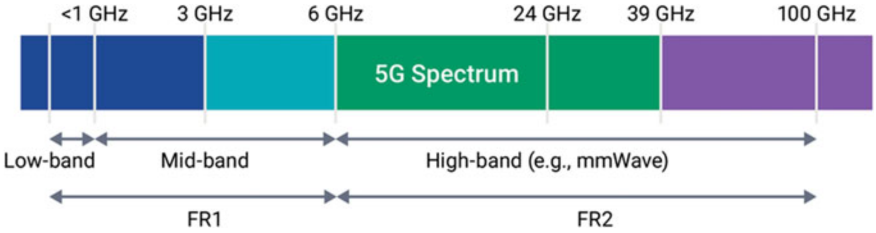


Fig. 1 Frequency band for 5G

defines the 5G frequency in the range of GHz that is from 3 to 300 GHz. The frequency band for 5G is split into two frequency bands, namely FR1 and FR2. The frequency band below 6 GHz is known as frequency range 1 (FR1), and the frequency band above 24 GHz is known as frequency range 2 (FR2) as shown in Fig. 1. Figure 1 also shows the three main frequency bands which are low band, mid-band, and high band. The frequency band below 1 GHz is called low frequency band, the range of frequency between 1 and 6 GHz is known as mid-frequency band, and the frequency above 6 GHz is known as the high frequency band.

Millimeter wave is the most enabling wave for the 5G network because of a high frequency band. Due to the short range of millimeter wave, different small cells are required to cover up the coverage area; therefore, different types of small cells are introduced like picocell and femtocell. The massive MIMO is used to deliver a more directed wave to each user. It has good directivity and sensitivity to obstacles.

5G provides good connectivity for ultra-dense networks and also provides ultra-low latency (approximately 1 ms), a high data rate in the range of GHz (from 1 to 10 GHz), high spectral efficiency, and high energy efficiency. Due to good connectivity, it makes everywhere a Wi-Fi zone. All the devices used in real life may be connected to the 5G network that is millions of devices connected to the 5G network per square kilometer. For the connection to the network, each device has its own specific IP address so IPv4 cannot provide the individual IP address to all connected devices because it can only provide IP address up to 2ˆ32 devices; therefore, 5G supports IPv6 to cover all the connected devices, and IPv6 can provide the IP address up to 2ˆ128 devices. 5G also supports worldwide wireless wave (WWWW). Section 2 overviews the background of the 5G technology and compares it with the previous technology like 3G and 4G. Section 3 explores the different issues and challenges of 5G networks like coverage area, spectral efficiency, latency, energy efficiency, and interference and compare different existing methodologies for resolving these issues and challenges.

2 Background

International Telecommunication Union (ITU) released the timeline and process which cover 5G technology like massive MIMO, heterogeneous networks (HetNet) composed of small cell base stations, and cloud radio access network (C-RAN). All these networks proposed to attained five times end-to-end latency, high data rate for high/low mobility, and expected to support ten times higher spectral efficiency. The concept of virtualization can be considered as a core network in 5G like network function virtualization (NFV), software-defined networking (SDN), and these networks can save on hardware and expand network flexibility [1].

International Telecommunication Union (ITU) splits 5G into three service categories, which are ultra-reliable low latency communication (URLLC), enhance mobile broadband (eMBB), and massive machine-type communication (mMTC) [2]. Figure 2 describes the three service categories of 5G and also explains the different features of 5G like high data rate, low latency, high mobility, ultra high density, and high complexity.

5G is the latest technology that is in the implementation stage worldwide. The data service is introduced from the third generation of the wireless network. Before the invention of the 5G technology, users achieve their required data by 3G and 4G networks which are not so fast as 5G but they fulfill users' requirements at a slow speed. The download speed of 3G slows as compared to 4G and 5G, and due to high latency in 3G, the buffering occurs most of the time while 4G has good speed and less latency as compared to 3G. Table 1 differentiates the three generations in a sense of different factors that are advanced in growing the technology from 3 to 5G. This

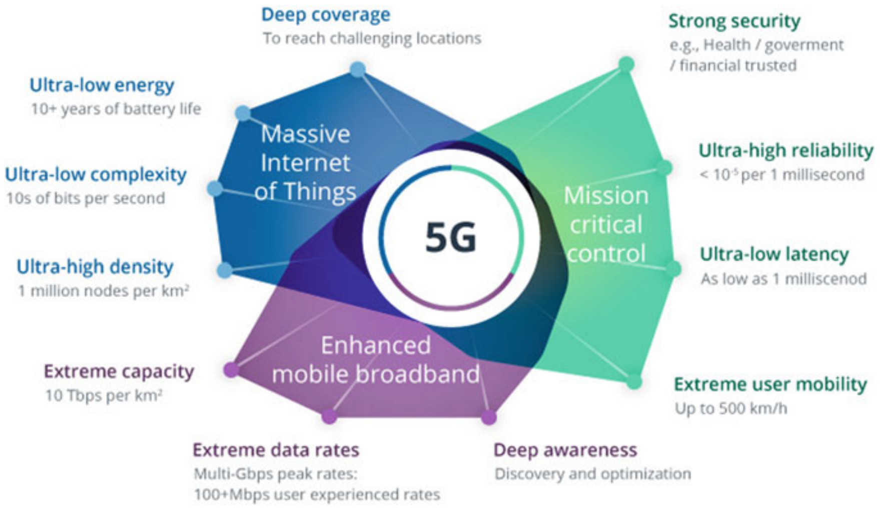


Fig. 2 5G service categories

Table 1 Comparison between 3G, 4G, and 5G

Comparison	3G	4G	5G
Introduced year	2001	2009	2018 or beyond
Frequency band	Low band	Mid-band	High band
Technology	WCDMA	LTE, WI-MAX	MIMO, mm wave
Access type	CDMA	CDMA	OFDM, BDMA
Bandwidth	25 MHz	100 MHz	30–300 GHz
Switching type	Packet switching expect for air interference	All packet switching	All packet switching
Speed	2 Mbps	Up to 1 Gbps	1–20 Gbps
Internet service	Broadband	Ultra broadband	Wireless world wide wave
Latency	100–500 ms	20–30 ms	< 10 ms
Handover	Horizontal	Horizontal/vertical	Horizontal/vertical
Core network	Packet network	Internet	Internet
Advantage	International roaming, high security	High speed handoff and global mobility	Extreme high speed, low latency
Applications	Mobile TV, GPS, video conferencing	Wearable devices, mobile TV, high speed application	Remote control vehicles, robots, medical procedures, high resolution video streaming

table includes the introduction year, frequency band, bandwidth, switching type, etc. Of different generations of the wireless network.

3 Issues and Challenges

5G has different challenges which are not practically achieved as expected theoretically. Some challenges like coverage area, spectral efficiency, latency, interference, energy efficiency, and power consumption are explained as follows.

I. Coverage area
II. Spectral efficiency
III. Latency
IV. Energy efficiency and low power consumption
V. Interference

3.1 Coverage Area

The 5G network is a heterogeneous network; it consists of different cells and node [3]. For 5G, a small cell is very useful because of its high user density and short wavelength due to its high frequency. There are mainly three types of cells described as a macrocell, microcell, and small cell. Small cell is also classified into two categories that are picocell and femtocell [3]. The deployment of such cells in a network will give a better result for coverage areas and connectivity.

Table 2 explains the coverage radius of different cells in a heterogeneous network, these cells provide, which type of communication takes place like indoor or outdoor or both. It also describes the consumption of power used for transmitting signals and the number of users in a cell that is supported by different cellular networks. These cells use a different type of backhaul; it may be wired or fiber or microwave.

Figure 3 explains the different areas, where different types of cell used for various applications. The femtocell serves as a home network, picocell provides the network to the large buildings or small outside area, microcell provides the network for urban, and macrocell supports the suburban region and provides good network service for outdoor communication.

To enhance the coverage of the 5G network, data channel and control channel should be considered. MIMO also enhances the coverage area along with the beam swapping and power boosting approach [4].

Table 3 shows some advantages and disadvantages of different techniques which are used to enhance the coverage area of the 5G network. The table represents the traditional method for achieving the coverage area, which is time consuming, cost inefficient, and required more resources as compared to the modern techniques like wireless sensor network, beam sweeping, and power boosting.

3.2 Spectral Efficiency

In 5G networks, for fast or ultra-fast delivery of data, spectral efficiency improvement is required to support smartphones and tablets [6]. Massive multiple input multiple output (MIMO) beamforming, zero forcing, and device-to-device communication (D2D) are few very useful techniques for improving spectral efficiency. Device-to-device communications have a very high spectral efficiency, but it is limited between two users only without involving the core network. Several multiplexing techniques are applied for the improvement of spectral efficiency like orthogonal frequency division multiplexing (OFDM), and it is also helpful to eliminate the inter-symbol interference (ISI) [4].

The massive MIMO hybrid beamforming system operates in a time division duplex (TDD). For achievable spectral efficiency, we study the combined effect of the quantized phase shifter, channel non-reciprocity, and channel estimation structure [7]. We can calculate the equation for spectral efficiency in both ideal and quantized

Table 2 Coverage area of different cell

Type of small cell	Coverage radius (m)	Indoor/outdoor	Transmit power (W)	No. of users	Backhaul type	Cost
Femto cell	10–50	Indoor	0.001–0.25	1–30	Wired/fiber	Low
Picocell	100–250	Indoor/outdoor	0.25–1	30–100	Wired/fiber	Low
Microcell	500–2500	Outdoor	1–10	100–2000	Wired/fiber/microwave	Medium
Macrocell	8000–30,000	Outdoor	10 to > 50	> 2000	Wired/fiber/microwave	High

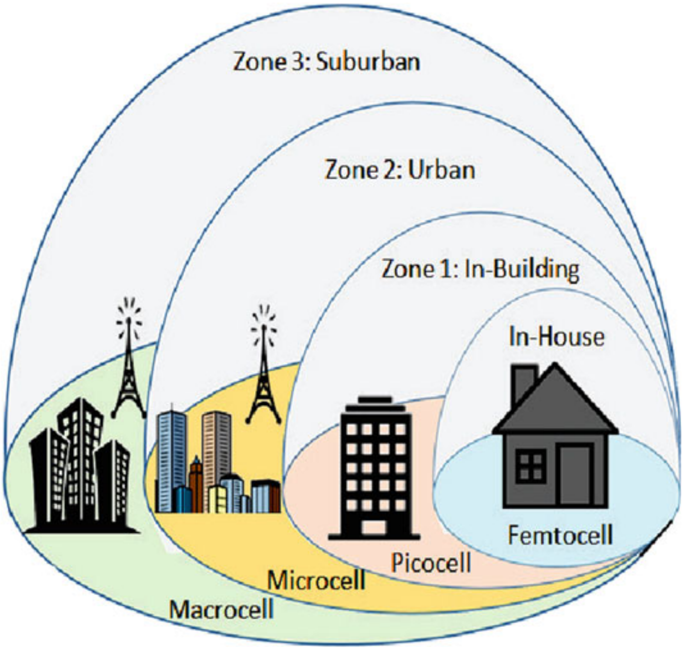


Fig. 3 Coverage area of different cell type

phase shifter cases and investigate the result [7]. The investigation result shows that the result of the quantized phase shifter is approximately close to the result of the ideal phase shifter which degrades the spectral efficiency [7]. Another technique for improving spectral efficiency is zero forcing. In a massive MIMO system, zero forcing allows accelerating performance through enabling an increase in dimension of antenna arrays which eliminate the processing problem and tolerance that is why a base station uses a huge number of antennas in multiple input multiple output (MIMO) system to achieve high throughput and more capacity [8]. Table 4 summarizes the two techniques for improvement of the spectral efficiency in the 5G network.

3.3 Latency

Latency is a time for end-to-end transmission of the signal. It plays a vital role in communication in the 5G network. It is greater than 100 ms in 3G and greater than 20 ms in 4G but in 5G, the requirement of latency is approximately 1 ms because 5G network can connect billions of devices, machines, and wearable devices to the network. Different automation devices, robots, and self-driving vehicles are innovated based upon 5G. Latency is the main concern for this achievement.

Table 3 Different techniques for improvement of the network coverage area

Techniques for coverage enhancement in 5G	Advantage	Disadvantage
Drive test	Give accurate coverage area. It is a traditional method. Use software tools for tracking the signal coverage [5]	Time consuming, costly process, and extra manpower required
Wireless sensor network	Easy to deploy at different places, no need to visit several times, less time consuming, and cost effective. This method is effective than the traditional drive test method. A network of sensors is formed, easy to calculate a coverage [5]	A hardware failure occurs on sensors due to weather changes. Power failures of sensor node Coverage is affected by faulty sensors
Beam sweeping	The user device receives continuous beams from the cell antennas and realizes a better network experience The multiple copies of the physical broadcasting channel (PBCH) and physical download control channel (PDCCH) are transmitted in the different beams in the interval of 2–5 ms. Eight beams can be supported for sub 6 GHz 5G NR system. The direction of the beam is specified and gives a more accurate coverage area [4]	It is quite challenging to achieve a high antenna gain More than eight beams cannot be supported for sub 6 GHz, 5G NR system
Power boosting	Power borrows from another subcarrier The total power of the carriers in each ODFM frame remains unchanged. It provides a better coverage area [4]	The power of the subcarrier, from which power is borrowed is reduced Only 3 dB power boost can be implemented without noticeable impact to the system

Table 4 Techniques for improvement of spectral efficiency

Techniques for improving spectral efficiency	Performance
Quantize phase shifter	In this technique, quantize phase shifter is used at beamformer, and the result of a quantized phase shifter is nearly equal to the ideal phase shifter which gives achievable spectral efficiency
Zero forcing	The zero forcing technique is used to mitigate the interference from the system which gives better spectral efficiency. With this technique, higher SIR is achieved that also increases the spectral efficiency. This technique is good for improving spectral efficiency

There are three service categories for low latency, and these are ultra-reliable low latency communication (URLLC), enhance mobile broadband (eMBB), and massive machine-type communication (mMTC) which are shown in Fig. 5 [9].

Enhance mobile broadband (eMBB) provides a high data rate from 10 to 20 Gbps in peak hour and in normal condition provides 100 Mbps. It is supported by macrocell, small cells, and provide high mobile connectivity about 500 Kmph. It also saves energy of the network by 100 times and applied to various applications like cloud computing, home broadband, television, AR, and VR.

Massive machine-type communication (mMTC) can allow connecting millions of devices per square kilometer and providing a network to large areas. It supports those devices which require a low data rate of about 1–100 Kbps and provide flexible machine-to-machine communication. The battery life is about 10 years for the devices which are driven by massive machine-type communication.

Ultra-reliable low latency communication (URLLC) is responsible for ultra-responsive connection and offers 1 ms on-air interface latency and 5 ms end-to-end latency between the user equipment and base station. It provides a low to medium data rate of about 50 Kbps to 10 Mbps and supports high speed mobility.

Figure 4 shows that the latency in 4G and 5G networks. From the figure, we can see that the latency is varying for 4G and 5G. For user equipment (UE) only 4 ms while for core network it is 1–2 ms in the 4 G network, but in 5G, latency is approximately 1 ms or below 1 ms for both UE and core network. This figure also shows that how the 4G cell can be divided into small cells toward the 5G network for better coverage and better efficiency for providing the best service to users.

Table 5 explains the latency requirement for different application area. The accepted packet data loss rate and also explain the key enabling networks to support the communications. These fields are supported by different types of communication environments.

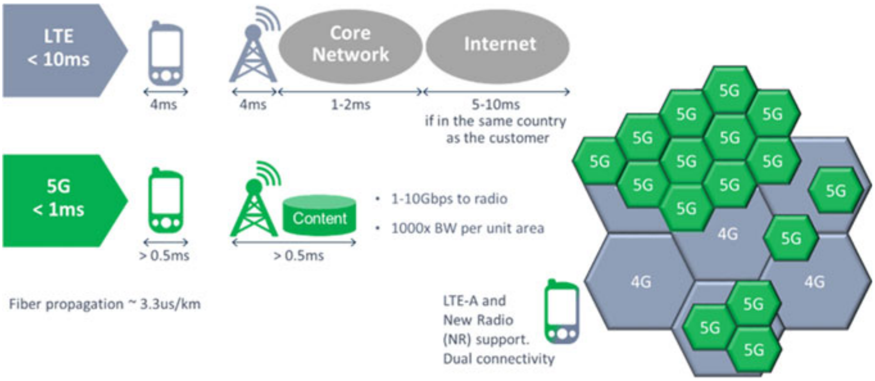
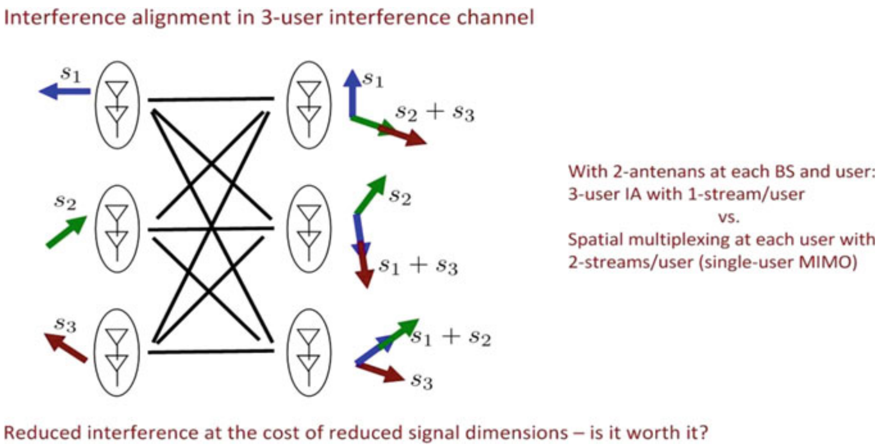


Fig. 4 LTE and 5G latency difference



Interference alignment in 3-user interference channel

With 2-antenans at each BS and user:
3-user IA with 1-stream/user
vs.
Spatial multiplexing at each user with
2-streams/user (single-user MIMO)

Reduced interference at the cost of reduced signal dimensions – is it worth it?

Fig. 5 Interference alignment for three users

Table 5 Highlights of 5G application and their requirements of ultra-low latency

Application	Expected latency (ms)	The acceptable packet loss rate	Key enabler
Factory automation	0.5–5	10^{-9}	The network is slicing and visual computing [10]
Intelligent transport system	100	10^{-5}–10^{-3}	Device-to-device and AI
Robotics	1	User defined	Haptic feedback
Virtual learning environment	5–10	User defined	Haptic communication
IOT/tactile Internet	1	10^{-9}	Haptic communication [2]
Virtual/augmented reality	1–4	10^{-4}	Haptic environment
E-health	1–10	User defined	Tactile internet and CODEC system [9]
Education and culture	5–10	User defined	Haptic communication [9]

3.4 Energy Efficiency and Low Power Consumption

The power consumption is a key enabling factor for 5G. In 5G networks, mostly small cells are deployed for delivering better service to the customers. Generally, the smaller cell (femtocell) is better for the home network, and it consumes very less power than the other cells for providing a good service in coordination with the base station. The power required for transmission in different cells (macro, micro, pico, and Femto) is described in Table 2. The base station consumes more power even there is no traffic to the network so the base station is always in active mode. 5G network resolve this issue; it makes it easy to reduce the power consumption in 5G for the base station. 5G allows the base station to go into deep sleep mode when no traffic on the network and suddenly resume in active mode even if any traffic load appears in the network so this property of 5G makes it more power efficient or energy efficient. Also, the same phenomenon continues in the case of 5G mobile phones and other 5G connected devices. The battery life of 5G phones and tablets is increased accordingly, and the average battery life becomes 10 years.

5G new radio network will consume very little power in very low traffic scenarios compared to LTE network when deployed in the same way and also provide the achievable energy saving with the extremely high traffic scenarios. With the increment in network capacity of LTE with NR microcell to additional LTE cell, the energy consumption reduced to 50%, while the LTE cell upgrade to micro-coverage to NR, the reduction in total energy consumption is up to70% [11].

3.5 *Interference*

The next-generation wireless communication network (5G) is an ultra-dense network, which provides a better signal to each user; for this purpose, massive MIMO beamforming antenna is used whose directivity is more than others and highly sensitive to signal. One of the main factors that affect the quality of the signal in 5G is interference. It may be inter-cell or intra-cell. There are different types of interference that occurs in 5G network like intercell interference, co-channel interference, adjacent channel interference, neighbor cell interference, and interference from other devices. There are very high interferences in 5G due to a huge number of connecting devices. The main challenge in the 5G network is to mitigate that interference from the network for better throughput and better communication. There are different techniques to mitigate the interference. Firstly, we apply the interference alignment, and after that, we apply the zero forcing or minimum mean square error to eliminate aligned interference.

Figure 5 represents the interference alignment process for three users. There are three transmitters and three receivers for transmitting and receiving signals. Signal S1 transmitted through transmitter1 and received by receiver 1, but except S1 other signals S2 and S3 are also coming to the reciever 1 from transmitter 2 and transmitter 3, so these two signals are interference for receiver 1 and aligned these signals in the same direction. These aligned signals have to eliminate. All the receivers have to receive only the signals from their corresponding transmitters.

The zero forcing pre-coding technique is also known as null steering. It is a method of spatial signal processing too in which multiple antennas at the transmitter end can cancel out the multiple user interference signals in a wireless communication network [12]. The precoder and decoder design based on the zero forcing technique is used to mitigate interference from the network at the receiver end [3]. Another method to eliminate the interference in 5G is a minimum mean square error (MMSE) which can be applied to the aligned interference [3]. Interference can also be minimized by using millimeter wave technology. For minimizing the interference, the route correction technique is applied, and it has three conditions for route correction. Route correction is applied when the base station predicts that the mobile users exceed the transmission range, the influence of the obstacle exceeds the standard, and the network traffic is too heavy [1].

Table 6 describes the different techniques which are used to mitigate the interference from the 5G network and compares these methodologies to each other. This comparison shows the zero forcing technique is better than others in sense of signal to interference ratio value and capacity.

Table 6 Different technique for mitigating interference and comparison

Techniques for interference mitigation	Explanation	Performance
Millimeter wave	This technique avoids the interference by using route correction means if any obstruction comes with a cell in the way of the customer, then it shift to another cell [1]	This quite ineffective if the neighbor cell is also busy or they have heavy traffic, then there is trouble shifting a load to other neighboring cells. This method is not so good as compared to others
Interference alignment	This technique is applied to align or equalize the interference at the receiver end to distinguish between signal and interference and separate them for avoiding interference [13]	This technique is quite effective than the millimeter wave technique
Zero forcing	This technique is very useful to mitigate the interference from the signal. This is applied in the form of a precoder and decoder. Calculate precoder matrix and interference suppression matrix to eliminate interference [3]	It gives a higher SIR (signal to interference ratio) than MMSE. Zero forcing has better performance and system capacities than MMSE because aligned interference is nullified completely
Minimum mean squared error (MMSE)	This is also applied to eliminate interference in the form of the mean square of the difference between receiving and transmitted signals. Interference can be reduced much more by using this method [3]	MMSE gives lesser SIR and fewer system capacities than zero forcing

4 Conclusion

In this paper, we highlight the different issues and challenges of the next-generation wireless communication network (5G network) and analyze different existing methodologies to resolve these issues and challenges from the 5G network to enhance the capacity and better network quality. The issues and challenges like coverage area, spectral efficiency, latency, power consumption, and interference are explored in this paper. This paper also overviews the basic characteristics and frequency range of the 5G network and compares fifth-generation with previous generations in sense of different terms, technologies, switching schemes, etc. As per the literature, different methodologies which are introduced earlier are analyzed, and some important conclusion has been put into tabular form in corresponding segments. There are some key technologies, which are applicable for the 5G network like massive MIMO, beamforming, small cell architecture, peer-to-peer communication, millimeter wave (mm wave), and visible light communication. After evaluating the different parameters in tabular form, our conclusion easily understands and compares or analyzes the issues and challenges in 5G communication and future work on the improvement.

5 Future Work

Many researchers and scientists are working in the field of next-generation wireless communication. Researchers are working to enhance the capacity of the heterogeneous network to provide a better experience in a sense of different issues like improving the coverage area and spectral efficiency, reduce interference, and low latency. This field has a lot of potential for young researchers who want to pursue their research based on future communication networks.

References

1. Wu, T.-Y., Chang, T.: Interference reduction by millimeter wave technology for 5G-based green communications. IEEE Access **4**, 10228–10234 (2017)
2. Siddiqi, M.A., Yu, H., Joung, J.: 5G ultra-reliable low-latency communication implementation challenges and operational issues with IoT devices. Electronics **8**(9), 1–18 (2019)
3. Imam, S., El-Mahdy, A.: Interference cancellation techniques in heterogeneous networks. Found. Comput. Decis. Sci. **43**(3), 153–180 (2018)
4. Liu, G., Hou, X., Huang, Y., Shao, H., Zheng, Y., Wang, F., Wang, Q.: Coverage enhancement and fundamental performance of 5G: analysis and field trial. IEEE Commun. Mag. **57**(6), 126–131 (2019)
5. Wang, H., Zhou, Y., Sha, W.: Research on wireless coverage area detection technology for 5G mobile communication networks. Int. J. Distrib. Sens. Netw. **13**(12), 1–11 (2017)
6. Al-Falahy, N., Alani, O.Y.: Technologies for 5G networks: challenges and opportunities. IEEE IT Prof. **19**(1), 12–20 (2017)

7. Chen, Y., Wen, X., Lu, Z.: Achievable spectral efficiency of hybrid beamforming massive MIMO systems with quantized phase shifters, channel non-reciprocity and estimation errors. IEEE Access **8**, 71304–71317 (2020)
8. Ali, A., Qureshi, I.A., Memon, A.L., Memon, S.A., Saba, E.: Spectral efficiency of massive MIMO communication systems with zero forcing and maximum ratio beamforming. Int. J. Adv. Comput. Sci. Appl. **9**(12), 383–388 (2018)
9. Al Amodi, A.M., Datta, A.: Towards on 5G services (Ultra-reliable low latency service; requirements, applications, and challenges). Int. J. Electr. Electron. Data Commun. **7**(6), 53–61 (2019)
10. Kelechi, A.H., Alsharif, M.H., Ramly, A.M., Abdullah, N.F., Nordin, R.: The four-C framework for high capacity ultra-low latency in 5G networks: a review. Energies **12**(18), 1–35 (2019)
11. Frenger, P., Tano, R.: More capacity and less power: how 5G NR can reduce network energy consumption. In: 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), pp. 1–5 (2019)
12. Sanenga, A., Mapunda, G.A., Jacob, T.M.L., Marata, L., Basutli, B., Chuma, J.M.: An overview of key technologies in physical layer security. Entropy **22**(11), 1–34 (2020)
13. Pappi, K.N., Karagiannidis, G.K.: Interference Mitigation Techniques for Wireless Networks, pp. 214–235. Cambridge University Press, Cambridge (2017)
14. Qamar, F., Hindia, M.H.D.N., Dimyati, K., Noordin, K.A., Amiri, I.S.: Interference management issues for the future 5G network: a review. Telecommun. Syst. **71**, 627–643 (2019)
15. Wang, J., Jin, A., Shi, D., Wang, L., Shen, H., Wu, D., Kishiyama, Y.: Spectral efficiency improvement with 5G technologies: results from field tests. IEEE J. Sel. Areas Commun. 1–8 (2017)
16. O'Connell, E., Moore, D., Newe, T.: Challenges associated with implementing 5G in manufacturing. Telecom **1**(1), 48–67 (2020)
17. Recioui, M.: Reducing Latency in 5G Network, pp. 1–89. Researchgate, UATL (2020)
18. Singh, S., Chawla, M.: A review on millimeter wave communication and effects on 5G systems. Int. Adv. Res. J. Sci. Eng. Technol. **4**(7), 28–33 (2017)
19. Ali, E., Ismail, M., Nordin, R., Abdulah, N.F.: Beam forming techniques for massive MIMO systems in 5G: overview, classification, and trends for future research. Front. Inf. Technol. Electron. Eng. **18**(6), 753–772 (2017)
20. Tripathi, A.K., Rajak, A., Shrivastava, A.K.: Role of 5G networks: issues, challenges and applications. Int. J. Eng. Adv. Technol. (IJEAT) **8**(6), 3172–3178 (2019)
21. Lakshitha, V., Rohith Reddy, K.V., Jayanth, G., Jagadeesh Raju, C., Manikandan, K.: 3G, 4G and 5G: a comparative study. Int. Res. J. Eng. Technol. (IRJET) **5**(3), 2911–2913 (2018)

IoT-Based Autonomous Energy-Efficient WSN Platform for Home/Office Automation Using Raspberry Pi

M. Chandrakala, G. Dhanalakshmi, and K. Rajesh

Abstract In this fast-paced world, every person likes to work in a faster way to satisfy their needs. IoT helps to accomplish this goal by connecting large number of devices as an automated system so that the extra works of humans are reduced. Smart home is one the familiar example for IoT technology. The huge development in IoT technology and the support of existing automation techniques are utilized in this research work to develop a home automation system. Raspberry pi-3 along with google assistance is incorporated in the proposed design to control the electrical devices in the home from anywhere. Proposed model is suitable for elder persons who cannot able to switch on or off the appliances so that it can be controlled over voice. Also it reduces the power consumption efficiently.

Keywords Raspberry Pi · Home automation · Internet of Things · Python · Wireless fidelity · Android phone · Raspbian operating system · Piezoelectricity · ThingSpeak · Wi-Fi shield

1 Introduction

Due to the emergence of communication technology, most of the appliances are automated nowadays. Most of the homes are evolving as smart homes by incorporating various technologies. Home automation introduces an increased computing power to control the electrical home appliances. In a smart home, more than one electrical device is controlled by using a single common device. User can control various electrical home appliances with the help of existing android-based mobile phone. Here, the control is transferred to the controller from the android mobile phone by incorporating the IoT concept. The objective of automation in home electrical management is to handle diverse control operations from a single device, and it controls the appliances from long distance, and at the same time, it is easily accessible. This automatic system not only provides a remote control but also helps in various

M. Chandrakala (✉) · G. Dhanalakshmi · K. Rajesh
Department of ECE, Siddartha Institute of Technology and Sciences, Hyderabad, Telangana, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_30

ways, which includes reduction in the power usage, saves money, and provides better environment.

In this venture, it plans to enhance the usability of controlling domestic computerization through webpage and server by utilizing Raspberry Pi. It concerns with altered control of light or any other residential machines. The client will communicate to Raspberry Pi through web through Wi-Fi organize. This framework is less exorbitant, permitting affordable e-home appliances. IoT or Web of Things is an up and coming innovation that allows the user to access the devices through Internet. Based on these, this research work proposed an IoT-based home automation model to control electrical devices through Internet.

Three loads are utilized in this framework to illustrate as home lighting and fan. Smart home consists of lights and other electronic gadgets controlled by a web-based home automation system. The devices can be controlled even from interior or exterior part of the home. Employing a Wi-Fi shield helps to act as an application for the Arduino to remove the requirement for wired association between Arduino board and web.

This venture discusses about the plan and improvement of enactment and further controls the home automation system through android. IoT is a promising solution for future along with the other trending technologies such as robotics or nanotechnology. The IoT is not new to this world; it is available with us for longer than 10 years. The Wi-Fi module ESP8266 (new ESP32) has really changed the world of IoT.

The Wi-Fi module used in the automation system is ESP8266, and it has integrated protocol for ICP/IP as SOC. It allows the microcontroller to access the network by hosting the application. Offload support is also available to control the network functions of other applications. The preprogrammed module with command set is used to control a modem.

Today automation is becoming a more popular phenomenon because of its less expensiveness. By connecting the personal gadgets, entertainment platforms and home appliances to a central automation unit (hub), it becomes easy to control everything from lights, temperature to music at any time by using the smart phone or a computer. The ThingSpeak Cloud platform creates a link between devices, objects and things inside campus, which can connect and work together for a common objective.

Another IoT platform that is widely used in various applications is ThingSpeak. It allows the user to visualize, aggregate and analyze the data streams in the cloud environment and provides instant visualization.

With the development of the Internet of Things technology, one of the most important features of the development of the information society is intellectualization. With the continuous improvement of people's living standards, the demand for enhancing the quality of life is gradually diversified. The appearance of the modern industry or intelligent kitchen will bring a new experience to people. People gradually need an industrial or domestic environment that can be closely connected with modern technology. With these demands, new concepts such as industrial monitoring, kitchen environment monitoring and intelligent kitchen control [1, 2] have come into our life. Therefore, this paper proposes a modern industry and intelligent kitchen model using

IoT. Another smart kitchen module reported in [3] is based on digitalized information and network, combined with smart phones and various sensors to realize intelligent management of industry and kitchens.

2 Literature Survey

Nowadays, everyone need a secured, reliable, user friendly and affordable peaceful lifestyle. Zigbee-based home automation system is reported by Gill et al. [1] controls various systems in a home. Raspberry Pi along with smart phone is used to control home appliance which is presented in research work [2]. IPv6 and 6LoWPAN-based automation system reported in Kovatsch et al. [3] research model performs better; however, the automation cost is quite high compared to existing techniques. Energy consumption is another important factor to be considered in home automation. The factors to be considered while designing are presented in the literature [4]. An automation system reported by Piyare et al. [5] used Arduino for simple and efficient home automation. Ramlee et al. [6] proposed a home automation system that integrates blue tooth and smart phone to support physically challenged persons as a fully automated system. The challenges in wireless sensor networks and its applications are reported in Rawat et al. [7] survey work. The power consumption monitoring application for home appliances is reported by Surya Devara et al. [8]. GSM, Internet and voice control-based automation system is reported by Yuksekkaya et al. [9] that used GSM module to control the SMS.

The thought of automation can be dated to October 2014, when Vikas Kumar, Nitish Bansal Gulam Hussain and Kunal Khivensara created the thought of accepts a further control. This venture expected which is exceptionally diverse than existing system. They were getting to execute with the assistance of specifically Wi-Fi/Wireless technology which fits the 802.11 WLAN charge benchmarks. The most asset of this framework is that it can be actualized with a more extensive range of not more than 200 m. It permits communicating with a brief and little setup without wired association. This system may well be amplified for an appropriate HVAC.

Aqeel-ur-Rehman et al., say that home automation is one of the developing system to change the life style of the people. Different types of home automation systems are available in the market. Some system targets the luxuries people, and others target only the people those who are needed like elderly people or physically challenged. The authors developed a new system to control various electrical appliances using wireless communication concept. This system is small in size and easy to install and maintain. Command identification is one of the important powerful interfaces between user and devices.

Elderly or physically challenged people need help to move from one place to another. Because of this reason home automation system was developed. The users can do their particular work by using commands to a Humanoid [10]. Al Shu'eili et al., explained about the development of home automation system. Home automation system provides sufficient help to the old people, physically challenged ones and those who live unaccompanied. This system deals with the overall design of the automatic home system. This system has been constructed by using low-power RF ZigBee. Research work provides better benefits by controlling multiple home appliances through a simple command message. This system was developed and tested in real time. In this test, 1225 command messages are tested. It recognizes 78.8% of the commands in proper manner [11].

Sitaram Pal, et al., describe the future usage of home automation. The home automation system is extremely useful for controlling different electrical home appliances. It controls certain tasks in automatic manner and regulates properly. This system is also cost-effective and reduces the usage of energy. Here, the user can send the control command message through their android phone. IoT concept creates the communication between user mobile and the microcontroller [12]. Neelima et al., designed a new system to control electrical appliances in home by using command. The main objective of the research work is to control the home electrical devices using smart phone with Arduino controller. This happened due to the growth of the communication technology. By using this technology, most of the houses become a smarter.

In conventional homes, switches are located on the wall. Manual operation is needed to on the switch. It is very difficult task for old age people and disabled. IoT-based home automation system eliminates these issues, and it is easy to control with single command or operation [13]. Neha et al., say that in current scenario, most of the people control the electrical devices manually. To control the electrical devices, more workforce is required. Due to this reason, the maintenance cost is also increased. To overcome this type of problem, devices are controlled by using wireless communication technology. Most of the devices are controlled by using IoT technology. This system is helpful for old people and physically challenged people those who are not able to move from one place to another to operate the switches on the wall.

This proposed system is controlled using commands. The main benefit of this proposed system is to reduce the manpower, time and problems which occur due to human negligence [14]. Sonali Sen et al., say that automation is a new and developing idea in twenty-first century. Due to the growth of current communication technology, smart phones are used to accomplish various tasks. Most of the applications are developed by using android operating system that processes the human command messages and gains the ability to control various devices. In this paper, authors have presented a new home automation system, which is controlled by using a command message. The entire activities are controlled by using an Arduino UNO microcontroller. The link created between the user and the device with the help of IoT concept [15].

Chandra Shakher Tyagi et al. explained the importance of computing environment. Currently, most of the people use their smart phones for accomplishing their day-to-day activities. With the help of mobile phones, people are performing various tasks. In this article, the authors have designed a new home automation system by using microcontroller and Bluetooth. This system is controlled by the human command commands. Various electrical devices are controlled by this proposed system. This system entirely replaces the existing wall switching system. The major benefit of the proposed system is to control various devices with the help of a single controller [16]. Abdul Aziz Md et al. have explained the importance of current communication technology called Internet of Things. This paper mainly concentrates on constructing a home automation system by using android mobile phone [17].

3 Existing System

In existing system, device controlling for home and office is done through manual mode of operations. It consumes time and complex in terms of manpower. All existing models like switching, Bluetooth, RF communication consume high power, and cannot be operated for longer distances.

To avoid all the disadvantages of existing systems, it is essential to develop a smart automation model with IoT using efficient Python programming and provide better control of home devices.

4 Proposed System

This proposed system is used to control the various electrical home appliances with the help of single controller. It is constructed by using Raspberry processor, android smartphone and IoT concept. The IoT creates the communication between controller and android mobile. This device is low in cost compared with other home automation systems. This proposed system can monitor and control devices through smart phone or tablet. This system is very helpful for elderly and disabled population.

It consists of two main parts. The first part is command identification system, and the second is wireless communication. Due to the growth using PC, Internet, smart devices and wireless communication technology, the user can easily access and manage home appliances from remote location. In the proposed system, command messages are used to control various appliances. The user can provide the command from the android-based phone. IoT technology creates a communication between the devices and user's phone. Using a single device, the user can control more than one device. Architecture of proposed smart automation model is illustrated in Fig. 1.

Home automation refers that a process that controls all the appliances in a home in an automated manner either by program or by human commands. Appliances like light, kitchen timer, fan, alarm and air conditioner with control techniques can be collectively termed as home automation. Devices can be controlled from anywhere through cloud under Wi-Fi or other Internet sources.

Earlier home automation system is developed with Arduino to control the devices with remote or by utilizes digital control and touch screens. However, it has few practical limitations while implementing in real time so to overcome the issues in existing systems, Raspberry Pi modules are introduced that has numerous advantages.

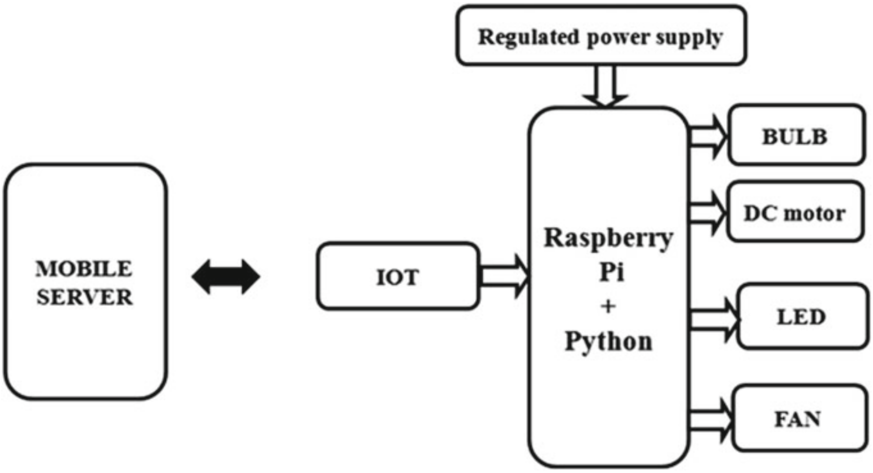


Fig. 1 Block diagram

It has an embedded Wi-Fi module as an in-built design so that web connection can be established easily. As of this, multiple Internet-based devices can be controlled efficiently. The home automation utilizing the IoT has been demonstrated to work by interfacing straightforward apparatuses to it, and the apparatuses were effectively controlled through web.

The outlined system forms agreeing to the necessity, for illustration exchanging on the light when the command is allowed. This will offer assistance to the client in order to urge an overview of different parameters within the home at any time and at any place. Low cost and adaptable home computerization system are developed by utilizing Raspberry Pi. By actualizing this sort of framework, the energy conservation has been guaranteed. The total control will be over the domestic machines from anywhere. This increases the comfort of humans and decreases the human efforts.

4.1 Raspberry Pi

The Raspberry Pi is a small single-board computer, which is especially designed for teaching purpose. It remains compatible with the operating system like Android, Linux, net BSD, RISCOS, Window-10 and ARM 64. It has a USB power port of 5 V, 3 A which can deliver full power to USB devices. Raspberry Pi comes with in-built CPU, RAM, Wi-Fi and Bluetooth.

Hardware Review
Left side 4 USB ports and one Ethernet port and at the bottom, a Power adaptor point, HDMI port and Audio Jack. Further, a micro SD card at the back of the PI and 40 GPIO pins to do real time projects, and other side it has two expansion slots.

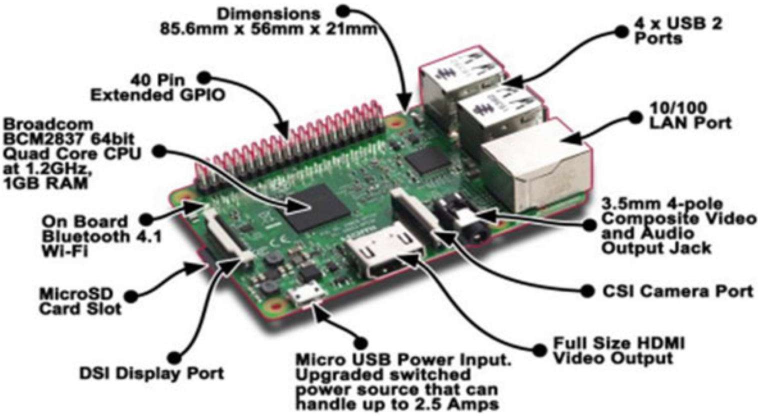

Dimensions
85.6mm x 56mm x 21mm
4 x USB 2 Ports
40 Pin Extended GPIO
10/100 LAN Port
Broadcom BCM2837 64bit Quad Core CPU at 1.2GHz, 1GB RAM
3.5mm 4-pole Composite Video and Audio Output Jack
On Board Bluetooth 4.1 Wi-Fi
CSI Camera Port
MicroSD Card Slot
Full Size HDMI Video Output
DSI Display Port
Micro USB Power Input. Upgraded switched power source that can handle up to 2.5 Amps

Fig. 2 Raspberry Pi

Raspberry Pi comes with

- CPU: 1.2 GHZ Quad-core ARM Cortex
- GPU: Broadcom VideoCore4
- RAM: 1 GB SDRAM
- 802.11n onboard wireless LAN
- In-built Bluetooth 4.0

In order to run the PI the corresponding OS must be installed on the SD card. Raspbian is the official supported OS for PI. The OS can be downloaded from Raspberry Pi official website.

https://www.raspberrypi.org/software/

To install Raspberry Pi operating system and other operating system to a micro SD card, an imager tool is available, namely Raspberry Pi imager. It can be downloaded from the Raspberry site (Fig. 2).

4.2 IoT Module

Internet of Things used for controlling any device or monitoring the device status through Internet. This proposed system uses this IoT module for taking the all parameters data and post into the cloud called server. ESP8266 modules as IoT module it can operate through Wi-Fi frequency concept (Fig. 3).

In various IoT applications, ESP8266 Wi-Fi module is widely used. There are several versions in this model such as ESP-01 to ESP-11 ESP-12E and WeMos D1 Mini. It can be used to develop a web server to handle HTTP requests, real input and interrupts, send mails and control outputs.

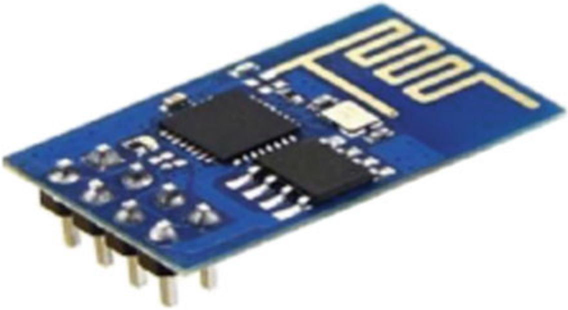


Fig. 3 ESP 8266

Fig. 4 16 × 2 LCD

4.3 LCD Display

LCD display is used to display the operation status. The basic display module can able to convert and display 16 characters per line as 16 × 2 display. 5 × 7 matrix format is used to display each character. This can display 32 characters having 2 columns. When each sensor is activated, corresponding massage will be displayed in 16 × 2 LCD modules. In this, we use four data pins; using this pins, we transfer the data from micro preprocessor to LCD (Fig. 4).

4.4 Buzzer

Buzzer is a speaker that is of small size and is also an audio signaling device. Piezoelectricity is the effect that is used here where the crystals will change its shape when the electricity is applied to it. Now, when the electricity is applied with the correct frequency, the buzzer makes a sound. The buzzers are used in various fields for the usage of alarm devices, timers, etc. The buzzer works as the following where the tone which is present sends some frequency of 1 kHz to a particular pin used, and delay is used to pause it for particular seconds, and then, buzzer stops the signal. This is continued for making a short beep sound from the buzzer (Fig. 5).

Fig. 5 Buzzer

4.5 Relay Module

Relay is used in the proposed design as switch that operates based on electromagnetic principle. Circuit connection can be established or closed through the relay units. It avoids manual switching operation, and it has two types such as small power and high power relay. Relay circuit includes an iron core that was surrounded by a control coil. Once the supply is given to the coil using the load contacts and switch a magnetic field will be produces to perform the operation (Fig. 6).

There are two main circuits in a relay. They are primary and secondary circuits. The primary circuit provides the control signal to operate the relay. The primary circuit is generally connected to a low-voltage DC supply. The secondary circuit is connected to the load which is being controlled.

Primary side has an electromagnetic coil, which generates electromagnetic field when current passes through it. At the end of the electromagnet, an armature is found. This is a small component which is pivoted when the electromagnet energizes, it attracts the armature. When the electro magnet is de-energized, the armature returns to its original position. Typically, a small spring is used to achieve this. A movable contactor is connected to the armature, when the armature attracted to the electromagnet, it closes and completes the circuit on the secondary side.

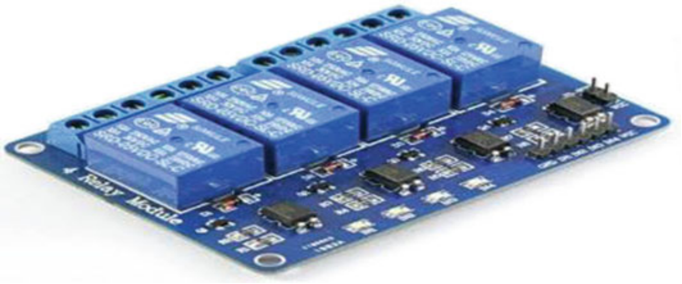


Fig. 6 Relay circuit diagram

4.6 Software

Python and machine learning-based algorithms used for the development of proposed framework. Python IDE used to write code and compile.

Raspbian operating system is used for developing the desired model.

5 Results and Discussion

The proposed home automation model performance is experimentally verified by the real-time testing. The figure illustrates web controller dashboard of the proposed home automation system.

Figure 7 illustrates that all input and output hardware modules relay, bulb, fan, motor which is controlled by input Wi-Fi module in build in Raspberry Pi 3 module. All the input and output modules operated with 5 V dc power supply. Figure 8 depicts graphical user interface (GUI) for user to switch ON or OFF home alliances from anywhere. The coding was developed for the proposed home automation system given in Fig. 7 and deployed to Raspberry Pi.

In Fig. 8, controller dashboard is designed with ON/OFF facilities. By clicking the controller tab, two buttons called ON and OFF will be displayed. Now, automation system becomes easy by controlling the bulb with ON/OFF button over the Internet. Application program interface call turns the state of the digital pin to HIGH or LOW. The state of this pin is further used to switch the relay either ON or OFF. This API call is activated when you press the button on the browser. When the relay is in OFF state, it breaks the circuit between the bulb and power supply. It results when the bulb goes OFF. When the relay is in ON state, it completes the circuit connection between bulb and power supply as a result of which bulb lights up. In the proposed home automation system, the controller is used as the dashboard of cloud Bolt IoT in which all actions of the hardware were considered. It takes the instructions, processes on it and gives the result of the instructions. When the relay is in ON state that means, it will activate the home appliances, and in similar way, when the relay is OFF, the controller will deactivate the product.

6 Conclusion

In olden days, the electrical appliances are controlled with help of switches. The switches are put on the wall itself. To operate the switches, man power is required. But the elderly people and disabled persons are not able to move from one location to another place. To avoid such kind of problem, a new system is constructed for home automation system. This system is controlled by sending command message from the user. The user provides the commands through their android mobile phone. Through

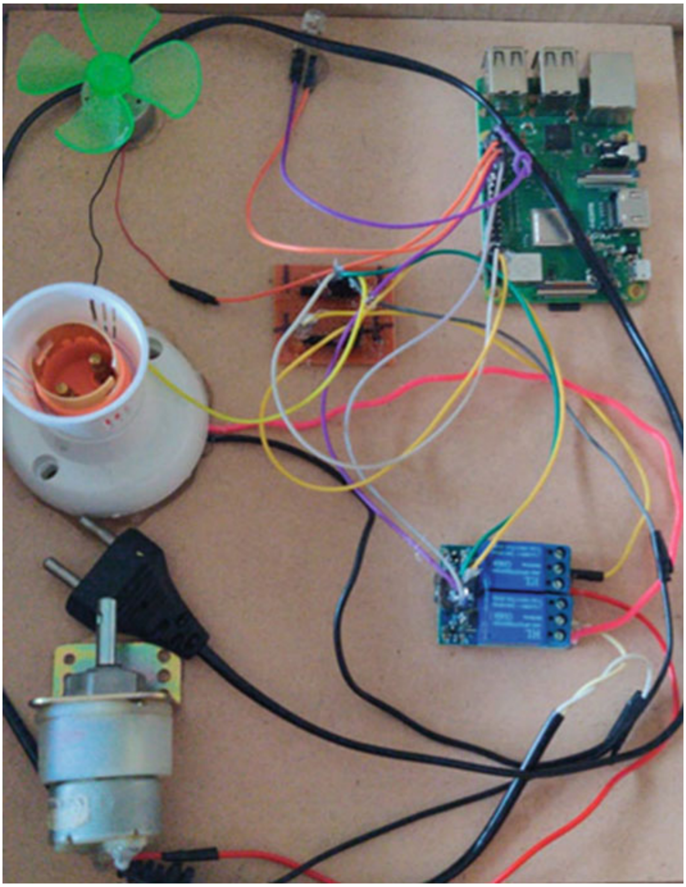


Fig. 7 Output hardware model

IoT technology, the command messages are transferred to the specific electrical devices. The proposed home automation system controlled by command has been successfully implemented and tested with real-time data. This system is developed by using Raspberry Pi processor with IoT module. This system can be used to control n number of devices with a single controller. Proposed model reduces the human efforts and eliminates the excess power consumption.

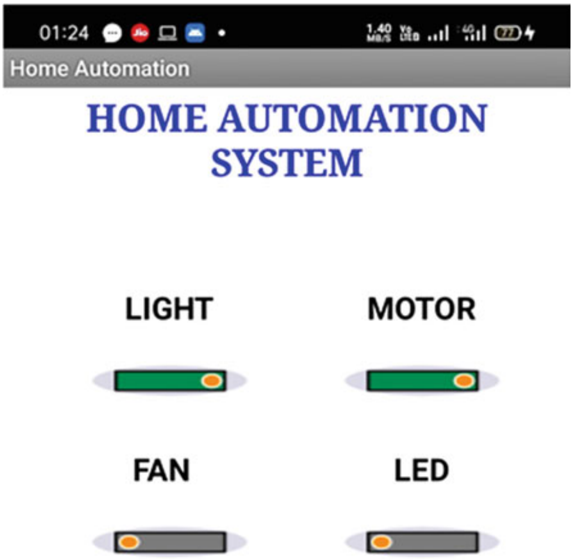


Fig. 8 Devices controlled ON

References

1. Gill, K., Yang, S.-H., Yao, F., Lu, X.: A zigBee-based home automation system. IEEE Trans. Consum. Electron. **55**(2) (May 2009)
2. Ashraf, I., Umer, M., Majeed, R., Mehmood, A., Aslam, W., Yasir, M.N., Choi, G.S.: Home automation using general purpose household electric appliances with raspberry Pi and commercial smartphone (September 22, 2020). https://doi.org/10.1371/journal.pone.0238480
3. Aravindh, J., Srevarshan, V.B., Kishore, R., Amirthavalli, R.: Home automation in IOT Using 6LOWPAN. Int. J. Adv. Comput. Eng. Network. ISSN: 2320-2106
4. Bernheim Brush, A.J., Lee, B., Mahajan, R., Agarwal, S., Saroiu, S., Dixon, C.: Home automation in the wild: challenges and opportunities. Microsoft Research. University of Washington
5. Piyare, R., Tazil, M.: Bluetooth based home automation system using cell phone 2011. In: IEEE 15th International Symposium on Consumer Electronics
6. Ridza Azri Ramlee, I.R., Tang, D.H.Z., Ismail, M.M.: System engineering and technology (ICSET). In: 2012 International Conference Smart Home System for Disabled People via Wireless Bluetooth
7. Rawat, P., Singh, K.D., Chaouchi, H., Marie Bonnin, J.: Wireless sensor networks: a survey on recent developments and potential synergies. J. Supercomputing
8. Kam, M., Suryadevara, N., Mukhopadhyay, S.C., Gill, S.P.S.: WSN based utility System for effective monitoring and control of household power consumption. In Conference Record—IEEE Instrumentation and Measurement Technology Conference (May 2014)
9. Yuksekkaya, B., Alper Kayalar, A., Bilgehan Tosun, M., Kaan Ozcan, M.: A GSM, Internet and speech controlled wireless interactive home automation system. IEEE Trans. Consum. Electron. **52**(3):837–843 (September 2006)
10. Aqeel-ur-Rehman, Arif, R., Khursheed, H.: Command controlled home automation system for the elderly or disabled people. J. Appl. Environ. Biol. Sci. **4**(8S), 55–64 (2014). ISSN: 2090-4274

11. Al Shu’eili, H., Gupta, G.S., Mukhopadhyay, S.: Command recognition based wireless home automation system. In: 2011 4th International Conference on Mechatronics (ICOM). IEEE (2011)
12. Pal, S., Chauhan, A., Gupta, S.K.: Command controlled smart home automation system. Int. J. Recent Technol. Eng. (IJRTE) **8**(3), 4092–4093 (2019). ISSN: 2277-3878
13. Neelima, J., Madhuri, S., Chaitanya, K., Anil, T., Mohana Rao, Ch.: Command control based home appliances using android devices on Arduino. Int. J. Electr. Electron. Eng. **9**(01), 1322–1325. ISSN: 2321-2045
14. Neha, S., Parvez, Md., Fatima, N.M., Marturkar, R.: Arduino based command controlled home appliances using Bluetooth. Int. J. Innov. Res. Comput. Commun. Eng. **5**(4), 459–462 (2017). ISSN: 2320-9798
15. Sen, S., Chakrabarty, S., Toshniwal, R., Bhaumik, A.: Design of an intelligent command controlled home automation system. Int. J. Comput. Appl. **121**(15), 39–42 (2015). ISSN: 0975-8887
16. Tyagi, C.S., Agarwal, M., Gola, R.: Home automation using command recognition and Arduino. Int. J. Recent Trends Eng. Res. 1–6. ISSN: 2455-1457
17. Aziz Md, A., Harshasri, K., Shanmukharao, K.: Cost effective command controlled home automation using IoT. Int. J. Eng. Res. Comput. Sci. Eng. (IJERCSE) **4**(3), 63–67 (2017). ISSN: 2394-2320

Detection of Early Depression Signals Using Social Media Sentiment Analysis on Big Data

Shruti S. Nair, Amritha Ashok, R. Divya Pai, and A. G. Hari Narayanan

Abstract Social media have become the new ‘reality’ for people as years go by and they have started linking their lives with these electronic devices. As a result, the increased chances of expressing themselves through media like Twitter, Instagram, Facebook, etc., have contributed to the study of depression analysis. The proposed paper predicts early signs of depression using supervised machine learning based on Naive Bayes, Decision tree, SVM, *k*-nearest neighbors on big data to find the accuracy on prediction.

Keywords Machine learning · Sentiment analysis · Decision tree · Big data · *k*-nearest neighbors

1 Introduction

This paper proposes an ascendable platform for the real-time processing for Twitter data. This framework is able to process a huge amount of Twitter data on a daily basis. This can help public health monitoring, surveillance and healthcare industries and also administrations may be supported. Twitter is one of the most common and rapidly rising microblogging and social media network services [1], where the registered users can share their voices and opinions as ‘tweets.’ Its popularity has increased, particularly among teenagers and young adults, as the platform allows its users to easily share their opinion and to get connected. In this paper, we have focused on the Twitter API as it is easily available and is open for all the public conversations which makes it easier to read and evaluate. The tweets are preprocessed in order to meet the requirements and filters using the big data approach. The tweets are then trained and tested and fed into various machine learning algorithms—Naive Bayes, Decision tree, SVM and *k*-nearest neighbors. This paper not only proposes the big data concept into the depression analysis of tweets but also compares various machine learning algorithms for their accuracy and completion time.

S. S. Nair (✉) · A. Ashok · R. Divya Pai · A. G. Hari Narayanan
Department of Computer Science and IT, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_31

Big data is a fast emerging and edge-cutting concept in the information technology industry and research areas. In the human lifestyle, social media has taken over a very wide area, [2] hence big data has also become very challenging to navigate and for the ease of usage. Enormous amount of digital data is produced on a daily basis through the invention of new technologies, IoT and smart devices. This data is unable to be managed by the traditional approaches like relational database management system (RDMS), as the data exceeds the limit of storage and inaccurate processing [3]. As a result of the failure of conventional methodologies, an open-source platform, Hadoop was invented to store and process data correctly and accurately for big data.

1.1 Depression and Sentiment Analysis

Sentiment analysis refers to natural language processing techniques used to classify the textual data whether it is positive, negative or neutral. In order to systematically define, extract, measure and analyze affective states and subjective knowledge, sentiment analysis which polarizes the data, with the methodology of NLP. It is also known as emotion AI, a commonly used classification tool. Depression is one of the most common psychological illnesses prevalent in the world. Studies reveal that a large number of people rely on social media for expressing their emotions through tweets, posts, blogs, etc. By the year 2020 over 264 million people around the world were found to be facing depression, as estimated by The World Health Organization (WHO). This is an emerging field of study by studying and training the data available from social networking sites and promoting mental health evaluations. In the paper [4], James and Gulden had proposed a sentimentor for Twitter data which utilizes Naive Bayes classifier so as to polarize the datasets and group the tweets into positive, negative or neutral sets.

1.2 Machine Learning and Bigdata

Machine learning (ML) is the field of computer algorithms that are enhanced by training and experience automatically. The algorithms used for the prsoposed model are Naive Bayes, Decision tree, SVM and *k*-nearest neighbors. Naive Bayes is one of the supervised classification algorithm, for conditional probability following the Bayesian theory [5].

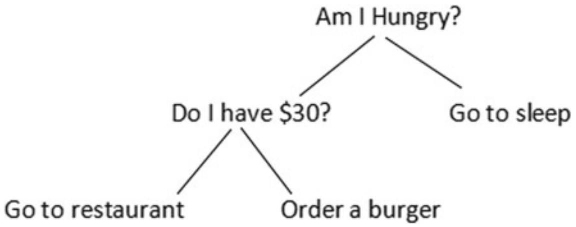
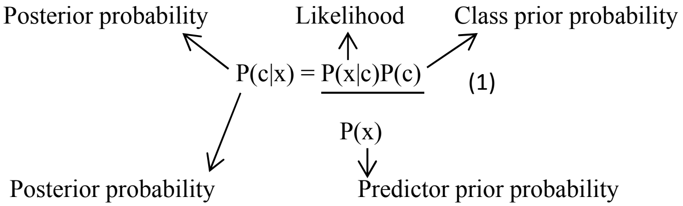


Fig. 1 Example of Decision tree



(1)

$$P(c \mid x) = P(x_1 \mid c) * P(x_2 \mid c) * \ldots * P(x_n \mid c) * P(c)$$

In this sense, Naive Bayes is a rational classifier with minimal storage and time-saving training [6].

Decision trees come under supervised machine learning, in which each leaf node represents a class and all the computation results in a final decision [7]. They are non-parametric learning method for regression as well as classification. Figure 1 shows an example of Decision tree of dividing classes on the basis of choices or preferences (yes or no).

SVM, also known as support vector networks, is an abbreviation for 'Support Vector Machine,' a natural language processing (NLP) technique which studies and classifies user feelings or intentions as positive, negative or neutral [8]. It generates hyperplane and finds data which are closest to the plane with separation.

k-nearest neighbors (KNN) are also a non-parametric algorithm (lazy learning) in which the training sets are used for regression as well as classification [9]. Unknown data comparison with training set is followed by the similarity calculation using Euclidean distance [10].

2 Literature Review

Through literature review, various machine learning and big data analysis papers were studied. Few papers on sentimental analysis were studied which concluded with positive, negative or neutral datasets using the Naive Bayes algorithm. Spencer and Uchyigitr [4] used ideologies like unigrams, bigrams and POS tags and the highest accuracy was proved to be the one using bigrams without POS against unigram. Another paper studied various text mining techniques proposed to different areas in multilingual format [11]. Their proposed framework includes data acquisition, preprocessing and other mining methodologies for natural language processing (NLP). The study also focused on ML classifiers. The study presents a relative study on the newer ideologies in text mining and its application in the evaluation level. The paper also discusses the evaluation process and results in more than 3 languages. In another paper focusing on sentiment analysis [12], they used fuzzy rule-based approach. They have done the paper with reference to tweets downloaded [12]. Readiness and ambivalence are dealt well with fuzzy-based systems. This model combines natural language processing (NPL) methodologies along with word classifier methods on the fuzzy rule to polarize the data as positive, negative or neutral sentiment class.

According to one paper on depression-related tweet analysis [13], two-third of the identified tweets were defined as the MDD diagnosis symptoms or depression associated communications or emotions. Shen et al. [14] along with their acquaintance worked on a learning model in order to detect early symptoms of depression on Twitter platform. They collected and studied a huge dataset from Twitter API to showcase and classify the usage of a normal and depressed social media user. Firstly, datasets were constructed, which are listed as: depression dataset, non-depression dataset, depression-candidate dataset [14]. Algorithms or methodologies used for the same are Naive Bayesian (NB), Wasserstein Dictionary Learning (WDL), multiple social networking learning (MSNL) and multimodal depressive dictionary learning (MDL). Another paper used deep learning approach on the early detection using keywords from Twitter on the registered user level, for performing comparative estimation of performance [15]. CNN-based models were found to outperform RNN-based models in their study. Balahur [16] introduced a strategy to normalize the language using the dataset arrangement and preprocessing, vocabulary was also generalized. Another paper completely focusing on emoticon expressions [17], studied the category, strength of emotions and those which illustrate feelings in text. Another paper [18] showed the rate of false information leasing in social media and finding the accuracy of the positive. RNN model is found to be the most accurate methodology while experimenting this among traditional methodologies and neural network. In [19] a paper for lexicon-based machine learning approach, it is proved that the sentiment analysis is directly relying on the lexicon method. The size and correctness of the lexicon approach are determining the working and accuracy of the analysis. In the paper 'Survey on Neural Network Architectures with Deep Learning' [20], deep learning networks are studied and classified into four category and again decomposed to briefly analyze each network. Various applications including face, pattern recognition are listed under the findings from the paper.

3 Proposed Model

In order to improve the accuracy of the early signs of depression detection on tweets using machine learning algorithms, we are putting forth the Apache Hadoop framework for big data preprocessing. The device or system should be compatible to implement the processing with a minimum of 8 GB RAM. In this model, Python 3.6.1 or higher is used with modules (Keras, TF, pandas, Numpy, sklearn and itertools). More than 10,000 data have been downloaded and tested for the proposed method.

Existing model have been found to be limited to comparatively lesser dataset and couldn't handle big data. The proposed model has been proved to be more accurate with big data analytics and preprocessing techniques as the Twitter data of over five years with more than 10,000 data is downloaded in real time. The study also compares the accuracy rates of four different classifiers to find out which among them results in faster as well as accurate output. As a result, the accuracy rates of proposed system are found to be higher compared to the existing machine learning algorithms.

The steps for working of proposed model are explained in steps and supported with a flowchart as in Fig. 2.

- A Twitter developers account is created from the official site (https://developer.twitter.com/en) with the basic needs as shown below:

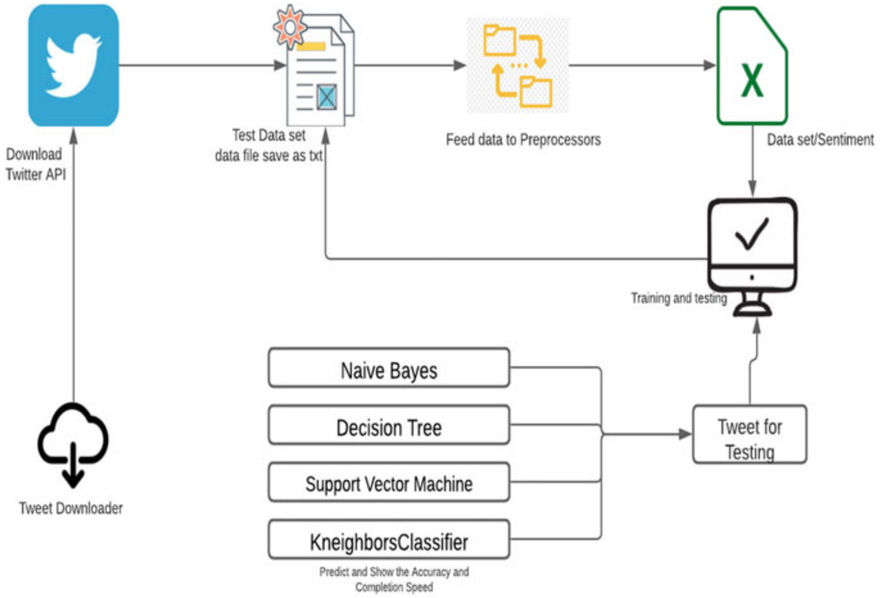


Fig. 2 Performance flowchart

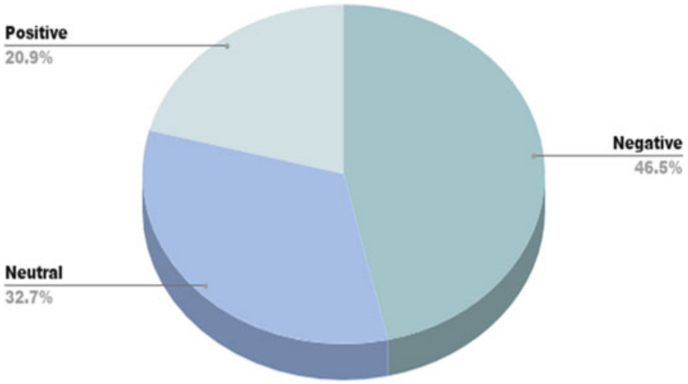


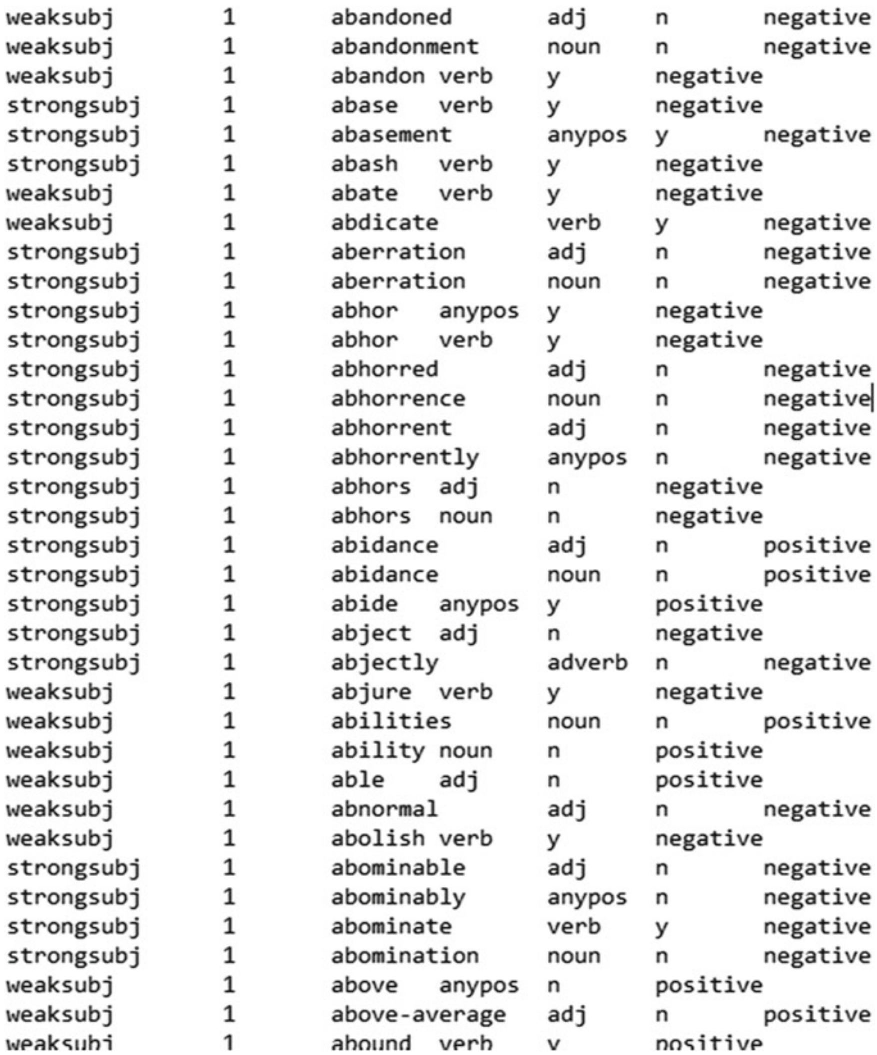
Fig. 3 Polarity pie chart of dataset

$$\text{Consumer\_key} =, \text{“consumer\_secret} = \text{”}, \text{“access\_token} = \text{”}, \text{“access\_secret} = \text{”} \tag{2}$$

- Once the Tweet download process from Twitter API and the dictionary is completed and read, preprocessor step starts. Words with their corresponding polarity are included in the dictionary.
- For every tweet, each word will be isolated, tokenized and polarized. Therefore, a tweet is calculated as the average of the polarity rates of the words,

$$\text{Average of Polarity} = \frac{\text{Sum of the polarities of the words}}{\text{Total number of words in the tweet}} \tag{3}$$

  The result of polarity of the dataset taken is represented in Fig. 3.
- The output of the preprocessing step is analyzed and stored in .xlxs file as the preprocessed sentiment dataset. Within it are listed, the sentiment tweet (ID) along with the corresponding polarity which is filtered by the keywords: positive, negative and neutral as in Fig. 4.
- As shown in Fig. 2, the training and predicting step follows the preprocessing. Each tweet is recovered based on their sentiment ID, once the proposed model starts to read the .xlsx file consisting of the dataset.
- These original data are then fed into the classifiers: Naive Bayes, Decision tree, SVM and k-nearest neighbors. More than 10,000 tweets were trained, tested and to compare the accuracy of the machine learning algorithms using the same dataset. As a result, In the console, the accuracy, completion time of each classifier will be listed.



```
weaksubj        1       abandoned       adj     n       negative
weaksubj        1       abandonment     noun    n       negative
weaksubj        1       abandon verb    y       negative
strongsubj      1       abase   verb    y       negative
strongsubj      1       abasement       anypos  y       negative
strongsubj      1       abash   verb    y       negative
weaksubj        1       abate   verb    y       negative
weaksubj        1       abdicate        verb    y       negative
strongsubj      1       aberration      adj     n       negative
strongsubj      1       aberration      noun    n       negative
strongsubj      1       abhor   anypos  y       negative
strongsubj      1       abhor   verb    y       negative
strongsubj      1       abhorred        adj     n       negative
strongsubj      1       abhorrence      noun    n       negative
strongsubj      1       abhorrent       adj     n       negative
strongsubj      1       abhorrently     anypos  n       negative
strongsubj      1       abhors  adj     n       negative
strongsubj      1       abhors  noun    n       negative
strongsubj      1       abidance        adj     n       positive
strongsubj      1       abidance        noun    n       positive
strongsubj      1       abide   anypos  y       positive
strongsubj      1       abject  adj     n       negative
strongsubj      1       abjectly        adverb  n       negative
weaksubj        1       abjure  verb    y       negative
weaksubj        1       abilities       noun    n       positive
weaksubj        1       ability noun    n       positive
weaksubj        1       able    adj     n       positive
weaksubj        1       abnormal        adj     n       negative
weaksubj        1       abolish verb    y       negative
strongsubj      1       abominable      adj     n       negative
strongsubj      1       abominably      anypos  n       negative
strongsubj      1       abominate       verb    y       negative
strongsubj      1       abomination     noun    n       negative
weaksubj        1       above   anypos  n       positive
weaksubj        1       above-average   adj     n       positive
weaksubj        1       abound  verb    y       positive
```

Fig. 4 Dictionary

4 Results

With this paper, sentiment analysis algorithms have been tested and shown to be more accurate with the use of big data concept. This study, supported by the proposed model proves better accuracy and faster completion time with the concept of big data analysis and machine learning algorithms together. The result is faster completion time and higher accuracy rates as compared to the existing machine learning algorithms used alone. Output of various algorithms compared using big data with the Apache Hadoop tool are mentioned. It is concluded that the Decision tree algorithm is having the highest accuracy rate with 97.3248% with a completion speed of 35.19268 s. Support vector machine (SVM) is identified as the second highest with an accuracy rate of 91.8869% with a completion speed of 1141.34724 s. Naive Bayes classifier produced an accuracy rate near to SVM, with 91.0245% with a time period of 7.952 s. The least accurate among the compared classifiers is *k*-nearest neighbors classifier with 81.8153% with a time period of 134.57 s. Table 1 shows the precision, recall and f_1 score along with accuracy of different classifiers (Fig. 5).

Table 1 Results generated for each classifier

Methods		Precision	Recall	F_1 Score	Accuracy
Naive Bayes	Positive	0.93	0.92	0.93	0.91
	Negative	0.91	0.63	0.73	
	Neutral	0.92	0.97	0.94	
Decision tree	Positive	0.93	0.94	0.94	0.97
	Negative	0.72	0.89	0.80	
	Neutral	0.98	0.93	0.95	
Support vector machine	Positive	0.95	0.92	0.93	0.91
	Negative	0.92	0.66	0.77	
	Neutral	0.90	0.99	0.94	
K-Neighbors	Positive	0.82	0.85	0.84	0.81
	Negative	0.98	0.18	0.31	
	Neutral	0.79	0.92	0.85	

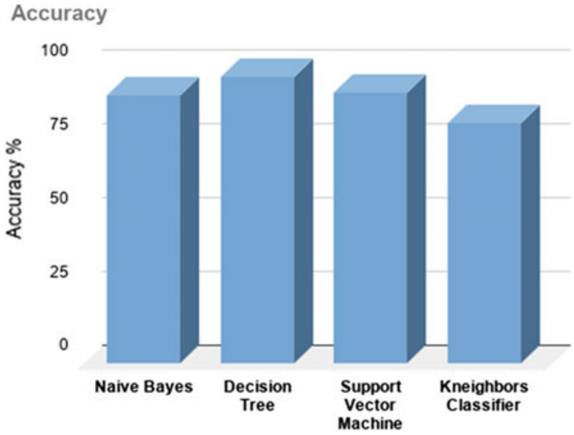


Fig. 5 Accuracy bar graph

5 Conclusion

The proposed methodology verifies the enhanced and better accuracy of Twitter data analysis on early depression detection with machine learning models using big data analysis. The accuracy rates of the considered classifiers (Naive Bayes, Decision tree, SVM, *k*-nearest neighbors) were found to be enhanced with big data using Hadoop. Among these, Decision tree is identified with the most accuracy rate on the big data. The data hence proposed is proved to be handy in the medical area, for the prevention of depression as well as suicide rates. For this paper, only Twitter data is considered for the early detection of depression. Any other social media data (Facebook, Instagram) can be downloaded and fed into the preprocessor for the same.

Currently, this model supports only the English language. In the future, this model can be extended for contextual segmentation and the usage of stopwords.

References

1. Gaikwad, A., Mokhade, A.: Twitter sentiment analysis using machine learning and ontology. Int. J. Innov. Res. Sci. Eng. Technol. **6**(1), 173–180 (2017). (International Conference on Recent Trends in Engineering and Science (ICRTES))
2. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good, the bad and the OMG! In: Fifth International AAAI Conference on Weblogs and Social Media (ICWSM) (2011)
3. Vasuki, M., Arthi, J., Kayalvizhi, K.: Decision making using sentiment analysis from twitter. Int. J. Innov. Res. Comput. Commun. Eng. (IJIRCCE) **2**(12), 7171–7177 (2015)
4. Spencer, J., Uchyigit, G.: Sentimentor: sentiment analysis of twitter data. In: 1st International Workshop on Sentiment Discovery from Affective Data (SDAD) (2012)
5. Dey, L., Chakraborty, S., Bose, B., Biswas, A.: Sentiment analysis of review datasets using Naïve Bayes and K-NN classifier. Int. J. Inf. Eng. Electron. Bus. (IJIEEB) **8**(4), 54–62 (2016)
6. Kaviani, P., Dhotre, S.: Short survey on Naive Bayes algorithm. Int. J. Adv. Eng. Res. Dev. (IJAERD) **4**(11), 607–611 (2017)
7. Hou, Y.: Decision tree algorithm for big data analysis. In: Proceedings of the 2018 International Conference on Transportation & Logistics, Information & Communication, Smart City (TLICSC), vol. 161, pp. 1951–6851 (2018)
8. Selamat, A., Zainuddin, N.: Sentiment analysis using support vector machine. In: 2014 International Conference on Computer, Communications, and Control Technology (I4CT), pp. 333–337 (2014)
9. Hota, S., Pathak, S.: KNN classifier based approach for multi-class sentiment analysis of twitter data. Int. J. Eng. Technol. (IJET) **7**(3), 1372–1375 (2018)
10. Tyagi, A., Sharma, N.: Sentiments analysis of twitter data using K-nearest neighbour classifier. Int. J. Eng. Sci. Comput. (IJESC) **8**(4) (2018)
11. Redhu, S., Srivastava, S., Bansal, B., Gupta, G.: Sentiment analysis using text mining: a review. Int. J. Data Sci. Technol. (IJDST) **4**(2), 49–53 (2018)
12. Vashishtha, S., Susan, S.: Fuzzy rule based unsupervised sentiment analysis from social media posts. Expert Syst. Appl. **138**(112834), 0957–4174 (2019)
13. Cavazos-Rehg, P.A., Krauss, M.J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., Bierut, L.J.: A content analysis of depression-related tweets. Comput. Hum. Behav. **54**, 351–357 (2016)
14. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., Zhu, W.: Depression detection via harvesting social media: a multimodal dictionary learning solution. In: Proceedings of the

Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), pp. 3838–3844 (2017)

15. Orabi, A.H., Buddhitha, P., Orabi, M.H., Inkpen, D.: Deep learning for depression detection of twitter users. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology, pp. 88–97 (2018)
16. Balahur, A.: Sentiment analysis in social media texts. Assoc. Comput. Linguist. (ACL) 120–128 (2013)
17. Aman, S., Szpakowicz, S.: Identifying expressions of emotion in text. In: Published by Text, Speech and Dialogue, 10th International Conference, TSD, pp. 196–205 (2007)
18. Haoxiang, W.: Emotional analysis of bogus statistics in social media. J. Ubiquit. Comput. Commun. Technol. (UCCT) **2**(03), 178–186 (2020)
19. Mitra, A.: Sentiment analysis using machine learning approaches (Lexicon based on movie review dataset). J. Ubiquit. Comput. Commun. Technol. (UCCT) **2**(03), 145–152 (2020)
20. 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)

Raspberry Pi-Based Heart Attack and Alcohol Alert System Over Internet of Things for Secure Transportation

G. Dhanalakshmi, K. Jeevana Jyothi, and B. Naveena

Abstract In this modern era, the number of accidents that occur around us is increasing unprecedently. The reasons for accidents might be the driver's health issues, drowsiness and alcohol consumption. To prevent accidents and improve the driver's safety, this research work has proposed a smart and secure transportation system. The proposed paper describes the design and implementation of smart and secure transportation system by using Raspberry Pi and Internet of Things. For alcohol detection, MQ3 sensor is used, and to detect heart attack, pulse sensor is used. Here, a gadget has been proposed with the intention to discover coronary heart assault via tracking the heart rate based on Internet of Things (IoT). Monitoring and alerting an individual, when the comatose gains cognizance by using the movement detection gadget. Two sensors are used to monitor the health of the patient. Pulse sensor is used to monitor the pulse rate functioning of heart. Alcohol sensor is used to detect alcohol percentage in air. By using these sensors, the physical condition of the driver can be detected. The system locks the vehicle if the driver is determined to be alcoholic or with improper pulse rate. Hence, it prevents the road accidents.

Keywords Raspberry Pi · Secure transportation · Alcohol detection · Internet of Things (IoT) · Python · Arduino board · Microcontroller · Sensors

1 Introduction

In developing countries like India, a day begins with the newspaper that contains at least a couple of headlines about the road accidents. The increasing fatal rate in our society is mostly due to the road accidents. The occurrence of accidents in India has many reasons such as less road quality maintenance, improper construction of the bridges, overcrowding on the busy lanes and many more. Apart from these, the underage driving, violating the government traffic regulations are the few other reasons. Most often, the youngsters involve in rash driving and drunken driving. Also, the loss of young lives causes a major loss to the country. The drunken drivers

G. Dhanalakshmi · K. Jeevana Jyothi · B. Naveena (✉)
Department of ECE, Siddartha Institute of Technology and Sciences, Hyderabad, Telangana, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_32

are more dangerous because they put the other innocent people lives in jeopardy. On average, 3972 innocent deaths occur every year who have been passengers, common pedestrians or cyclists, or sober drivers of different vehicles. More than 10 innocent victims are killed each day because of the alcohol consumption of other riders. To save a driver's life, several devices are developed inside a vehicle, which include an airbag system and antilock braking system. The losses due to road accidents are comprised of treatment cost, productivity loss, disability, deaths, etc. Around 1.35 million people across the globe lose their lives every year due to road accidents. In developing nations like India, the death toll due to road traffic accidents has been increasing every year. This alarming situation leads the government to impose strict laws and heavy fine amount to restrain from this trend though these steps are not sufficient to reduce the deaths caused by accidents. The road accidents in India are mainly caused by violating the speed limit, driving on both sides on a one-way, multitasking while driving, accidents caused due to stray animals and drunken driving. Considering all the reasons for road accidents and drunken driving is one of the most dominating reasons. In India, drunken driving is a severe and punishable offense in which any drunken driver, if found with more than 30 mg of alcohol in 100 ml of blood, can be fined up to INR 10,000 or/and imprisoned up to 6 months. The available measures taken by the government and other road traffic-related agencies are not at all sufficient to mitigate the accidents caused by drunken driving. Manual procedures are highly time-consuming and detective in nature, whereas this problem requires some preventive measures to be deployed in conjunction with the latest technologies. The victim also faces financial trouble due to the unavailability of third party insurance, health insurance and other health facilities. A trustless computing technology like blockchain can play a great role in assuring the right treatment to the right person. Internet of Things (IoT) technology in the current healthcare era consists of gadgets, services and wireless sensors that find physiological values which can either be a wearable device or a sensor that streams information to physical, and also to cloud-based servers. Comfortable systematic overseeing of affected person's physiological symptoms has the ability to augment the basic medical exercise, especially in emerging nations that have a depravity of healthcare professionals. Now-a-days healthcare system is drastically shifting from reactive responses to deal with the intermittent situations by using an active methodology wherein the patient is categorized by early detection, prevention and better management in healthcare facilities. On this schema, healthcare situation and tracking and wellbeing control have a major contribution toward one's fitness-care. This device is especially crucial in nations having a widespread getting older populace, like India wherein these devices can be utilized to noticeably upgrade the entire first-class of existence. The proposed device is basically designed to be utilized in healthcare facilities or even at home to monitor and measure various vital parameters of the patient. The results thus obtained are recorded by using the Arduino board and are sent to a server, which is then stored into a database. With respect to this, the results are also displayed on a website and the doctor assigned to the patient can login to the patient's database and monitor continuously. In this paper, a model of basic human body monitor system using Arduino microcontroller and an online server is used. For a healthy person, ordinary heart rate is 60 to one hundred bpm (beats per minute). Athlete's heartbeat usually varies from 40 to 60 bpm relying upon their fitness. If someone's heartbeat is constantly over a hundred beat consistent with minute, then the man or woman is stated to have higher coronary heart fee, which is likewise infamous as tachyarrhythmia. It can diminish the performance of heart by letting down the amount of blood pumped through the body can result in chest pain and lightheartedness. With the development in generation, it is straightforward to reveal the patient's coronary heart charge even at domestic range. IoT is dexterity of network mechanism to mind and acquire information from global ubiquitously and then proportion the data across internet to anywhere, and it may be controlled for a few tenacities. The values from the sensors are sent to a server, which displays the values on the webpage and is also stored in an online database.

ESP8266, which is a Wi-Fi module, is utilized for the transmission between Arduino and server. The software that is used is developed using Raspberry Pi, HTML and Java programming language. The physiological values measured are then taken into account and are updated in specific intervals. These values can be viewed from anywhere in the world by using a web-enabled tool. Now, when the physiological parameters are greater than the threshold values, then the physician gets an alert. This system is targeted at patients, who need frequent monitoring of their physiological parameters by a health expert or a caretaker. This system has the ability to send alert notifications either through IOT or e-mail if the parameters cross the threshold values this system thereby helps in taking the appropriate action at the right time which could save the person from any critical health issues. Also, the system can help in limiting the hospital bills which may arise when a patient is hospitalized. Such a system helps in the patient from getting frequently admitted to the hospital when the damage is beyond control.

2 Literature Survey

A versatile shape that plays ongoing exam of physical data to reveal human nicely being situations in any specific circumstance (for example, direction of each day physical video games, in medical hospital conditions). A well-known searching, non-prevent studies need to likewise be finished on cellular phones. Downsizing of low-manage microelectronics and a long manner off structure is being revamped into right proper in a large wide interior that reinforces the idea that healthcare agencies in sufferers and beneficial experts. The flexible shape which plays nonstop assessment of physiological records to expose shows display human beings beneficial circumstances. The performance of progressing examination of physical data that is used to evaluate one's prosperity circumstances. Improved in extending the capacity with respect to trinkets and lacking regards with steady assessment. Reduced thought is now focused on the headway in assessment methodologies for assessing the current prosperity commission of watched people. Lacking regards can be achieved using sensor disillusionments or got by invalid regard substitution [1]. A job dependent on astute versatile consideration structure with prepared instrument in incessant consideration circumstance is produced and completed. The jobs in this structure infuse various components which are the healthcare experts, medical staff, caretakers, and also the patients. Each of the jobs requires that the person must use a mobile device. For instance, a cell phone that can be used to make a contact the server with a focus to the extent that the person can travel without any limitations. This device is programmed in the back-end to deal with any sort of crisis and then sends alarms subsequent to receiving any crisis messages. The after effect of this system is unusual and the schema subsequently educates the healthcare experts. When the patient travels to any place, the mobile phone attached to him is connected with Bluetooth to the device, and when the patient is suffering from any uneasiness, the system will know [2]. This paper uses a few physiological parameters to determine the health condition of the patient, for example, ECG values, pulse values, internal heat level and others, then understanding the zone of social meeting of patients within crisis facility conditions. The blend of electronic systems and remote frameworks provides a beneficial technique which gives rise to noninvasive and also unavoidable organizations mentioned by the current technological administration circumstances. Recent inventions in the field of nano-technology have led to a spark in the utilization of electronic measures to estimate the physiological parameters in a noninvasive manner. Exchanges drives in remote sensor frameworks (WSNs) offer a monetarily clever response for help. It is incredibly cutting-edge and precise that doesn't invoke a few important essentials which are required in the present and also future technology in this area of expertise [3]. The essentialness skilled perpetual thrives watching, the model gathering, idiosyncrasy thrift communication and comprehensive distinguishing to reduce the risk of private transmission, taking care of and affirming the data. To evaluate those techniques, show off the inventions of approximately a couple of gadgets of significance upgrades related to electricity and also capability conditions, which also assists in comprehending the capacity of extended haul nonstop fitness looking at in which genuinely considered one of a type physiological signal are stuck, broke down and located for later time is imagined as a way to empower an active and all-encompassing way to address medicinal offerings. Those intending to execute and power effectiveness with significant enhancements in figuring ordinary sign managing has gotten manageable. Biomedical sensors have been used for prosperity confirmations for quite a while and various signs to separate the information which are a characteristic of a patient's condition. Inertness is the time between a mis-happening and the time that is taken by a doctor or a healthcare expert to react. This scheme is an aggregation to diminish all out vitality utilization significantly more [4]. Road traffic accidents—the leading motive of death by means of injury. Road traffic accidents are predictable and preventable, but good statistics are essential to recognize how road safety interventions and era can be effectively transferred from developed international locations where they've tested effective. Awareness of the consequences. D. Selvathi proposed a drowsiness detection system using deep neural network for vehicle drivers in road accident avoidance system that can help to prevent the accidents by the detection of the drowsiness of the person who was driving the vehicle. It can be used to send an alert sound so that the person drowsiness will be removed slowly by the alert sound [5], since the invention of the wheels, road accidents are the most unwanted event to happen to an individual. However, accidents are quite often; unfortunately, people don't obey the traffic rules and never learn from their and other mistakes. Out of all the reasons for road accidents, drunken driving proves to be a dominating reason across the world. Many researchers have paid attention, particularly in mitigating these mishaps and used a series of tools and techniques. Initial systems were a detective and not able to predict and prevent such accidents in advance. The availability of high-speed internet and the development of IoT technology have opened another gate for finding the solution of these research problems [6].

3 Existing System

In existing system, there is no safety for the drivers in case of health issues. Hence, many accidents occur due to drowsiness, alcohol consumption and heart attacks. The disadvantage of past proposed framework is that it gives bogus alert by recognizing each individual inside the vehicle and stops the motor by paying little heed to the driver. An unexpected stop of the vehicle makes mishaps in the streets. It is evident from the referred literature that road accident prevention has remained a fairly researched area for researchers and academicians. Still, the existing solutions offer various limitation in terms of hardware, technology and other concerns. However, it can be concluded that there is a lack of solution that is accurate and efficient.

4 Proposed System

The proposed paper helps to design and implement a smart and secure transportation system using Raspberry Pi and Internet of Things to alert the alcohol detection using MQ-3 sensor and heart attack detection of the driver using pulse sensor. A gadget has been developed with the intention to discover coronary heart assault via tracking the heart charge based on Internet of Things (IoT). The proposed drunk driving prevention model involves four main modules, i.e., MQ3 sensor module, IoT module, machine learning module, recommendation or alert generation module, as shown in Fig. 1. MQ3 sensor module has proved to be useful in gas leakage detection, which consists of an alcohol sensor that continuously detects the alcohol gas in the air, and heartbeat sensor is used to detect the heart rate and heart attacks. It receives input from the breath of the vehicle driver, which may or may not be drunk. The output of this module is supplied to the IoT device module.

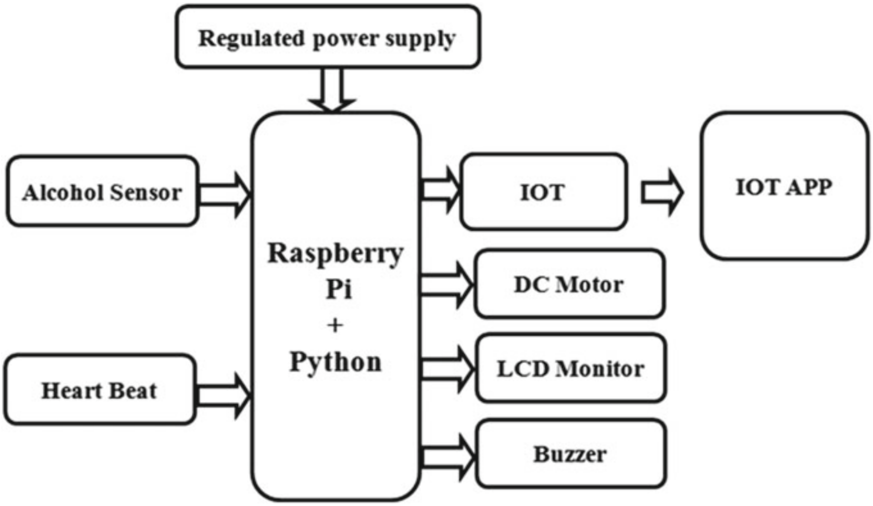


Fig. 1 Block diagram

In this paper, the problem of drunken driving and driver health issues (like heart attack), which is a major cause of road accident deaths is addressed by proposing a technology for intervention solution. The proposed drunk driving prevention model joins together two important technology of the world, i.e., Internet of Things (IoT). This model is capable of delivering the response as a value of the amount of alcohol in each sample. These samples can have values that vary continuously as per the content of alcohol. The level of alcohol checked as per the BAC level that best fits this problem is those that apply regression. This system manages a govern of identifying the breath; if a driver is failed, the sensor recognizes the level of liquor in the driver's breath. On the off chance that it crosses a set limit, an alarm message will be produced by the framework, and the vehicle engine stops in a split second and posts the data into IoT module. This system is an innovation to reduce such cases. This system is for the safety of the people who are outside the vehicle as well as inside the vehicle. The system dependably screens the driver's breath by putting it on the driver directing, or dashboard and drivers' breath can be consistently seen by it.

The system detects the level of alcohol in the breath of the driver then our system will start its working and locks the engine so that the vehicle fails to start. In the second condition, if the driver isn't drunk while he starts the vehicle and the motor is begun, he/she drinks yet. At the same time, driving, the sensor recognizes alcohol in his breath and stops the vehicle so the auto would not quicken any further, and the driver can guide it to the roadside. In this system, we utilize a microcontroller interfaced with an alcohol sensor alongside an LCD screen and an MQ3 sensor for recognition of alcohol, here the alcohol sensor constantly monitors the breath of the user and sends signals to microprocessor. The microprocessor Raspberry Pi detects the alcohol as per the given limit (its threshold value), the sensor sends the signal, and it will display a message on the LCD screen and also stops the engine working and locks the vehicle. The product needs a push button to start the engine. On the off chance that the system recognizes the alcohol level is high at the time of starting the vehicle, and the vehicle won't start. If alcohol is detected after the vehicle starting point, the system bolts the vehicle around then, and in the second case, with the assistance of the Torque converter, it switches the Gears naturally, so the driver does not need to work a grasp lever to switch gears.

This system will continuously monitor the heartbeat status using pulse sensor which is installed in driver seat belt if any fluctuations or abnormal heartbeat, it will automatically alert the IoT framework and locks the engine so that vehicle fails to start same as alcohol sensor in case of heart attacks.

4.1 Features Provided by the Project

This project provides a solution to avoid these accidents up to certain extent. Accident prevention system using alcohol detection and heart attack detection serves as the solution for accident prevention with the following features:

- The proposed system prevents accidents by detecting the consumed alcohol level using alcohol sensor and detects heart attack by using pulse sensor and takes necessary measures to avoid accidents.
- This system provides the alert system facility along with the automatic shutdown facility of the vehicle.
- This project deals with the detection of the driver's state unlike other intelligent systems which work on the mechanical aspects of the vehicle.
- This project provides a cost effective, reliable and feasible system for accident prevention.

4.1.1 Raspberry Pi Controller

The Raspberry Pi controller is a small single board computer designed, especially for teaching purpose. It is compatible with operating system like Android, Linux, net BSD, RISCOS, Window-10, ARM 64, etc. It has a USB power port of 5 V, 3 A which can deliver full power to USB devices. The CPU is 1.5 GHz 64/32 bit quad-core ARM cortex A 72, provided with four RAM slots, two with inbuilt Wi-Fi and other two without Wi-Fi. An SD card is also used to run the Raspbian Buster OS. The input power is rated at 5 V DC, 3 A (minimum), a power-bank with a 5 V DC, 2.4 A rating can be used, and if no extra peripherals are connected to the Raspberry Pi controller, this has to provide either through a USB-C connector or GPIO header pin (Fig. 2).

The Raspberry Pi foundation is working on yet another model of the popular Raspberry Pi controller boards, as the Raspberry Pi 3 Model B board. The new

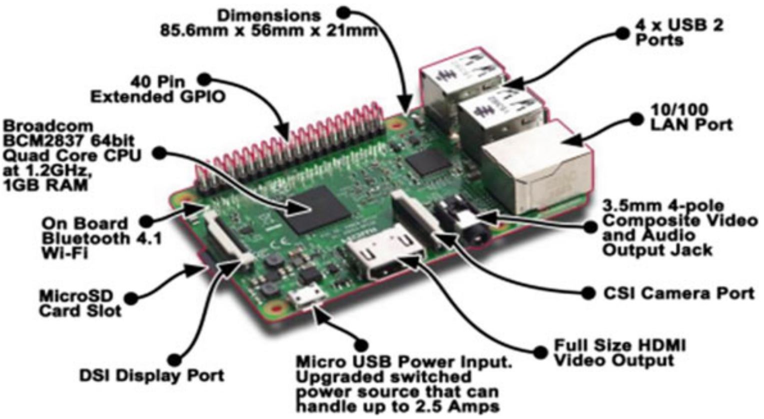


Fig. 2 Raspberry Pi

board looks very similar to that of Raspberry Pi 2 Model B, with added on-board Bluetooth 4.0 and Wi-Fi 802.11 b/g/n. Let's play "spot the difference" with an RPi 2 (Raspberry Pi 2) at the top and an RPi 3 (Raspberry Pi 3) under. On the top left corner, we'll find the Wi-Fi/BT chip antenna, and on the right of the 40-pin connectors two through holes, probably a RUN header can be found on RPi 2 for reset where the chip antenna is now placed on RPi 3. So, the through holes are not new, but they've just moved them. All the connectors have exactly the same placement between the two versions when compared. On board's other side, just above the micro SD slot, the wireless module can be found, with a J5 connector soldered. J5 is a JTAG connector, so it will not be soldered probably with the version that ships. As that for RPi 2, they've used the same Elpida B8132B4PB-8D-F RAM chip (1 GB). So, although we can't be 100% confident, the RAM appears to be the same, and the processor is still connected to a similar USB to Ethernet chip, so probably they might have kept the same architecture, except for the CPU core. So, built-in Wi-Fi and Bluetooth and 64-bit ARM cores (likely Cortex A53) are the only major changes on RPi 3. The hardware-related data for this device is SoC—Broadcom BCM2837 ARMv8 supports for 64bit, quad-core Cortex A53 processor @ 1.2 GHz, and 1080 pixel 30 H.264 high-profile decoder. Accelerated Open VG, with a dual-core Video Core IV GPU @ 400 MHz supporting OpenGL ES 2.0. It is capable of 1.5 G texel/s, 1 G pixel/s, or 24 Giga FLOPs with a DMA infrastructure and texture filtering. The storage details include system memory with 1 Giga Byte LPDDR2 and micro SD slot for additional memory. Video and audio output through HDMI 1.4 connectivity, composite video port and 4-pole stereo audio connectivity, Ethernet port of 10/100 M, provision for Bluetooth 4.1 LE and Wi-Fi protocol 802.11 b/g/n up to 150 Mbps.

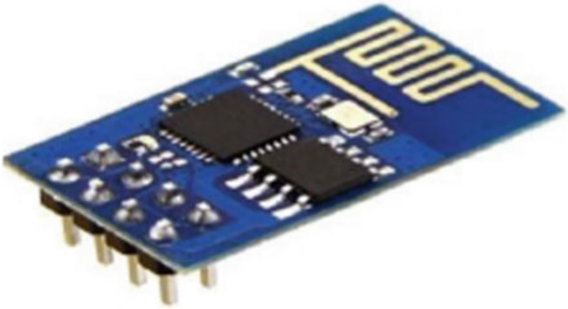


Fig. 3 ESP 8266

4.1.2 IoT Module

Internet of Things used for controlling any device or monitoring the device status through internet. This proposed system we use this IoT module for taking the all parameters data and post into the cloud called server. ESP8266 modules as IoT module it can operate through Wi-Fi frequency concept (Fig. 3).

The ESP8266 is a Wi-Fi module mostly used in IoT applications. It consists of several versions like ESP-01 to ESP-11 ESP-12E and WeMos D1 Mini. We can use this module for creating a web server, send HTTP requests, send emails, control output and read input and interrupts.

4.1.3 LCD Display

This is an electronic display device which is used for a liquid crystal display to produce an image. It is a common basic display module that is used in DIYs and also in the circuits and it will also translate a 16 × 2 display having 16 characters per line. It also consists of two lines. Each character is displayed in 5 × 7-pixel matrix format. Liquid Crystal Display (LCD) is used to display the parameters for obtaining the status of the proposed system. This can display 32 characters having 2 columns. When each sensor is activated corresponding message will be displayed in 16 × 2 LCD modules. In this we use four data pins using these pins we transfer the data from micro preprocessor to LCD (Fig. 4).

4.1.4 Buzzer

Buzzer is a speaker that is of small size and is also an audio signaling device. Piezoelectricity is the one effect that is used here where the crystals will change its shape when the electricity is applied to it. Now, when the electricity is applied with the correct frequency, the buzzer makes a sound. The buzzers are used in various fields for the usage of alarm devices, timers, etc. The buzzer works as the following where the tone which is present sends some frequency of 1 kHz to a particular pin used and

Fig. 4 16 × 2 LCD

Fig. 5 Buzzer

delay is used to pause it for particular seconds and then buzzer stops the signal. This is continued for making a short beep sound from the buzzer (Fig. 5).

4.1.5 Relay Module (Heartbeat Sensor)

Heart rhythm can be tracked two ways: in one case, the pulse is regulated manually at the fingertips or at the neck and in another; a Heartbeat tracker can be used. Throughout this project we developed an Arduino and Heartbeat Signal Heart Rate Monitor System. A functional heart beating machine can be used to locate the Heartbeat Machine Concept, the Heartbeat Sensor operating and the Arduino-based heart rate monitoring device (Fig. 6).

A coronary heart price reveal (HRM) is a private tracking tool that allows one to measure/show heart price in real time or report the coronary heart rate for later have a look at. It is largely used to gather heart fee information while performing

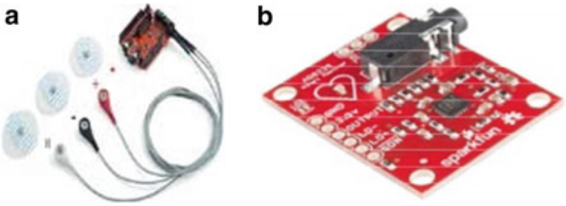


Fig. 6 **a** ECG sensor. **b** Heart rate

diverse forms of physical workout. Measuring electrical heart statistics is referred to as Electrocardiography.

4.1.6 Alcohol Sensor

The sensor used here is MQ3, which has ability to detect alcohol percentage in breath. Researchers have already found the relation of alcohol content in blood to the alcohol content in breath. By using this relation, it is possible to calibrate the alcohol sensor (Fig. 7).

This unit is utilized to distinguish the level of alcohol at the threshold value. The simple yield of which is connected to an Arduino board. The MQ-3 gas sensor has delicate material which is SnO_2, this has less conductivity in earth free air. At the point when the focus on alcohol gas exists, the sensor's conductivity is high alongside the gas concentration rising.

MQ-3 sensor is designed for detecting gases like Alcohol, Benzene and LPG, etc. MQ-3sensor has high sensitivity toward alcohol and has a very good resistance to disturb of gasoline, smoke and vapor. This sensor provides an analog output based on alcohol concentration. When the alcohol gas exists in the air, the sensor's conductivity gets higher.

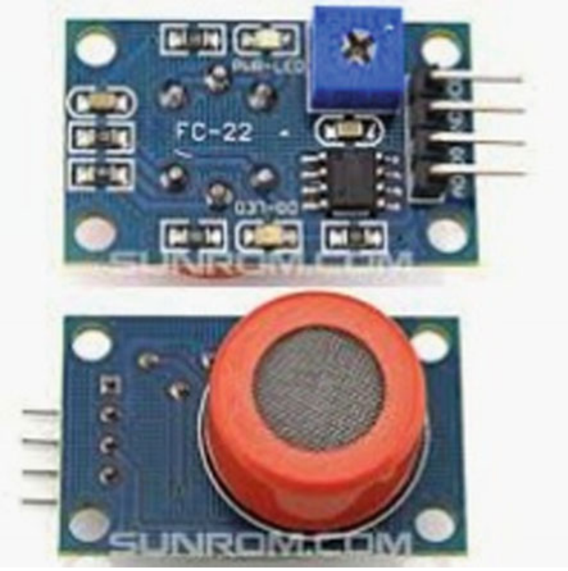


Fig. 7 Alcohol sensor

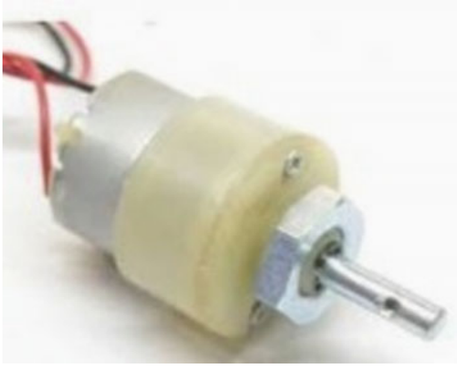


Fig. 8 Motor

4.1.7 DC Motor

DC motor works on the principal that when a current-carrying conductor is placed in a magnetic field then it experiences a torque then it will start. This is known as motoring action. If current direction in the wire is reversed, then the direction of motor rotation also reverses.

In our project DC motor is used as engine starter which would be connected to crankshaft of vehicle's engine, the speed of a dc motor is corresponded to the supply voltage, so if we reduce the supply voltage, the motor will at low speed. The speed controller work by varying the average voltage sent to the motor. This voltage is depending upon the alcohol sensor (MQ3). That means when the alcohol sensor sensed the alcohol percentage less than 40 mg per 100 ml, the motor will run. When the sensor sensed the alcohol percentage above 40 mg/100 ml, the motor will stop (Fig. 8).

4.1.8 Software

The proposed framework is built/developed using Python and machine learning-based algorithms. Python IDE is used to write code and compile. Raspbian operating system is used for developing the desired model.

5 Results and Discussion

The performance evaluation of this proposed alcohol detection system is done by performing real-time testing. The figure illustrates web controller dashboard of the proposed for security and driver safety. Below hardware model is integrated with alcohol sensor, heartbeat sensor along with output actuators relay and dc motor to execute real-time prototype system. We executed and results verified successfully (Fig. 9).

- While testing in real-time 2 sensors ((1) alcohol sensor and (2) pulse sensor) are used. Both sensors are connected and results are sent to Raspberry Pi.
- **Testing alcohol sensor**: If alcohol % is greater than fixed level (39 mg/100 ml) it will display alert on LCD, red LED blinks and the DC motor (which is assumed as a vehicle engine) stops running.
- **Testing pulse sensor**: If person/driver pulse changes drastically below or above the referenced level (60–100), it will display alert (driver heartbeat) on LCD, red LED blinks and the DC motor (which is assumed as a vehicle engine) stops running.

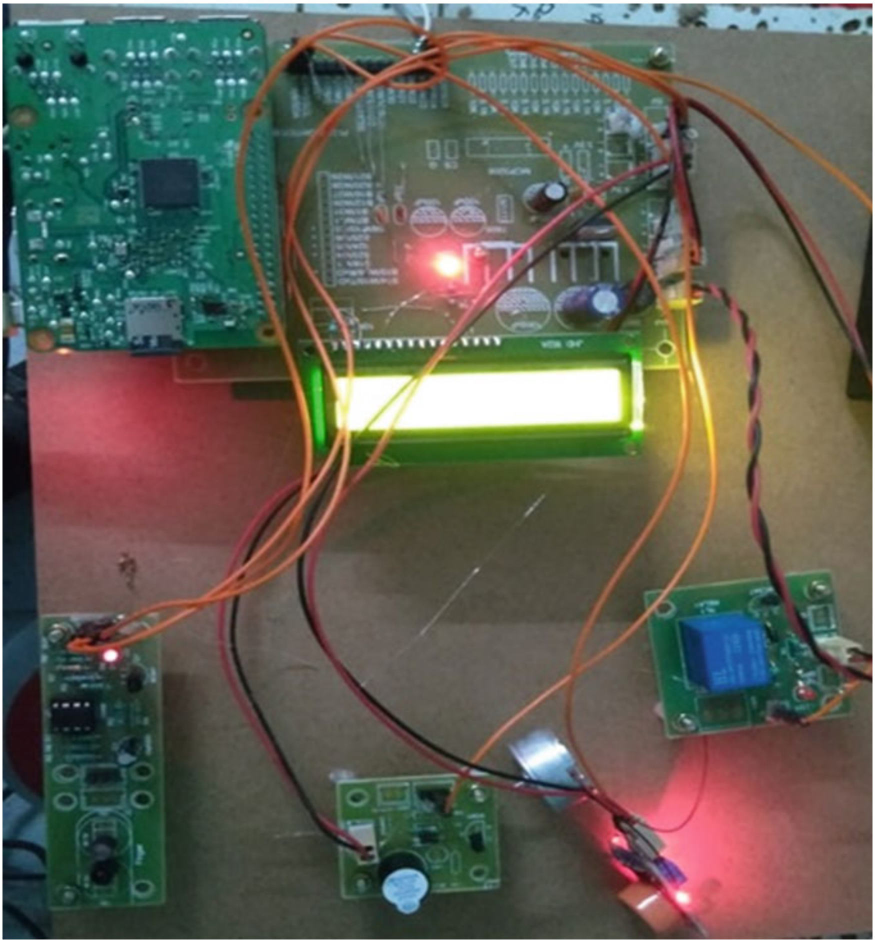


Fig. 9 Output hardware model

6 Conclusion

The values of these parameters were analyzed and alerted in this proposed system. It automatically alerts in IoT module in case of heart attack, vehicle halting by detection of alcohol, and accident alert system is developed. This model is subjected to testing by using a DC motor as a reference for a car. This model is observed to be successful in delivering the alert messages within a fraction of seconds to the IoT server during alcohol detection and accident detection. The efficiency of the model is best as compared to the existing models. It automatically alerts in IoT module in case of heart attack, vehicle halting by detection of alcohol, and accident alert system is developed. This model is subjected in testing by using DC motor as a reference for a car. This model is observed to be successful in delivering the alert messages within the fraction of seconds to the IoT server during alcohol detection and accident detection. The efficiency of the model is best as compared to the existing models.

6.1 Future Scope

- In the future, we can improve the device by including a GPS tracker to trace the location of the vehicle.
- As the system has limitation like it can only be implemented in the cases when car windows are shut as airflow may disturb the alcohol detection level of MQ-3 sensor, this case can be considered for upgrading the sensor as future scope. It can be concluded that this accident prevention system using alcohol sensor and GPS module is a cost effective, reliable, power efficient and feasible solution for prevention of accidents.
- The resistance value of MQ3 alcohol sensor used in this project is different for different concentration of gases. Therefore, the sensor is needed to be calibrated carefully and its sensitivity should be adjusted using potentiometer. Temperature and humidity should also be considered while using alcohol sensor.

References

1. Gina, T.N., Rahmani, M.J.A., Westerlund, T., Liljeberg, P., Tenhunen, H.: Fog computing in healthcare internet-of-things: a case study on ECG feature extraction. In: IEEE International Conference on Computer and Information Technology, pp. 1–8 (2015)
2. Tsai, C., Lai, C., Chiang, M., Yang, T.: Data mining for internet of things: a survey. IEEE Commun. Surv. Tutorials **16**(1), 77–97 (2014)
3. Liu, B., Li, J., Chen, C., Tan, W., Member, S., Chen, Q., Zhou, M.: Efficient motif discovery for large-scale time series in healthcare. IEEE Trans. Industr. Inf. **11**(3), 583–590 (2015)
4. Rolim, C.O., Koch, F.L., Westphall, C.B., Werner, J., Fracalossi, A., Salvador, G.S.: A cloud computing solution for patient's data collection in health care institutions. In: ETELEMED'10.

Second International Conference on eHealth, Telemedicine, and Social Medicine, 2010, pp. 95–99. IEEE (2010)

5. Yang, S., Gerla, M.: Personal gateway in mobile health monitoring. In: 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), Mar 2011, pp. 636–641
6. Wang, L., Yang, G.-Z., Huang, J., Zhang, J., Yu, L., Nie, Z., et al.: A wireless biomedical signal interface system-on-chip for body sensor networks. IEEE Trans. Biomed. Circuits Syst. **4**(2), 112–117 (2010)

Automatic Classification of Music Genre Using SVM

Nandkishor Narkhede, Sumit Mathur, and Anand Bhaskar

Abstract The growing number of music content online has opened up new possibilities for the introduction of successful digital knowledge access services known as music referral systems that help user groups in searching, finding, sharing, and creating. The music recovery approach based on specific similarity information combines several similarity features, including audio and contextual similarities, such as tone format features and melodic details. Audio classification is very important for recovering audio files quickly. To get the best results from audio classification, it is important to choose the best feature set and follow the best analysis method. Support vector machines (SVMs) are implemented by learning from input samples to classify music into separate classes of music genres. The SVM study excelled in the music category classification.

Keywords SVM · ZCR · STE · LPC · RMS · MFCC · MIR and genre

1 Introduction

Information retrieval (IR) requires very little discipline and music information retrieval (MIR) requires different approaches than other field subjects. Prior to the development of the Internet, musical compositions for libraries were arranged alphabetically and were technologically advanced. Around the world of digital music, numerous studies are being conducted and how the user experience can be improved. Many unlabeled music files can be downloaded, cached, and contain incorrect or suspicious tags [1]. Automated classification of genres, however, is not an easy task to do as music develops in a short period of time [1, 2]. In addition, developments in audio and video signal processing and data exploration have resulted in a comprehensive study of music signal analysis, such as content-based music retrieval, music genre classification, duet analysis, music interpretation, and music information retrieval and music instrument identification and classification [1]. Identification

N. Narkhede (✉) · S. Mathur · A. Bhaskar
Sir Padampat Singhania University, Udaipur, Rajasthan, India
e-mail: narkhede.nandkishor@spsu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_33

techniques for musical instruments include many applications such as detecting and analyzing solo lines, retrieving audio and video, music dictation, playlist creation, group of sound background, analyzing video scenes and tagging [1].

Advanced music libraries are gaining a reputation for being professional archives and private music collections. The number of people interested in audio libraries is also increasing due to improvements in Internet access and network bandwidth [1]. But warehouses are backbreaking with a large music archive and this is time consuming, especially when classifying audio style by hand. Music is divided into genres and subspecies, based not only on sound but also on lyrics [1, 3]. This interferes with the classification. To further complicate matters, the concept of music style may change over time [4]. For example, rock songs done 5 decades ago are very different now.

2 Related Work

This section describes the work done by researchers in the similar area. Xu et al. [5] applied a multi-layer classifier based on support vector machine to classify the music genre and achieved an accuracy of 93.14%. Mutiara et al. [6] used several kernels of nonlinear support vector machines (SVM) for classification of music genres extracting feature sets related to timbre, rhythm, tonality, and LPC from music files and achieved accuracy of 76.6%. In [7], authors used polynomial SVM to classify music using MFCC features and polynomial SVM classifier to achieve an accuracy of 78%. Aryafar et al. [8] performed automatic music genre classification using spasity-eager SVM and obtained accuracy of just 37%. Kyaw and Renu [9] used multi-layer SVM for music genre classification and obtained the accuracy of 93%.

3 Proposed System

The outline of the proposed system is shown in Fig. 1. To build a dataset of input audio signals, we considered the GTZAN dataset from the Marsyas site, which contains 1000 music signals in ten different categories [10]. All audio tracks in the GTZAN dataset are.au format, 16-bit, 30 s long, 22,050 Hz mono file. Figure 2 represents the signal given as input. These audio music signals are filtered using average or mean filters. As shown in Fig. 3, this process results in the amplitude normalization and Gaussian noise elimination in the audio signal.

The segmentation process divides the audio signal into voice and silent frames. For partitioning, we used ZCR and STE as time domain properties as shown in Figs. 4 and 5, respectively, and frequency domain properties, spectral flux as depicted in Fig. 6 and spectral skewness [11]. Below are all the features used in the segmentation

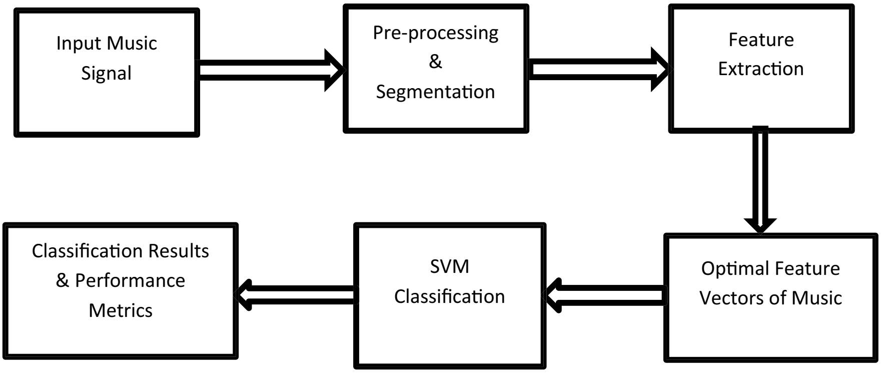

Input Music Signal
Pre-processing & Segmentation
Feature Extraction
Optimal Feature Vectors of Music
SVM Classification
Classification Results & Performance Metrics

Fig. 1 System schema

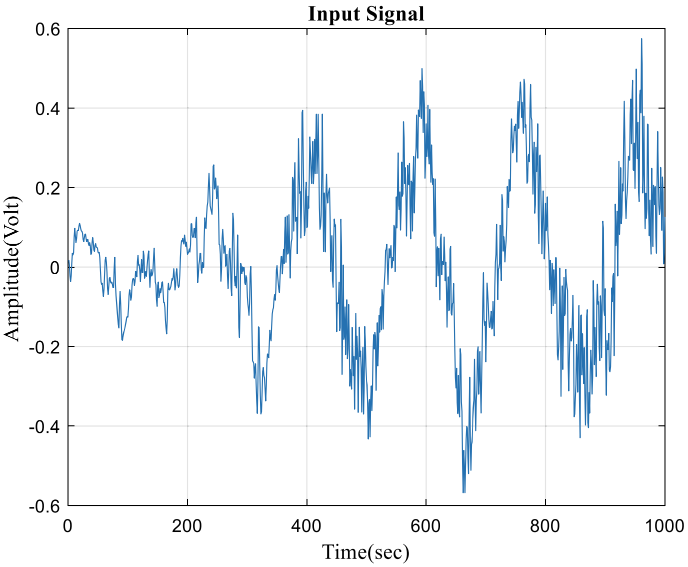


Fig. 2 Input signal

process to find the voiced segment in the input signal. The voiced segment is shown in Fig. 7.

Short-time energy [12]: Representation of Amplitude Differences. It is calculated using

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \tag{1}$$

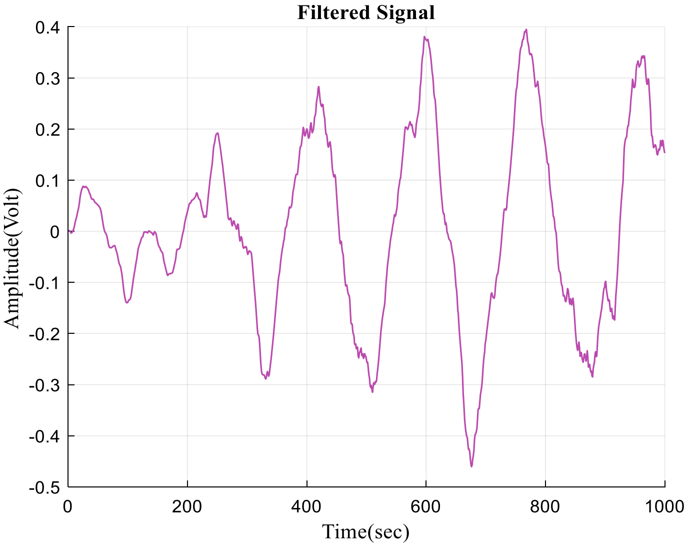


Fig. 3 Filtered signal

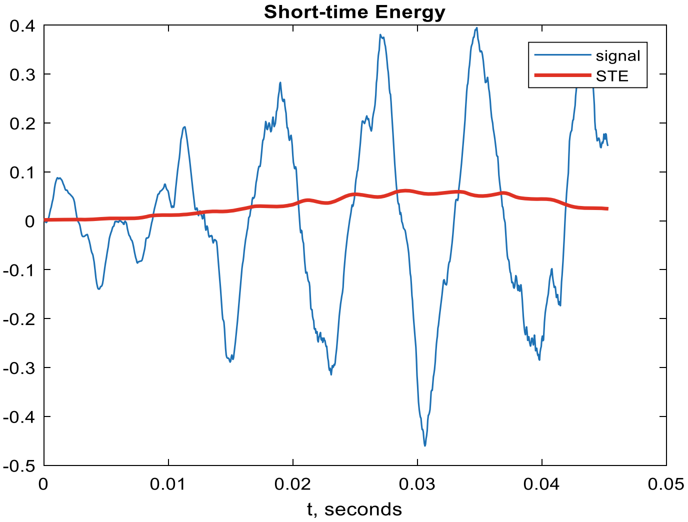


Fig. 4 Short-time energy

Zero crossing rate [12]: The ZCR in the signal reflects the mark change rate. It uses rectangular window function for measurement. Measured using this formula:

$$Z_n = \sum_{m=-\infty}^{\infty} |\mathrm{sgn}[y(m)] - \mathrm{sgn}[y(m-1)]|w(n-m) \tag{2}$$

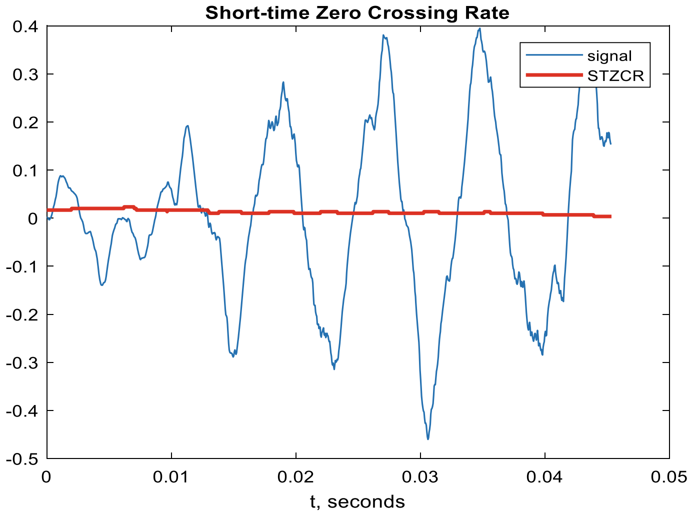


Fig. 5 Short-time ZCR

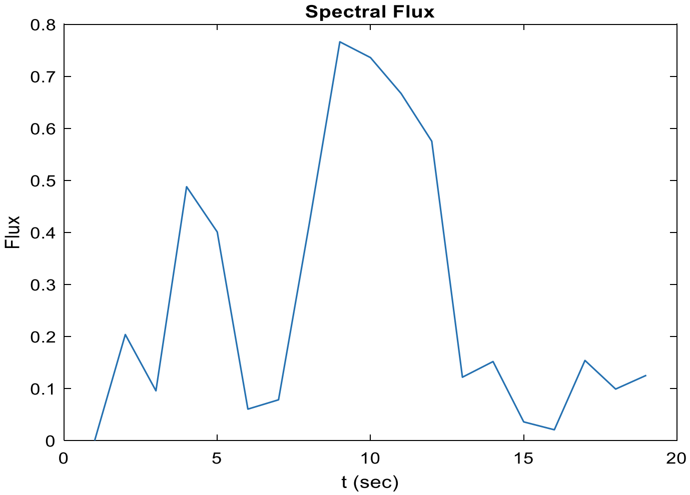


Fig. 6 Spectral flux

Spectral skewness [12]: The pitch portrayed in the music signal is skewness. In the upper and lower parts of the spectrum, the curve represents more energy [12]. Spectral skewness of input signal is −0.034447.

Spectral flux [12]: The spectral flux (SF) is the magnitude of the average spectrum difference between two consecutive frames in the provided clip [12].

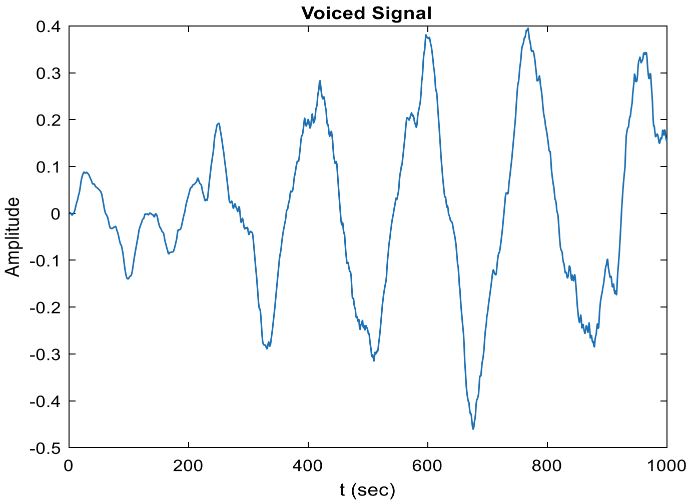


Fig. 7 Voiced segment

The next step in the proposed research is extraction of features from input signal in three separate domains. In the time domain, we use autocorrelation method to measure RMS, ZCR, and pitch salience ratio.

Root mean square (RMS) [12]: The RMS represents the square root of the mean audio amplitude over a given time period [12]. It checks the sound of the audio frame.

$$\mathrm{RMS}_j = \sqrt{\frac{1}{N}\sum_{m=1}^{N} x_j^2(m)} \tag{3}$$

Pitch saliency ratio [12]: This is the ratio of silent frames to maximum frames in the music signal [12]. If RMS < 10%, the frame is silent.

We calculated the characteristics of a frequency domain such as bandwidth, spectrogram, frequency centroid, spectral centroid, and pitch.

Bandwidth [12]: This refers to the frequency range of the signal containing data [12]. It is calculated according to the equation:

$$B_j = \sqrt{\frac{\int_0^{\omega_0} (\omega - \omega_c)\left|X_j(\omega)\right|^2 \mathrm{d}\omega}{\int_0^{\omega_0} \left|X_j(\omega)\right|^2 \mathrm{d}\omega}} \tag{4}$$

The bandwidth of the input signal is shown in Fig. 8.

Spectrogram [12]: This is a three-dimensional illustration as depicted in Fig. 9. The *X*-axis represents the properties of time. The *Y*-axis shows the frequency components of the audio signal. Dark region refers to the strength of an audio signal at that frequency [13]. The spectrogram divides the signal into overlapping segments, each

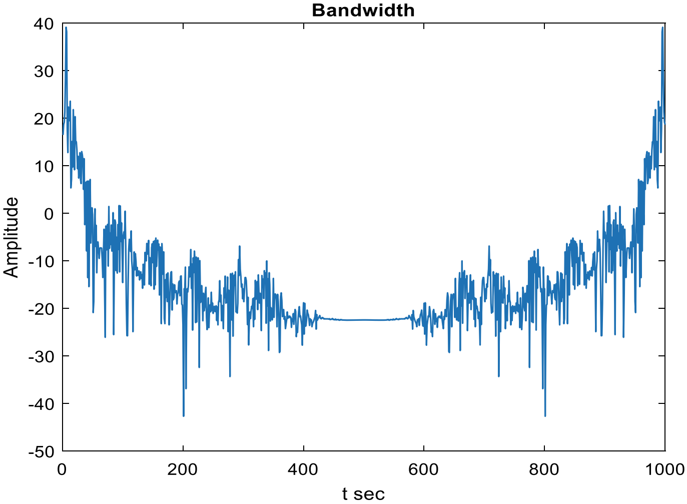


Fig. 8 Bandwidth of input signal

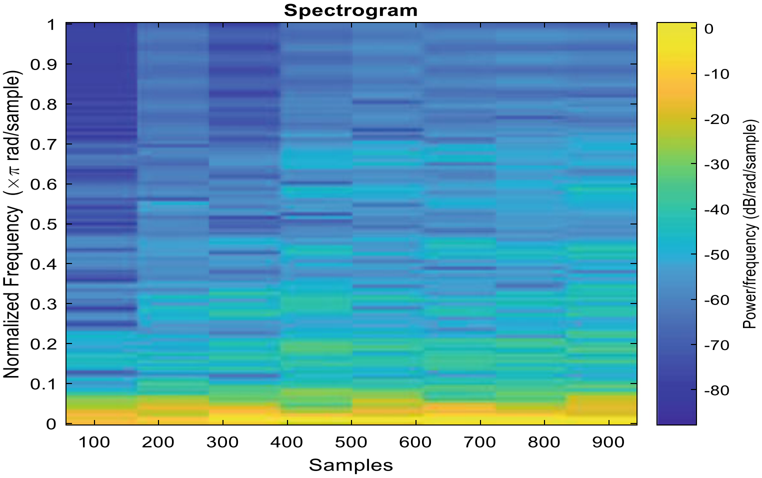


Fig. 9 Spectrogram of input signal

segment is filtered by a Hamming window, and the output is provided using N-point DFT [12].

Frequency Centroid [12]: It maintains signal brightness. It is computed using equation:

$$\omega_c j = \frac{\int_0^{\omega_0} \omega \left|X_j(\omega)\right|^2 \mathrm{d}\omega}{\int_0^{\omega_0} \left|X_j(\omega)\right|^2 \mathrm{d}\omega} \tag{5}$$

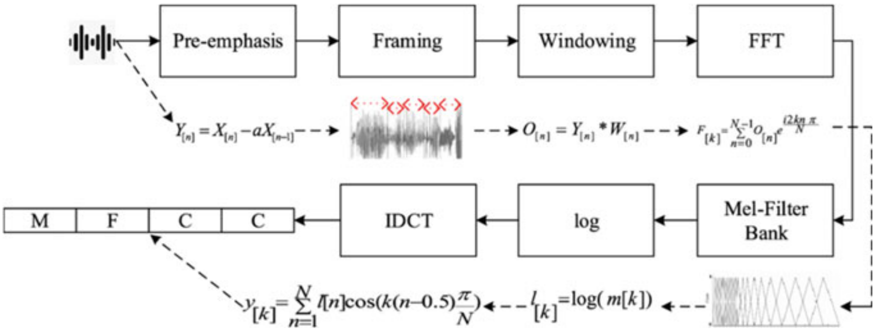

Pre-emphasis
Framing
Windowing
FFT
$Y_{[n]} = X_{[n]} - aX_{[n-1]}$
$O_{[n]} = Y_{[n]} * W_{[n]}$
$F_{[k]} = \sum_{n=0}^{N-1} O_{[n]} e^{\frac{i2kn\pi}{N}}$
M F C C
IDCT
log
Mel-Filter Bank
$y_{[k]} = \sum_{n=1}^{N} l[n] \cos(k(n-0.5)\frac{\pi}{N})$
$l_{[k]} = \log(m[k])$

Fig. 10 MFCC coefficient estimation process

Spectral centroid [12]: It deals with the chromatic variation of sound, i.e., the high frequency components of the spectrum. It is calculated using the formula:

$$C_r = \frac{\sum_{k=1}^{N/2} f[k]|X_r[k]|}{\sum_{k=1}^{N/2} |X_r[k]|} \tag{6}$$

Pitch [12]: Pitch or tone refers to the basic wavelength of the human voice [12]. Input signal pitch is 22.050 kHz.

In coefficient domain, we computed Mel Frequency Cepstral Coefficients (MFCCs). Firstly, speech data is emphasized, then emphasized data is framed according to a defined time. Then, it is applied a Hamming windowing function. Next, discrete Fourier transform is carried out to the data. Logarithm is applied to the processed data by applying mel-scale. Finally, MFCC data is obtained by applying inverse discrete cosine transform. The entire process of MFCC coefficient estimation is shown in Fig. 10.

4 SVM Classification

The "support vector machine" (SVM) is an inspected machine learning algorithm that can be used for classification or regression challenges. However, it is mostly used in classification problems [14]. Of the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is the number of feature vectors we have) and the value of each feature is the value of the specific coordinate. Next, we classify it by finding a hyperplane that separates the two classes as in Fig. 11. The support vector is a compilation of individual observations. The SVM classifier is the boundary that best separates the two classes (hyperplane/line). [14] The kernel approach plays a key role in correctly classifying a new object (test case) from the available examples (train cases). The kernels use a collection of mathematical operations to change the order of real objects. The process of rearranging an object is called the mapping

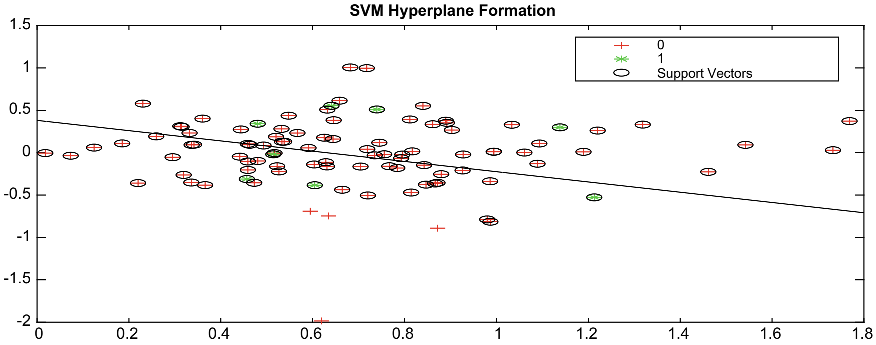
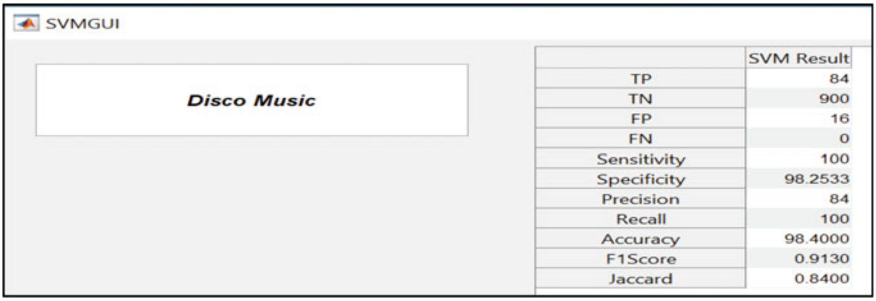


Fig. 11 SVM hyperplane

process [15]. The function of the kernel refers to the dot product of the input data points, which are transformed [16] and mapped to the high dimensional function space. Feature vectors input into the SVM classification include STE, ZCR, pitch, spectral flux, spectral centroid, and three MFCC modules. 70% of total signals in dataset are taken as training samples and remaining 30% are taken as test samples. There are 10 different genres of music in the used dataset. The linear kernel function is used for the experiment.

5 Result Analysis

The input signal can be also classified as blues, classical, country, rock, reggae, jazz, metal, hiphop, pop, disco, etc., depending upon which genre signal is taken. In this paper, we show the classification of input signal as disco. The performance metrics of SVM classifier computed for the experiment with respect to this input signal are shown in Fig. 12. SVM learning in the proposed system demonstrated better results



Disco Music

	SVM Result
TP	84
TN	900
FP	16
FN	0
Sensitivity	100
Specificity	98.2533
Precision	84
Recall	100
Accuracy	98.4000
F1Score	0.9130
Jaccard	0.8400

Fig. 12 Classification of music genre and performance metrics of SVM classifier

in classification of music genres with an accuracy of 98.4% that is comparatively higher.

6 Conclusion

In this paper, we demonstrated the classification of music genres using SVM classifier with linear kernel function. The music signals from GTZAN dataset were taken for experiment. The feature vectors from three different domains like time, frequency and cepstral domain were computed and given as the inputs to the classifier. The SVM classifier outperformed well giving the classification accuracy of 98.4% which is higher as compared to accuracies obtained by researchers in literature review.

References

1. Thiruvengatanadhan, R.: Music genre classification using SVM. Int. Res. J. Eng. Technol. (IRJET) **05**(10), 1059–1061 (2018)
2. Joder, C., Essid, S., Richard, G., Member, S.: Temporal integration for audio classification with application to musical instrument classification. IEEE Trans. Speech Audio Process. **17**(1), 174–186 (2009)
3. Serwach, M., Stasiak, B.: GA-based parameterization and feature selection for automatic music genre recognition. In: Proceedings of 2016 17th International Conference on Computational Problems of Electrical Engineering, CPEE (2016)
4. Van Dijk, L.: Bachelorthesis Information Science: Finding Musical Genre Similarity Using Machine Learning Techniques, pp. 1–25. Radboud Universiteit Nijmegen (2014)
5. Xu, C., Maddage, N.C., Shao, X., Cao, F., Tian, Q.: Music genre classification using support vector machines
6. Mutiara, A.B., Refianti, R., Mukarromah, N.R.A.: Musical genre classification using support vector machines and audio features. TELKOMNIKA **14**(3), 1024–1034
7. Patil, N.M., Nemade, M.U.: Music genre classification using MFCC, K-NN and SVM classifier. Int. J. Comput. Eng. Res. Trends **4**(2), 43–47 (2017)
8. Aryafar, K., Jafarpour, S., Shokoufandeh, A.: Automatic musical genre classification using sparsity-eager support vector machines. In: 21st International Conference on Pattern Recognition (ICPR 2012), 11–15 Nov 2012, Tsukuba, Japan
9. Kyaw, L.Y., Renu: Using support vector machine for music genre classification
10. http://marsyas.info/downloads/datasets.html
11. Pradeep Kumar, D., Sowmya, B.J., Chetan, K.G.S.: A comparative study of classifiers for music genre classification based on feature extractors. IEEE, pp. 190–194 (2016)
12. Patil, N.M., Nemade, M.U.: Content-based audio classification and retrieval using segmentation, feature extraction and neural network approach. In: Bhatia, S., Tiwari, S., Mishra, K., Trivedi, M. (eds.) Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing, vol. 924. Springer, Singapore. https://doi.org/10.1007/978-981-13-6861-5_23
13. Yandre, M.G.C., Oliveira, L.S., Silla, Jr., C.N.: An evaluation of convolutional neural networks for music classification using spectrograms. Appl. Soft Comput. 1–39 (2016)
14. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

15. https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989
16. Mutiara, A.B., Refianti, R., Mukarromah, N.R.A.: Musical genre classification using support vector machines and audio features. TELKOMNIKA **14**(3), 1024–1034 (2016)

Crime Rate Prediction Based on K-means Clustering and Decision Tree Algorithm

Jogendra Kumar, M. Sravani, Muvva Akhil, Pallapothu Sureshkumar, and Valiveti Yasaswi

Abstract The major cause of crimes that initiate nuisance for society in many ways is human behavior disorder. In many countries, the crimes and accidents are seriously monitored. Crime analysis is a process, which completely analyses the patterns and trends over a period of time. Nowadays, different sources of crime data provide a greater opportunity for performing large analysis in the research community. The proposed work analyses the crime under different locations (considering latitude and longitude) and different time periods. The proposed work predicts the crime type and gives an input about the date and location. The application is developed as a Windows application by using TKinter-Python for crime prediction. Machine learning concepts and implementation are used here for performing crime analysis and prediction, which aid the ease of understanding the data in multiple ways and further predict it with good accuracy. The algorithms used behind this work are K-means and decision tree algorithm. The proposed work aims at analyzing the data mining concepts for clustering and classifying the crime prediction. The results show that the classification method outperforms in terms of detection and accuracy. Experimented results are evaluated for error calculation, and they are further analyzed in this study.

Keywords Crime analysis · Crime prediction · Data mining · Decision tree · K-means algorithm · Machine learning · Classification · Clustering

1 Introduction

Around the world, every day, the crime rate is growing at an unprecedented rate. As a result, security among the society has become a big threat, and it increases the complexity on continuous monitoring paradigm. Robbery, burglary, arson, murder, and trafficking are the histories growing every year. Though the victims cannot be predicted, the place of crime and the probability for crime in particular place can be

J. Kumar (✉) · M. Sravani · M. Akhil · P. Sureshkumar · V. Yasaswi
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_34

predicted. Though the prediction cannot be with full accuracy, however, the results can be helpful in analyzing and focusing on reducing crime rates. Criminology is one of the most important domain, where the data mining concepts can give the best results in reduction of crime rates. These days, the data is highly available, which can be exploited for further study and avoiding crimes at certain places.

Understanding the characteristics of crime is important for further analysis. Finding crime patterns is a major study in countries like England, and they have extracted some of the patterns from the offenders. These data included the place of crime, property used for crime, day of week, etc. Finding crime patterns with above data was accurate, and it has given provable accuracy. The data mining concepts like clustering and classification are also useful for police patrol to perform further analysis. Crime detection requires lots of intelligence to solve the problem by integrating the possible machine learning techniques. Spatial data mining plays a major role, which is highly available and affordable for a few years. Huge amount of dataset is available for exploring the crime details, thus making an effective tool for identifying the crime pattern remains as a major challenge.

The increased problem in this area of study may include storing and analyzing the crime information. Some form of incomplete data is available at high ranges; thus, it may affect the accuracy of predictions. Recording crime data should follow multiple methods; thus, inconsistency is a major problem. Law enforcement department must share all the available data. Investigating and collecting relevant information takes much human effort and consumes more time. Moreover, the data will have the solved crimes, where the patterns are available. However, the nature of crime may change over a period of time, and thus, it may become as an unknown pattern. These challenges in crime prediction have motivated to propose a machine learning method to solve the aforementioned challenges (Fig. 1).

In the proposed system, we have done crime data analysis with many parameters and factors including daily arrests, monthly arrests, number of domestic violence, top 5 monthly, weekly, and daily crime are visualized.

Using decision tree algorithm and K-means clustering algorithm, we are predicting the type of crime for the given latitude and longitude.

In the following chapters, related works are discussed in Sect. 2. Section 3 discusses the proposed methodology and crime prediction implementation in detail. In Sect. 4, evaluations of decision tree algorithm and K-means are discussed. Finally, on the last section, work Conclusion and directions for further enhancements are discussed.

2 Related Work

In the literature, existing works on crime prediction and analysis are done with various data mining techniques and machine learning algorithms. Some of the existing works are discussed in detail in this chapter.

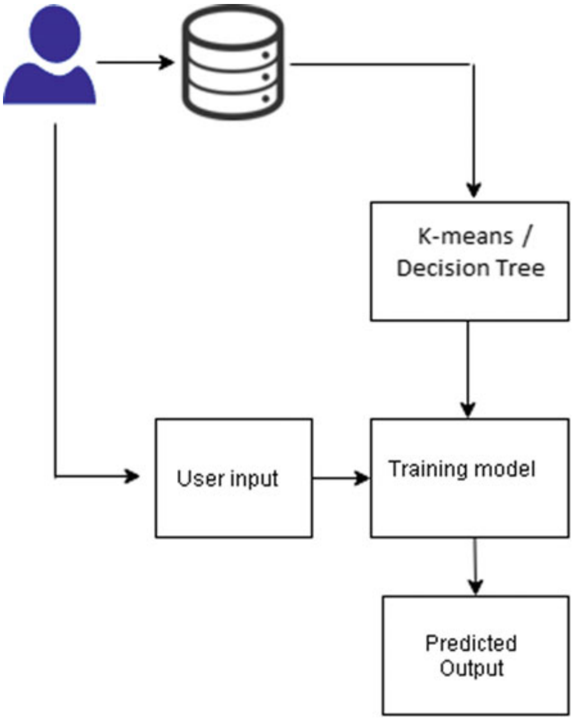


Fig. 1 Overview of crime prediction

One of the data mining concepts, association rule mining is applied in [1] for mining profile information from log data. Concept hierarchies and belief factor were running in background to collect the profile information. The traditional process of profile filtration namely data filter and data conversion were used; then, association rule is applied to get the required profiles.

Time series analysis is discussed in [2] for clustering approach to find the hot spot of crime. The dataset with spatiotemporal data was considered. The dataset created with a certain number of questionnaire with precision and context information is collected, and then, performance of system was studied with machine learning techniques for finding the hot spots.

Few more interesting studies were done on spatiotemporal crime spot detection. The work on [3] considered urban crime prediction in China, and the method used here was ARIMA model. This model has predicted the situation of crime in a particular place and time period.

The work [4] studied spatiotemporal patterns of crimes on urban crime dataset. The authors developed a transfer learning model which is integrated with urban crime prediction in New York City (NYC) dataset over the period of 2012 to 2013. The study considered meteorological data and point of information to study the crime pattern.

Classification approach on crime prediction UCI crime dataset was studied in [5]. The author evaluated the performance of support vector machine (SVM) and random forest (RF) classification models. The study proved that the SVM algorithm outperformed in crime prediction through its effective classifications.

The author in [6] studied crime occurrence using spatiotemporal data through seasonal auto regressive technique SARIMA, and the spatiotemporal patterns are evaluated through long short-term model (LSTM) of machine learning algorithms. Through experimental results, the author proved that LSTM performed good.

The inference from the existing study concludes that there is a high demand for highly effective model to get optimized results for crime prediction [9]. The existing work discussed above was studied enormously on applying data mining techniques like association rule algorithms for crime prediction [10]. However, there is a method which needs to be effective on prediction, which should be an effective model for spatiotemporal dataset [11].

3 Proposed Work

The proposed work is an application, which enables the user to give geographical input such as latitude and longitude values [12]. The application enables to classify crime in as many types as possible. Here in dataset, there are thirty types considered. The work implemented in Python with TK inter as Windows application; the libraries and packages installed are scikit-learn, pandas, and matplotlib [13]. The dataset used for implementation is Chicago Crime dataset from 2011 to 2015 dataset with more than 2 lakh instances. Decision tree and K-means algorithm are applied for crime prediction. The application used to identify crime types and complete visualization is also done [14].

Dataset Details

Crime dataset with incidents of crime monitored in City of Chicago over the period of 2001–2015 is considered for the study [15]. Data from Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) is considered in this study. Privacy of crime victims is ensure by not declaring their contact or addresses (Fig. 2).

The details of crime dataset and its attribute names and description are given in Table 1.

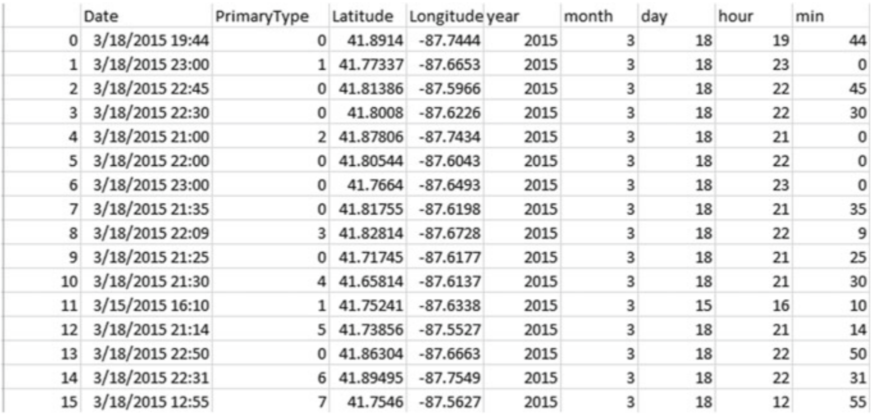
Data preprocessing includes data visualization and cleaning. Data cleaning includes converting time stamp to split year, month, day, hour, minute, and seconds. The dataset is split into train and test set and given to machine learning algorithm of *K*-means and decision tree to train the models. The dataset is split as below mentioned.

X_Train represents the following input.

data_train[['Latitude', 'Longitude', 'year', 'month', 'day', 'hour', 'min']].

Y_Train represents the following input.

data_train [['primarytype']]).



	Date	PrimaryType	Latitude	Longitude	year	month	day	hour	min
0	3/18/2015 19:44	0	41.8914	-87.7444	2015	3	18	19	44
1	3/18/2015 23:00	1	41.77337	-87.6653	2015	3	18	23	0
2	3/18/2015 22:45	0	41.81386	-87.5966	2015	3	18	22	45
3	3/18/2015 22:30	0	41.8008	-87.6226	2015	3	18	22	30
4	3/18/2015 21:00	2	41.87806	-87.7434	2015	3	18	21	0
5	3/18/2015 22:00	0	41.80544	-87.6043	2015	3	18	22	0
6	3/18/2015 23:00	0	41.7664	-87.6493	2015	3	18	23	0
7	3/18/2015 21:35	0	41.81755	-87.6198	2015	3	18	21	35
8	3/18/2015 22:09	3	41.82814	-87.6728	2015	3	18	22	9
9	3/18/2015 21:25	0	41.71745	-87.6177	2015	3	18	21	25
10	3/18/2015 21:30	4	41.65814	-87.6137	2015	3	18	21	30
11	3/15/2015 16:10	1	41.75241	-87.6338	2015	3	15	16	10
12	3/18/2015 21:14	5	41.73856	-87.5527	2015	3	18	21	14
13	3/18/2015 22:50	0	41.86304	-87.6663	2015	3	18	22	50
14	3/18/2015 22:31	6	41.89495	-87.7549	2015	3	18	22	31
15	3/18/2015 12:55	7	41.7546	-87.5627	2015	3	18	12	55

Fig. 2 Crime data

Table 1 Attributes and its description

Attribute	Description
ID	Case ID
CaseNumber	Case number
Date	Date of crime
Block	Block ID
PrimaryType	Crime type
Des	Description
LocDes	Location description
Lat	Latitude of crime place
Lon	Longitude of crime place

As the crime type is predicted, the Y column/target column is set as crime type.

Figures 3 and 4 show the visualization of crime dataset. The data visualization has most significance in understanding crime's nature and study more about the crime type and according to geographical location. This is useful for complete understanding of the nature of dataset. In visualization part, crime visualization based on crime type, location, weekly crime, monthly crimes, and top 5 crimes based on weekly, daily, and monthly is plotted for analysis.

The implementation is carried out with below methodologies, and it is also represented as architecture in Fig. 3. The work has the following benefits.

- The dataset is visualized to understand the pattern completely with time, crime type, location, etc.
- The effective learning model is built using decision tree and *K*-means algorithm
- The application ensures that user can give input and get live predictions of crime type in given location and over a period of time

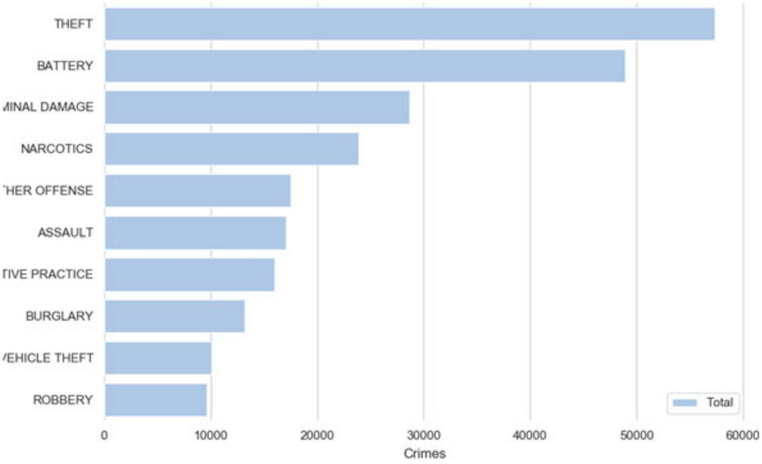
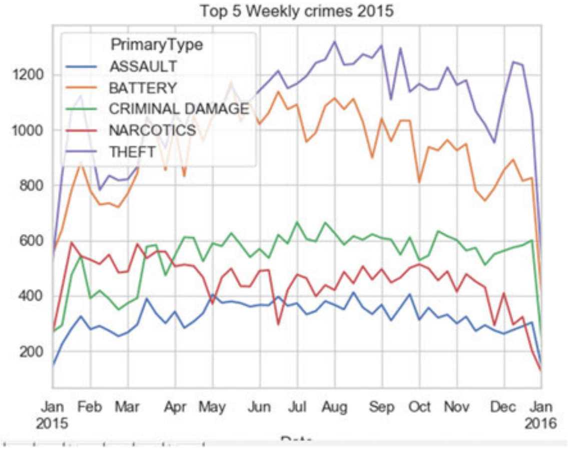


Fig. 3 Visualizing of Crime dataset by type

Fig. 4 Visualizing of crime dataset by type-by week



The following methodologies were carried out on implementation.

a. Chicago dataset collected from uci.edu repository.
b. Visualizing dataset as graph for week, month, and day wise
c. Split data as X-train, Y-train, X-test, and Y-test
d. Apply machine learning classification algorithm decision Tree and K-means algorithm
e. Train the ML model
f. Trained model is tested with 20% dataset from test set
g. Give single input from user through Windows application interface and predict crime type

The crime type prediction is done by getting test input from user through graphical user interface (GUI). The latitude, longitude, and timestamp should be given as input to predict the crime type in that location. After getting the user input, the preprocessing of data is done for the required format.

Figure 5 represents system architecture Crime prediction process. Chicago, CLEAR dataset is considered for implementation with decision tree algorithm and *K*-means algorithm.

a. Decision Tree Algorithm

Decision tree is one of the traditional methods used for dataset classification. Class label prediction is started from the root of the tree for decision making. Root attribute is compared with the instance value. According to the compared value with branch, it is forwarded to next node. The tree is constructed such that branches are results and leafs are decision made by algorithm.

Pseudo-code for Decision Tree Algorithm

```
Step 1:    Entropy for crime dataset is calculated
```

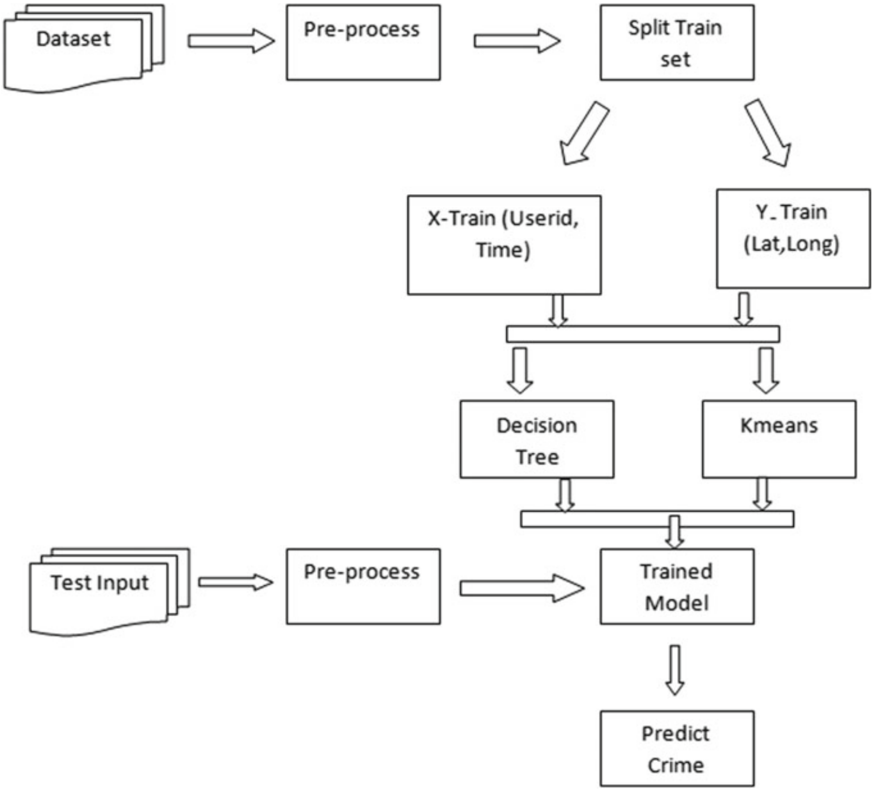


Fig. 5 Architecture—crime prediction

```
Step 2:  for an attribute or a feature:
    Entropy is calculated with categorical values.
    Average entropy is considered for available attribute.
    Gain for attribute is calculated.
Step 3:  Sort and find highest value of gain
Step 4:  Repeat above till tree is formed
```

b. *K*-means Clustering Algorithm

Clustering is more suitable for unsupervised datasets, and thus, it is interesting to go for clustering on a crime dataset. This creates or partitions data into groups, which is called clusters [7]. Thus, the data identified through this model is useful for predicting crime type.

Pseudo-code for *K*-means Clustering Algorithm

```
Step 1.  Get a predefined $k$ value, where $k$ is the number of cluster to form.
Step 2   On the cluster center, there are k points chosen.
Step 3   Euclidean distance function is used to match the given object/record on
         the cluster center.
Step 4   The centroid value is calculated for all record in all clusters.
Step 5   Repeat steps 2, 3, and 4 till the points are matched to cluster.
```

The Windows application is created in Python TK inter, where user can able to get results from interface [8]. The application home page is given in Fig. 6. This has minimum input such as geographic location input in terms of latitude, longitude, date, and time. User can able to get the crime type as result from the learned machine learning model.

The crime types considered in this dataset is around 31 type are 'Battery,' 'Other Offense,' 'Robbery,' 'Narcotics,' 'Criminal Damage,' 'Weapons Violation,' 'Theft,' 'Burglary,' 'Motor Vehicle Theft,' 'Public Peace Violation,' 'Assault,' 'Criminal Trespass,' 'Crim Sexual Assault,' 'Interference With Public Officer,' 'Arson,' 'Deceptive Practice,' 'Liquor Law Violation,' 'Kidnapping,' 'Sex Offense,' 'Offense Involving Children,' 'Prostitution,' 'Gambling,' 'Intimidation,' 'Stalking,' 'Obscenity,' 'Public Indecency,' 'Human Trafficking,' 'Concealed Carry License Violation,' 'Other Narcotic Violation,' 'Homicide,' and 'Non-Criminal.'

4 Results and Discussions

The proposed work is implemented in Python 3.6.4 or Anaconda 3 with a few important libraries namely scikit-learn, pandas, and matplotlib. The crime dataset is applied machine learning algorithm such as decision tree and K-means algorithm. We used these machine learning algorithms and identified crime categories possible in particular location, with specified date and time. The decision tree algorithm achieves good accuracy (Fig. 7).

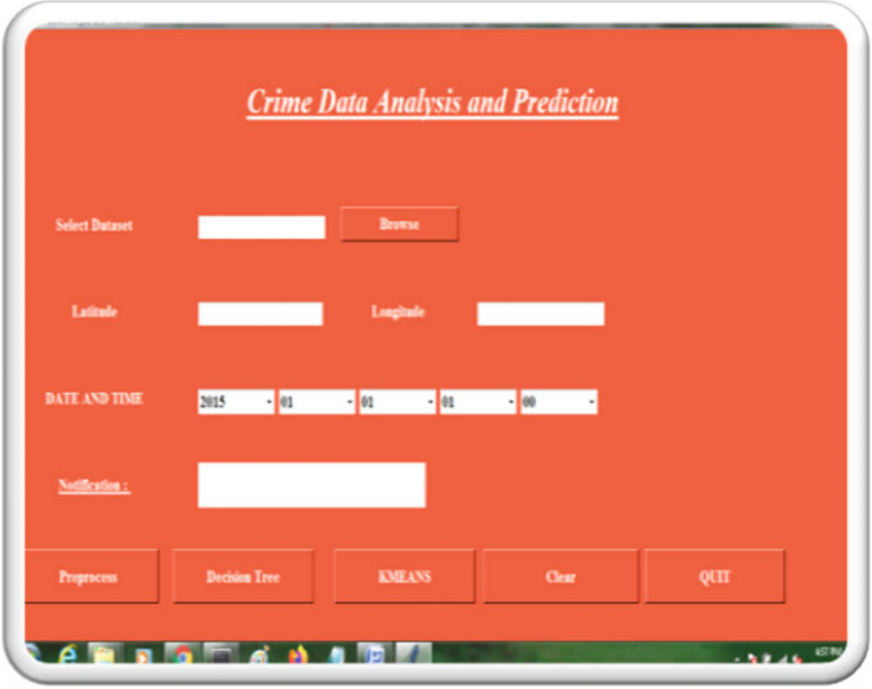
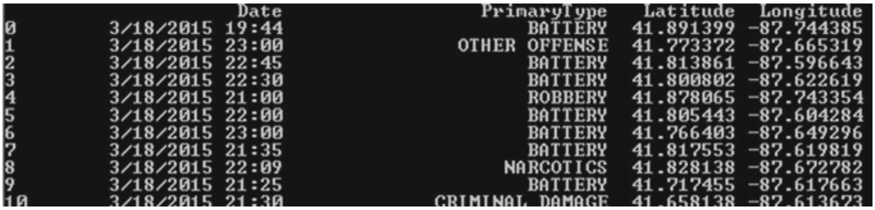


Fig. 6 Application homepage

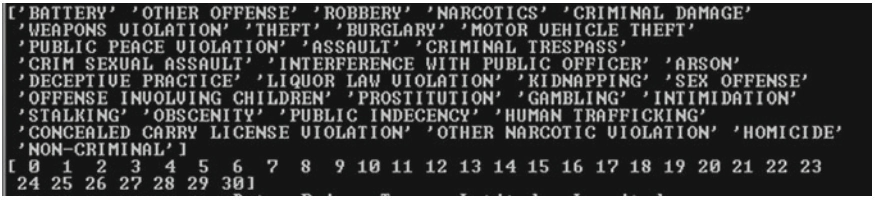


```
                Date                     PrimaryType  Latitude  Longitude
0         3/18/2015 19:44                    BATTERY  41.891399 -87.744385
1         3/18/2015 23:00              OTHER OFFENSE  41.773372 -87.665319
2         3/18/2015 22:45                    BATTERY  41.813861 -87.596643
3         3/18/2015 22:30                    BATTERY  41.800802 -87.622619
4         3/18/2015 21:00                    ROBBERY  41.878065 -87.743354
5         3/18/2015 22:00                    BATTERY  41.805443 -87.604284
6         3/18/2015 23:00                    BATTERY  41.766403 -87.649296
7         3/18/2015 21:35                    BATTERY  41.817553 -87.619819
8         3/18/2015 22:09                  NARCOTICS  41.828138 -87.672782
9         3/18/2015 21:25                    BATTERY  41.717455 -87.617663
10        3/18/2015 21:30            CRIMINAL DAMAGE  41.658138 -87.613623
```

Fig. 7 Preprocess results to clean data

There are a total of 31 crime types available in the dataset, whereas accuracy may degrade due to the high number of classes. The output of class conversion to numeric value is shown in Fig. 8.

The machine learning algorithm decision tree is trained using given features from crime dataset. The dataset is considered 80% for training and 20% as test set. The *X*-train values are such as lat and long of geographical data, date, and time, and *Y*-train is class value of crime type, which is considered as multi-class 0–30. Once the algorithm is trained, *X*-test is given as input and algorithm accuracy and error values are evaluated. The accuracy achieved by algorithm is provided in Figs. 9 and 10.



```
['BATTERY' 'OTHER OFFENSE' 'ROBBERY' 'NARCOTICS' 'CRIMINAL DAMAGE'
 'WEAPONS VIOLATION' 'THEFT' 'BURGLARY' 'MOTOR VEHICLE THEFT'
 'PUBLIC PEACE VIOLATION' 'ASSAULT' 'CRIMINAL TRESPASS'
 'CRIM SEXUAL ASSAULT' 'INTERFERENCE WITH PUBLIC OFFICER' 'ARSON'
 'DECEPTIVE PRACTICE' 'LIQUOR LAW VIOLATION' 'KIDNAPPING' 'SEX OFFENSE'
 'OFFENSE INVOLVING CHILDREN' 'PROSTITUTION' 'GAMBLING' 'INTIMIDATION'
 'STALKING' 'OBSCENITY' 'PUBLIC INDECENCY' 'HUMAN TRAFFICKING'
 'CONCEALED CARRY LICENSE VIOLATION' 'OTHER NARCOTIC VIOLATION' 'HOMICIDE'
 'NON-CRIMINAL']
[ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 24 25 26 27 28 29 30]
```

Fig. 8 Results of preprocess of class label conversion

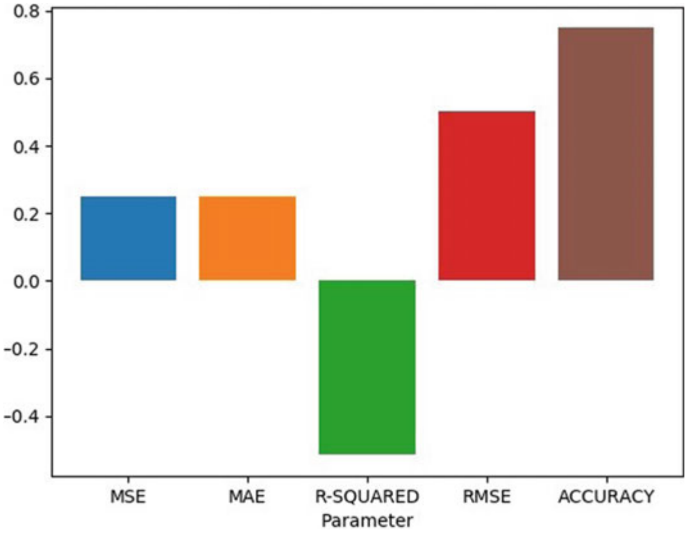


Fig. 9 Crime prediction through decision tree

The evaluation metrics calculated are accuracy, mean square error (MSE), mean absolute error (MAE), *R*-Squared error, and root mean square error (RMSE) are evaluated and shown above. Experimental evaluations show that the proposed system model of decision tree and *K*-means clustering outperforms in detection of crime type.

5 Conclusion

Crime is a behavior disorder prevailing in all countries and growing over the period of time. The earlier detection of the crime may consume a lot of human intelligence. However, data mining concepts and machine learning models are widely used these days for accuracy prediction of crime type. In this work, decision tree and K-means algorithm are proposed for crime prediction. Decision tree is a classification algorithm and *K*-means is a clustering algorithm, where both are implemented

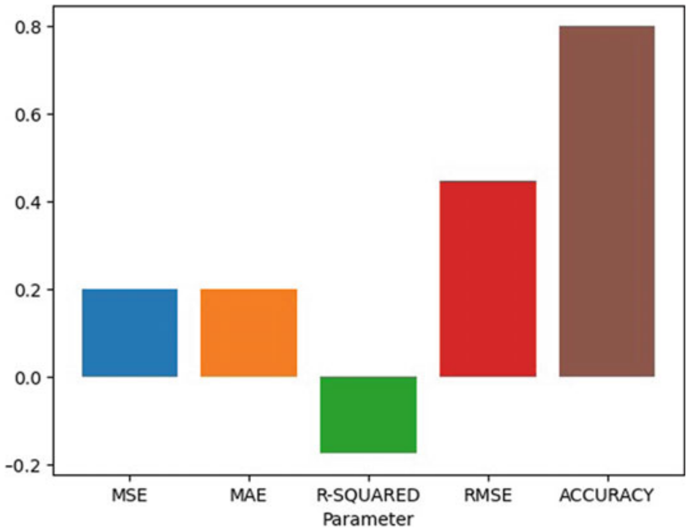


Fig. 10 Crime prediction through K-means algorithm

and evaluated for crime prediction. The work also proposed with a windows interface, wherein user gains the ability to given geographical details such as latitude, longitude, date, and time to get the crime predicted. This application detects crime as type 0–30 different classes. Experimental results show that the system outperforms in detection, even though considered as multi class problem. In future, we are interested to use deep learning models such as convolutional neural networks (CNN) or deep neural networks (DNN).

References

1. Abraham, T., de Vel, O.: Investigative profiling with computer forensic log data and association rules. In: IEEE International Conference on Data Mining. Proceedings, Maebashi City, Japan, 2002, pp. 11–18. https://doi.org/10.1109/ICDM.2002.1183880
2. Butt, U.M., Letchmunan, S., Hassan, F.H., Ali, M., Baqir, A., Sherazi, H.H.R.: Spatio-temporal crime HotSpot detection and prediction: a systematic literature review. IEEE Access **8**, 166553–166574 (2020). https://doi.org/10.1109/ACCESS.2020.3022808
3. Li, Z., Zhang, T., Yuan, Z., Wu, Z., Du, Z.: Spatio-temporal pattern analysis and prediction for urban crime. In: 2018 Sixth International Conference on Advanced Cloud and Big Data (CBD), Lanzhou, China, 2018, pp. 177–182. https://doi.org/10.1109/CBD.2018.00040
4. Zhao, X., Tang, J.: Exploring transfer learning for crime prediction. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 2017, pp. 1158–1159. https://doi.org/10.1109/ICDMW.2017.165
5. Zaidi, N., Mustapha, A., Mostafa, S., Razali, M.: A classification approach for crime prediction (2020). https://doi.org/10.1007/978-3-030-38752-5_6.
6. Ibrahim, N., Wang, S., Zhao, B.: Spatiotemporal crime hotspots analysis and crime occurrence prediction (2019). https://doi.org/10.1007/978-3-030-35231-8_42

7. Kumar, J.: Hybrid image segmentation model based on active contour and graph cut with fuzzy entropy maximization. Int. J. Appl. Eng. Res. **12**(23), 13623–13637 (2017). ISSN 0973-4562 © Research India Publications. http://www.ripublication.com
8. Rohini, D.V., Isakki, P.: Crime analysis and mapping through online newspapers: a survey. In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, India, 2016, pp. 1–4. https://doi.org/10.1109/ICCTIDE.2016.7725331
9. Murray, A., Grubesic, T.: Exploring spatial patterns of crime using non-hierarchical cluster analysis (2013). https://doi.org/10.1007/978-94-007-4997-9_5
10. McDowall, D., Loftin, C., Pate, M.: Seasonal cycles in crime, and their variability. J. Quant. Criminol. **28** (2011). https://doi.org/10.1007/s10940-011-9145-7
11. Kitchenham, B., Mendes, E., Travassos, G.: Cross versus within-company cost estimation studies: a systematic review. IEEE Trans. Softw. Eng. **33**, 316–329 (2007). https://doi.org/10.1109/TSE.2007.1001
12. Kapoor, P., Singh, P., Cherukuri, A.K.: Crime Data set analysis using formal concept analysis (FCA): a survey (2020). https://doi.org/10.1007/978-981-15-0372-6_2
13. Chauhan, C., Sehgal, S.: A review: Crime analysis using data mining techniques and algorithms. In: International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 21–25. https://doi.org/10.1109/CCAA.2017.8229823
14. Anand, J.V.: A methodology of atmospheric deterioration forecasting and evaluation through data mining and business intelligence. J. Ubiquit. Comput. Commun. Technol. (UCCT) **2**(02), 79–87 (2020)
15. Chakrabarty, N.: A regression approach to distribution and trend analysis of quarterly foreign tourist arrivals in India. J. Soft Comput. Paradigm (JSCP) **2**(01), 57–82 (2020)

Comparative Study of Optimization Algorithm in Deep CNN-Based Model for Sign Language Recognition

Rajesh George Rajan and P. Selvi Rajendran

Abstract The fundamental part of the neural network is the learning rate, and the strategy of adopting the learning process in a neural network is carried out using optimization algorithms or optimizers. This optimization algorithm helps us produce better results to the model by changing the parameters like bias and weights, i.e., it helps us maximize or minimize the error function and depends on the learnable parameters. In this paper, we examine how an End-to-End CNN model named ASLNET recognizes the alphabets of the American sign language using various optimizers such as Stochastic Gradient Descent (SGD), Root-Mean-Square propagation (RM-Sprop), Adaptive Gradient Algorithm (Adagrad), Adaptive Delta (Adadelta),Adaptive Moment Estimation (Adam), Adam with Nesterov Momentum (Nadam), LookAhead and Rectified Adam (RAdam). To avoid the overfitting issues, traditional data augmentation techniques are used to compare our model with data augmentation and without augmentation with these optimizers. Among these, LookAhead and RAdam are the most recently developed. The experiment is conducted on 2 NVIDIA TESLA P100 GPUs of batch size 64, and the investigation was based on benchmark ASL Finger Spelling dataset.

Keywords Optimization algorithms · Deep CNN · Finger Spelling dataset · Sign language recognition

1 Introduction

Every researcher has been looking for fast and stable optimization algorithms. Remarkably, stochastic gradient-based optimization, such as stochastic gradient descent (SGD), has made a considerable achievement in several areas despite its simplicity. There were mainly two types of optimization algorithms called first-order and second-order optimizations. The optimization algorithm widely utilized for the first order is gradient descent. To improve deep learning performance, we use many optimization algorithms to estimate the weights of connections between

R. G. Rajan (✉) · P. S. Rajendran
Hindustan Institute of Technology and Science, Chennai, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_35

nodes. In the area of science and engineering, stochastic gradient-based optimization has a significant role. The optimization of some objective function needs to be maximized or minimized with respect to its parameters that can be put in various issues in these aspects. When the objective function is distinguishable from its parameters, the gradient descent is a relatively effective way of optimization, since the first-order partial derivatives calculation with respect to all parameters is of the same computational complexity during function evaluation. An efficient optimization technique that has been utilized in a variety of machine learning methods, including recent advancements in deep learning [1–4].

In deep learning, adaptive gradient approaches have been typically utilized. While stochastic gradient descent (SGD) has been one of the most common algorithms for many years, training deep neural networks has trouble overcoming severe problems such as ill-conditioning and time requirements for large-scale datasets. Manual tuning of the learning rate is needed, and it is hard to parallelize. Thus, the difficulties of SGD allowed more complex algorithms to be invented. The optimization algorithms used by deep learning are currently changing their learning speeds through training. Basically, for each parameter, the adaptive gradient strategies change the learning rate.

2 Optimization Algorithms

In general, optimization algorithm is a parameter estimation technique that was executed repeatedly by evaluating different values until an optimum solution is found. There were different types of optimization algorithms. They were as follows:

2.1 Stochastic Gradient Descent (SGD)

Stochastic gradient descent (SGD) considers a specific dataset to execute every iteration, i.e., a batch size of one [5]. The dataset is shuffled randomly and chosen for performing the iteration 'k'. It is one of the faster techniques and performs only one update at a time, where $x(i)$, $y(i)$ are the training minibatches of 'm' examples, ε_k is the learning rate, and $J(\theta)$ is the gradient of loss function—$J(\theta)$ with respect to parameters—'θ'.

$$\hat{g} \leftarrow \frac{1}{m}\nabla_\theta \sum\nolimits_i L\big(f\big(x^{(i)};\theta\big), y^{(i)}\big)$$
$$\theta \leftarrow \theta - \epsilon_k \hat{g}$$

2.2 Adagrad

It is the algorithm for gradient-based optimization, and it modifies the learning rate to the parameters, providing smaller updates for regularly occurring features and more extensive updates for the infrequent features [5]. Dean et al. [6] found that the improvement and robustness of SGD, Adagrad is used for training large-scale neural nets. It merely allows the learning rate η to adapt based on the parameters. Adagrad adjusts the general learning rate η for each parameter $\theta(i)$ at each time step t based on the previous gradients estimated for $\theta(i)$. The primary issue with Adagrad is the decay of the learning rate.

2.3 AdaDelta

An extension of the Adagrad is AdaDelta to rectify the decaying learning rate. It limits the window of accumulated past gradients to some fixed size w, rather than collecting all previously squared gradients. Another thing about AdaDelta is that it does not even have to assign a learning rate by default [7].

2.4 RMSProp

Hinton [3, 8] also suggests an adaptive learning rate approach to address the rapidly declining rate of Adagrad. RMSprop often splits the learning rate by an exponentially decreasing average squared gradient. Hinton [3] suggests that γ must be chosen to 0.9, whereas a better default value is 0.001 for the learning rate η.

2.5 Adam

Adam depicts for adaptive moment estimation. Adam is another approach that assesses the value of adaptive learning for each parameter [9]. Besides keeping an exponentially decaying average of previously squared gradients such as AdaDelta, it also holds an exponentially declining average of previously gradients $M(t)$, AdaDelta.

2.6 AdaMax

AdaMax is an adaptive form of stochastic gradient descent and an Adam version based on the norm for infinity. Unlike the SGD, AdaMax offers the significant advantage of being far less sensitive to choosing the hyper-parameters [9].

2.7 NAdam

It is also an optimization method used for noisy gradients or high curvatures gradients. The learning process is strengthened by summing up the exponential decay for the previous and current gradient of the moving averages [10].

2.8 LookAhead

LookAhead was motivated by the latest developments in the understanding of loss surfaces of deep neural networks. This algorithm calculates weight updates by looking ahead to the 'fast weights' sequence generated by some other optimizer. LookAhead increases the reliability of learning, reduces the variance of its internal optimizer and has little memory and computation costs [11].

2.9 RAdam

It is a novel modified form of Adam, by bringing a term called 'warm-up' to resolve adaptive learning rate variance [12]. The adaptive learning rate undesirably varies in the early stage of model training due to the usage of a limited number of training samples. To minimize this variation, it is necessary to use lower learning rates during the first few epochs ('heuristic warm-up') for each training cycle.

3 Image Data Augmentation Based on Image Manipulations

A huge amount of labeled data is needed to train a proper convolutional neural network to obtain the best performance. The annotating process is a time-consuming task and costly process. The conventional method implies various transformation techniques, is able to introduce various variations in the images and keeps all the recognizable features.

Table 1 Augmentation techniques applied on dataset

Training dataset	
Parameters	Arguments
Shear range	0.2°
Rescale	1/255
Rotation	25°
Horizontal flip	True
Center-cropped	True
Height shift factor	0.1
Width shift factor	0.1

This section describes the geometric augmentation techniques.

(a) **Flipping**: Flipping an image means reversing the images across the horizontal or vertical axis and obtaining a mirror image. This approach is simple to implement. The vertical flip is also equivalent to rotating the images by 1800.

(b) **Rescale**: The image can be either inward or outward. It resizes the images according to a factor called the scaling factor (SF), i.e., reconstruction of the image according to SF.

(c) **Rotation**: This method augmentation is done by rotating the image either left or right on an axis between 1° and 25°.

(d) **Crop**: Apart from scaling these operations, random cropping is done by randomly selecting a sample from the original image.

(e) **Color augmentation**: In this augmentation, it deals with altering the color properties of an image by changing its pixel values. This augmentation consists of brightness, contrast, saturation and hue operations. In this case, the brightness operation is performed, and the resultant image becomes either darker or brighter compared to the original input.

(f) **Shear**: It is a bounding box transformation with the help of a transformation matrix. This operation can be done either horizontally or vertically with a shearing factor (Table 1).

4 Dataset

In this work, we took RGB images of ASL Finger Spelling benchmark dataset [13]. It was obtained from five different users and consists of RGB and depth images having similar lighting and background. There were 95,697 images, and out of this, 70% of images, i.e., 66,987 used for training and 28,710 for testing. This dataset comprises 24 static signs except for the letters *j* and *z* because these two letters require temporal and spatial relations. The ASL Finger Spelling benchmark dataset is the dataset used in this proposed work. It contains both color and depth images collected from five different users and contains twenty-four static signs, except the letters *j* and *z*, as both temporal and spatial relationships are needed for these two letters. Figure 1

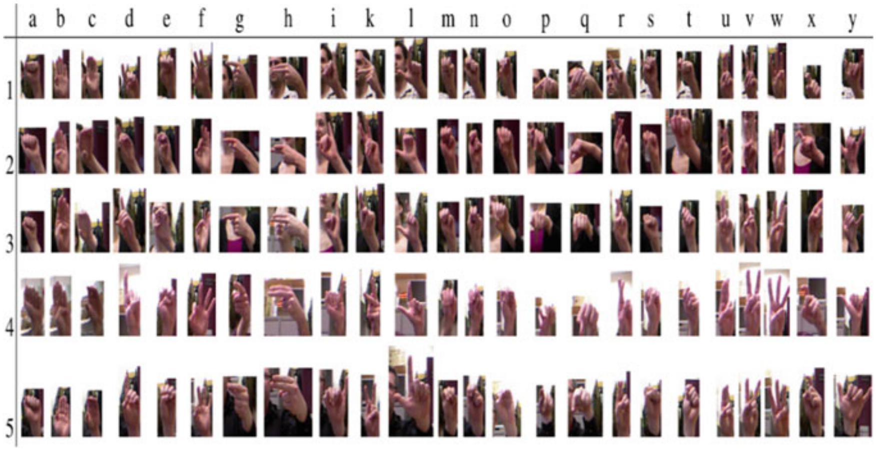


Fig. 1 ASL Finger Spelling dataset ımages

provides a subset of the dataset. There are 95,697 images in total, and approximately 4000 images are contained in each alphabet for each user. With the help of Kinect, five people with non-identical lighting conditions and background conditions record the ASL dataset. About ~500 non-identical hand gesture images are present in the ASL dataset. This dataset is also challenging since it has different backgrounds and different illumination.

5 ASLNET Model Architecture

This research focuses on the issues of ASL alphabet recognition using computer vision method. A deep neural network based on CNN is used to generate a model called ASLNET that can identify signs. The system is based on a supervised model, and the entire process split into CNN training and testing. The schematic presentation of the ASLNET model is displayed in Fig. 2. The input image is 64 × 64 pixels in size, and the first kernel is (5,5) in size, with 32 distinct filters. The kernel size (3,3) was used in the second layer, and there were 64 output filters in that layer, and the pool size (2,2) was used for all the pooling layers,and in the first drop out layer, we used 25% dropout in the convolution and then 35% dropout in the dense layers. We set the initial random weight of Keras using Glorot normal initializer, also called Xavier normal initializer to the filter.

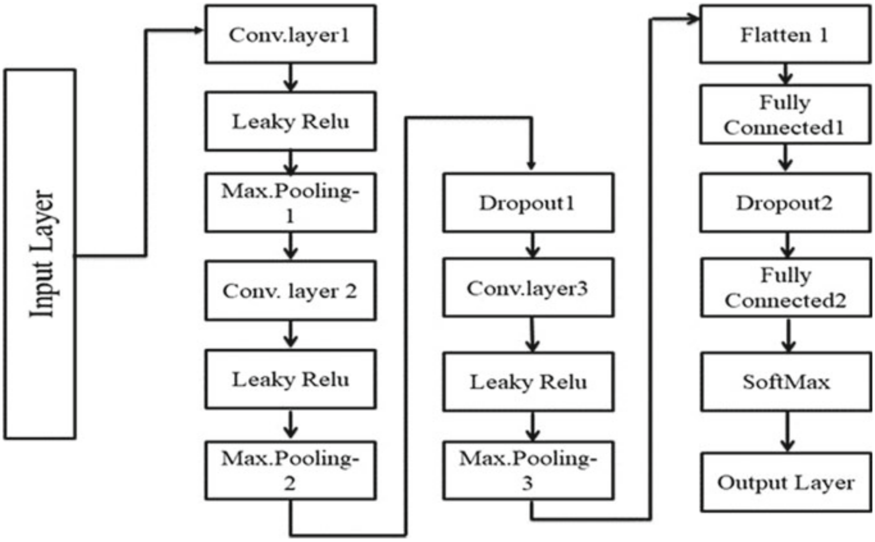


Fig. 2 Proposed ASLNET model

6 Results and Discussion

The model was implemented using Python 3.3., Keras [14] and Tensor flow [15], an open-source software library as the backend. The dataset is split into training and testing part and in the training phase, and 70 percent of the images from each alphabet were alternatively chosen and provided to the CNN input along with the respective labels. We run up to 30 epochs and follow the existing learning rate for different optimization techniques after using data augmentation techniques on a total of 1,82,700 images for training and 78,328 images for testing. For without augmentation techniques, a total of 60,900 for training and 26,100 for testing. The experiments are carried out in IBM Minsky Server having 20 core CPU and 2 NVIDIA Tesla P100 GPUs.

In this study, we analyze the different optimization algorithms for determining the parameters of a deep CNN model named ASLNET with and without augmentation techniques. The nine optimization algorithms do not show a greater difference in their loss. For ADAM, LookAhead and RAdam, the accuracy performance is better when compared to other optimization algorithms (Fig. 3; Table 2).

7 Conclusion

In this paper, the most widely used optimization algorithms in deep learning were investigated on ASLNET, a CNN model used to evaluate other American sign

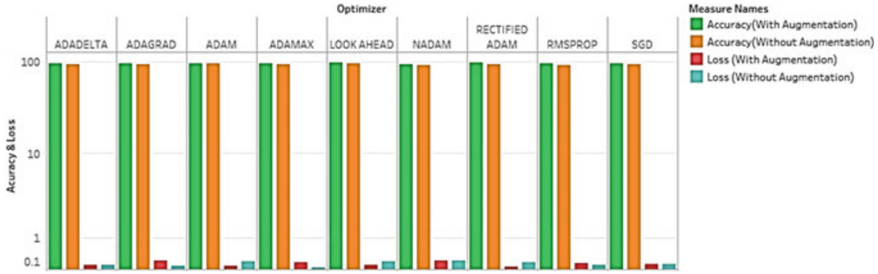


Fig. 3 Accuracy and loss of different optimizers

Table 2 Augmentation techniques applied on dataset

Optimizer	Loss (with augmentation)	Loss (without augmentation)	Accuracy (with augmentation)	Accuracy (without augmentation)
SGD	0.2329466886554	**0.215821157889**	95.7143583774	**93.6982224103**
ADAGRAD	0.3186287783284	**0.187144388551**	96.1034476757	**94.7599800365**
ADADELTA	0.2048942783651	**0.203609842563**	96.6355443000	**93.8574173589**
RMSPROP	0.2425326260323	**0.198891775963**	95.2540695667	**92.8700996733**
ADAM	0.1748025993615	**0.282200689524**	96.1464107036	**95.9958700259**
ADAMAX	0.2668547177342	**0.129936859850**	96.0817148685	**94.2274568002**
NADAM	0.3070191636399	**0.317115879231**	93.1768029935	**92.9580027586**
LOOK AHEAD	0.1984439280768	**0.298936510562**	97.8668651275	**97.1745892200**
RECTIFIED ADAM	0.1545992606690	**0.259885248632**	98.7330396175	**94.0018675924**

To distinguish between the table with loss and the accuracy obtained without the use of augmentation techniques

language letters. The comparison of the model with and without data augmentation using the above-mentioned different optimizers is presented in this article. This work is done on benchmarked Finger Spelling dataset, and from the above result, LookAhead, Rectified Adam and Adam are much better than other optimization algorithms. Because of adaptive approaches dominance and statistical efficiency, they appear to provide remarkable outcomes on the tasks. Still, research to discover better adaptive approaches continues in the field of deep learning.

References

1. Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., et al.: Recent advances in deep learning for speech research at microsoft. In: ICASSP 2013

(2013)

2. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
3. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Process. Mag. IEEE **29**(6), 82–97 (2012)
4. Graves, A., Mohamed, A.-R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649. IEEE (2013)
5. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
6. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Senior, A., Tucker, P., Yang, K., Le, Q.V., et al.: Large scale distributed deep networks. In: Advances in Neural Information Processing Systems, pp. 1223–1231 (2012)
7. Zeiler, M.D.: Adadelta: an adaptive learning rate method (2012). arXiv preprint arXiv:1212.5701
8. Tieleman, T., Hinton, G.: Lecture 6.5—RMSProp, COURSERA: neural networks for machine learning. Technical report (2012)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
10. Dozat, T.: Incorporating nesterov momentum into Adam (2016)
11. Zhang, M.R., Lucas, J., Hinton, G., Ba, J.: Lookahead optimizer: k steps forward, 1 step back (2019). arXiv preprint arXiv:1907.08610
12. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond (2019). arXiv preprint arXiv:1908.03265
13. Pugeault, N., Bowden, R.: Spelling it out: real-time ASL fingerspelling recognition. In: Proceedings of the 1st IEEE Workshop on Consumer Depth Cameras for Computer Vision, jointly with ICCV'2011 (2011)
14. Chollet, F.: Keras, 2015. Available: https://keras.io/
15. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous distributed systems (2016)

Cardinal Correlated Oversampling for Detection of Malicious Web Links Using Machine Learning

M. Shyamala Devi, Uttam Gupta, Khomchand Sahu, Ranjan Jyoti Das, and Santhosh Veeraraghavan Ramesh

Abstract The problem with malicious websites is growing day by day as it leads to the black listing of websites. The unauthorized websites are gathering the user's database information and their assets. Few of the URLs are completely used as a host webpage to publish unrelated web content that signifies cyber-attacks. Cracking the presence of malicious website still pertains as open task due to the lack of web characteristics for malicious and benign websites. To overcome this problem, we are using machine learning techniques for detecting the malicious content and web links. Backgrounding the above, this paper used malicious webpage dataset extracted from UCI dataset repository for predicting the level of mushroom edibility. The categorization of malicious webpage classes is achieved in five ways. Firstly, the dataset consisting of 21 features with 1781 records and is preprocessed with encoding, feature scaling and missing values. Secondly, raw dataset is fitted to all the classifiers with and without the presence of feature scaling and the performance is analyzed. Thirdly, the cardinality free malicious dataset is fitted to all the classifiers with and without the presence of feature scaling and the performance is analyzed. Fourth, the correlated free malicious dataset is fitted to all the classifiers with and without the presence of feature scaling and the performance is analyzed. Fifth, the oversampled malicious dataset is fitted to all the classifiers with and without the presence of feature

M. S. Devi (✉)
Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
e-mail: shyamaladevim@veltech.edu.in

U. Gupta · K. Sahu · R. J. Das
Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
e-mail: vtu10607@veltechuniv.edu.in

K. Sahu
e-mail: vtu10767@veltechuniv.edu.in

R. J. Das
e-mail: vtu10703@veltechuniv.edu.in

S. V. Ramesh
Electronics and Communication Engineering, Sri Sairam Engineering College, Chennai, Tamilnadu, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_36

scaling and the performance is analyzed with precision, recall, accuracy, running time and F-score. Implementation analysis portrays that the decision tree classifier for raw, Cardinality reduced dataset tends to retain the accuracy with 97.1% before and after feature scaling. The random forest classifier with correlated free dataset tends to retain the 96.6% accuracy before and after feature scaling. Decision tree classifier for oversampled dataset tends to retain the accuracy with 98.4% before and after feature scaling. From the above analysis decision tree classifier is found to be more efficient in its accuracy with all raw, cardinality free and oversampled dataset.

Keywords Machine learning · Cardinality · Correlation · Accuracy · Classification and oversampling

1 Introduction

The performance of detection phishing websites can be received with different other machine learning strategies. Machine learning techniques outcomes effectively on the bigger database resources for the detection of malicious webpages. The actors of the system can be secured with the existence of malicious webpages dynamically while they operates on the system. A technique that might be used towards developing associate in nursing anti-phishing URL software by warning the users with respect to possible security problems.

2 Literature Review

In this paper, the creators have assessed numerous administered batch learning classifiers to experimentally compare assortment of classifiers and affirm the one that yields the only execution inside the drawback of identifying phishing [1]. In this, the authors have performed multi-layered authentication to detect phishing websites based on feature vector and prevention with the objective of predicting whether a page is phishing or legitimate accurately. The methodologies used in this approach are text-based detection, visual similarity-based detection and feature-based detection [2]. In this paper, the authors have performed a survey on different malicious webpage detection techniques to find the maximum accuracy in malicious webpage detection. This requires training of a large dataset for more accuracy [3].

In this paper, the creators have proposed a classification to demonstrate that not as it take care of approximately the grammatical nature of the URL, but it takes care of the synonym meaning of the powerfully changing URLs. After implementing supervised machine learning concepts such as random forest model, it was found that convolutional neural network (CNN) got the best results to fabricate the classification model with specifically described feature sets [4]. In this paper, the authors have proposed a malicious URL detecting method based on proposed URL behaviors and attributes [5]. In this paper, a paradigm for detecting and countering the manually induced concept drifts. It was found that after data collection and feature extraction and implementing algorithms like gradient boosting and neural network algorithms, it achieved a superior accuracy [6].

In this paper, the authors have proposed a method to protect a user from browsing malicious content. The methodologies used are the data preprocessing pipeline, word embedding vectors, domain-specific engineered features, re-sampling techniques and classifiers employed [7]. In this paper, the creators have actualized shrewdly pernicious URL discovery with feature examinations. Firstly, the dataset is utilized to memorize the XGBoost classifier, which features a detection exactness of 99% [8].

In this paper, a framework that analyzes URLs in arrange activity that is too competent of altering its location models to adjust to modern noxious substance. Stream model training (MS) one side class perceptron algorithm is implemented [9]. In this project, the creators display their discoveries on the strategies of identifying phishing websites. Information mining calculations together with classifier calculations are utilized in arrange to realize a palatable result. As for the included choice calculation, gain ratio property and help quality are chosen [10]. In this paper, the creators have proposed an outfit approach which employs developmental thinking to realize improvement of classification precision within the location of malevolent Web pages [11]. In this paper, a strategy to classify an URL as either malevolent or generous by taking care of course awkwardness is proposed [12]. In this paper, a multi-layer demonstrate for identifying noxious URL is proposed. The channel can straightforwardly decide the URL by preparing the limit of each layer channel when it comes to the edge [13]. In this paper, they have proposed a demonstration that is competent of adjusting to the energetic behavior of the phishing websites and in this way learn the highlights related with phishing site discovery [14]. In this paper, the creators have attempted to plan a convolutional gated-recurrent-unit (GRU) neural organize for the location of pernicious URLs location based on characters as content classification [15].

3 Overall Proposed Architecture

3.1 Dataset Preparation

The malicious webpage dataset is extracted from UCI database machine store. The dataset description is shown below in Table 1.

3.2 Proposed System

The overall workflow is appeared in Fig. 1.

Table 1 Dataset features

Features	Range of values
1. URL	Sample (M0_109, B0_2314, B0_911)
2. URL Length	Numerical values = 16–250
3. Number of Special characters	Numeric values = 5–44
4. Character set	Char value format of ASCII types
5. Server details	Server name = sample (Apache/2.4.10) Heptu web
6. Content Length	Numerical values = 0–649, 264
7. Country details	Any country
8. State details	Any state
9. Registration date	Day, month, year along with hours, minutes
10. Date Updated details	Day, month, year along with hours, minutes
11. Tcp Conversation	Numerical value = 0–1195
12. Distance Remote TCP	Numerical value = 0–708
13. Remote IPS details	Numerical value = 0–1717
14. Application Bytes	Numerical value = 0–2,362,907
15. Source application Packets	Numerical value = 0–17
16. Remote Application Packets	Numerical value = 0–1285
17. Source Application Bytes	Numerical value = 0–2,060,013
18. Remote Application Bytes	Numerical value = 0–2,362,907
19. Application Packets	Numerical value = 0–1200
20. DNS Query details	Numerical value = 0–20
21. Target Type	Numerical value = 0 or 1 (0 = benign, 1 = malicious)

The paper contributions is given below.

(i) Firstly, the dataset consisting of 21 features with 1781 records and is preprocessed with encoding, feature scaling and missing values.

(ii) Secondly, raw dataset is fitted to all the classifiers with and without the presence of feature scaling and the performance is analyzed.

(iii) Thirdly, relationship of all features is done to find the high cardinality features and are removed from the dataset. The cardinality free malicious dataset is

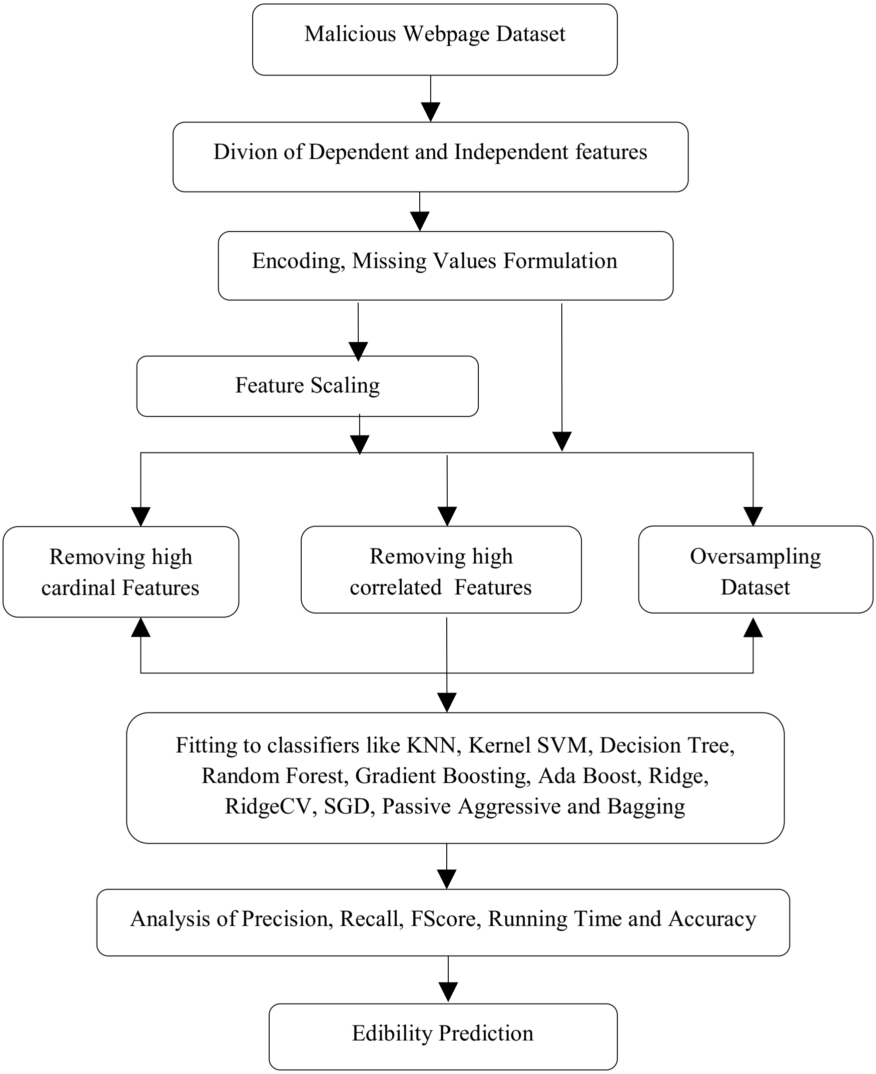

Malicious Webpage Dataset
Divion of Dependent and Independent features
Encoding, Missing Values Formulation
Feature Scaling
Removing high cardinal Features
Removing high correlated Features
Oversampling Dataset
Fitting to classifiers like KNN, Kernel SVM, Decision Tree, Random Forest, Gradient Boosting, Ada Boost, Ridge, RidgeCV, SGD, Passive Aggressive and Bagging
Analysis of Precision, Recall, FScore, Running Time and Accuracy
Edibility Prediction

Fig. 1 Overall architecture flow

fitted to all the classifiers with and without the presence of feature scaling and the performance is analyzed.

(iv) Fourth, relationship of all features is done to find the high correlated features and are removed from the dataset. The correlated free malicious dataset is fitted to all the classifiers with and without the presence of feature scaling and the performance is analyzed.

(v) Fifth, target class is observed to have non-sampled data with 87.9% benign websites and 12.1% were malicious. So the dataset is oversampled with

RandomOversampler to equalize the target class. The oversampled malicious dataset is fitted to all the classifiers with and without the presence of feature scaling and the performance is analyzed with precision, recall, accuracy, running time and *F*-score.

4 Results and Discussion

4.1 Implementation Setup

The mushroom dataset extracted from the UCI machine learning data store is used for implementation. The dataset contains 20 independent features and 1 'Type' target feature. The code is drafted with python under Anaconda Navigator with Spyder IDE. The dataset is split with 80:20 for training and testing dataset.

4.2 Dataset Exploratory Analysis

The distribution and correlation of the features appears in Fig. 2.

The high cardinality and correlation of the features appears in Fig. 3.

The categorical and the numerical data type of the features appears in Fig. 4.

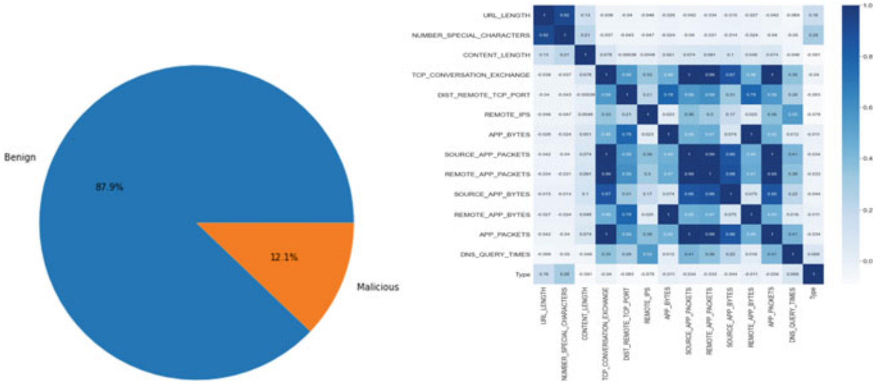
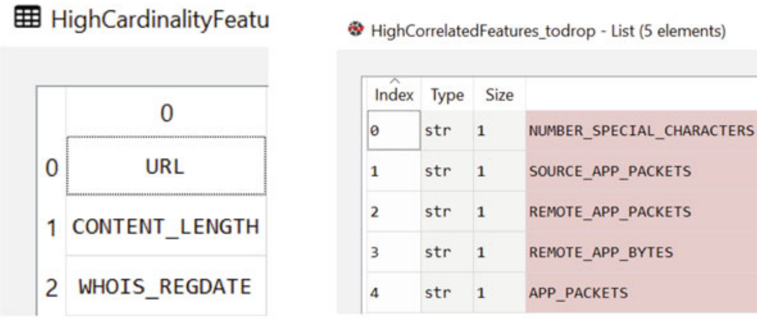


Fig. 2 Target class and correlation analysis



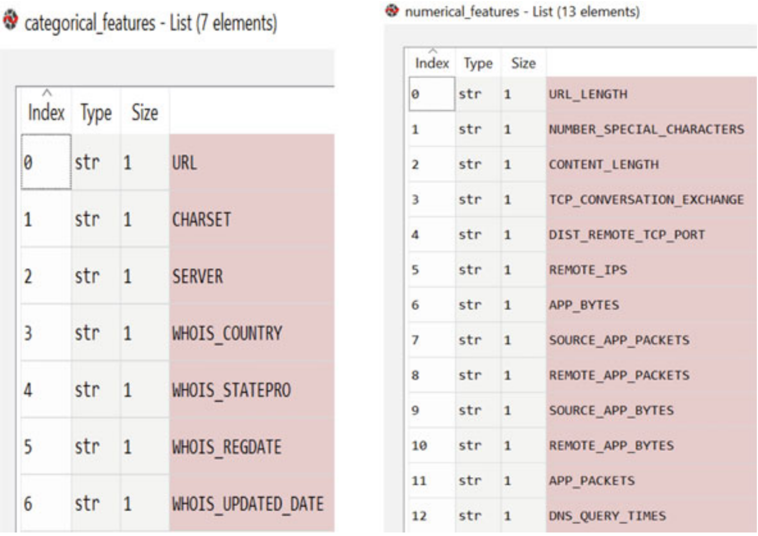
HighCardinalityFeatu

	0
0	URL
1	CONTENT_LENGTH
2	WHOIS_REGDATE

HighCorrelatedFeatures_todrop - List (5 elements)

Index	Type	Size	
0	str	1	NUMBER_SPECIAL_CHARACTERS
1	str	1	SOURCE_APP_PACKETS
2	str	1	REMOTE_APP_PACKETS
3	str	1	REMOTE_APP_BYTES
4	str	1	APP_PACKETS

Fig. 3 High cardinality and high correlated Features



categorical_features - List (7 elements)

Index	Type	Size	
0	str	1	URL
1	str	1	CHARSET
2	str	1	SERVER
3	str	1	WHOIS_COUNTRY
4	str	1	WHOIS_STATEPRO
5	str	1	WHOIS_REGDATE
6	str	1	WHOIS_UPDATED_DATE

numerical_features - List (13 elements)

Index	Type	Size	
0	str	1	URL_LENGTH
1	str	1	NUMBER_SPECIAL_CHARACTERS
2	str	1	CONTENT_LENGTH
3	str	1	TCP_CONVERSATION_EXCHANGE
4	str	1	DIST_REMOTE_TCP_PORT
5	str	1	REMOTE_IPS
6	str	1	APP_BYTES
7	str	1	SOURCE_APP_PACKETS
8	str	1	REMOTE_APP_PACKETS
9	str	1	SOURCE_APP_BYTES
10	str	1	REMOTE_APP_BYTES
11	str	1	APP_PACKETS
12	str	1	DNS_QUERY_TIMES

Fig. 4 Categorical and numerical features in the dataset

4.3 Classifier Analysis Before and After Feature Scaling

The raw dataset is fitted to all the classifier like logistic regression, KNN, Kernel SVM, decision tree, random forest, gradient boosting, AdaBoost, Ridge, RidgeCV, SGD, Passive Aggressive and Bagging classifier with and without the presence of feature scaling. The performance analysis is done and the analysis is depicted in Tables 2 and 3 with the readings appears in Figs. 5 and 6.

Table 2 Performance metrics of raw dataset before feature scaling

Classifier	Precision	Recall	*F*Score	Accuracy	Running time (ms)
Logistic Regression	0.843	0.885	0.836	0.885	0.108
KNeighbors	0.915	0.922	0.917	0.922	0.203
Kernel SVM	0.783	0.885	0.831	0.885	0.377
Gaussian Naive Bayes	0.856	0.266	0.294	0.266	0.031
Decision Tree	0.972	0.972	0.972	0.972	0.016
Extra Tree	0.942	0.941	0.941	0.941	0.016
Random Forest	0.968	0.966	0.964	0.966	0.055
Gradient Boosting	0.966	0.966	0.964	0.966	0.857
AdaBoost Classifier	0.956	0.958	0.956	0.958	0.484
Ridge Classifier	0.947	0.947	0.941	0.947	0.044
Ridge ClassifierCV	0.947	0.947	0.941	0.947	0.135
SGD Classifier	0.737	0.571	0.644	0.571	0.035
Passive Aggressive	0.812	0.818	0.815	0.818	0.057
Bagging Classifier	0.956	0.958	0.956	0.958	0.377

Table 3 Performance metrics of raw dataset after feature scaling

Classifier	Precision	Recall	*F*Score	Accuracy	Running time (ms)
Logistic Regression	0.949	0.950	0.945	0.950	0.062
KNeighbors	0.945	0.947	0.945	0.947	0.409
Kernel SVM	0.954	0.955	0.952	0.955	0.344
Gaussian Naive Bayes	0.892	0.625	0.692	0.625	0.031
Decision Tree	0.972	0.972	0.972	0.972	0.031
Extra Tree	0.942	0.941	0.941	0.941	0.016
Random Forest	0.968	0.966	0.964	0.966	0.063
Gradient Boosting	0.963	0.964	0.962	0.964	0.851
AdaBoost Classifier	0.956	0.958	0.956	0.958	0.455
Ridge Classifier	0.940	0.941	0.934	0.941	0.031
Ridge ClassifierCV	0.938	0.933	0.920	0.933	0.125
SGD Classifier	0.959	0.961	0.959	0.961	0.047
Passive Aggressive	0.944	0.947	0.945	0.947	0.062
Bagging Classifier	0.956	0.958	0.956	0.958	0.297

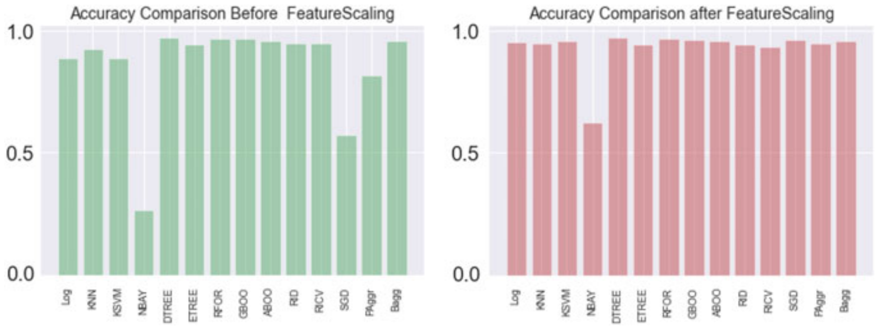


Fig. 5 Accuracy comparison before and after feature scaling

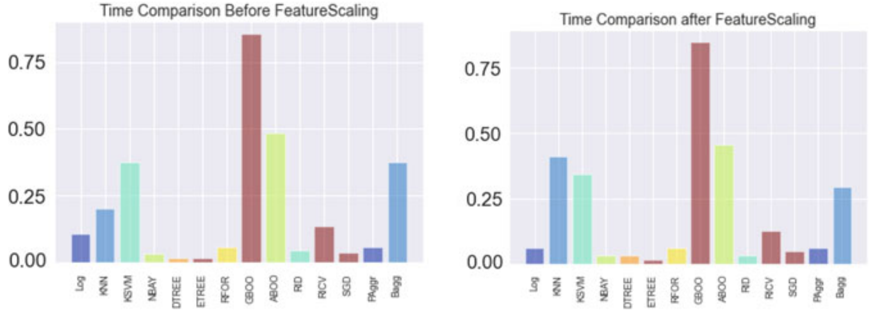


Fig. 6 Time comparison before and after feature scaling

4.4 Cardinality Free Analysis Before and After Feature Scaling

The high cardinality features like “Url, Content_Length, Whois_regdate” are removed from the dataset and cardinality free dataset is applied to all the classifiers with and without the presence of feature scaling. The performance analysis for the cardinality free dataset is done, and the analysis is depicted in Tables 4 and 5 with the readings appear in Figs. 7 and 8.

4.5 Correlated Free Analysis Before and After Feature Scaling

The high correlated features like “Number_Special_Characters, App_Packets, Source_App_Packets, Remote_App_Packets, Remote_App_Bytes” are removed from the dataset and correlated free dataset is applied to all the classifiers with and

Table 4 Performance metrics of cardinality free dataset before feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running time (ms)
Logistic Regression	0.843	0.885	0.836	0.885	0.109
KNeighbors	0.915	0.922	0.917	0.922	0.241
Kernel SVM	0.783	0.885	0.831	0.885	0.378
Gaussian Naive Bayes	0.856	0.266	0.294	0.266	0.028
Decision Tree	0.972	0.972	0.972	0.972	0.033
Extra Tree	0.942	0.941	0.941	0.941	0.016
Random Forest	0.968	0.966	0.964	0.966	0.062
Gradient Boosting	0.969	0.969	0.968	0.969	0.899
AdaBoost Classifier	0.956	0.958	0.956	0.958	0.429
Ridge Classifier	0.947	0.947	0.941	0.947	0.049
Ridge ClassifierCV	0.947	0.947	0.941	0.947	0.156
SGD Classifier	0.814	0.731	0.766	0.731	0.047
Passive Aggressive	0.802	0.860	0.827	0.860	0.047
Bagging Classifier	0.950	0.952	0.950	0.952	0.322

Table 5 Performance metrics of cardinality free dataset after feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running time (ms)
Logistic Regression	0.949	0.950	0.945	0.950	0.047
KNeighbors	0.945	0.947	0.945	0.947	0.391
Kernel SVM	0.954	0.955	0.952	0.955	0.344
Gaussian Naive Bayes	0.892	0.625	0.692	0.625	0.024
Decision Tree	0.972	0.972	0.972	0.972	0.031
Extra Tree	0.942	0.941	0.941	0.941	0.016
Random Forest	0.968	0.966	0.964	0.966	0.063
Gradient Boosting	0.969	0.969	0.968	0.969	0.846
AdaBoost Classifier	0.953	0.955	0.954	0.955	0.453
Ridge Classifier	0.940	0.941	0.934	0.941	0.031
Ridge ClassifierCV	0.938	0.933	0.920	0.933	0.141
SGD Classifier	0.952	0.952	0.952	0.952	0.031
Passive Aggressive	0.944	0.947	0.945	0.947	0.047
Bagging Classifier	0.951	0.952	0.951	0.952	0.328

without the presence of feature scaling. The performance analysis for the cardinality free dataset is depicted in Tables 6 and 7 with readings appear in Figs. 9 and 10.

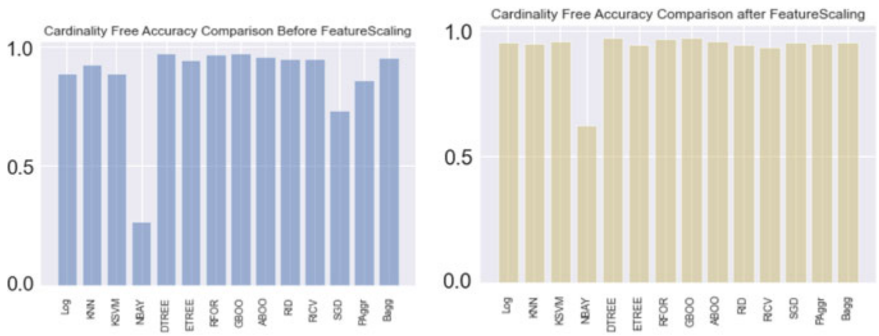


Fig. 7 Cardinality free accuracy comparison before and after feature scaling

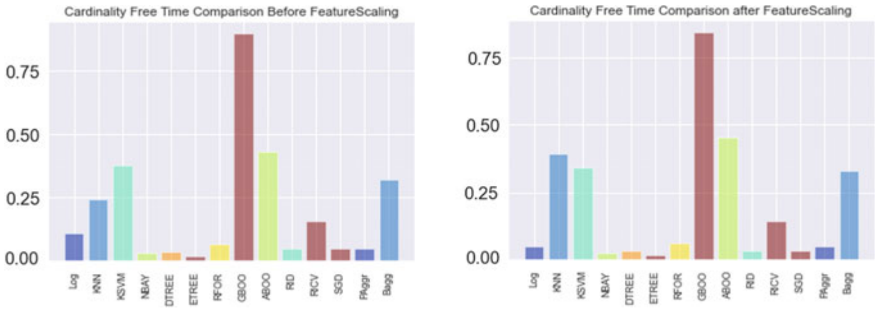


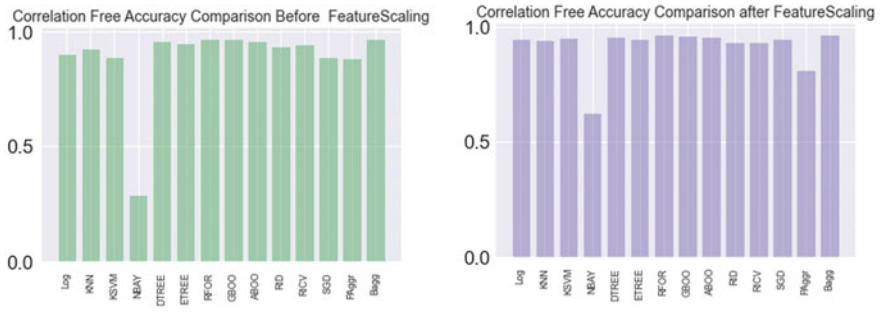
Fig. 8 Cardinality free time comparison before and after feature scaling

Table 6 Performance metrics of correlated free dataset before feature scaling

Classifier	Precision	Recall	*F*Score	Accuracy	Running time (ms)
Logistic Regression	0.882	0.899	0.875	0.899	0.110
KNeighbors	0.915	0.922	0.917	0.922	0.248
Kernel SVM	0.783	0.885	0.831	0.885	0.412
Gaussian Naive Bayes	0.861	0.289	0.326	0.289	0.032
Decision Tree	0.955	0.955	0.955	0.955	0.024
Extra Tree	0.949	0.947	0.948	0.947	0.018
Random Forest	0.966	0.966	0.964	0.966	0.077
Gradient Boosting	0.963	0.964	0.962	0.964	0.851
AdaBoost Classifier	0.954	0.955	0.954	0.955	0.451
Ridge Classifier	0.929	0.933	0.924	0.933	0.053
Ridge ClassifierCV	0.939	0.941	0.935	0.941	0.131
SGD Classifier	0.783	0.885	0.831	0.885	0.043
Passive Aggressive	0.783	0.882	0.830	0.882	0.096
Bagging Classifier	0.963	0.964	0.962	0.964	0.279

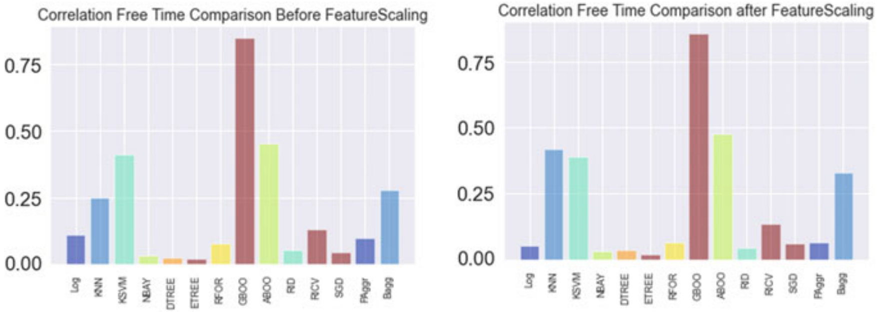
Table 7 Performance metrics of correlated free dataset after feature scaling

Classifier	Precision	Recall	*F*Score	Accuracy	Running time (ms)
Logistic Regression	0.944	0.947	0.944	0.947	0.053
KNeighbors	0.941	0.944	0.942	0.944	0.420
Kernel SVM	0.947	0.950	0.947	0.950	0.389
Gaussian Naive Bayes	0.892	0.625	0.692	0.625	0.034
Decision Tree	0.955	0.955	0.955	0.955	0.037
Extra Tree	0.949	0.947	0.948	0.947	0.020
Random Forest	0.966	0.966	0.964	0.966	0.064
Gradient Boosting	0.960	0.961	0.958	0.961	0.858
AdaBoost Classifier	0.956	0.958	0.957	0.958	0.476
Ridge Classifier	0.929	0.933	0.924	0.933	0.044
Ridge ClassifierCV	0.929	0.933	0.924	0.933	0.135
SGD Classifier	0.947	0.947	0.947	0.947	0.060
Passive Aggressive	0.903	0.810	0.838	0.810	0.064
Bagging Classifier	0.966	0.966	0.964	0.966	0.332


Correlation Free Accuracy Comparison Before FeatureScaling

Correlation Free Accuracy Comparison after FeatureScaling

Fig. 9 Correlated free accuracy comparison before and after feature scaling


Correlation Free Time Comparison Before FeatureScaling

Correlation Free Time Comparison after FeatureScaling

Fig. 10 Correlated free time comparison before and after feature scaling

Table 8 Performance metrics of oversampled dataset before feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running time (ms)
Logistic Regression	0.780	0.724	0.710	0.724	0.207
KNeighbors	0.917	0.912	0.912	0.912	0.434
Kernel SVM	0.568	0.529	0.461	0.529	4.407
Gaussian Naive Bayes	0.763	0.577	0.490	0.577	0.048
Decision Tree	0.985	0.984	0.984	0.984	0.047
Extra Tree	0.982	0.981	0.981	0.981	0.021
Random Forest	0.994	0.994	0.994	0.994	0.113
Gradient Boosting	0.974	0.973	0.973	0.973	1.582
AdaBoost Classifier	0.971	0.971	0.971	0.971	0.799
Ridge Classifier	0.937	0.935	0.934	0.935	0.073
Ridge ClassifierCV	0.937	0.935	0.934	0.935	0.242
SGD Classifier	0.244	0.494	0.326	0.494	0.075
Passive Aggressive	0.652	0.652	0.651	0.652	0.177
Bagging Classifier	0.974	0.973	0.973	0.973	0.433

4.6 Oversampling Analysis Before and After Feature Scaling

The target class is observed to have non-sampled data with 87.9% benign websites, and 12.1% were malicious. So the dataset is oversampled with RandomOversampler to equalize the target class. The oversampled dataset is applied to all the classifiers with and without the presence of feature scaling. The performance analysis for the oversampled dataset is depicted in Tables 8 and 9, and the readings appear in Figs. 11 and 12.

5 Conclusion

This paper attempts to make the analysis of performance with respect to raw data, cardinality free dataset, correlated free dataset and oversampled dataset. Experimental results show that the decision tree classifier for raw and cardinality reduced datasets tends to retain the accuracy with 97.1% before and after feature scaling. The random forest classifier with correlated free dataset tends to retain the 96.6% accuracy before and after feature scaling. Decision tree classifier for oversampled dataset tends to retain the accuracy with 98.4% before and after feature scaling. From the above analysis, decision tree classifier is found to be more efficient in its accuracy with all raw, cardinality free and oversampled datasets. The future work is to adapt the various dimensionality reduction methods to analyze the performance efficiency of the classifier algorithms.

Table 9 Performance metrics of oversampled dataset after feature scaling

Classifier	Precision	Recall	*F*Score	Accuracy	Running time (ms)
Logistic Regression	0.956	0.955	0.955	0.955	0.153
KNeighbors	0.914	0.909	0.909	0.909	0.778
Kernel SVM	0.967	0.966	0.966	0.966	1.407
Gaussian Naive Bayes	0.842	0.768	0.756	0.768	0.049
Decision Tree	0.985	0.984	0.984	0.984	0.048
Extra Tree	0.982	0.981	0.981	0.981	0.026
Random Forest	0.994	0.994	0.994	0.994	0.095
Gradient Boosting	0.974	0.973	0.973	0.973	1.592
AdaBoost Classifier	0.971	0.971	0.971	0.971	0.769
Ridge Classifier	0.930	0.928	0.928	0.928	0.056
Ridge ClassifierCV	0.919	0.919	0.919	0.919	0.213
SGD Classifier	0.932	0.931	0.931	0.931	0.089
Passive Aggressive	0.951	0.950	0.950	0.950	0.089
Bagging Classifier	0.982	0.981	0.981	0.981	0.408

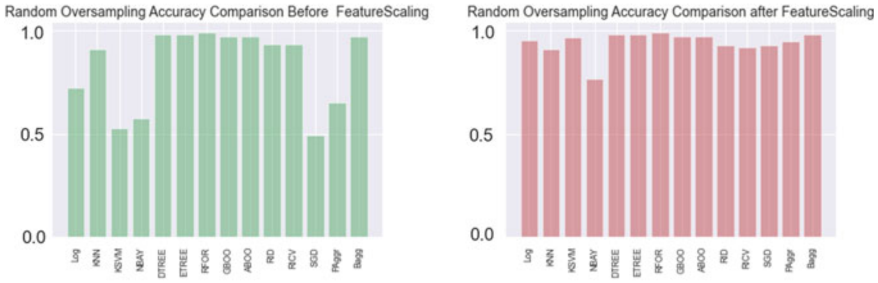


Fig. 11 Oversampled accuracy comparison before and after feature scaling

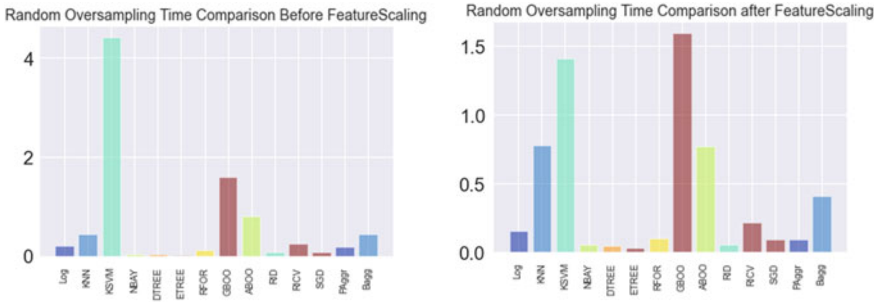


Fig. 12 Oversampled time comparison before and after feature scaling

References

1. Basnet, R.B., Doleck, T.: Towards developing a tool to detect phishing URLs: a machine learning approach. In: Proceedings of IEEE International Conference on Computational Intelligence & Communication Technology, pp. 220–223 (2015)
2. Selvan, K., Muthuraman, V.: Detection of phishing web pages based on features vector and prevention using multi layered authentication. Int. J. Pure Appl. Math. **119**, 564–573 (2018)
3. Chamidah, N., Wasito, I.: Fetal state classification from cardiotocography based on feature extraction using hybrid K-means and support vector machine. In: Proceedings of International Conference on Advanced Computer Science and Information Systems, 25 Feb 2016
4. Patil, D, Patil, J.: Survey on malicious web pages detection techniques. Int. J. U- and E-service Sci. Technol. **8**, 195–206. https://doi.org/10.14257/ijunesst.2015.8.5.18.(2015)
5. Jagannathan, D.: Cardiotocography—a comparative study between support vector machine and decision tree algorithms. Int. J. Trend Res. Dev. **4**(1) (2017)
6. Vanhoenshoven, F., Napoles, G., Falcon, R., Vanhoof, K., Koppen, M.: Detecting malicious URLs using machine learning techniques. In: Proceedings of Computational Intelligence (SSCI), Athens, Greece, pp. 1–8 (2016). https://doi.org/10.1109/SSCI.2016.7850079
7. Silva, R.M., Almeida, T.A., Yamakami, A.: Towards web spam filtering using a classifier based on the minimum description length principle. In: Proceedings of 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, pp. 470–475 (2016). https://doi.org/10.1109/ICMLA.2016.0083
8. Singhal, S., Chawla, U., Shorey, R.: Machine learning & concept drift based approach for malicious website detection. In: Proceedings of International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, pp. 582–585 (2020). https://doi.org/10.1109/COMSNETS48256.2020.9027485
9. Crisan, A., Florea, G., Halasz, L., Lemnaru, C., Oprisa, C.: Detecting malicious URLs based on machine learning algorithms and word embeddings. In: Proceedings of International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, pp. 187–193 (2020)
10. Chen, Y.C., Ma, Y.W., Chen, J.L.: Intelligent malicious URL detection with feature analysis. In: Proceedings of IEEE Symposium on Computers and Communications (ISCC), Rennes, France, pp.1–5 (2020). https://doi.org/10.1109/ISCC50000.2020.9219637
11. Gabriel, A.D., Gavrilut, D.T., Alexandru. B.I., Stefan, P.A.:Detecting malicious URLs: a semi-supervised machine learning system approach. In: Proceedings of Symbolic and Numeric Algorithms for Scientific Computing, pp. 233–239 (2016)
12. Aydin, M., Butun, I., Bicakci, K., Baykal, N.: Using attribute-based feature selection approaches and machine learning algorithms for detecting fraudulent website URLs. Computing and Communication Workshop, p. 774–779 (2020).
13. Sajedi, H., Allameh, F.: Detection of malicious web pages by evolutionary ensemble learning. Int. J. Hybrid Intell. Syst. 51–59 (2017)
14. Manjeri, A.S., Nair, P.C.: A machine learning approach for detecting malicious websites using URL features. In: Proceedings of International conference on Electronics, Communication and Aerospace Technology, pp. 555–561 (2019)
15. Yang, W., Zuo, W., Cui, B.: Detecting malicious URLs via a keyword-based convolutional gated-recurrent-unit neural network. IEEE Access **7**, 29891–29900 (2019)

Simulation of Speckle Noise Using Image Processing Techniques

Noor H. Rasham, Heba Kh. Abbas, Asmaa A. Abdul Razaq, and Haidar J. Mohamad

Abstract The image noise is considered as one of the significant problems in scientific applications. The simulation of the speckle noise within a standard image is studied using the presented algorithm. Different speckle noise ratios were added, with per cent (0.01–0.06), to simulate noise within different images. This added noise based on the mathematical equations to simulate the behavior of this type of noise. The main work divided into two steps; the first step is the classification method which based on the minimum distance and it used to classify the tested image with different homogenous areas and compare it with the corresponding noise images in the same location. In the first step, the knowledge of understanding the behavior of speckle noise achieved. The second step is the statistical criteria namely mean and standard deviation which is used to calculate the speckle factor (SF) to know the effect of noise within the image. In this step, the effect of the noise is obvious by checking the statistical values of SF. The behavior of the speckle noise is well-described and recognize based on the presented algorithm and method.

Keywords Speckle noise · Additive noise · Multiplicative noise · Noise simulation algorithm · Speckle factor · Mean and standard deviation criteria

1 Introduction

Noise is undesirable data that distorts the image and makes it almost impossible to analyze. Studying noise, therefore, helps to understand its effect on the image or to decide the best ways to eliminate it from the pictures [1]. This aims to strengthen and recreate the original image data with the least potential damage. The additive noise generated as white random points which do not depend on the signal and

N. H. Rasham · H. Kh. Abbas
Department of Physics, College of Science for Women, University of Baghdad, Baghdad, Iraq

A. A. Abdul Razaq · H. J. Mohamad (✉)
Department of Physics, College of Science, Mustansiriyah University, Baghdad, Iraq
e-mail: Haidar.mohamad@uomustansiriyah.edu.iq

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_37

have a density of fixed points [2]. In the form of a linear relationship, its statistical approximation defined by the uniform distribution and Gaussian distribution [3, 4].

A random value of noise is applied to the exact light value for a given pixel in the optical image data process that transforms light into a continuous electronic current [5]. Multiplicative noise is a random noise that depends on the signal. The high noise located in the bright areas of the picture and the lower the noise at the lower the intensity of light. This implies that the relationship between noise and amplitude is a linear relationship [6].

Speckle can be described as an artifact of destructive interference and its magnitude depends on the relative phase between two returned echoes that overlap. Like other imaging techniques that use coherent sources [7, 8], such as laser, radar (SAR) and ultrasound images where acoustic waves are vulnerable to speckle corruption, which can be eliminated without affecting the significant image information [9–11]. Speckle varies in the sense that it is a deterministic artifact from other forms of noise, meaning that two signals or images obtained under precisely the same conditions will undergo exactly the same pattern of speckle corruption, except where any or any of the situations vary, the pattern of speckle corruption will be different. In the high-intensity field, the speckle texture is normally kept [12, 13].

There are many previous studies in the area of reducing speckle noise in ultrasound images. Anjali Kapoor and Taranjit Singh [14], the method of speckle noise reduction was addressed, as well as the different filters that are widely used for speckle removal. Things that affect have different features because the information material inside the image is preserved by a few filters and few filters have a smoothing mechanism. Similarly, others are better at sensing edges and can be applied for the segmentation of tumors, etc. Diwakar and et al. [15] analyzed initial noisy CT images are a threshold in the Shearlet domain using the bayes shrinkage law. The suggested solution is compared to current approaches, and it is observed that the suggested technique outperforms existing approaches in terms of visual clarity, image quality index (IQI), and peak signal-to-noise ratio (PSNR). Experimental assessment reveals that the suggested method (i) effectively eliminates noise within CT type, (ii) preserves edge and structural data, and (iii) retains medically applicable evidence. Hyunho Choi and Jechang Jeong [16] proposed a method focused on speckle reduction of anisotropic diffusion (SRAD) and a Bayes threshold in the wavelet domain. SRAD is used as advanced step in this method, and the Bayes threshold useful in removing remaining noise in the acquired images. As compared to conventional filtering techniques, the suggested methodology demonstrated superior performance in terms of peak signal-to-noise ratio (average = 28.61 dB) and structural similarity (average = 0.778). In depth, 27 techniques defined by Duarte-Salazar et al. [17], which primarily concentrate on smoothing or removing speckle noise in medical ultrasound images. The aim of this research is to stress the importance of enhancing this smoothing and elimination, which is related to many processes addressed in other studies (such as the identification of regions of interest). In addition, the definition of this set of techniques makes it possible for more precise scope analyses and analysis to be carried out and initially researched covers some classical approaches, such as spatial filtering, diffusion filtering, and wavelet filtering. Following that it describes recent techniques in the field of machine learning focused on image recognition, which are not yet commonly recognized, but are incredibly significant, along with many existing and hybrid systems in the field of speckle noise filtering. Eventually, five full-reference (FR) loss measures, which are common in filter evaluation processes, are defined, as well as a methodology for compensating between FR and non-reference (NR) metrics, which can improve the characterization of filters by taking into account the perceptual accuracy information provided by NR metrics of their conduct.

In this study, noise simulation done using tested images (without noise), then different ratios of multiplication noise (speckle noise) were added with per cent (0.01–0.06). The supervised classification method based on the minimum distance used to classify the tested image into different areas and compare it with the corresponding noise images, calculate both the mean and the standard deviation to calculate the number of looks to know the effect of noise on the image. This method describes the type of noise within the image.

2 Theoritical Background

2.1 Noise Simulation Model

One of the key purposes of this simulation is to investigate the behavior of speckle noise that occurs as a result of the deployment of coherent devices for imagery. An additive noise can be written as [18]:

$$w(x, y) = s(x, y) + n_a(x, y) \tag{1}$$

where *s(x,y)* original signal, $n_a(x, y)$ additive noise, the corrupted image *w(x,y)*, and *(x,y)* pixel position.

The multiplicative noise distribution's statistical approximations are similar to the distribution of Poisson or speckle noise. The law is accompanied by a multiplicative noise [19]:

$$w(x, y) = s(x, y) \times n_m(x, y) \tag{2}$$

where $n_m(x, y)$ is multiplicative noise function.

Speckle results, nevertheless, are usually regarded as an undesirable noise that degrades the quality of the picture, but this undesirable noise is often known to be a knowledge carrier for rough surface properties. When coherent electromagnetic wave (EMW) occurs on a rough surface relative to the wavelength of the EMW event, speckle generation system can be described by [20]:

$$u(x, y, t) = A(x, y)\mathrm{e}^{i2\pi vt} \tag{3}$$

where *u(x, y, t)* represents the incident EMW, ν is the frequency of EMW, and *A(x, y)* is the complex amplitude given by:

$$A(x, y) = |A(x, y)|\mathrm{e}^{i\phi(x,y)} \tag{4}$$

where *ϕ(x,y)* represents a random phase, whereas the reflected wave has phases with random distributed over the primary interval $[-\pi, \pi]$, the complex reflected amplitude can bewritten in the following form [21, 22]:

$$A = \sum_{k=1}^{c} |a_k| e^{i\phi_k} \tag{5}$$

$$= \sum_{k=1}^{c} |a_k| \cos\phi_k + i \sum_{k=1}^{c} |a_k| \sin\phi_k \tag{6}$$

where, a_k and ϕ_k amplitude and the phase of the k-th scattering area, respectively, and c the summation of scatters [23].

$$I(x, y) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \left| u(x, y; t) \right|^2 \mathrm{d}t = \left| A(x, y) \right|^2 \tag{7}$$

Hence, Eq. (7) can be used to represent the speckle pattern, in which the intensity is given by:

$$I = |A|^2 \tag{8}$$

By assuming that $|a_k| = 1$, the concentration can be characterized as follows [24]:

$$I = \left[\left(\sum_{k=1}^{c} \cos\phi_k \right)^2 + \left(\sum_{k=1}^{c} \sin\phi_k) \right)^2 \right] \cdot f \tag{9}$$

where f is the conservative factor to reduce the expected losses in intensity, and keeping the energy constant. This factor can be computed by assuming an initial value "$f_0 = 1$", after that create a speckle pattern and calculate the mean "μ", the value of the conservative factor, finally, determined "$f = 1/\mu$". The speckle factor "SF" of each simulated image can be estimated as follows [25];

$$\mathrm{SF} = \mu^2/\sigma^2 \tag{10}$$

where σ^2 and μ represent the variance and the mean of the speckle distribution function, respectively.

The simplest classification method of supervised methods is the minimum distance technique which computes the mean vectors for each class. For each type, the

Euclidian nice ways from any mysterious pixel to the linear combination. Except where an edge is indicated, each pixels are labeled to the nearest class at that point. In other image characterization implementations, this approach is extremely arbitrary. When the number of training samples is small, the benefit of this technique is not only that it is a very simple and computationally efficient approach, but it also offers greater exactness than other classifiers [26].

$$\mathrm{MD} = \left| \sum_{c=1}^{nc} (I_b(x, y) - \mu(b)) \right| \rightarrow \text{class } c \tag{11}$$

where MD represents a minimum distance between the pixel and mean of *a* class, *c* the index of class with value (1 to *nc*), *nc* the classes number, *I(x, y)* intensity values, and μ the image band mean (b) in class (c).

3 Method

3.1 Noise Simulation Algorithm

The noise is simulated with the standard image (House and Sinai Desert) as in the following algorithm.

Noise Simulation Algorithm

Input: standard images (House, Sinai Desert)

Output: noise image, histogram additive and multiplicative noise, plot relationship between mean and standard deviation, number of look (NL)

Start algorithm

Step1: read standard image(img) img = imread(img);

Step2: convert color image to gray image I = rgb2gray(img);

Step3: display image(img, I).

Step4: add speckle noise to standard image per cent (0.01–0.06).

JJ = imnoise(I, ‘speckle’).

Step5: calculate additive and multiplicative noise;

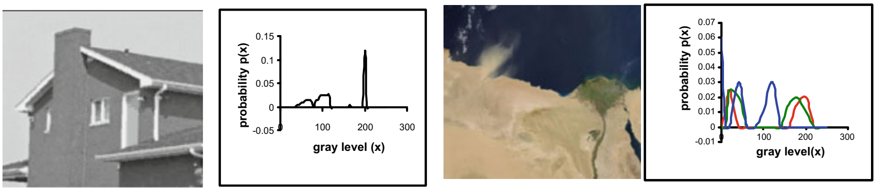
N_additive noise = Ji − I; N_multiplicative = Ji./I * 100; where Ji image noise i (noise value); I gray image.

Step6: plot histogram to noise image, additive and multiplicative noise for noise image imhist(Ji);imhist(N_additive), imhist(N_multiplicative);

Step7: extract four bLook from different position in standard image and dropping the same position in noisy image then classified image using minimum distance method.

Step8: calculate mean and standard deviation for each block.

M(i) = mean2(Ti); ST(i) = std2(Ti); where Ti: number of bLook, i:block number, M: mean of each block, ST: standard deviation of each block.



(a) (b)

Fig. 1 The tested images. **a** House image with histogram, **b** Desert Sinai image with histogram

Step 9: plot relationship between mean and standard deviation to find number of look (SF) = M^2/ST^2.

$Slp(K) = M(i)^2/ST(i)^2$;

End algorithm

There are two types of images were used. The first image is 8-bits (256 × 256) grayscale images (0–255) (House Image) which is well known. It's histogram is shown in Fig. 1a. The second image is colored image with 24 bits (600 × 800) (Sinai desert) and gray intensity for each beam ranging between (0–255) and the histogram is shown in Fig. 1b.

4 Results

4.1 Speckle Noise

The simulated speckle noise with the tested images (House and Desert) is present in Fig. 2a, b. The same values of the noise (with per cent from 0.01 to 0.06) are added to the tested images and plot the histogram to check the noise behavior for the additive and multiplicative noise. The change is noticeable for the image by increasing the noise value as well as the histogram. The algorithm shows the change in the histogram is important because it is a direct indicator of the type of noise. This can be confirmed by another step, i.e., run the minimum distance method (Fig. 3).

4.2 Minimum Distance

The image areas determined and homogeneous areas extracted from each target. Then, the characteristics of these targets determined, represented by the mean and the standard deviation of each target, to classify them using the minimum distance method as shown in Fig. 4. The purpose is to check the noise in the same location for different noise ratio. Then, these data are plotted to show the noise behavior.

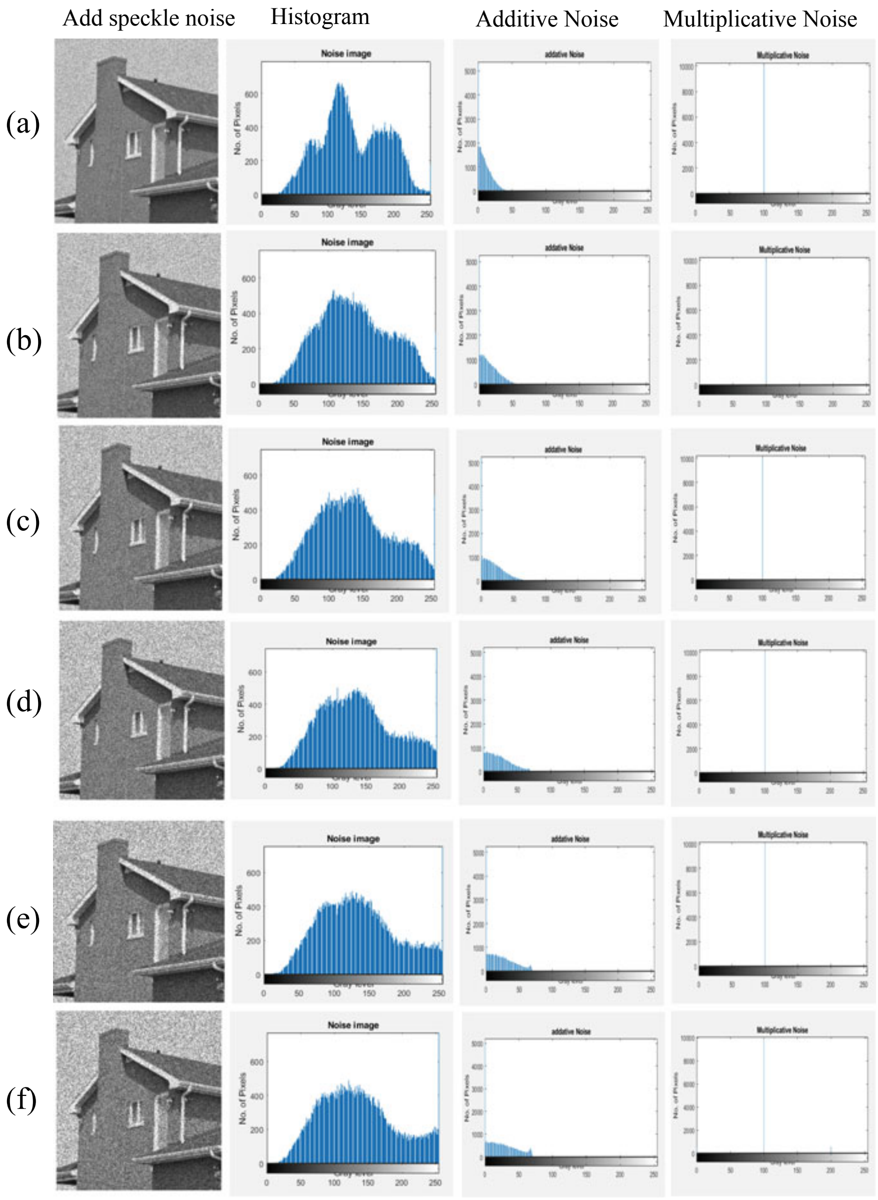


Fig. 2 The simulated images with Speckle noise ratio. **a** 0.01, **b** 0.02, **c** 0.03, **d** 0.04, **e** 0.05, **f** 0.06 with House image, additive and multiplicative histogram

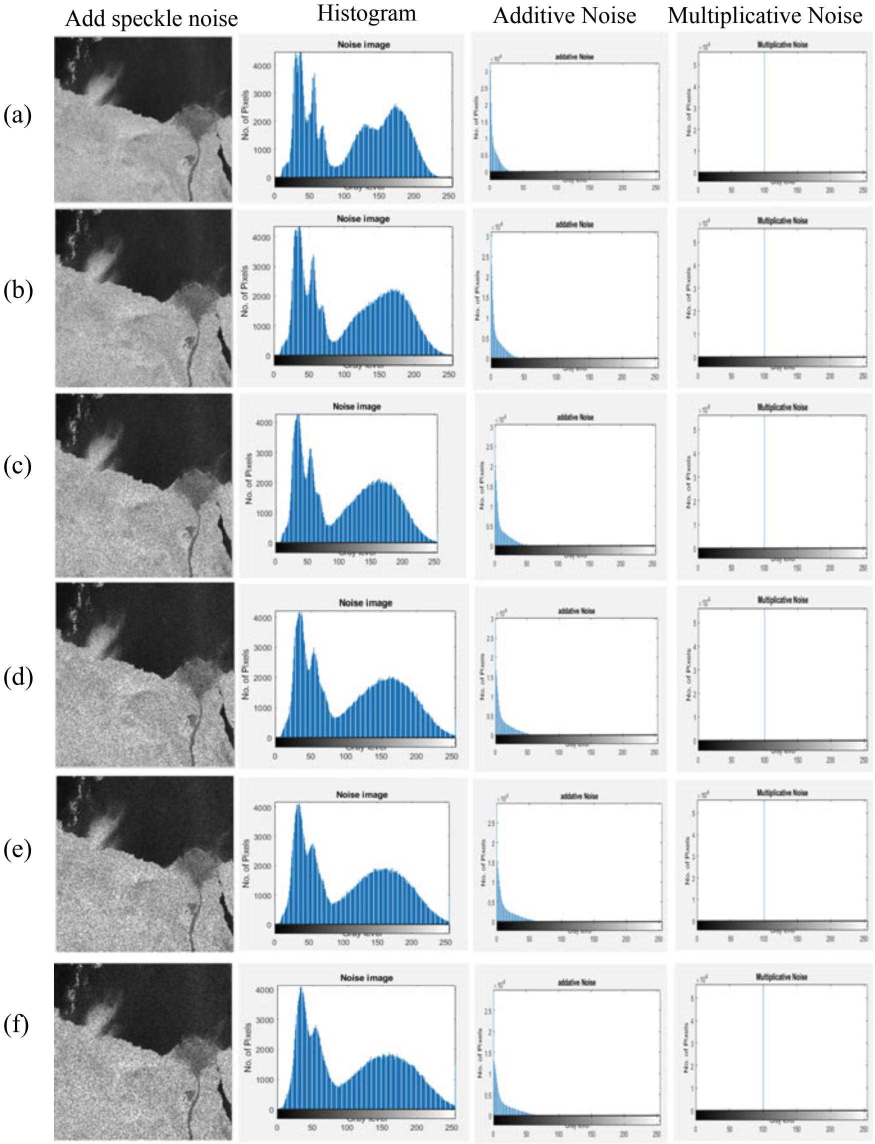


Fig. 3 The simulated images with Speckle noise ratio. **a** 0.01, **b** 0.02, **c** 0.03, **d** 0.04, **e** 0.05, **f** 0.06 with desert Sinai image, additive, and multiplicative histogram

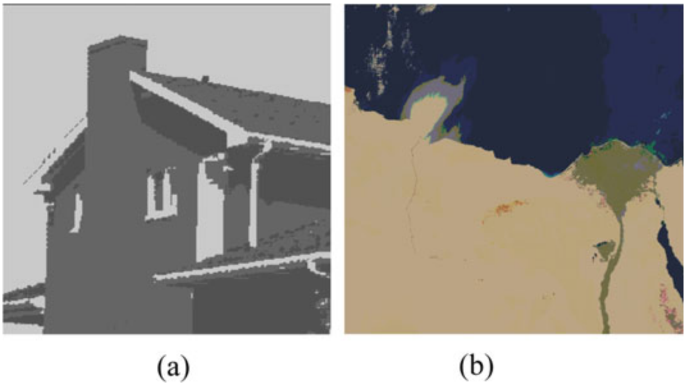


Fig. 4 The classified images using MD for **a** house and **b** desert Sinai

The classification process based on four blocks chosen manually for different locations within the image. The statistical criteria, like mean and standard deviation, are calculated for the four blocks. The same position, for the blocks, considered to measure the noise for different noise ratio. The block locations dispatch in Fig. 5.

Speckle factor (SF) plotted in Fig. 6 describes the relationship between mean and standard deviation for each noise ratio. This behavior is linear and increased with an increase in the noise. Because the mean is almost constant and the STD is changing with the increase in the noise.

Tables 1 and 2 show the SF behavior for the noise in the image, which increases by increasing the noise ratio. This means SF is increase with increase the noise in the image and with the distortion in the image. This considered an indicator of the noise

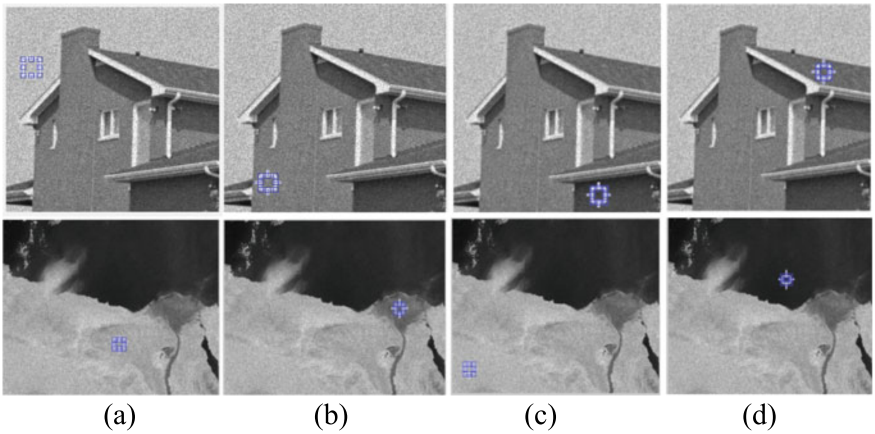


Fig. 5 The block locations in the House and Desert Sinai images with **a** class1, **b** class2, **c** class3, and **d** class 4

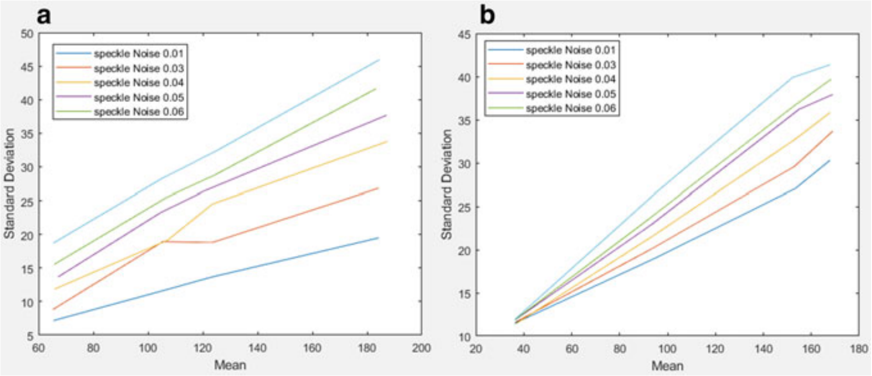


Fig. 6 Mean and standard deviation for each speckle noise for **a** house image, **b** desert Sinai image

Table 1 Statistical criteria's speckle noise for house image

Speckle noise images	Classes	Mean($\boldsymbol{\mu}$)	Standard deviation ($\boldsymbol{\sigma}$)	SF $= \frac{\mu^2}{\sigma^2}$
0.01	a	65.4926	7.1855	12.3193
	b	105.6533	11.6609	
	c	123.0657	13.6443	
	d	184.3000	19.5048	
0.02	a	65.2978	8.8386	18.0059
	b	105.4178	18.9417	
	c	123.3391	18.7934	
	d	184.2425	26.8446	
0.03	a	65.6838	11.8310	21.9456
	b	107.2089	19.1707	
	c	123.4464	24.5107	
	d	187.4425	33.7766	
0.04	a	67.0919	13.6672	24.0155
	b	104.4222	23.2168	
	c	120.6194	26.4431	
	d	187.0900	37.6826	
0.05	a	65.9007	15.5926	26.0484
	b	106.6800	25.4857	
	c	124.3183	28.7794	
	d	183.3000	41.6410	
0.06	a	65.5515	18.7612	27.2047
	b	105.4222	28.3512	
	c	124.5467	32.2911	
	d	184.5350	45.9659	

Table 2 Statistical criteria's speckle noise for Desert Sinai image

Speckle noise images	Classes	Mean (μ)	Standard deviation (σ)	SF = $\frac{\mu^2}{\sigma^2}$
0.01	a	36.4498	11.5180	18.7693
	b	94.0300	18.9296	
	c	153.4333	27.0661	
	d	167.7214	30.2873	
0.02	a	36.6364	11.6314	22.0580
	b	93.4103	20.1171	
	c	152.8676	29.5777	
	d	168.9928	33.6894	
0.03	a	36.7368	11.4350	24.4066
	b	94.2705	21.6703	
	c	153.2519	32.7344	
	d	167.8611	35.8415	
0.04	a	36.4104	11.8817	26.0637
	b	93.8942	23.0439	
	c	154.8954	36.2138	
	d	169.0939	37.9454	
0.05	a	36.4817	11.9153	27.8211
	b	92.6689	23.6630	
	c	152.8176	36.6106	
	d	168.4554	39.7364	
0.06	a	36.3275	11.9328	29.4598
	b	94.1415	26.3350	
	c	152.2019	39.9039	
	d	168.0408	41.3927	

level in the image. Moreover, this factor should be measured in the homogenous region.

5 Conclusions

The noise simulation is presented in this work for the speckle noise. In the classification method, there are four blocks extracted to calculate mean and STD for these blocks for all images with different noise ratio (in the same position). The relationship between the mean and STD used to know the type of noise if it is multiplicative or additive. The relationship linearly increases, which means the noise is multiplicative.

SF behavior should be in the homogeneous region because the mean is usually constant, therefore we rely on the $\boldsymbol{\sigma}^2$ value. The noise is high when $\boldsymbol{\sigma}^2$ value is low and vice versa. This should be in the homogeneous region because the edges increase the $\boldsymbol{\sigma}^2$ value.

Moreover, the histogram of the image changing in the additive noise, while it fixed in the histogram of the multiplicative noise. This means that the type of noise is multiplicative.

It is recommended to apply the suggested method with a different type of images and sizes.

References

1. Ashour, A.S., et al.: Light microscopy image de-noising using optimized LPA-ICI filter. Neur. Comput. Appl. **29**(12), 1517–1533 (2018)
2. Gravel, P., Beaudoin, G., De Guise, J.A.: A method for modeling noise in medical images. IEEE Trans. Med. Imaging **23**(10), 1221–1232 (2004)
3. Fekriershad, S., Tajeripour, F.: Color texture classification based on proposed impulse-noise resistant color local binary patterns and significant points selection algorithm. Sens. Rev. (2017)
4. Sanamzadeh, M., Tsang, L., Johnson, J.T.: 3-D electromagnetic scattering from multilayer dielectric media with 2-D random rough ınterfaces using T-matrix approach. IEEE Trans. Antennas Propag. **67**(1), 495–503 (2018)
5. Faraji, H., James MacLean, W.: CCD noise removal in digital images. IEEE Trans. Image Process. **15**(9), 2676–2685 (2006)
6. Boyat, A.K., Joshi, B.K.: A review paper: noise models in digital image processing. arXiv preprint arXiv:1505.03489 (2015)
7. Mugunthan, S.R.: Concept of Li-Fi on smart communication between vehicles and traffic signals. J. Ubiquit. Comput. Commun. Technol. **2**, 59–69 (2020)
8. Kumar, T.S.: Video based traffic forecasting using convolution neural network model and transfer learning techniques. J. Innov. Image Process. (JIIP) **2**(03), 128–134 (2020)
9. Mateo, J.L., Fernández-Caballero, A.: Finding out general tendencies in speckle noise reduction in ultrasound images. Expert Syst. Appl. **36**(4), 7786–7797 (2009)
10. Rueda-Clausen, C.F., Morton, J.S., Davidge, S.T.: Effects of hypoxia-induced intrauterine growth restriction on cardiopulmonary structure and function during adulthood. Cardiovasc. Res. **81**(4), 713–722 (2009)
11. Kang, J., Lee, J.Y., Yoo, Y.: A new feature-enhanced speckle reduction method based on multiscale analysis for ultrasound b-mode imaging. IEEE Trans. Biomed. Eng. **63**(6), 1178–1191 (2015)
12. Guan, F.D., et al.: Anisotropic diffusion filtering for ultrasound speckle reduction. Sci. China Technol. Sci. **57**(3), 607–614 (2014).
13. Lai, D.: Independent component analysis (ICA) applied to ultrasound image processing and tissue characterization (2009).
14. Kapoor, A., Singh, T.: Speckle reducing filtering for ultrasound images. Int. J. Eng. Trends Technol. (IJETT), **37**(5) (2016)
15. Diwakar, M., Lamba, S., Gupta, H.: CT image denoising based on thresholding in shearlet domain. Biomed. Pharmacol. J. **11**(2), 671–677 (2018)
16. Choi, H., Jeong, J.: Speckle noise reduction for ultrasound images by using speckle reducing anisotropic diffusion and Bayes threshold. J. Xray Sci. Technol. **27**(5), 885–898 (2019)
17. Duarte-Salazar, C.A., et al.: Speckle noise reduction in ultrasound images for improving the metrological evaluation of biomedical applications: an overview. IEEE Access **8**, 15983–15999 (2020)

18. Wu, S., Zhu, O., Xie, Y.: Evaluation of various speckle reduction filters on medical ultrasound images. IEEE Eng. Med. Biol. (2013)
19. Ullah, A., Chen, W., Khan, M.A., Sun, H.G.: A new variational approach for multiplicative noise and blur removal. https://doi.org/10.1371/journal.pone.0161787 (2017)
20. Massonnet, D., Feig, K.L.: Radar interferometry and its application to changes in the Earth's surface. Rev. Geophys. **36**(4), 441–500 (1998)
21. Lopes, A., Nezery, E., Touzi, R., Lanr, H.: Structure detection and statistical adaptive speckle filtering in SAR images. Int. J. Remote Sens. **14**(9), 1735–1758 (1993)
22. Baraldi, A., Parmiggiani, F.: A refined gamma map SAR speckle filter with improved geometrical adaptivity. IEEE Trans. GE-33, **5**, 1245–1257 (1995)
23. Kennie, T.J.M., Mathews, M.C.: Remote sensing in civil engineering. Surrey University Press, Halsted (1985)
24. Michael Hord, R.: Digital image processing of remotely sensed data. Academic Press, INC., New York (1982)
25. Alzuky, A.A.D.: Quantitative analysis of synthetic aperture radar (SAR) images. Ph.D. thesis, College of science, university of Baghdad (1998)
26. Ougiaroglou, S., Evangelidis, G., Dervos, D.A.: A fast hybrid classification algorithm based on the minimum distance and the k-NN classifiers. https://doi.org/10.1145/1995412.1995430 (2011)

Wi-Fi-Based Indoor Patient Location Identifier for COVID-19

A. Noble Mary Juliet, N. Suba Rani, S. R. Dheepiga, and R. Sam Rishi

Abstract Nowadays, more number of people are affected by COVID by other infected people. The coronavirus spreads through an infected person when he/she talks, coughs or sneezes in front of others. Before getting COVID test results, infected person may be moving along with normal people in shopping malls, education campus and industries. It is difficult to identify uninfected people who are moving along with the infected people. This paper proposed location identification model which is used to track and locate the infected person or recovered person inside the building. The location sensing model in wireless network and global positioning system (GPS) cannot be used to track users inside the buildings. The proposed system uses Wi-Fi-based model for monitoring patients and identifying the location of infected people in different floors of the building. RSSI technology is used for tracking the mobile device that is carried by patients in the indoor environments. Then, it finds the other uninfected people near them by using their smartphones.

Keywords Wi-Fi · Access point · Received signal strength indicator · Indoor positioning system · Smartphone

A. Noble Mary Juliet (✉) · N. Suba Rani
Department of Computer Science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India
e-mail: cse.julie@drmcet.ac.in

N. Suba Rani
e-mail: suba@drmcet.ac.in

S. R. Dheepiga
Sopra Steria, Kanchipuram, Tamil Nadu, India

R. Sam Rishi
Department of Information Technology, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_38

1 Introduction

The COVID virus spreads from one person to others through droplets from an infected person, when they talk, coughs or sneezes in front of others. Also catching people who have close contact with someone who has COVID-19. Nowadays, more people are affected by COVID disease because of accessing any medium which is accessed by infected person near them. In any small or structured organization, it is easy to identify people who are having contact with an infected person. But in the case of a shopping mall, education campus or any unstructured organization, it is difficult to identify those people. Therefore, in order to overcome this problem, a system was designed that can easily track a person's position in a large organization. There is a great demand to provide people with suitable indoor positioning solutions. The recent technology development in the smartphone market and location-based services has led to the development of positioning systems with relatively high accuracy. Smartphones are used to identify location of person in outdoor environment. It uses the global positioning system (GPS), which works by satellites, positioned a thousand miles from the ground. But efficiency is reduced while approaching the indoor. GPS signals are reduced because of more obstacles like walls, rooms, closed areas with glasses and different indoor complex structures. This restriction increases the difficulty of designing indoor tracking system using GPS.

Different technologies like wireless signal transmission are used for tracking or locating a person or a device in the building. In recent technological updation, Wi-Fi plays a vital role in location recognition. As compare to other exiting infrastructures, it gives easy deployment, cost effective and improved accuracy. Recent smartphones are designed to handle different applications including indoor and outdoor location positioning capabilities using different sensors like GPS, Wi-Fi and Bluetooth transceiver. GPS-based person tracking system needs internet connection and GPS working only defined maps. Our proposed system based on Wi-Fi access points do not need Internet connection. Wi-Fi coverage area of access point or mobile phones is usually 150–200 feet. This is because access points are usually used, and their locations are optimized for data communication. Accuracy of this is based on the walls, doors, number of access points or people. Smartphone sensors are able to handle floor level location identification. Received signal strength indicator (RSSI) is also used to find position without any additional hardware and is easy to estimate. Wi-Fi access points and mobile devices are used to determine the RSS value. This paper aims to identify and track the infected person and uninfected person in precise location of mobile devices in an indoor environment. Healthcare executives have been working hard to resolve issues such as facility safety, employee satisfaction, the quality of care provided to patients and high costs and inefficiencies that affect their bottom line. This paper describes the indoor patient location identifier using Wi-Fi.

The rest of this paper is arranged as follows. Section 2 shows previous related works in indoor positioning system. Section 3 presents the implementation approaches of indoor location identifier using Wi-Fi. Section 4 presents experimental and simulation results of the proposed model. Finally, Sect. 5 draws conclusions with directions of future work.

2 Related Works

For the indoor location tracking system, the most frequently used technology is Wi-Fi. This is a standard and less expensive technology used by many people, with basic components. The technology is compatible with electronic devices that use radio waves to transmit information in air. Wi-Fi tools such as smartphones usually communicate over 2.4 GHz, but nowadays, 5 GHz channels are used as they have less vibration, less interference, faster speeds and more stable performance. With Wi-Fi technology, different forms of position estimation and location determination are done. Autonomous smartphone-based Wi-Fi positioning system by using access points localization and crowdsourcing [1] method is used for automatic access points and their propagation parameters estimation by employing an indoor navigation. Other than Wi-Fi, there are several wireless standards available like RFID, Bluetooth and Infrared. Both Wi-Fi and Bluetooth are frequent accessing medium used for indoor positioning in campus premises. A survey in indoor fingerprint positioning based on Wi-Fi [2] discussed indoor positioning system and algorithm using Wi-Fi fingerprint indoor positioning. A Study on room-level accuracy of Wi-Fi fingerprinting-based indoor localization systems [3] discussed about feasibility of room-level location detection in small and large organization and focused on examining the quality of room-wise detection and accuracy of the fingerprinting method that is applied along with standard Wi-Fi radio infrastructure.

The signal blocking problem caused by obstacles existed inside the building. An improved Wi-Fi trilateration-based method for indoor positioning system [4] resolved the above problem by improving received signal strength measurement. The accuracy of indoor location identification system depends on the participating user and their location. Practical location validation in participatory sensing through mobile Wi-Fi hotspots [5] proposed a location validation system (LVS) that provides secure positioning from location-spoofing attacks, and also, the user location is verified with the help of Wi-Fi in their smartphones and accepting connections from nearby smart devices and locating their position inside the sensing area. For indoor patient location tracking, certain methods are used such as time of arrival, angle of arrival and received signal strength and fingerprint. Time of arrival is a parameter used for knowing the exact time that a signal was sent from the base station to the user. It specifies a circle of possible locations in the two-dimensional area. This circle has its middle point at the base station, and its radius corresponds to the distance. The angle of arrival of a signal is the direction from which the signal such as radio, optical or acoustic is received. The received signal strength (RSS) is the strength of a received signal measured at the receiver's antenna. This parameter calculates the distance between the transmitter and the receiver. Yadav et al. [6] proposed a TKBE algorithm that handles the signal fluctuations and drift errors using a fuzzy-logic Kalman filter.

It also discussed RSSI-based localization depending on the environment and proved that accuracy decreases with the change of the environment and the fading effect. Gang et al. [7] proposed indoor location-based services system. It used hybrid indoor positioning methods based on Bluetooth beacons, geomagnetic field, sensors and smartphone cameras and can be used for indoor location-based applications. Also, it proved that the performance of each positioning method got the preferred accuracy for some conditional inputs.

3 Proposed Model

In this work, we propose a mechanism that can monitor the patient's indoor location within a certain period of time. The mechanism will automatically maintain a database containing data of all people including infected people, which will indicate the time and place of their visits. This data is intended to be used by people for safety measures. This recommendation assumes that the sensors used to detect Wi-Fi strength are constantly carried by the people. Alternatives may be sensors built into smartwatches or any Wi-Fi-based devices. This paper deals with smart devices; thus, it traces person who are hold smartphones or smart devices. It is assumed that when a person enters into the organization, Wi-Fi connection will be established with his or her smartphone. Access points in the organization update current user's information and upload it into cloud database with regular interval.

Wi-Fi access points carried out this process by receiving the signals from the user devices. It estimates the signal strength between the Wi-Fi access point and the smartphone using RSS measurements. Some threshold values will be used while handover between access points. The access points select the strongest signals with specified range. In this proposed system, user devices are need not to be established a connection with access point, but it sends a bacon to the nearest access points. This feature helps the access point to gather signals from mobile devices without crossing security process. Access points are designed to maintain a table that includes mobile phone number in its specified range. If the device leaves its range, then the user data will be removed from current table. This table will be updated into cloud database regularly. Figure 1 shows the block diagram of the proposed system. Modules present in the proposed system are as follows:

3.1 Calibration Phase

In this module, access point information like access point location, id and MAC address are stored in a table called access_points_table. This table is maintained by the server. A software component is installed in all access points which helps the access points to handle beacon signals from user's smartphones. This component includes a user table which is used to store the data of people who are connecting

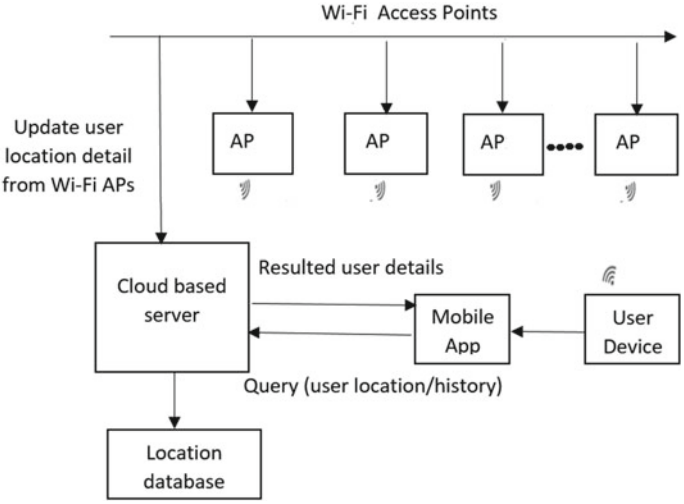


Fig. 1 Block diagram of proposed system

with access points through beacon signals and is called user_registration_table. This table includes phone number, signal strength, access point, time in and time out of each smartphone connected with access points. This table updates with regular time interval. It is assumed that a piece of software component must be installed in user's smartphones who are all entering the building or campus. And they are insisted to switch on their Wi-Fi connection. To complete this process, the user must complete registration process with server. The server will be maintaining user_entry_table which includes name, mobile number, address, time in and time out. After successful verification, login credentials are allocated to the user and allows to enter into the organization. This process will take some time for a new user but existing will be allowed directly with login in process. After successful registration or login, it starts tracking the user. That is, the Wi-Fi is enabled all the time, and the application sends the MAC address, the connected access point and other details to the database at regular intervals.

3.2 Location Identification Phase

This module helps user to find their current location. If the user is new to the campus, they are not aware about campus means, and this phase helps them to find their current location by just clicking the current location button. It displays the current location of the user. The current location is identified by searching for the nearest access point based on the strength of the received signal. It also displays the nearest location using *access_points_table* from cloud database.

3.3 User Roaming Monitoring Phase

This module helps to find the registered users in a particular location by access point id and find the user's history by their phone number stored in access points. Whenever the people enter into one access point coverage, a new entry is added into user_entry_table in access points. When they are moving to another place, the out time is recorded and updated in the user_registration_table in the server. A new entry will be added in the next nearest access point based on RSS. This system provides information about moving history of any person with the help of user_entry_table and user_registration_table in cloud database. Nowadays, a greater number of people are affected by COVID by other infected people. Before getting COVID test results, infected person may be moving inside the campus along with the normal people. If the test result is positive, then we must find out the other people who are near them or moving along with them. This module is able to handle this situation. If we enter a mobile number of an infected person, then it will display moving history of that person and also the display the other people who are staying along with them.

3.4 Alerting Phase

This phase will find out if any person who had recovered from COVID, and it gives an alert to people about the recovered person in their location. Then, the others may take safety measures. It is assumed that the infected person's phone number is collected from government resource. When a recovered person and non-infected person is in the same Wi-Fi range, the uninfected person will be warned.

4 Experiment and Evaluation

Experiment Setup

In this project, the experiment is conducted inside our college campus. More than 25 number of available access points are randomly placed in different areas like laboratory, classrooms and administration office. The positioning software is developed using android studio and installed in the smartphones or smart devices like tablet. This software connected with nearest access points receives the RSS signal within distance of 25 m. The RSS value is calculated between the smart device and the access points signal strength at a particular distance as shown in Table 1. The test was done with fifty persons moving to ten different places in the campus. This proposed system produces 90% accuracy as shown in Fig. 2.

The measurements have been carried out using Wi-Fi Internet during the movement of people inside the building. The system performance is demonstrated through the IEEE 802.11. The range of the RSSI will increase because the human behaviour

Table 1 RSS value at particular distance

S. no.	Distance (m)	RSS (dBm)
1	1.5	−50
2	5	−63
3	9	−70
4	14	−75
5	25	−83

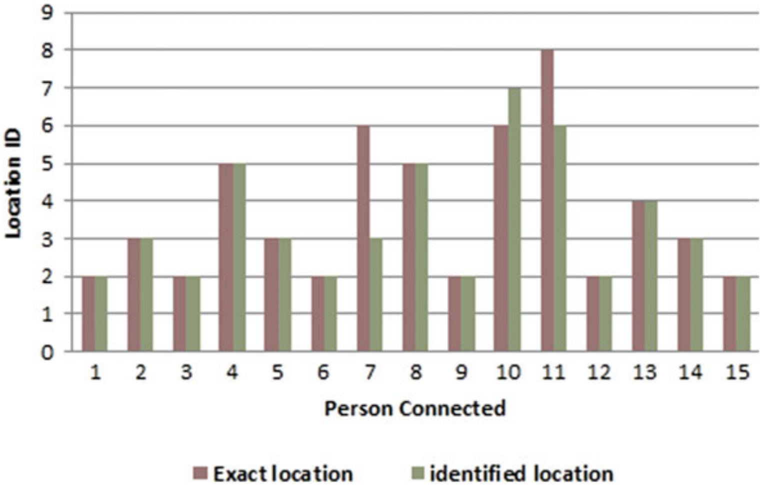


Fig. 2 Accuracy of proposed system

impacts the propagation path of the real signal of APs, which causes the RSSI to increase/decrease instantly. The received signals (RSS) is in negative form. If the signal strength is high, then the RSS value will be close to zero. Figures 3 and 4 show the sample screenshots of the proposed system.

5 Conclusion

More number of people are moving around and accessing common resources and amenities in large campus like shopping malls and educational institutes. Maintaining physical records for finding people who are gathering or accessing common resources in one particular location is very difficult. As recent technologies in wireless commination and smartphones are improving, the paper focuses more on COVID patient's location identification using smartphone or smart devices using Wi-Fi without GPS model. In this paper, we have proposed an indoor location identifier based on Wi-Fi for ensuring the high accuracy and reduced accessing time. First, we have proposed

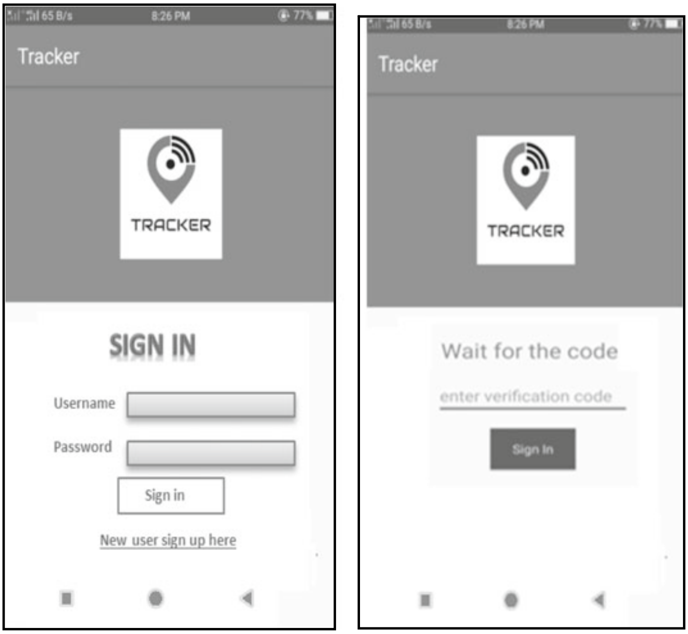
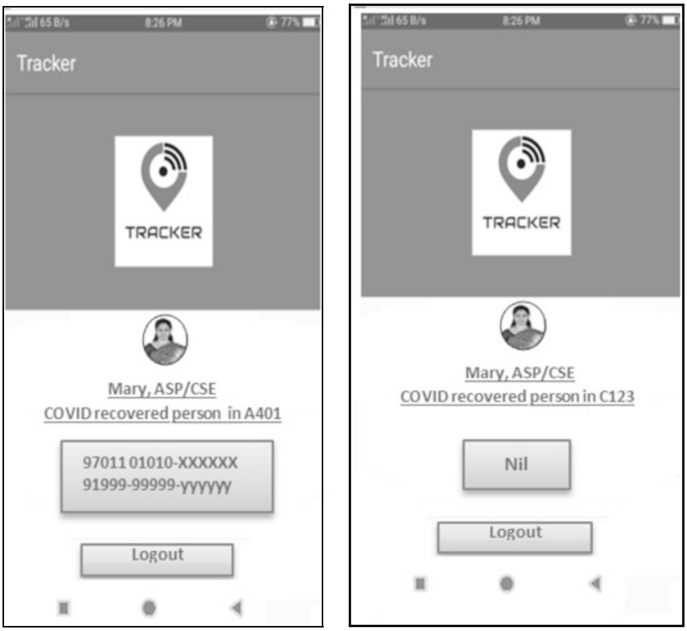


Fig. 3 Registration page of proposed system



Mary, ASP/CSE
COVID recovered person in A401

97011 01010-XXXXXX
91999-99999-yyyyyy

Mary, ASP/CSE
COVID recovered person in C123

Nil

Fig. 4 Infected person in current location of proposed system

calibration phase, which authenticates stores user information in access points using Wi-Fi capability of modern smartphone with RSS signal. Furthermore, we have introduced location identification phase which is used to find the current location of a user, and also, they can start a guided tour from their current location. We have also proposed a positioning phase to find the registered users in a particular location by access point id and to find the user's history by their phone number stored in the access points. Results conclude that the proposed system is an efficient model, produces more accuracy and applicable to large organization with Wi-Fi-based campus. The future goal of the system is to provide efficient and improved accuracy in indoor location tracking system using recent technologies such as Bluetooth, GSM and RFID, and increasing more access points with large coverage area will improve positioning accuracy. Since the smartphone sensors are being developed, we can extend this mechanism by using different sensors like accelerometers, gyroscopes and magnetometers to provide more information about the users and their locations in the indoor environment.

References

1. Zhuanga, Y., Syedb, Z., Georgyb, J., El-Sheimya, N.: Autonomous smartphone-based WiFi positioning system by using access points localization and crowdsourcing. Pervasive Mobile Comput. **18** (2015)
2. Xia, S., Liu, Y., Yuan, G., Zhu, M., Wang, Z.: Indoor fingerprint positioning based on Wi-Fi: an overvie. Int. J. Geo-Inf. **6**, 135–160 (2017)
3. Çabuk, U.C., Dalkılıç, F., Dağdeviren, O.: A study on room-level accuracy of wi-fi fingerprinting-based ındoor localization systems. Celal Bayar Univ. J. Sci. **15**(1), 17–22 (2019)
4. Rusli, M.E., Ali, M., Jamil, N., Din, M.M.: An ımproved ındoor positioning algorithm based on rssı trilateration technique for Internet of Things. In: International Conference on Computer and Communication Engineering (2016)
5. Restuccia, F., Saracino, A., Martinelli, F.: Practical location validation in participatory sensing through mobile WiFi hotspots. In: 17th IEEE International Conference on Trust, Security and Privacy ın Computing and Communications (2018)
6. Yadav, R.K., Bhattarai, B.G., Gang, H.S., Pyun, J.Y.: Trusted K Nearest Bayesian estimation for ındoor positioning system. IEEE Access **7**, 51484–51498 (2019)
7. Gang, H.S., Pyun, J.Y.: A smartphone indoor positioning system using hybrid localization technology. Energies **12**, 3702 (2019)
8. Kim, S., Ha, S.H., Saad, A., Kim, J.H.: Indoor positioning system techniquesand security. In: Forth International Conference on e-Technologies and Networks for Development (2015)

Enabling Identity-Based Data Security with Cloud

Arya Sundaresan, Meghna Vinod, Sreelekshmi M. Nair,
and V. R. Rajalakshmi

Abstract In disseminated stockpiling organizations, customers store data indirectly onto the cloud and comprehend the data giving to others. Far away information uprightness taking a gander at is proposed to the attestation the respectability of the informational index aside in the cloud. In some fundamental appropriated amassing frameworks, for example, the EHRs structure, the cloud document may give touchy data. At the point when the cloud report is shared, the delicate information ought not to be known to anybody. Bringing the whole shared report can comprehend the inclusion of the collaborating records; however, it does not permit different gatherings to utilize this shared record instructions to recognize data offering to sensitive information stowing away in distant data uprightness exploring still has not been researched up to now. To address this issue, this paper proposes a distant data genuineness surveying plan that recognizes data offering to sensitive information concealing. To accomplish high security, this paper proposes security calculations for producing the private key and for making record label name for relating information. The entire information's are encoded twice prior to shipping off cloud. Just the information proprietor can see the entire information from cloud. Personality-based trustworthiness examining ensure that different clients could just see the required information by concealing delicate data safely.

Keywords Blind · Integrity auditing · sha256 · File tag · Auto private key generation

1 Introduction

To develop identity-based integrity audit process and secure cloud storage sensitive information, conveyed registering is the latest development in the field of flowed figuring. It gives distinctive on the web and on-demand benefits for data accumulating, network organizations, stage organizations, etc. Distributed processing has

A. Sundaresan (✉) · M. Vinod · S. M. Nair · V. R. Rajalakshmi
Department of Computer Science and IT, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_39

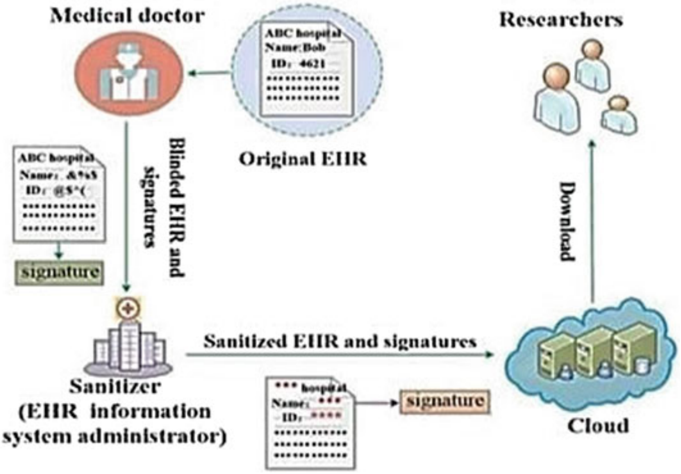


Fig. 1 EHRs example

actually wandered acclaim and shaped into a huge example in IT. We play out an especially purposeful review of appropriated registering and explain the particular challenges glancing in this paper.

Figure 1 shows pictorial description for EHRs. EHRs are divided into two sections, each of which contains classified information. Firstly, personal sensitive information such as patient's name and id. Secondly is organizational sensitive information such as hospital name. Sanitizer can be regarded as administrator of the EHR information system in a hospital. Personal sensitive information should not be exposed to sanitizer, and it should also be not exposed to the cloud and the shared users. EHR of client's data is generated by the respective doctors which is send to sanitizer for storing in EHR that normally contain sensitive information regarding the client.

To retain the privacy of client, the doctor will blind the client's confidential facts of each EHR before exposing it to sanitizer, and hence, these messages are stored into EHR.

When doctor needs the EHR, sends request to the sanitizer. From EHR, the sanitizer downloads the blinded EHR and send them back to doctor. At last, the doctor recovers the original EHR from the respected blinded EHR. At the moment of uploading and sharing this EHR to cloud for research purpose, the sanitizer needs to sanitize the data blocks corresponding to the patient's sensitive information of the EHR.

Wildcards are used to replace data blocks. Sanitizer transforms data signatures into valid sanitized EHRs. Sanitizer does not need to interact with doctors. At final, sanitized EHRs and their corresponding signatures are uploaded to the cloud by sanitizer. Through this method, EHR can be shared and used by researchers with sensitive information of EHRs hidden. Sanitizer is essential because of below mentioned reasons. After the data blocks corresponding to confidential information are blinded, the contents of data sets can become messy code. Sanitizer can unify the format by using wildcard to modify the contents of the data block. Sanitizer sanitizes data blocks containing personal details, protecting the organization's privacy. Sanitizer, as the EHR administrator, can download the blinded EHR and give it to the doctor, who can then restore the initial EHR from the blinded. Sanitizer can sanitize bulk EHRs and upload it to the cloud at a fixed time period.

2 Background and Related Works

2.1 Previous Research

Clients are troubled with an enormous volume of information to store information locally. A great many associations and individuals need to store cloud-based information. As a result of the equipment breakdown, human mistake and programming bug in the cloud, the information store on the cloud is ruined or lost. Some new security dangers to information proprietors are brought about by distributed storage. For some critical security penetrates, most cloud clients do not need distributed storage. The classification of their re-evaluated records is an essential worry of cloud clients. Numerous different elements can contribute towards significant information debasement. To begin with, providers of cloud administrations are not totally trusted. As a result, for financial purposes, the cloud specialist organization can erase phenomenal or unreached information with the goal that it can save space for different records to store for extra costs. Second, because of cloud worker glitch, the executives' slip-ups or foe assaults, the put away information might be undermined. Notwithstanding, a cloud specialist organization can intentionally disguise information misfortune occasions to protect a decent standing. The cloud has been a significant danger for distributed storage for information uprightness and spillage.

2.2 Present Research

'Third-Party Auditor' set up to confirm regularly, on the customer, the integrity of cloud data for decrease these computing. The user-side burden [1] model called 'Proof of retrievability' (POR) was introduced, and functional system was proposed. In this method, it is possible to recover the information recorded on cloud to ensure these confidentiality on the information. Pseudorandom feature and BLS signature dependent [2]. Wang et al. suggested a remote data integrity auditing method for privacy preserving with the use of random maski technique to preserve data privacy [3]. Build a 'remote data integrity auditing scheme' supporting data privacy protection. Worku et al. used a different random masking technique. Compared with the scheme, this system achieves higher effectiveness. To decrease these user-side computing burden of signature generation [4], design a lightweight remote data integrity auditing system. The TPM allows users to generate signatures under this scheme to support data dynamics [5] (Fig. 2).

Identity-based integrity for secure cloud storage, in order to facilitate the sharing of data effectively with confidential information hidden in auditing, our framework designed achieves these objectives: ‘Accuracy’ correctness of the ‘private key’ is to confirm that those the private key will pass user verification when PKG gives the user an accurate key. Auditing ‘correctness’ to make sure that the evidence it provides will pass the inspection of the TPA cloud which will properly store the sanitized data of the customer. ‘Sensitive information’ to confirm the important information which is not found in a register infected and that all important information is cannot be revealed (Fig. 3).

File was ready to share without exposing sensitive information. Here, first, the user blinds the information blocks, such as the file’s private sensitive data, and generates the corresponding signatures. Such signatures are used to guarantee the authenticity

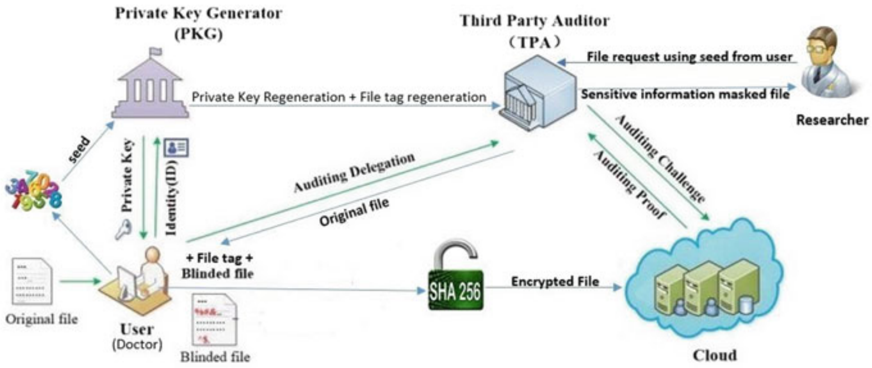
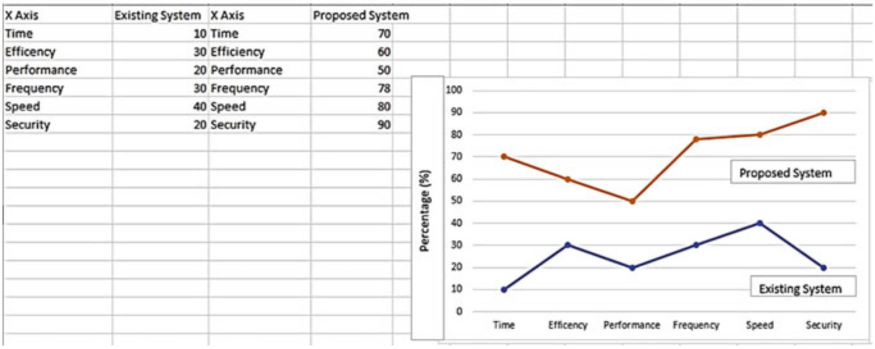


Fig. 2 Proposed system architecture



X Axis	Existing System	X Axis	Proposed System
Time	10	Time	70
Efficency	30	Efficiency	60
Performance	20	Performance	50
Frequency	30	Frequency	78
Speed	40	Speed	80
Security	20	Security	90

Fig. 3 Comparison between existing and proposed system

of the file and verify the integrity of the file. This blind file and its corresponding signatures are then sent by the user to the sanitizer. After receiving the message from the user, the sanitizer sanitizes these blinded data blocks and therefore the data block, such as the confidential details of the company, then converts for the sanitized paper, sanitized information blocks sign to legal. Finally, this 'Sanitized file' and its associated sign sent into the cloud by the sanitizer.

In order to achieve dignity, these signatures will not be used to verify the authenticity of the sanitized audit stage. To check the credibility of the sanitized file contained inside the cloud. And then, with the auditing evidence of information ownership, the cloud reacts to the TPA. Hence, the file containing sensitive information is often shared for research purpose without revealing sensitive information.

3 Experiment and Result Analysis

This procedure uses double encryption to hide confidential information. In this way, the files stored in the cloud are also exchanged and can be used by anyone, as long as the sensitive information is protected by a private key. Block-level principles are used to store data randomly and maintain protection. Five kinds of different organizations are involved in the device model: 'cloud', 'consumer', 'sanitizer', 'private key generator' (PKG) and 'third party auditor' (TPA).

- Cloud: The cloud provides the user with massive data space for storage. Every user can upload their data into cloud and share the data through cloud storage user —A user can be a member of a company that there are a huge amount records must recorded. 'Disinfectant': These disinfectant are responsible for disinfecting the blocks of information within a file, such as sensitive information (sensitive personal information and, therefore, sensitive information of the organization), making these block signatures legitimate.
- Private Key Generator: Other organizations trust the PKG. It is responsible for creating public parameters of the system and thus for the user's private key consistent with his identity ID for the user consistent with his identity ID.
- Third Party Auditor: A public verifier could be the TPA. It is responsible for checking, on behalf of users, the accuracy of the information stored inside the cloud.

3.1 Algorithm

TagGen, Challenge, Proof and Verify. • Configuration (1k) → (params, mpk, msk): This algorithm takes k, which is a security parameter, as input and generates the public system parameters, the mpk master public key and the master secret.

Extract (mpk, msk, params, ID) → skID: This algorithm takes the pp system parameters, msk 'master secret key', and user ID as input. skID of the user's private key is issued.

- PSKGen (params, IDu, IDp) → u: This algorithm accepts as input the parameters of the system parameters, user ID IDu, proxy server ID. Result is proxy signing key u.
- TagGen (F, u) → σ: This algorithm takes a file F and a proxy signature u as input. Computes and generates the corresponding knowledge block label σ = (σ1, σ2, …, σn).
- Task (Finfo) → C: This algorithm takes a knowledge file as input and generates a set of tasks.
- Test (F, C, σ) → P: This algorithm takes the file F, task set C, from the validation labels TPA and σ as input and generates a response test P 11
- Verify (C, P, params, mpk, Finfo) → 0/1: This algorithm takes C call set, P answer test, system parameters, mpk master public key, Finfo knowledge file information as input and output as result audit trail 0 or 1.

3.2 Performance of Different Processes

Private key generation and personal key verification spend almost an equal amount of time, which is almost 0.31 s, As reflected in Fig. 4, time spent producing these signature is 1.476 s. Time for the verification of the signature is 2.318 and 0.041 s of classified information sanitization, respectively. So it can be assumed that the signature authentication spends the longest time in these processes, and therefore, the shortest time is spent on confidential information sanitization (Fig. 5).

For different block numbers from 0 to 1000, we produce signatures, in our experiment, supplemented by an interval of 100, to test the efficiency of signature generation and signature verification. As reflected in Fig. 6, the price of time producing sign

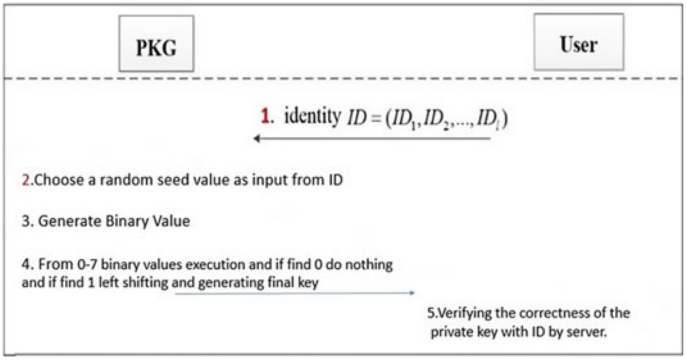

PKG
User
1. identity $ID = (ID_1, ID_2, \ldots, ID_i)$
2.Choose a random seed value as input from ID
3. Generate Binary Value
4. From 0-7 binary values execution and if find 0 do nothing and if find 1 left shifting and generating final key
5.Verifying the correctness of the private key with ID by server.

Fig. 4 Private key generator

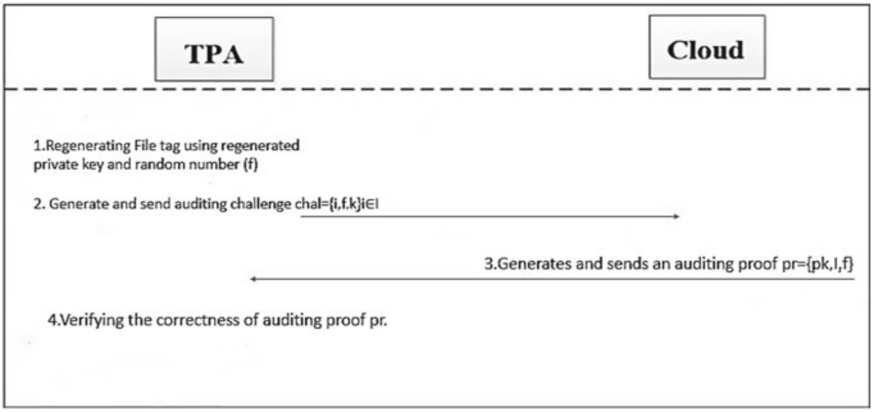

TPA
Cloud
1.Regenerating File tag using regenerated private key and random number (f)
2. Generate and send auditing challenge chal={i,f,k}i∈I
3.Generates and sends an auditing proof pr={pk,I,f}
4.Verifying the correctness of auditing proof pr.

Fig. 5 Third party auditor

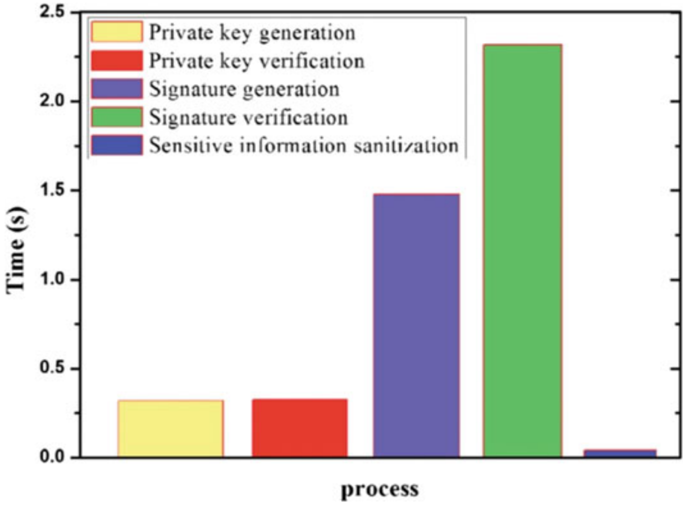


Fig. 6 Performance of different processes

and thus confirmation of the signature increase in quantity on info block. In those moments to produce signatures, the time varies between '0.121' and '12.132' s. The time of authentication of the signature varies from '0.128' to '12.513' s (Fig. 7).

The cloud file may contain some confidential data in the 'electronic health records' (EHRs) framework, when the cloud file is exchanged, the data should not be known to anyone, and it will be unable to use. To guarantee the credibility, these information recorded within the cloud. Identity-based integrity auditing confirms that the opposite users could only view the needed data by hiding sensitive information securely. Disadvantage of the prevailing system is safety issue. Signature algorithm addresses

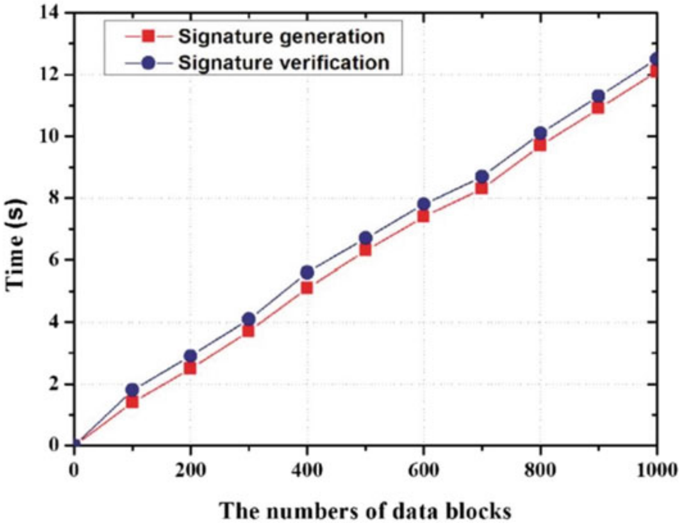


Fig. 7 In the 'signature generation', 'signature verification' process

the downside of the current method. It supports block less verifiability, which permits the verifier to see the integrity of information without uploading the entire data from the cloud. Identity-based cryptography is supported by simplifying the complicated management of certificates (Fig. 8).

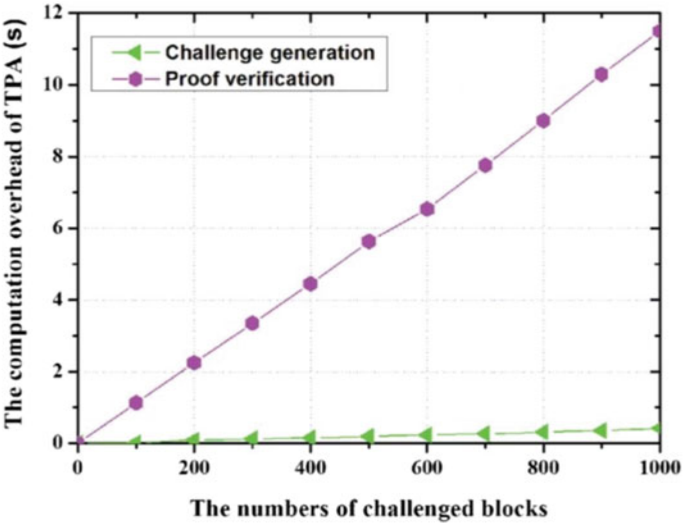


Fig. 8 In stage honesty auditing, the computing overhead of TPA

4 Conclusion

Other than secured distributed storage, a character-based information trustworthiness review measure frameworks are proposed to encourage information offering to mystery touchy data. The PKG produces the client's private key viable with his personality ID in our plan. The client ought to check the accuracy of the private key that has been acquired. This client should utilize a blinding variable to dazzle the information blocks like the private delicate data of the main document when there is a requirement for the client to transfer information to the cloud to secure private touchy information of principal record from the sanitizer. At the point when needed, by utilizing this blinding component, the client can recuperate the main document from the blinded one. This client at that point uses the constructed signature calculation for the blinded record to get marks. It will be typical for these marks to confirm the authenticity of this blinded record. Also, the client produces a document label that will be utilized to guarantee that the name of the record identifier and a couple of check esteems are right. The purchaser additionally gauges a change esteem that is typically used to change over sanitizer marks. At last, the client sends the blinded record, its comparing marks, and subsequently the change an incentive to the sanitizer going with the document. At the point, when the above messages are genuine, the sanitizer cleans the visually impaired information impedes first in a viable configuration, and similar to the secret data given by the association, it disinfects data blocks to secure the protection of the association and changes its comparing marks into substantial ones for the sterilized record with the estimation of the change. At long last, this transfers the sterilized documents to the cloud with the important marks. The cloud creates a test confirmation viable with the TPA challenge if the information uprightness reviewing task is performed. Through checking, if the test evidence is precise, TPA will check the validity of the cleaned document that was put away in the cloud.

References

1. Ateniese, G., Bums, R., Curtmola, R., Kissner, L., Peterson, Z., Song, D.: Provable data possession at untrusted Stores. In: Proceedings of the 14th ACM Conference on Computer and Communications Security, ser.CCS '07, pp. 598–609 (2007)
2. Juels, B., Kaliski, S.: Pors: proofs of retrievability for large files. In: Proceedings of the 14th ACM Conference On Computer and Communications Security, ser.CCS'07, pp. 594–597 (2007)
3. Wang, C., Chow, S.S. M., Wang, Q., Ren, K., Lou, W.: Privacy-preserving public auditing for secure cloud storage. IEEE Trans. Comput. 62(2), 362–375
4. Worku, S.G., Xu, C., Zhao, J., He, X.: Secure and efficient privacy-preserving public auditing scheme for cloud Storage. Comput. Electr. Eng. **40**(5), 1703–1713 (2014)
5. Shen, W., Yu, J., Xia, H., Zhang, H., Lu, X., Hao, R.: Light-weight and privacy preserving secure cloud auditing scheme for group users via the third party medium. J. Netw. Comput. Appl. **82**, 56–64 (2017)

6. Ren, K., Wang, C., Wang, Q.: Security challenges for the public cloud. In: Internet Computing IEEE , vol. 16(1), pp. 69–73, Jan 2012
7. Shacham, H., Waters, B.: Compact proofs of retrievability. J. Cryptol. **26**(3), 442–483 (2013)
8. Guan, C., Ren, K., Zhang, F., Kerchbaumn, F., Yu, J.: Symmetric-key based proofs of retrievability supporting public verification. In: Computer Security—ESORICS, pp. 203223 (2015)
9. Wang, Q., Wang, C., Ren, K., Lou, W., Li, J.: Enabling public auditability and data dynamics for storage security in cloud computing. IEEE Trans. Parallel Distrib. Syst. **22**(5), 847–859 (2011)
10. Yu, J., Hao, R., Xia, H., Zhang, H., Cheng, X., Kong, F.: Intrusion-resilient identity-based signatures: concrete scheme in the standard model and generic construction. Inf. Sci. **442**, 158–172 (2018)
11. Yu, J., Wang, H.: Strong key-exposure resilient auditing for secure cloud storage. IEEE Trans. Inf. Forensics Secur. **12**(8), 1931–1940 (2017)
12. Yu, J., Ren, K., Wang, C., Varadharajan, V.: Enabling cloud storage auditing with key-exposure resistance. IEEE Trans. Inf. Forensics Secur. **10**(6), 11671179 (2015)
13. Sun, J., Fang, Y.: Cross-domain data sharing in distributed electronic health record systems. IEEE Trans. Parallel Distr. Syst. **21**(6), 754–764 (2010)
14. Wang, B., Li, B., Li, H.: Oruta: privacy-preserving public auditing for shared data in the cloud. In: 2012 IEEE Fifth International Conference on Cloud Computing, pp. 295–302, June 2012
15. Yang, G., Yu, J., Shen, W., Su, Q., Fu, Z., Hao, R.: Enabling public auditing for shared data in cloud storage supporting identity privacy and traceability. J. Syst. Softw., **113**(C), 130–139 (2016)
16. Yu, Y., Au, M.H., Ateniese, G., Huang, X., Susilo, W., Dai, Y., Min, G.: Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage. IEEE Trans. Inf. Forensics Secur. **12**(4), 767–778 (2017)
17. Ateniese, G., Chou, D.H., de Medeiros, B., Tsudik, G.: Sanitizable signatures. In: Proceedings of the 10th European Conference on Research in Computer Security, ser. ESORICS'05, pp. 159–177. Springer, Berlin (2005)
18. Ateniese, G., de Medeiros, B.: On the key exposure problem in chameleon hashes. In: Security in Communication Networks, pp. 165–179. Springer, Berlin, 2005.
19. Zhang, Y., Yu, J., Hao, R., Wang, C., Ren, K.: Enabling efficient user revocation in identity based cloud storage auditing for shared big data. IEEE Trans. Dependable Secure Comput. (2018)
20. Raj, J.S.: Improved response time and energy management for mobile cloud computing using computational offloading. J. ISMAC **2**(01) (2020)
21. Haoxiang, W., Smys, S.: MC-SVM based work flow preparation in cloud with named entity identification. J. Soft Comput. Paradigm (JSCP) **2**(02) (2020)
22. https://images.app.goo.gl/Qd4AmGgXZYFdUd8P6
23. https://images.app.goo.gl/LBL86rfKtkcVVEPa9
24. https://images.app.goo.gl/GzMCP1EV9XJQJ73Q6
25. https://images.app.goo.gl/znwNkaXiE1Nb6jxp7
26. https://images.app.goo.gl/pVCTFusTRRnT2KAN6
27. https://images.app.goo.gl/gnnFE57vRdxEWr8z7
28. https://images.app.goo.gl/fAYePUhqBtV19Qv96

A Study and Review on Image Steganography

Trishna Paul, Sanchita Ghosh, and Anandaprova Majumder

Abstract Steganography is the science that involves encrypting data in a suitable multimedia carrier, such as image, audio, and video files. The main purpose of image steganography is to hide the data in images. This means that it encrypts the text in the form of an icon. Steganography is done when there is communication takes place between sender and receiver. In a day of data transfer over the network, security is paramount. Before the development of stenography, data security is a major research concern for researchers. Steganography is gaining importance due to the rapid development of users on the Internet and secret communication. In this paper, we discuss about various type of existing image steganography techniques and analyze the advantages and disadvantages of different types of image steganography techniques.

Keywords Steganography · Multimedia carrier · Communication · Data transfer · Security · Image steganography

1 Introduction

The challenges in protecting individuals' privacy are becoming more difficult as digital communication technology progresses and computing capacity and storage rises. The degree to which people value privacy varies from one person to the next. To protect personal privacy, numerous methods have been investigated and developed. The most noticeable is possibly encryption, followed by steganography. Encryption is sensitive to noise and is commonly observed, whereas steganography is not. Steganography is a widely used technique that manipulates information to hide their existence. Although steganography provides good security, the term stenography comes from the Greek words Stegano's (in enclosure) and Graptos (written) which

T. Paul (✉) · S. Ghosh · A. Majumder
Department of Computer Science and Engineering (CSE), Dr. B.C. Roy Engineering College,
Fuljhore, Durgapur, West Bengal 713206, India

A. Majumder
e-mail: anandaprova.majumder@bcrec.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_40

literally translates “cover writing.” Stenography is commonly called ‘hidden’ contact. Steganography means hiding the presence of messages in other messages (audio, video, image, and communication). Multiple media such as images, audio, video, etc., are used as cover media by today’s stenography systems since digital images are frequently sent via email or distributed via other Internet communication applications. This is not the same as saving a message’s original material. In simple terms, it is the same as hiding information in other records [1–3].

1.1 Types of Steganography

Various stenographic techniques have been used to achieve protection depending on the type of core object.

Image Steganography: In stenography, covering as an image is referred to as a picture stenography. Typically, this technique uses pixel intensity to hide information [2].

Text Steganography: It consists of hiding data contained in text files. In this way, every text message’s ninth letter conceals sensitive information. In text files, there are several ways to hide data. These techniques are (a) Method based on format, (b) random and statistical method, and (c) linguistic method. Since a text file is not a suitable medium for steganography, it is one of the most challenging methods. It can be used to hide records. Knowledge is hidden in electronic texts and records and is one of the most critical of these technologies (e-documents). Another example is the use of text to hide information on web pages [4, 5].

Audio Steganography: Audio steganography is when audio is used as a carrier to conceal information. Audio steganography is accomplished using digital audio formats like WAVE, MIDI, MPEG AVI, or stenography. (a) Low-bit encoding, (b) phase coding, and (c) spread range are different techniques of audio steganography [4].

Video Steganography: It is a process used in digital video formats to hide some kind of files or information. As a carrier for printed content, video (a series of images) is used. Typically, the value of the discrete cosine transformation (DCT) (e.g., 6.668 to 7) that is used to hide the information that is visible to the human eye in each picture in the video varies. Steganography for video uses H.264, Mp4, MPEG, AVI, or other video formats [4].

Network Steganography: Network protocols such as TCP, UDP, ICMP, and IP are also used for application of steganography. Network protocol stenography is when you take key artifacts and use the protocol as a carrier. In the OSI network layer model, stencils can be retrieved from unused header bits in the TCP/IP field via encrypted channels [2]. Different steganography types are shown in Fig. 1 as follows.

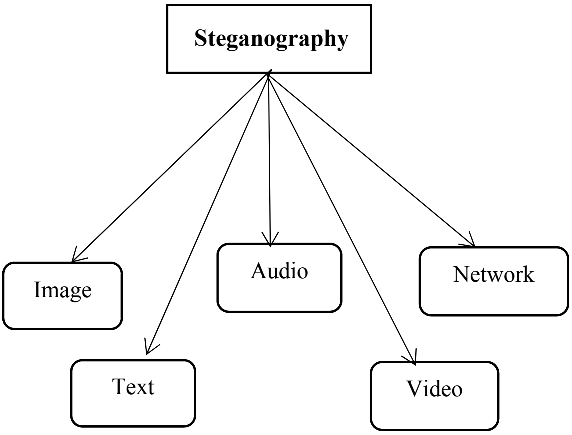


Fig. 1 Types of steganography diagram

2 History

Wax Table: People wrote hidden messages on wood in ancient Greece and then covered it with wax [6].

Shove Head: It was also used back in ancient Greece. The slave's head was shaken and secret messages were written on his skull. Then, the slave's hair was allowed to grow and the secret message came to the recipient after shaving his head again [6].

Invisible Ink: Encrypted messages were written using invisible ink which only appeared when the message-carrying paper was heated. As invisible inks, liquids like milk, vinegar, and fruit juice were used [6].

Morse Code: Hidden messages had been written on the yarn in Morris code. The fabric that the carrier wore was made of wool. Furthermore, at a television conference, Jeremiah Denton turned a blind eye to the Morse code for spelling the word "torture" This prompted the US military to ensure that in North Vietnam, US POWs were tortured [6].

3 Image Steganography Materials and Process

Stego-Key: The stego key is the key used to encrypt information in a coating and then take out the information. It can be a password or a digit provided with the support of a pseudorandom number generator for determining possible embedding locations [6].

Message: It is information that must be concealed in some type of digital media [6].

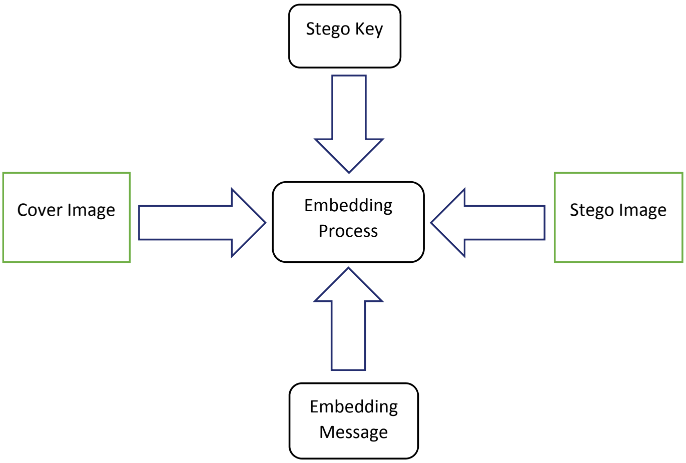


Fig. 2 Image steganography process

Cover Image: It is the medium by which messages like images, audio, video, and other digital media are transmitted [6].

Stego Image: The cover image is the stego image that has a hidden message hidden inside it. It is used to retrieve the secret message at the receiver location [6].

Usually, image steganography is a secret method of knowledge sketching and produces a stego image. This image of stigma was then sent via a well-known channel to the other party, where the other party does not realize that there is a secret message in this image of stigma. With or without a stego key, the secret message after the stego picture can be retrieved until the end of the received message [7]. The process of image steganography is shown in Fig. 2 as follows.

4 Image Steganography Techniques

The techniques of image steganography can be broken down into subsequent domains.

4.1 Spatial Domain Method

LSB and level encoding were used to modify the cover image and hidden data in the spatial domain. To begin, the cover image is decomposed into bit planes. After which, the LSB is replaced with hidden data suit. The most common steganographic technique is LSB substitution. Since it will not impact the value of the original pixel, this substitution definition involves embedding at the lowest weighting bit [8, 9].

Local domain techniques are broadly categorized into:

LSB (Least Significant Bit): To hide data, this method is commonly used. Embedding is done with bits of sensitive data in this form, removing the LSB of image pixels. After embedding, the image obtained is very close to the real picture as the shift in the least significant bit of the image pixel does not make any difference to the image [4, 9].

PVD (Pixel Value Differencing): Two consecutive pixels are chosen in this method to embed the data. Testing the difference between two consecutive pixels and determining whether the two pixels belong to an edge region or a flat area decide the payload [4, 10].

GLM (Gray Level Modification): Previously suggested, the procedure was developed in 2004 by Potdaret al. This method used to map the data (not embedding or hiding it) by altering the grey level values. This technique uses weird and even number concepts for map figures within an image. From the math function, pixels are selected from a given cover image. The gray surface values of these pixels are tested and compared to the bitstream that has to happen mapped in the picture [7].

PCM (Parity Checker Method): This technique made use of the notions of even and odd parity and the parity checker. Even parity refers to the presence of an even number of 1 s in the pixel value, while odd parity refers to the presence of an odd number of 1 s in the pixel value [11].

4.2 Transform Domain Method

To conceal data, the transform domain employs MSB. Because of its picture freedom, this technique is commonly used. Since it focuses on parts that have not changed, such as image editing, cutting, or resizing, transform domain is more powerful than LSB. In both harmful and illegal compression images, transform domain works best. Techniques for transforming domains are [2, 12]:

DFT (Discrete Fourier Transformation)

A discrete Fourier transform is a strictly discrete transform that transforms discrete time indicators into multiple frequencies of multiple frequencies in this technique. These techniques are changing a limited list of evenly spaced patterns of an event, including a list of rules for a complete set of complex sinusoids arranged by their frequency. It can be said that the sampling function is often converted from its real domain to a frequency domain along the line with time or position [13].

DCT (Discrete Cosine Transformation): DCT is one approach to convert signal to initial frequency ingredients. It displays an image as a summary of the sinusoids, different frequencies and dimensions [12].

DWT (Discrete Wavelet Transformation): This is a numerical instrument for figuratively dissolving an image. Useful for this signal processing which is non-stationary. The change is based on small waves of varying frequency and duration, called wavelets. Wavelet transformation is based on small waves of varying frequency and duration. The wavelet transform gives you image frequency as well as spatial clarity [12, 14].

4.3 Distortion Technique

By distorting the signal, this method is used to store hidden data. The encoder process modifies the cover image in a series of steps, and the decoder phase uses a hidden key to decipher the encrypted information back to the real information with the hidden data [13].

During the decoding process, distortion methods necessitate details about the original cover, and the code of conduct functions to measure the discrepancy between the initial cover picture and the new cover picture and the altered cover image in order to recover the encrypted message. The encoder modifies the cover image in a number of ways. As a result, information is known as signal distortion storage. The stego object is generated using this method by making various changes to the cover image. This collection of adjustments is made to match the coded message that must be transmitted. The message is encoded using pseudonyms and pixels chosen at random. The message bit returns ‘1’ if the stego-image at the given message pixel differs from the cover image, otherwise it returns ‘0’ [6].

4.4 Masking and Filtering

The data is hidden by labeling an image in this technique. When watermarks become a part of the picture, this method is advantageous. Rather than hiding the data in the noisy part of the image, it will be embedded where it is more important. Watermarking methods are more integrated into the picture and can be used without fear of destroying it. For the cover picture, the hidden message is more relevant. In 24-bit and greyscale images, this technique is used [4, 13].

5 Analysis of Different Domain and Techniques of Image Steganography

Domain	Techniques	Advantages	Disadvantages
Spatial	LSB	It is used for data insertion purposes It is really easy to execute [2]	Recovered quickly by an unauthorized individual [2]
	PVD	Strong embedding capability and exceptional stego-image imperceptibility [2]	Two connecting pixels divide the cover image into non-overlapping blocks, in each section, in each block (pair), for embedded data and adjustments to different pixels [2, 10]
	GLM	It has a low computational complexity and a large capacity for knowledge concealment [7]	Include binary data. It is necessary to map from one to one The modification of the image causes data loss Embedding capacity is limited [7]
	PCM	For message insertion and retrieval, there is an odd and even parity Recovery of message bits from all locations is permitted [7]	The capability of payloads is low [7]
Transform	DFT	Changes can be applied to the entire image [2]	Such types of approaches are computationally complex [2]
	DCT	Peak signal-to-noise ratio is high (PSNR)[2]	Noticeable secret data artifact [2]
	DWT	It is only useful for binary images [2]	It is not useful for color image support [2, 14]
Masking and filtering	(i) Process LSB is more efficient. The data is unaffected by the compression of the image [15] (ii) Data is hidden in parts of the image that are visible [2]		These techniques are limited to 24 bits and can only be used on grayscale images [2]

6 Conclusion

We reviewed several articles on steganography methods in this research paper. Steganography is an ancient and robust technique used in a variety of applications, including confidential data sharing. When used in conjunction with cryptography, steganography becomes more powerful. The advantages and disadvantages of various image stenography methods are discussed. It is impossible to foresee the best route. Message concealment can be achieved effectively using LSB, according to recent local domain techniques.

Acknowledgements I would like to express my gratitude to my supervisor, my friend, Ms.Sanchita Ghosh (Student of Dr. B C Roy Engineering College), who guided me throughout this paper. I would also like to thanks my professor Mrs. Anandaprova Majumder (Asst. Professor of Dr. B C Roy Engineering College), who advised and helped me to finalized my project.

References

1. Nosrati, M.: An introduction to steganography methods, Aug 2011 (2016)
2. Hussain, M., Hussain, M.: A survey of image steganography techniques (2013)
3. Chanu, Y.J., Tuithung, T., Manglem Singh, K.: A short survey on image steganography and steganalysis techniques. In: 2012 3rd National Conference on Emerging Trends and Applications in Computer Science, Shillong, India, pp. 52–55 (2012). https://doi.org/10.1109/NCETACS.2012.6203297
4. Kour, J.: Steganography techniques—a review paper, vol. 9359, no. 5, pp. 132–135 (2014)
5. Singh, P., Chaudhary, R., Agarwal, A.: A novel approach of text steganography based on null spaces. IOSR J. Comput. Eng. (2012) academia.edu
6. Tiwary, A.: Different image steganography techniques : an overview. Int. J. Comput. Eng. Appl. 0–13 (2019)
7. Hashim, M.M., Rahim, M.S.M., Alwan, A.A.A review and open issues of multifarious image steganography techniques in spatial domain. J. Theor. Appl. Info. Technol. **96**(4), 956–977 (2018)
8. Rakhi, & Gawande, S.: A review on steganography methods. IJAREEIE, **2**(10), 4635–4638 (2013)
9. Hashim, M.M., Rahim, M.S.M., Johi, F.A., Taha, M.S.: Performance evaluation measurement of image steganography techniques with analysis of LSB based on variation image formats. Int. J. Eng. Technol. (2018) uruk.edu.iq
10. Rawat, P., Pandey, A.K., Singh Kushwaha, S.: Advanced image steganographic algorithms and breaking strategies. Int. J. Comput. Appl. (IJCA) concern (2014)
11. Rajkumar, Rishi, R., Batra, S.: A new steganography method for gray level images using parity checker. Int. J. Comput. Appl. **11**(11), 18–24 (2010). https://doi.org/10.5120/1627-2188
12. Sharma, S., Kumar, U.: Review of transform domain techniques for image steganography. Int. J. Sci. Res. ISSN (Online Index Copernicus Value Impact Factor) **4**(5), 194–197 (2015). https://doi.org/10.13140/RG.2.1.4797.1928
13. Arya, A., Soni, S.: A literature review on various recent steganography techniques, pp 143–149 (2018)

14. Nag, A., Biswas, S., Sarkar, D., Sarkar, P.P.: A novel technique for image steganography based on DWT and Huffman encoding. Int. J. Comput. Sci. Secur. (2011)
15. Chandramouli, R, Memon, N.: Analysis of LSB based image steganography techniques. In: Proceedings of 2001 International Conference on Image Processing (Cat. No. 01CH37205), pp. 1019–1022 (2001)

Fault Detection in SPS Using Image Encoding and Deep Learning

P. Hari Prasad, N. S. Jai Aakash, T. Avinash, S. Aravind, M. Ganesan, and R. Lavanya

Abstract Satellite power system (SPS) is considered as the core of the satellite, where the faults occurring here adversely have an impact on the health of the satellite, thereby affecting the mission. This can be avoided by early detection of the faults occurring in the SPS. This work proposes a model to classify the faults present in the SPS using 2-dimensional convolutional neural network (2-D CNN) by encoding the multivariate time series data present in the ADAPT dataset into images. Encoding is done by using the methods such as Markov transition field (MTF), Gramian angular summation field (GASF), recurrence plot (RP), and spectrogram. Promising results were obtained using the GASF and 2-D CNN combination, which have yielded a test accuracy of 87.5%. The precision, recall, F1 score, and AUC score were 0.89, 0.854, 0.865, and 0.94, respectively.

Keywords SPS · ADAPT dataset · MTF · GASF · RP · Spectrogram · 2-D CNN

1 Introduction

Fault detection is a vital component that should be examined for high cost and safety-concerned assets. Satellite power systems (SPS) require early detection of faults that would help in the prevention of any unexpected abnormal events. If these anomalies are unattended, they can have irreversible impacts and can lead to the failure of the whole mission since SPS is the core of the satellite [1]. An SPS primarily consists of solar panels and rechargeable batteries, and abrasion of these can cause anomalies, thereby resulting in subsequent failures of the battery, solar cell, or array. This, in turn, will lead to the loss of communication with the satellite or heating up of the satellite which may even lead to the explosion of the satellite.

The SPS is very complex which makes it difficult to implement mathematical diagnostic models whereas the data-driven diagnosis is highly efficient. It can be implemented with the test data and the telemetry data as it is easily obtained [2]. Usually,

P. Hari Prasad (✉) · N. S. Jai Aakash · T. Avinash · S. Aravind · M. Ganesan (✉) · R. Lavanya
Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: m_ganesan1@cb.amrita.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_41

satellite telemetry data consists of data values called mnemonics [3]. Each mnemonic is represented as a time-imprinted discrete series classified into two types—univariate and multivariate [4]. In univariate time series, a single parameter depends on the time scale, whereas in the multivariate time series, multiple parameters depend on the time scale. Satellite telemetry data falls under multivariate time series data. It has affluent and intricate details that can be used to cognize the relationship between each parameter.

The model-based diagnosis of faults in SPS has been widely implemented. Mengshoel et al. [5] have proposed a method that would overcome the limitations of modeling and real-time reasoning by introducing autogeneration of Bayesian networks, and it is then compiled into arithmetic circuits. Feiyi and Jinsong [6] have compared all the merits and drawbacks of various methods. He states that it is hard to get the probabilities of Bayesian networks; he also has obtained a false positive rate of 25% and false-negative rate of 7% in the method of Testability Engineering and Maintenance System (TEAMS), and in the other methods, either the model is difficult to understand or takes a long time for isolation. Ocak and Loparo [7] have detailed a new bearing fault detection and diagnosis scheme that gave out better accuracies but the window size must be large.

The homogeneities between the time series data can be used to detect the faults with methods such as dynamic time warping (DTW). DTW has high intricacy and can produce pathological results, which in turn leads to meaningless calibration, where a single instance of time series overlaps onto a bigger section of other time series. The extraction of features from the time series data such as discrete Fourier transform (DFT) and wavelet transform is also used. The other form of strategy is machine learning (ML) and 1-dimensional convolutional neural network (1-D CNN). ML needs a lot of good quality dataset to get trained and takes a long time for highly complex problems. 1-D CNN does not encode the position and orientation of an object for predictions. They are likely to lose all their internal data regarding the position and orientation, and they may not be able to handle the information as they route all the information to the same neurons.

The dataset is utilized from advanced diagnostic and prognostic testbed (ADAPT) developed by NASA Ames research center. The data is acquired from the electrical power system, which simulates the functions of an SPS. Several experiments were run on the testbed with different configurations, and the sensor data were also recorded. Faults were induced by physical and software means into the power system, and data were recorded with respective fault information. The data is a multivariate time series (MTS) data sampled at a rate of 2 Hz.

The complexity in diagnosing and localizing the fault in an SPS arises as the signals from the sensors are in the form of MTS. This limitation can be overcome by the advantageous models that are present in deep learning. Lv et al. [8] have stated in their work that deep learning has the highest average fault classification rates compared to sparse representation, support vector machine (SVM) [9], random forest, and structure SVM. This shows that rather than the prevalent methods, deep learning gives better accuracy and profoundly analyses the data.

Deep learning is a subconcept of ML which is inspired by the activities of the human brain. Deep learning algorithms try to extract the high-level features from the given dataset, and it is also suitable for unstructured data. These characteristics are helpful to analyze faulty data. The conventional preprocessing methods such as fast Fourier transform (FFT), sliding window, and wavelet transform work with univariate time series. Whereas, MTS requires exceptional preprocessing methods as the data comprises abundant circumstantial information. To overcome this, in this paper, we have implemented various image encoding techniques.

Yang et al. [10] have given out a comparison of various preprocessing methods such as DTW, combDTW, STKG-SVM-K3, and STKG-IF-NB-SVM+M and has obtained an error rate of 2.01%, 2.01%, 1.23%, and 2.23%, respectively. But whereas appending GASF and GADF have given out the value of 1.06 and 1.57, respectively, with the wafer dataset. This states that encoding the MTS into images reduces the error rates and gives out a better accuracy. By aggregating all the MTS data into a single image helps us produce information from the noise received while collecting the data and also helps us perceive the correlation between the variables present. Encoding MTS into images is an affluent representation of data, so understanding co-occurrence and latent states of data becomes easier. Since encoding is largely motivated by polar coordinates transformations, we can easily distinguish the information from noise and take dominance by using the relations rather than switching the space. The methods of Markov transition field (MTF), Gramian angular summation field (GASF), recurrence plot (RP), and spectrogram have been used.

2 Methodology

The proposed method for fault classification in SPS uses 2-dimensional convolutional neural network (2-D-CNN) to classify the ADAPT data into faulty or normal conditions. However, before the classification process, the data is windowed using a nonoverlapping window and encoded into images using preprocessing methods like MTF, GASF, RP, and spectrogram. The concepts of the preprocessing methods and CNN are explained in the following subsections (Fig. 1).

2.1 Spectrogram

Spectrograms are two-dimensional graphs where frequency and time are the two dimensions, where the third dimension is represented by colors. It is a visual way of representing the signal strength over time at various frequencies. It is an exemplary representation of signals by which the energy of the signals can be analyzed with short-time Fourier transform (STFT). STFT which is also termed as time-dependent Fourier transform is used to analyze nonstationary signals with the help of steady-state analysis, assuming short-term stationarity [11]. x_n is a discrete time-domain

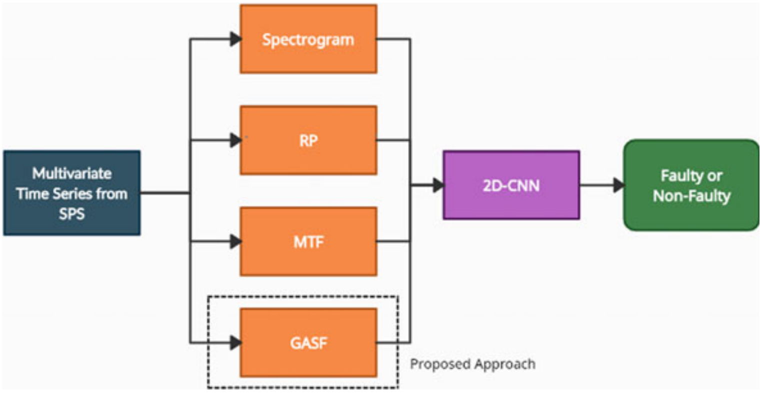


Fig. 1 Workflow

signal that is sampled where $n = 0, 1, 2, \ldots, n - 1$, where n is the length of the signal. Let $x(n)$ be the discrete time-domain signal that is used for sampling. Here, n can take the values 0, 1, 2, … $N - 1$ where n is the sampling point number in the time domain, and N denotes the length of the signal. The framed signal $x(n)$ is represented as $x_n(m)$, $n = 0, 1, 2, \ldots, N - 1$ where n denotes the frame number, m denotes the time sequence number of the frame synchronization, and N denotes the frame length. The equation of STFT is denoted as in Eq. (1).

$$X(n, k) = \sum_{m=0}^{N-1} w(m)x(m)\mathrm{e}^{-f\frac{2\pi km}{N}} \tag{1}$$

where $w(m)$ is the hamming window and is given in Eq. (2)

$$w(m) = 0.54 - 0.46\cos\frac{\pi m}{N} \tag{2}$$

To find the result of the estimation of the short-term amplitude spectrum, we use the DFT $|X(n, k)|$. The power spectral function is given in Eq. (3)

$$p(n, k) = |X(n, k)|^2 = (X(n, k)) \times (\mathrm{conj}(X(n, k))) \tag{3}$$

2.2 Gramian Angular Field (GAF)

Gramian Angular Field [12] encodes the MTS data into images by rendering MTS into a polar coordinates-based matrix. This characteristic of the Gram matrix helps us preserve temporal dependency as the time dimension is automatically encoded in the matrix from top left to bottom right as time increases. First, we normalize the MTS data which is of form $X = \{x_1, x_2, x_3 \ldots x_n\}$ by rescaling the values such that it falls under the interval of $[-1, 1]$. This can be done by using Eq. (4).

$$\tilde{x}_i = \frac{(x_i - \max(X)) + (x_i - \min(X))}{\max(X) - \min(X)} \tag{4}$$

This rescaled value is converted to polar coordinates. The value of the time series is computed as the angle and its corresponding timestamp is computed as the radius. This conversion is done using Eq. (5)

$$\begin{cases} \emptyset = \arccos(\tilde{x}_i), \ -1 \le \tilde{x}_i \le 1, \ \tilde{x}_i \in \tilde{X} \\ r = \frac{t_i}{N}, \qquad\qquad t_i \in \mathbb{N} \end{cases} \tag{5}$$

where t_i is a timestamp and N is a constant.

GAF can encode the time series in two different ways—GASF and GADF. GASF uses cosine functions as shown in Eq. (6), whereas GADF uses the sine function as shown in Eq. (7). Using these trigonometric functions allows us to maneuver the temporal correlation between different time intervals.

$$\text{GASF} = \begin{bmatrix} \cos(\emptyset_1 + \emptyset_1) & \ldots & \cos(\emptyset_1 + \emptyset_n) \\ \cos(\emptyset_2 + \emptyset_1) & \ldots & \cos(\emptyset_2 + \emptyset_n) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \cos(\emptyset_n + \emptyset_1) & \ldots & \cos(\emptyset_n + \emptyset_n) \end{bmatrix} \tag{6}$$

$$\text{GADF} = \begin{bmatrix} \sin(\emptyset_1 + \emptyset_1) & \ldots & \sin(\emptyset_1 + \emptyset_n) \\ \sin(\emptyset_2 + \emptyset_1) & \ldots & \sin(\emptyset_2 + \emptyset_n) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \sin(\emptyset_n + \emptyset_1) & \ldots & \sin(\emptyset_n + \emptyset_n) \end{bmatrix} \tag{7}$$

The GASF matrix can be simplified as shown in Eq. (8).

$$\text{GASF} = \widetilde{x' \cdot \tilde{x}} - \sqrt{1 - \tilde{x}^2}' \cdot \sqrt{1 - \tilde{x}^2} \tag{8}$$

And the simplified equation of the GADF matrix is shown in Eq. (9)

$$\text{GADF} = \sqrt{1 - \tilde{x}^2}' \cdot \tilde{x} - \tilde{x}' \cdot \sqrt{1 - \tilde{x}^2} \tag{9}$$

As we can see in the matrix, it is symmetrical by the main diagonal, and it stores the original values or the angular information due to which it is easy to compute the temporal dependency, as $\cos(\theta)$ is monotonic when $\theta \epsilon [0, \pi]$. However, GAF is large due to the augmentation $n \rightarrow n^2$. This can be overcome by implementing piecewise aggregation approximation [13], where the data of n dimensions is reduced to N dimensions by dividing the data into N equisized frames, then this transformation produces a piecewise steady approximation of the original sequence.

2.3 Markov Transition Field (MTF)

In MTF, the time series of the signal X is divided into N quantile bins. For every instance x_i in the series, a respective quantile n_i is assigned. The Markov transition probabilities tend to protect the information present in the time domain sequentially [12]. So, the Markov transition probability M_{ij} is the transition probability between two quantile bins n_i and n_j. With this, we build the $M \times M$ Markov transition matrix as shown in the matrix below (10). So Q_{ij} in the matrix denotes the frequency at which the quantile n_j is followed by n_i. This allocates the self-transition probability in the diagonal. Therefore, it is non-symmetrical and preserves the data in the temporal range. Unlike GAF, it cannot revert to original data.

$$M_{ij} = \begin{bmatrix} q_{ij}|x(1) \in n_i, x(1) \in n_j & \dots & q_{ij}|x(1) \in n_i, x(m) \in n_j \\ q_{ij}|x(2) \in n_i, x(1) \in n_j & \dots & q_{ij}|x(2) \in n_i, x(m) \in n_j \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ q_{ij}|x(m) \in n_i, x(1) \in n_j & \dots & q_{ij}|x(m) \in n_i, x(m) \in n_j \end{bmatrix} \tag{10}$$

2.4 Recurrence Plot (RP)

RP, through phase space, helps in visualizing the trajectory's nature and recurrence. The trajectory is represented in an abstract mathematical space [14]. RP is calculated based on the recurrence matrix. The recurrence matrix is defined as shown in (11)

$$R_{i,j} = \begin{cases} 1 : \vec{x}_i \approx \vec{x}_j, \\ 0 : \vec{x}_i \not\approx \vec{x}_j, \end{cases} \quad i, j = 1, \dots, N \tag{11}$$

These systems are developed by a series of these vectors.

With this, the signals are encoded into an image. Here, n is the number of states that are considered and $x_i = x_j$ implies equality till an error ε. At two different instances i and j, the recurrence matrix analyzes the states of a system. This is designated by one in the matrix if the states are alike, i.e., $R_{i,j} = 1$, whereas if the states are dissimilar, it is designated as 0, i.e., $R_{i,j} = 0$. With these characteristics of RP, we can easily study the nonlinear parameters of a system and deriving correlation and mutual information become facile. The extraction of efficient information from limited and non-static data is very simple with RP where the other methods fail.

2.5 Convolutional Neural Network (CNN)

The used architecture of CNN is shown in Fig. 2. The model consists of 3 blocks of Convolution (Conv2D) and Maxpooling2D layers followed by a Flatten and 2 Dense layers. 2D Convolutional layers take a three-dimensional input and pass a filter over the input. The pixel value in the current filter is determined, and the dot product is estimated, and then, it is passed onto the next layer. Maxpooling2D layer is used to reduce the size of the tensor; it outputs the maximum value in the determined window. Maxpooling2D layer helps in reducing the amount of computation [15]. Rectified linear unit (ReLU) activation function is used for the convolution and a dense layer, and the output layer uses a sigmoid activation function [16]. ReLU activation function is a piecewise linear function which outputs the input if the input is positive and zero if the input is negative. This activation function is used as it is computationally inexpensive and achieved better performance. Whereas, the sigmoid activation function always outputs a value between [0, 1]; this nature can be used to determine the prediction using a threshold value, therefore it is used in the output layer. Additionally, a Dropout layer and l2 regularizers are used to prevent overfitting. The hyperparameters like the number of layers, the number of neurons, and epochs were determined experimentally for optimizing the performance of the model.

3 Results

Each instance of the time series was assigned to specific quantile bins and the Markov transition probabilities were computed, and with this, the Markov transition matrix was formed as shown in (10). The attributes present in the matrix were encoded into an image and then was fed into 2-D CNN. The data window considered here was 20. The training accuracy was found to be 87.50%, and the testing accuracy was found to be 83.93%. The precision, recall, F1 score, and AUC score were 0.842, 0.828, 0.832, and 0.87, respectively (Tables 1 and 2).

The time–frequency analysis of the MTS data was done using STFT. The MTS was divided into consecutive frames, and then, STFT analysis was done using Eq. (1). The sampling rate was assigned to be 2 Hz, this was done on a non-overlapping data window of 10. The training accuracy was found to be 76.79%, and the testing accuracy was found to be 73.21%. The precision, recall, F1 score, and AUC score were 0.840, 0.687, 0.677, and 0.73, respectively.

Fig. 2 CNN architecture

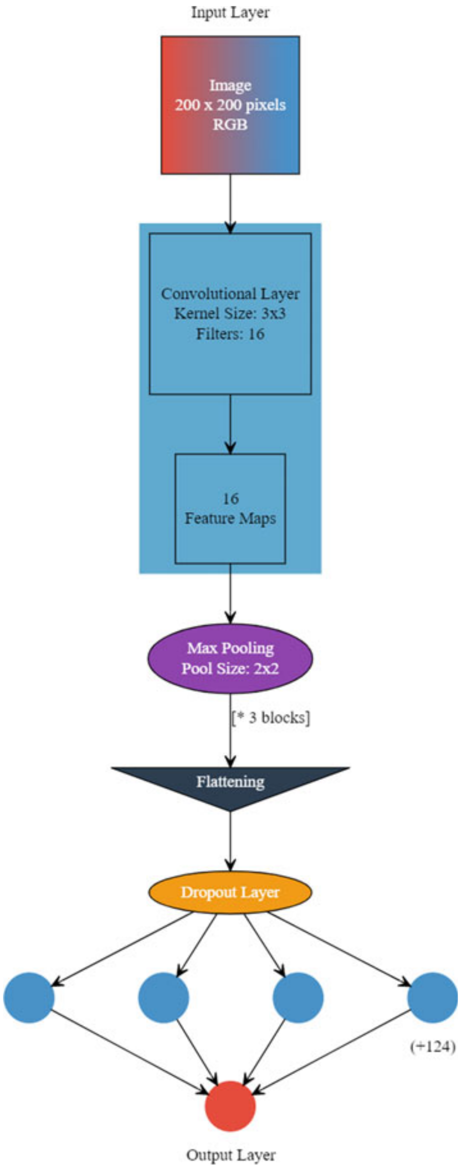


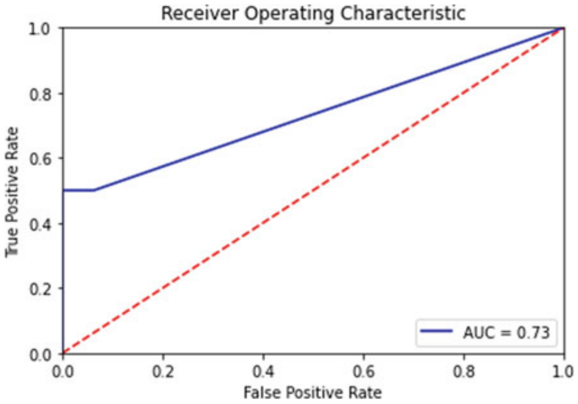
Table 1 Comparison of accuracy

Method	Train accuracy (%)	Test accuracy (%)
MTF	87.50	83.93
GASF	88.39	87.50
RP	89.29	87
SPEC	76.79	73.21

Table 2 Model evaluation

Method	Precision	Recall	F1-score
MTF	0.842	0.828	0.832
GASF	0.893	0.854	0.865
RP	0.912	0.847	0.862
SPEC	0.840	0.687	0.677

Fig. 3 ROC for spectrogram



The MTS data when encoded into images produces unique images for every univariate time series data which results in the accumulation of multiple images, so it is necessary to bind these images vertically and then transfer them into 2-D CNN. The considered non-overlapping data window is 20.

GASF gave a training and testing accuracy of 88.39% and 85.50%, respectively. The classification metrics obtained in GASF were precision—0.893, recall—0.854, F1 score—0.865, and AUC score—0.94.

The number of instances a trajectory visits the same space was calculated using Eq. (10). The non-overlapping data window used here is 20, and the image depiction of the changing nature of the system was obtained. The achieved training accuracy is 89.29% and testing accuracy of 87% was obtained. The classification metrics-precision, recall, F1 score, and AUC score were 0.912, 0.847, 0.862, and 0.93, respectively (Figs. 3, 4, 5, 6, 7, 8, 9 and 10).

4 Conclusion

In the proposed framework, the MTS data was encoded into images using MTF, GASF, RP, and Spectrogram. The encoded images were fed into 2-D CNN and the fault classification was done, and then, the accuracies were predicted. RP and the GASF obtained a result of 87% and 87.5%, respectively, but GASF performed well

Fig. 4 ROC for MTF

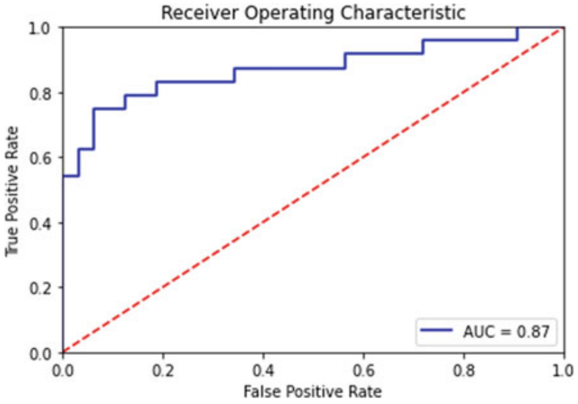


Fig. 5 ROC for GASF

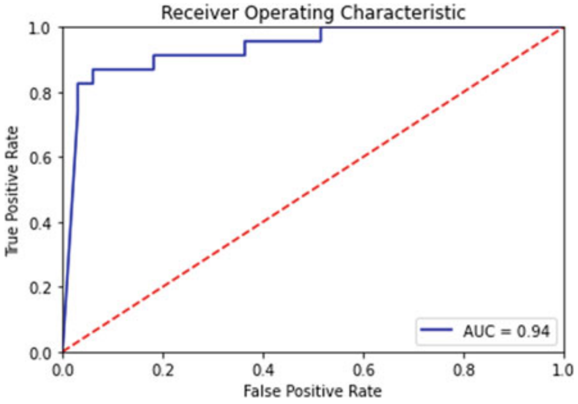


Fig. 6 ROC for RP

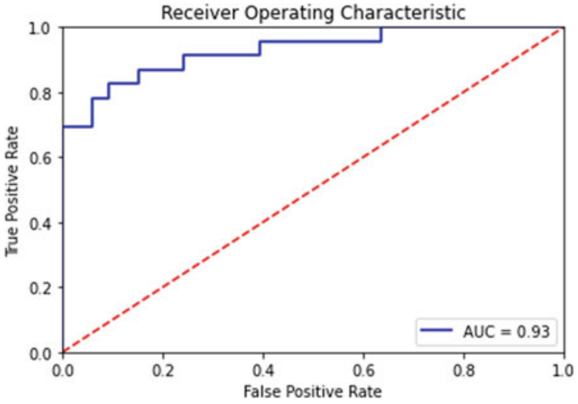


Fig. 7 Confusion matrix for spectrogram

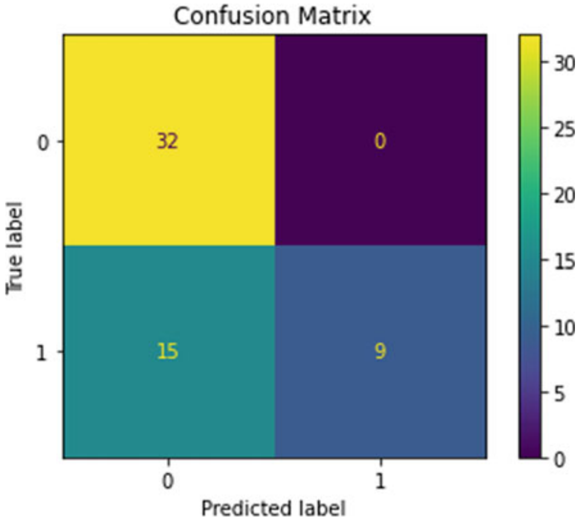
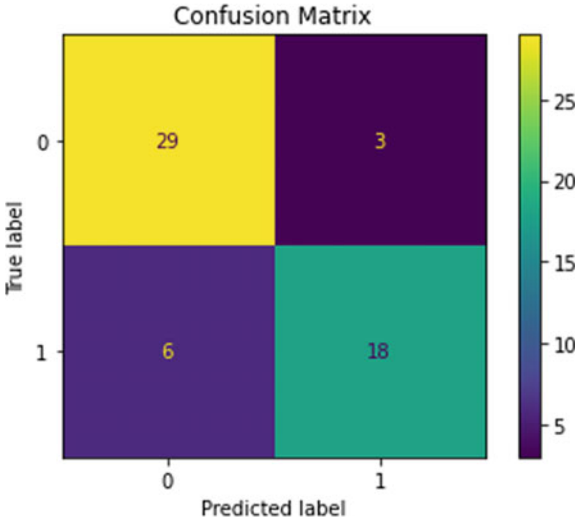


Fig. 8 Confusion matrix for MTF



on all the window sizes and has a better AUC score compared to RP. This shows that GASF outperforms the other models. The preprocessing done has proved to be unique and has helped in increasing the accuracy than the conventional preprocessing methods. Though the proposed model works fine for binary and multiclass classification, it is difficult to classify the occurrence of simultaneous faults with image encoding techniques.

Further scope of the work would be to perform other imaging techniques. Multiclass classification and fault localization can be implemented. Proctoring and health

Fig. 9 Confusion matrix for GASF

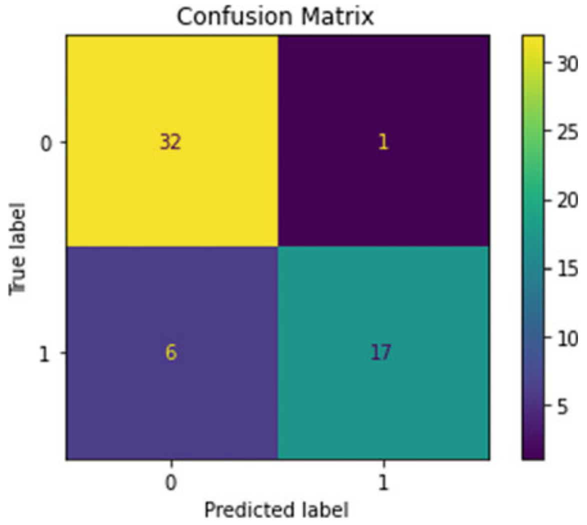
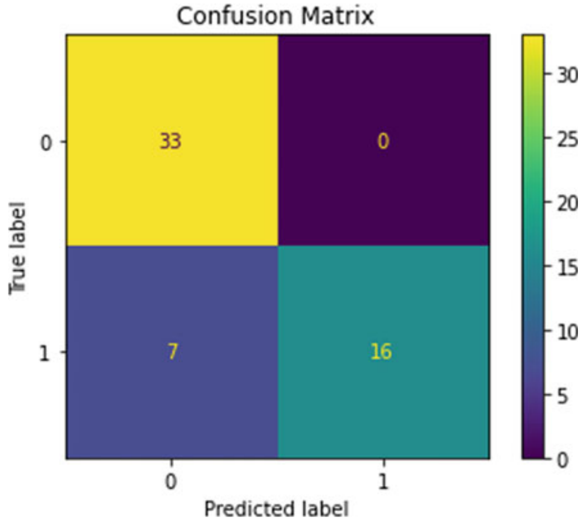


Fig. 10 Confusion matrix for RP



management of the SPS can be done by monitoring the health condition and by proctoring the faults occurring in the SPS.

References

1. Landis, G.A., Bailey, S.G., Tischler, R.: Causes of power-related satellite failures. In: 2006 IEEE 4th World Conference on Photovoltaic Energy Conference, vol. 2, pp. 1943–1945. IEEE (2006)
2. Suo, M., Zhu, B., An, R., Sun, H., Xu, S., Yu, Z.: Data-driven fault diagnosis of SPS using fuzzy Bayes risk and SVM. Aerosp. Sci. Technol. **84**, 1092–1105 (2019)
3. Crowley, N.L., Apodaca, V.: Analysis of satellite telemetry data. In: 1997 IEEE Aerospace Conference, vol. 4, pp. 57–67. IEEE (1997)
4. Ganesan, M., Lavanya, R., Nirmala Devi, M.: Fault detection in SPS using convolutional neural network. Telecommun. Syst. 1–7 (2020)
5. Mengshoel, O.J., Darwiche, A., Cascio, K., Chavira, M., Poll, S., Uckun, N.S.: Diagnosing faults in electrical power systems of spacecraft and aircraft. In: AAAI, pp. 1699–1705 (2008)
6. Feiyi, R., Jinsong, Y.: Fault diagnosis methods for advanced diagnostics and prognostics testbed (ADAPT): a review. In: 2015 12th IEEE International Conference on Electronic Measurement and Instruments (ICEMI), vol. 1, pp. 175–180. IEEE (2015)
7. Ocak, H., Loparo, K.A.: A new bearing fault detection and diagnosis scheme based on hidden Markov modeling of vibration signals. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), vol. 5, pp. 3141–3144. IEEE (2001)
8. Lv, F., Wen, C., Bao, Z., Liu, M.: Fault diagnosis based on deep learning. In: 2016 American Control Conference (ACC), pp. 6851–6856. IEEE (2016)
9. Manohar, N., Sharath Kumar, Y.H., Rani, R., Hemantha Kumar, G.: Convolutional neural network with SVM for classification of animal images. In: Emerging Research in Electronics, Computer Science and Technology, pp. 527–537. Springer, Singapore (2019)
10. Yang, C.-L., Yang, C.-Y., Chen, Z.-X., Lo, N.-W.: Multivariate time series data transformation for convolutional neural network. In: 2019 IEEE/SICE International Symposium on System Integration (SII), pp. 188–192. IEEE (2019)
11. Yu, W., Huang, S., Xiao, W.: Fault diagnosis based on an approach combining a spectrogram and a convolutional neural network with application to a wind turbine system. Energies **11**(10), 2561 (2018)
12. Wang, Z., Oates, T.: Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, vol. 1 (2015)
13. Keogh, E.J., Pazzani, M.J.: Scaling up dynamic time warping for data mining applications. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 285–289 (2000)
14. Marwan, N., Carmen Romano, M., Thiel, M., Kurths, J.: Recurrence plots for the analysis of complex systems. Phys. Rep. **438**(5–6), 237–329 (2007)
15. Saiharsha, B., Diwakar, B., Karthika, R., Ganesan, M.: Evaluating performance of deep learning architectures for image classification. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 917–922. IEEE (2020)
16. Vinayan, V., Anand Kumar, M., Soman, K.P.: Capturing discriminative attributes using convolution neural network over ConceptNet numberbatch embedding. In: Emerging Research in Electronics, Computer Science and Technology, pp. 793–802. Springer, Singapore (2019)

A Comparative Study of Information Retrieval Models for Short Document Summaries

Digvijay Desai, Aniruddha Ghadge, Roshan Wazare, and Jayshree Bagade

Abstract The judicial system has evolved tremendously over the past years. Thousands of cases are registered daily and stored in the form on documents which are used by lawyers whenever required. Lawyers are important stakeholders in judicial system and constantly study multiple cases during their work. Manual retrieval of this information from a collection is very difficult. This is where the information retrieval system comes in picture. This article is a brief comparison of various information retrieval models which are currently being used. It includes the Boolean model, TF-IDF model, vector space model, Okapi BM25 model and fuzzy search models. Each of these models is tested on three datasets, and their results were noted. The experimental results unfold that the Okapi BM25 model outperformed the other models in the case study. The results also show that document pre-processing plays an important role in the effectiveness of the query-document matching.

Keywords Information retrieval · TF-IDF · Okapi BM25 · Vector space model · Fuzzy set theory

1 Introduction

After the increase in the use of the internet over the last decade, an incredible amount of information is available and easily accessible through the internet. Information retrieval (IR) systems are necessary in order to manage and retrieve the required information effectively. An information retrieval system has the ability to store, represent, access and organize information items. A set of keywords known as queries are required to search the information required by the user. Queries are what the people are searching for in the system. These keywords give a brief description of the information.

D. Desai (✉) · A. Ghadge · R. Wazare · J. Bagade
Vishwakarma Institute of Information Technology, Kondwa, Bk, Pune, Maharashtra 411048, India

J. Bagade
e-mail: jayashree.bagade@viit.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_42

Indian judicial system dates back to colonial times from those days till date the system has evolved tremendously as of now thousands of cases are registered on a daily basis. Lawyers are the most important stakeholders of this judicial system, and they constantly need to study a high number of cases in their daily life. Unavailability of a system which is scalable, easily accessible and has a fast search mechanism from a corpus of documents has always been the need. With the advent of technology, especially in the field of AI and ML several judicial systems are switching to technology-enabled solutions for easing their process and making it fast. This case study intends to develop such a system for which it is necessary to make a choice of an IR model which would best serve the purpose. IR retrieval models are compared find out the best of them which shows overall good performance for different datasets available.

In order to design such a system, a benchmark technique needs to be identified. In this paper, four algorithms, namely Boolean, TF-IDF, vector space, okapi bm25 model are compared and evaluated to find out which model provides better results. The models are tested on three different datasets.

2 Related Works

In this case study, it is inferred that most of the existing approaches used are TF-IDF, okapi bm25, vector space, fuzzy search. Below is a list of some of the approaches for information retrieval systems.

Aguilar et al. [1] proposed comparative system for feature extraction from educational contents in which they used different techniques BM25, LSA, Doc2Vec and LDA on their dataset, and it concluded by saying that each one showed better behaviour for each type of content provided to them, and thus, there wasn't any specific model which completely outperformed the other. Dai et al. [2] proposed five models that were compared in their study, namely TF-IDF, cosine similarity, okapi bm25, KL-divergence retrieval model, indri model.

From which their study showed that the indri model which is a statistical language model outperformed vector space modelling approach (cosine similarity) in which they used *Dirichlet prior's* method for smoothing SLM.

Svore and Burges [3] showed a machine learning approach for improving the bm25 model. In which they have proposed a new model, namely $LambdaBM25_f$ model.

The model optimizes directly for the chosen target IR evaluation measure and avoids the necessity of parameter tuning, yielding a significantly faster approach.

Jimenez et al. [4] proposed improving TF-IDF factors in bm25 by using collection term frequency. In which BM25-CTF obtained improvements in all measures, particularly in mean average precision (mAP) and geometric mean average precision (gMAP). Kural et al. [5] used a clustering approach for information retrieval systems and concluded that there wasn't any difference between generic models and

Table 1 Comparison with previous works

	a	b	c	d	e
Aguilar et al. [1]	✔	✔	–	✔	✔
Dai et al. [2]	✔	–	✔	✔	✔
Svore and Burges [3]	–	✔	–	✔	–
Jimenez et al. [4]	✔	✔	–	✔	–
Kural et al. [5]	–	–	–	–	–
Rekha [6]	✔	✔	✔	✔	–
Bhatia et al. [7]	–	–	–	–	✔
Proposed	✔	✔	✔	✔	✔

clustering-based generic models. This paper inferred that clustering seems more efficient as a rejection aid than a selection aid. Another paper which shows a clustered-based model for the IR system proposed by Rekha [6]. In which they have shown how to classify retrieved documents which help to regroup the relevant ones. They have said that this increases the effectiveness of retrieval by providing to users at least one cluster with a precision higher than the one obtained without classification.

Instant fuzzy search using probabilistic-correlation based ranking, proposed by Rekha [6]. In which they have experimented with instant fuzzy keyword search, instant fuzzy multi-keyword search, probabilistic-correlation based relevance ranking and by using the first two algorithms they have computed their answers based on answers of previous queries and also determined an answer check by using the third algorithm, namely probabilistic-correlation based relevance ranking. Bhatia et al. [7] conducted a survey on information retrieval models/concepts. In which they have studied the Boolean model, ranking algorithms, vector space model and after that proposed the need of personalized information retrieval systems based on user needs and also discussed the concept of evolutionary algorithms.

Table 1 shows comparison among research papers and our work based on these criteria

(a) Do they consider different datasets
(b) Are all the dataset general in nature (not corresponding to a particular domain only)
(c) Is the system solely based on IR models
(d) No ontology must be defined with respect to a model to improve search results
(e) Studied more than one model.

3 Methodology

Representation, storage, organization and access to information items are the basic pillars of the information retrieval system. As shown in Fig. 1, in the first step, a corpus is stored on a remote pc or a database the documents are further pre-processed. In the

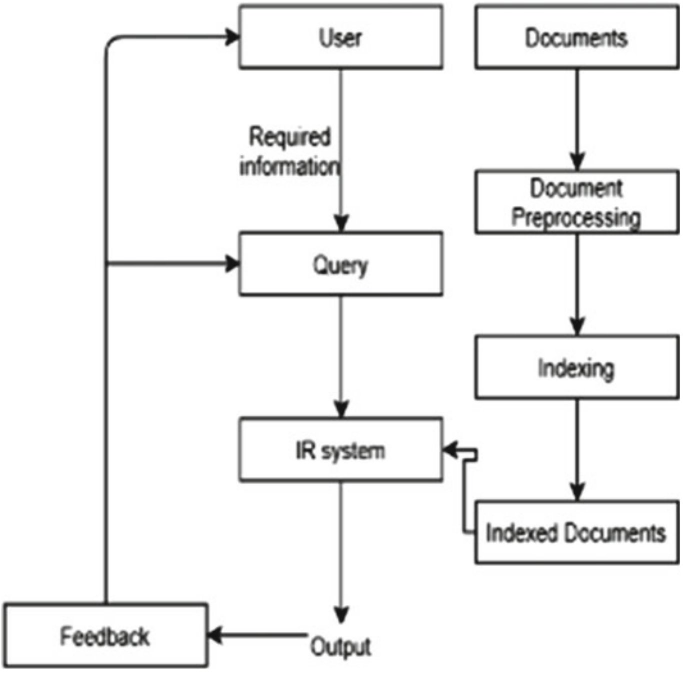


Fig. 1 General structure of IR models

next step, indexing is carried out on the documents based on their document serial number. Indexing is used to store the documents in a specified format which is used for retrieval of the documents. These indexed documents are further passed to the IR system. In an IR system, after pre-processing, these words are stored in the form of an inverted index. Now comes the part of users actually interacting with the system in which user enters queries into the system after which the queries are pre-processed. Finally, query matches are performed and the retrieval algorithm ranks the document based on user query. The top '*n*' number of documents is shown by the IR system. This process as a whole makes the information retrieval model.

3.1 Information Retrieval System Algorithms

3.1.1 Boolean Algorithm

This is the basic model of information retrieval systems. The Boolean retrieval model uses logical functions for processing the query to retrieve the required data from the database. This is the easiest approach for finding a document in a database. This creates a term-document matrix which stores the terms and represents their presence inside the document with the help of 1 (if present)/0 (if not present). The model is basically based on Boolean algebra and set theory. The terms are retrieved by using a combination of logical AND (Union) and logical OR (Intersection) operators.

Consider a query issued by user having two words 'A' and 'B' then:

A AND B: represents the documents that contain both A and B
A OR B: represents the documents that contain either A or B
NOT A: represents the documents that do not contain A.

3.1.2 TF-IDF Model

TF-IDF (term frequency-inverse document frequency) is a numerical statistic which tells how important a certain word is to a document in a collection of corpuses. It is used as a weighting model in IR systems. The principle of TF-IDF model is that the more often a term occurs in a document, the more representative it is of this document. All the terms in a query as well as document are weighted by the heuristic TF*IDF weighting formula.

The term frequency (***tf***$_{i,j}$) is computed with respect to the *i*th term and *j*th document:

$$\mathrm{tf}_{i,j} = \frac{n_{i.j}}{\sum_{j=0}^{x} n_{kj}} \tag{1}$$

where $n_{i,j}$ are the occurrences of the *i*th term in the *j*th document, *x* is total number documents and n_{jk} are the occurrences of *k*th terms in the *j*th document.

The inverse document frequency (***idf***$_i$) is defined as the number of documents in a collection that contain a term *t*. The idf gives document level statistics rather than a collection wide one.

Inverse document frequency (***idf***$_i$) takes into consideration the *i*th terms and all the documents in the collection and is represented as,

$$\mathrm{idf}_i = \log\frac{|D|}{|\mathrm{df}_i|} \tag{2}$$

where *D* is the total number of documents in collection and df_i the document frequency of the *i*th term in the document.

3.1.3 Okapi BM25 Model

This model is based on the probability of retrieving the relevant and irrelevant data is matched. This is known as the probability theory of data.

BM25 is a ranking function which is used to estimate relevance of documents to a given search query. Okapi best match 25 (BM25) was developed as a way of building a probabilistic model which is sensitive to term frequency and size of the document.

In the current TF-IDF there's no problem with idf in practice, but if there are invariable sizes of each document inside the corpus then tf creates a problem. So, for tuning the tf best match 25 provides k and b parameters and also changes tf part. This TF-IDF style formula has consistently outperformed other formulas in standard benchmarks over the years:

$$\mathrm{BM}_{25} = \mathrm{tf}^{*} \cdot \log_2\left(\frac{N}{\mathrm{df}}\right) \tag{3}$$

where

$$\mathrm{tf}^{*} = \frac{\mathrm{tf} \cdot (k+2)}{\left(k \cdot \left(1 - b + b \cdot \left(\frac{\mathrm{DL}}{\mathrm{AVDL}}\right)\right) + \mathrm{tf}\right)} \tag{4}$$

where tf = term frequency, DL = document length, AVDL = average document length k and b are tuning parameters.

The general settings for BM25 are: $b = 0.75$ and $k = 1.75$.

3.1.4 Fuzzy Set Model

Definition: A fuzzy set **A** of a universe of discourse **U** is characterized by a membership function $\mu_{\mathrm{A}}{:}U \rightarrow [0, 1]$ which associates each element u of U in a member $\mu_{\mathrm{A}}(u)$ in the interval [0, 1].

Consider a query term that defines a fuzzy set and that each document has some degree of membership with the set (in the interval [0, 1]). This interpretation of the retrieval process is termed as fuzzy theory. Moreover, document queries are represented through sets of keywords which are only partially related to the real semantic contents of the documents and queries.

Common problem with IR models is that the query-document matching is approximate matching which induces a sense of vagueness in the system, which can be dealt with using a fuzzy approach. In which each term is associated with a fuzzy set and each document has a degree of membership in this fuzzy set. The key idea is to introduce a degree of membership associated with the elements of a set which varies from 0 to 1.

3.1.5 Vector Space Model

Vector space model or term vector model comes under the type of algebraic model. In this each document and query are represented in vector forms.

$$d_i = \left(t_{1,i}, t_{2,i}, t_{3,i}, t_{4,i}, t_{5,i}, t_{6,i}, t_{7,i}\right) \tag{5}$$

$$q = \left(t_{1,q}, t_{2,q}, t_{3,q}, t_{4,q}, t_{5,q}, t_{6,q}, t_{7,q}\right) \tag{6}$$

where d_i be the ith document and q be the respected query entered by user. If there exists any term inside the document, it value in vector is nonzero and weighting schema used in this model is TF-IDF.

E.g.: Consider a word to be the term then the vector's dimensionality is equal to the number of words.

In Fig. 2, a graph among two documents and a single query is shown. The angle between q vector and doc2 vector is termed as theta, whereas alpha is an angle between doc1 and q.

Here, the relevance ranking of documents is calculated by comparing derivation of angle between each document vector and required query vector where query is represented with the same dimension as that of the document vector.

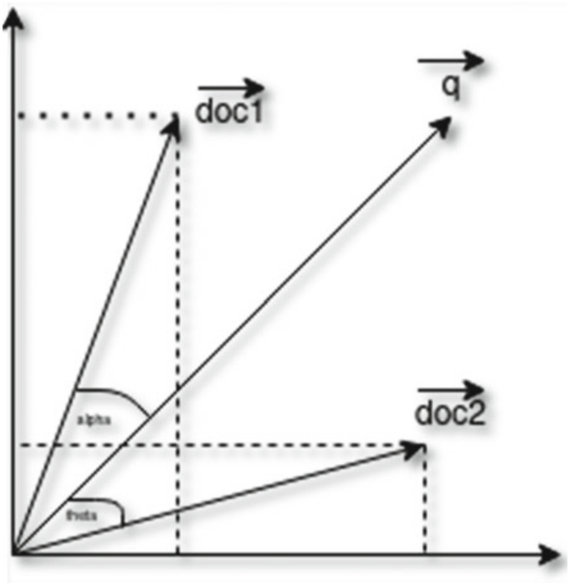
$$\cos\theta = \frac{\mathrm{doc}_2 \cdot q}{\|\mathrm{doc}_2\| * \|q\|} \tag{7}$$

where $\mathrm{doc}_2 \cdot q$ is the dot product of both the vectors as shown in Fig. 2, $||q||$ is the norm of vector q and $||\mathrm{doc}_2||$ is the norm of vector doc_2.

The norm of any vector v (say) is calculated as follows

$$\|v\| = \sqrt{\sum_{j=1}^{n} v_j^2} \tag{8}$$

Fig. 2 Graph among two documents



So, the final formula derived using (5), (6), (7) and (8) is:

$$\cos\theta = \frac{\text{doc}_2 \cdot q}{\|\text{doc}_2\| * \|q\|} = \frac{\sum_{p=1}^{n} t_{p,i} * t_{p,q}}{\sqrt{\sum_{p=1}^{n} t_{p,j}^2}\sqrt{\sum_{p=1}^{n} t_{p,q}^2}} \tag{9}$$

4 Building an IR Model

See Fig. 3.

4.1 *Pre-processing*

The corpus (documents) needs to be pre-processed for which it goes through a series of steps as shown in Fig. 3. Firstly, tokenization (task of converting raw text into word tokens) of text is carried on the corpus then it is converted into lowercase letters followed by a series of punctuation, then stop words removal, and finally, the text is lemmatized.

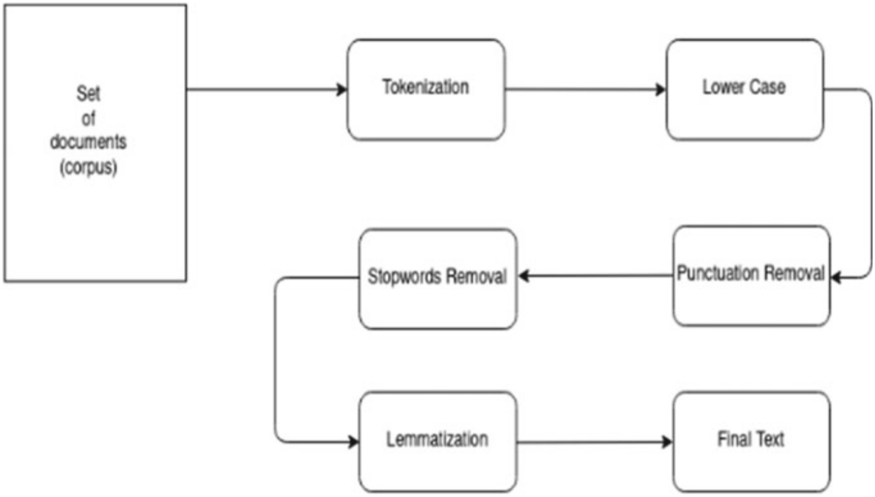


Fig. 3 Overall pre-processing of the text before creation of an inverted index

4.1.1 Removing Stop Words

There are some frequently occurring words in a dataset. These words are called stop words. They usually do no contribute much in the uniqueness of dataset. These words can be removed to reduce the indexing storage.

Example:

Document 1—The first document.

In the provided example, ‘the’ is a stop word which is most frequently used and carries no additional information for the specific document. So, these words are not required for indexing and also consume a lot of space. So, they will be removed completely before indexing the documents. The comparison will also be done without removing the stop words in order to check how much performance difference do these words make on the retrieval of the documents.

4.1.2 Stemming and Lemmatization

A corpus (document) may consist of different forms of a word which depicts similar meaning, but their form changes due to grammatical use in a sentence example: (fly and flying), (see and saw), (result, resulting and results), these all words depict a certain word, and other are just noun or plural forms of that word. The purpose lemmatization and stemming are to decrease inflectional and derivationally related forms of the words to a common base form. Stemming refers to a process that chops off or removes the ends of words. It also includes the removal of derivational affixes.

Lemmatization is the process of doing things properly with the use of a vocabulary and morphological analysing the words, in order to remove inflectional endings only and convert the word to its base or dictionary form, which is known as the lemma. The only difference between stemming and lemmatization is the way both the algorithm works in case of stemming the algorithm focuses on removing either prefix or suffix from a word to reduce it to a root word example: Studying -> Study, Studies -> Studi. In case of lemmatization, the word is reduced or transformed to its root lemma with the help of predefined vocabulary dictionary for example words ‘Studying’ and ‘Studies’ both will be reduced to word ‘Study’. This basically means that stemming takes the input word and chops off the letters of different forms of the same word. On the other hand, lemmatization finds the meaning of the word and will always give a valid word as a final output.

There are several algorithms available for stemming. The most popular stemming algorithm used is the *Porter’s algorithm*; it consists of 5 phases of word reductions, applied sequentially within each phase there are various conventions to select rules, like choosing the rule from each ruling group which can be applied to the lengthiest suffix. A lemmatizer on the other hand is a tool from natural language processing which does full morphological analysis which produces very modest analysis for retrieval though it can be slow compared to stemming algorithms.

Sentence: learning information retrieval system was fun.

Stemmed Sentence output: learn inform Retrieve System was fun.

Lemmatized Sentence output: learn information retrieval system was fun.

4.2 Building of an Inverted Index

This step involves indexing the processed documents and storing them on the machine. Indexing is done on the basis of document serial number (*docID*).

As shown in Fig. 4, each word from the document gets indexed as per their document's serial number. Further, multiple entries in a single document as well as different documents are merged which leads to creation of inverted index.

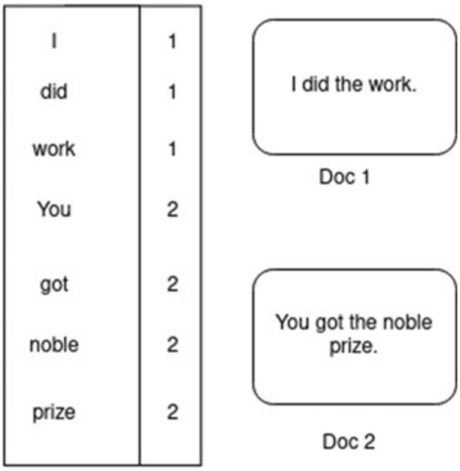
For each and every term *w*, a list of all documents containing the term *w* must be stored. As identified, each document by *docID* which is document serial number. Linked lists are used for this purpose.

As shown in Fig. 5, the left side elements represent dictionaries and the other side ones are postings list. Each term inside the dictionary points with the help of a pointer to its respected *docID* in which its present already. Moreover, the postings are sorted for the ease of operation. There are two main types of inverted index, namely:

1. Record-level inverted index.
2. Word level inverted index.

A record-level inverted index is used. This is the formation of inverted index data structure which is the backbone of search engine algorithms. There are slight modifications in the inverted index depending on each IR model. This inverted index

Fig. 4 Indexing after pre-processing of two documents



I	1
did	1
work	1
You	2
got	2
noble	2
prize	2

I did the work.

Doc 1

You got the noble prize.

Doc 2

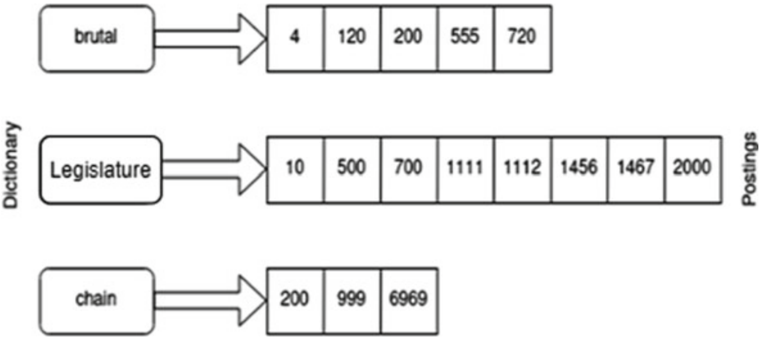


Fig. 5 Inverted index representation

is used by the retrieval algorithm to retrieve documents from the dataset to give the required search results.

4.3 Query Formulation and Matching

Query formulation refers to the process during which the user-issued query is transformed into a structured query representation that is used by the IR system to find the relevant documents. It is a process during which the original keyword query issued by the user is transformed into a structured query representation that is consumed by the search engine.

It modifies the user entered keyword query in order to better represent the actual intent of the query. This tweaked query is then used as an input to the IR system's ranking so-called query-document matching retrieval algorithm. Query terms are also pre-processed depending on the IR system which includes tokenization, stemming and stop words elimination.

Query-document matching is done to estimate the relevance of a document to the user-issued query. Every IR model has its own ranking function and retrieval algorithms for finding the similarity between the query and the documents. Once the values are calculated, the documents are ranked on the basis of most to least relevant documents and are displayed to the user.

4.4 Evaluating the Model

Evaluation the performance of the model is done by using mean average precision. During evaluating, the TF-IDF, Okapi BM25 model and the vector space model are tested to find the best model among them.

Mean average precision (mAP): This metric tells us about the number of items in the top-K results which are relevant. Mathematically,

$$\mathrm{AP} = \frac{p@1 + p@2 + p@3 + \cdots + p@n}{n} \tag{10}$$

where AP = Average Precision and n = Total number of queries.

Mean of the AP values over all queries is termed as mean average precision (mAP).

5 Results

While evaluating these models, three different datasets, namely Cranfield dataset, CACM collection dataset, CISI dataset are used for testing. The Cranfield dataset is derived from research papers over the period of 1945–1962 which consist of articles from aeronautic research. Cranfield dataset comes in two forms: 1400 collection and the 200 collection, 1400 collection is used here. The dataset consists of 1400 documents and 225 queries.

CACM dataset is a slightly bigger dataset which consists of 3204 documents and 64 queries. It is based on bibliographic information from articles of communication of the ACM dated 1958–1979 and consists of titles and abstracts from the journal.

CISI stands for Centre for Inventions and Scientific Information which is the name of the organization which created the dataset. The CISI dataset consists of about 1460 documents and 112 associated queries.

The Boolean model is very simple and convenient to build. But, the retrieval of documents can be carried only in two states: True or False. The Boolean model was left out as it is a very basic model and does not retrieve relevant results to the user queries as expected. The fuzzy model was also not included while testing as it only helps to get better results mostly when any misspelled queries are present in the dataset used for testing. During the testing, it is observed that each model acts differently depending on the datasets.

The TF-IDF model retrieves documents based on term frequency of a document multiplied by the inverse document frequency. But, the highest TF-IDF words of a document may not make sense with the topic of the document. This method of retrieval is further improved by the vector space model which introduces the cosine similarity function which further improves the performance of retrieval of documents. TF-IDF is a probability of just a simple term occurring, whereas vector space is a covariant probability. The vector space model unlike TF-IDF model does not decide the ranking of documents solely based on the weighting factor but also considers similarity factor between two words in a document by calculating cosine similarity of two document vectors it can be determined that how similar they are with respect to a given query and thus rank the documents accordingly.

However, the drawback of the vector space model is that it does not consider the length of the document due to which short documents are favoured over long documents while ranking the documents. This is because the long documents have poor similarity values. On the other hand, bm25 being a probabilistic model overcomes this drawback by tuning the tf parameter of the TF-IDF model using *k* and *b* (tuning parameters of BM25 formula), and also it is proven by results obtained in Table 2. Also, the Okapi BM25 model works best and gives more consistent results on the datasets used for testing.

In some cases, removing the stop words may not help in increasing the performance of the model as these words might help to retrieve the documents in a more effective way. This was justified by the performance of the models as shown in Table 3. The results of the CISI dataset (Table 4) show us the limitations of these models in the form of low mean average precision values. This is because of the diversity present in the documents of CISI dataset. This can be improved by using concepts like document clustering and classification algorithms in addition to these models. Keeping a record of which documents does the user opens after asking a query can also help in increasing the accuracy of the models.

The comparison with related works is given in Table 5. Aguilar et al. [1] have

Table 2 Results of Cranfield dataset

Name of model	Mean average precision
Okapi BM25 with stop list	0.5298
Okapi BM25 without stop list	0.5408
Vector space model with stop list	0.4219
Vector space model without stop list	0.4267
TF-IDF	0.3672

Table 3 Results of CACM dataset

Name of model	Mean average precision
Okapi BM25 with stop list	0.4289
Okapi BM25 without stop list	0.4401
Vector space model with stop list	0.2959
Vector space model without stop list	0.2664
TF-IDF	0.1995

Table 4 Results of CISI dataset

Name of model	Mean average precision
Okapi BM25 with stop list	0.2887
Okapi BM25 without stop list	0.2851
Vector space model with stop list	0.1977
Vector space model without stop list	0.1725
TF-IDF	0.1186

used the title, description and the keywords as well of the document to retrieve the documents. This helps to increase the performance of the models. The datasets used for evaluation also contribute to the variation in performance of the models. Also, the evaluation method used by them is by calculating the f1 score. In contrast, the datasets used in evaluation in this paper consists only of the document content for retrieval. The evaluation method used by this paper is MAP which is a more effective way of evaluation of information retrieval models. These are the reasons for the lower performance of our models.

Dai et al. [2] the authors have used the CACM dataset which is the same as used by this paper. Lemmatization and stop words removal are not carried out which shows a lower performance of BM25 as compared to our proposed models. Also, the method of evaluation used is different which shows a variation in the accuracy.

Svore and Burges [3] and Jimenez et al. [4] again have a low performance of the BM25 model as compared to the model used in this paper. This is because the differences in the removal of stop words are different and the lemmatization methodologies used by them.

Table 5 Comparison with state of art methods

References	Evaluation accuracy					Dataset used	Methodology for calculating precision/recall
	TF-IDF without stop list	Vector space model without stop list	Vector space model with stop list	Okapi BM25 without stop list	Okapi BM25 with stop list		
Aguilar et al. [1]	–	–	–	–	0.7290	MCRPC	F1 score using precision and recall
Dai et al. [2]	0.2532	0.3057	–	0.3095	–	CACM	Average precision
Svore and Burges [3]	–	–	–	–	0.3698 (MP@10)	Real-world Web-scale data collection	Mean NDCG@L
Jimenez et al. [4]	–	–	–	–	0.2606	TREC 8-1	MAP (mean average precision)
Proposed solution	0.3672	0.4267	0.4219	0.5408	0.5298	Cranfield	MAP (mean average precision)
	0.1995	0.2664	0.2959	0.4401	0.4289	CACM	MAP (mean average precision)
	0.1186	0.1725	0.1977	0.2851	0.2887	CISI	MAP (mean average precision)

6 Conclusion

In this paper, evaluation of different information retrieval models is carried out. Several trials on three datasets are done which had totally different characteristics. This paper concludes that the Okapi BM25 model works is the best and gives more consistent results than the other models used in the comparison. It also explains the limitations of these algorithms which can be further improved by clustering the indexed documents. This conclusion is very important as it will help to provide a base for designing an efficient full text-document based search engine. This will also help to develop domain-specific search engines and recommender systems.

References

1. Aguilar, J., Salazar, C., Velasco, H., Monsalve-Pulido, J., Montoya, E.: Comparison and evaluation of different methods for the feature extraction from educational contents. Computation **8** (2020)
2. Dai, S., Diao, Q., Zhou, C.: Performance comparison of language models for information retrieval. IFIP Adv. Inf. Commun. Technol. **187** (2005)
3. Svore, K.M., Burges, C.J.C.: A machine learning approach for improved BM25 retrieval. In: International Conference on Information and Knowledge Management, Proceedings (2009). https://doi.org/10.1145/1645953.1646237
4. Jimenez, S., Cucerzan, S.P., Gonzalez, F.A., Gelbukh, A., Dueñas, G.: BM25-CTF: improving TF and IDF factors in BM25 by using collection term frequencies. J. Intell. Fuzzy Syst. **34** (2018)
5. Kural, Y.B., Robertson, S., Jones, S.: Clustering information retrieval search outputs (1999). https://doi.org/10.14236/ewic/irsg1999.9
6. Rekha, J.U.: Instant fuzzy search using probabilistic-correlation based ranking. Indian J. Sci. Technol. (2020). https://doi.org/10.17485/ijst/v13i11.2020-32
7. Bhatia, P.K., Mathur, T., Gupta, T.: Survey paper on information retrieval algorithms and personalized information retrieval concept. Int. J. Comput. Appl. **66** (2013)
8. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to information retrieval (2008). https://doi.org/10.1017/cbo9780511809071
9. Cranfield collection. http://ir.dcs.gla.ac.uk/resources/test_collections/cran/
10. Robertson, S.: Microsoft Cambridge at TREC-9: filtering track (2001)
11. Soergel, D.: TREC: Experiment and Evaluation in Information Retrieval (Book Review). Digital Libraries and Electronic Publishing (2006)
12. CACM collection. http://ir.dcs.gla.ac.uk/resources/test_collections/cacm/
13. CISI (a data set for information retrieval). https://www.kaggle.com/dmaso01dsta/cisi-a-dataset-for-information-retrieval
14. Singhal, A.: Modern information retrieval: a brief overview. Bull. IEEE Comput. Soc. Tech. Comm. Data Eng. **24** (2001)
15. Pannu, M., James, A., Bird, R.: A comparison of information retrieval models. In: Proceedings of WCCCE 2014: The 19th Western Canadian Conference on Computing Education—In-Cooperation with ACM SIGCSE (2014). https://doi.org/10.1145/2597959.2597978
16. Ponte, J.M., Croft, W.B.: Language modeling approach to information retrieval. SIGIR Forum (1998). https://doi.org/10.1145/3130348.3130368
17. Xue, G.R., et al.: Optimizing web search using web click-through data. In: International Conference on Information and Knowledge Management, Proceedings (2004). https://doi.org/10.1145/1031171.1031192

18. Amo, P., Ferreras, F.L., Cruz, F., Rosa, M.: Smoothing functions for automatic relevance feedback in information retrieval. In: Proceedings—International Workshop on Database and Expert Systems Applications, DEXA, vol. 2000, Jan 2000
19. Trotman, A., Puurula, A., Burgess, B.: Improvements to BM25 and language models examined. In: Proceedings of the 2014 Australasian Document Computing Symposium, pp. 58–65. Association for Computing Machinery (2014). https://doi.org/10.1145/2682862.2682863
20. Joby, P.P.: Exploring devops: challenges and benefits. J. Inf. Technol. **1**(01), 27–37 (2019)
21. Chen, J.I.Z., Lai, K.-L.: Data conveyance maximization in bilateral relay system using optimal time assignment. J. Ubiquitous Comput. Commun. Technol. (UCCT) **2**(02), 109–117 (2020)

Network Attack Detection with QNNBADT in Minimal Response Times Using Minimized Features

S. Ramakrishnan and A. Senthil Rajan

Abstract Internet is a medium of globally interconnected independent networks. Though it was created to interconnect government research laboratories in 1994, it has witnessed phenomenal growth and has expanded to service millions of users in governments, academia and public/private organizations for multitude of purposes. Internet has been evolving continuously. Internet has also evidenced many attacks on its networks called cyberattacks. As Internet evolves, adversaries also evolve in their attacking techniques making it imperative to guard networks from attacks. In spite of firewalls, AVs (Anti Viruses) and other defense mechanisms, there is an implicit need to monitor deliberate proliferations. IDSs (Intrusion Detection Systems) are techniques that help monitor networks and raise alarms on finding damaging proliferations. This also implies IDSs need to be quick in their assessments of malicious behavior on the network. This paper proposes a NN (Neural Network) based IDS that can quickly respond to attacks by analyzing low-level network details. The proposed scheme is evaluated on the In CIRA-CIC-DoHBrw-2020 dataset where it averagely scores above 90% in accuracy when benchmarked on different sample sizes.

Keywords IDSs · NNs · Machine learning · Feature selection · Dimensionality reduction · CNNs · Network security

1 Introduction

ARPA (Advanced Research Projects Agency), the originator of Internet in 1969, developed this medium to share information and resources among researchers using remote logins, file transfers and electronic mail. It is a common knowledge that Internet is growing rapidly at an unprecedented rate. The growth of the Internet accelerated with the invention of HyperText Transfer Protocol (HTTI) and World Wide Web (WWW), rise in number of Web sites. Internet growth can be measured with the number of hosts, domain names and devices that use this medium. Figure 1 depicts projected growth of the Internet and its related devices.

S. Ramakrishnan (✉) · A. Senthil Rajan
Department of Computational Logistics, Alagappa University, Karaikudi, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_43

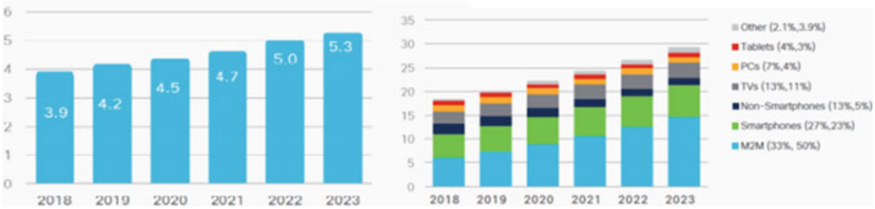


Fig. 1 Global Internet users and devices [1]

This astronomical growth can be attributed to growths in social web and mobile technology evolutions. Mobile technological innovations have resulted in global users of the Internet. Internet allows display of web pages (Personal/Official) with minimal investments and creates a large base. Blogging, online commerce, CC (Cloud Computing), online shopping malls, digital transaction and social media have made the Internet a powerful platform. It is the major source of information for millions. These operations are based on networks their backbone which handle terabytes of data daily with important implications on deployed networking technologies. The open nature of the Internet has also invited adversaries into its frameworks making data security a major issue. Attackers are also growing smarter with technology. Technological advances have been driving economy with leading trends like mobile payments, ecommerce, cloud computing and Big Data. They have also been drivers for increased cyberattacks and risk for users and businesses. The nature cyberthreats are also becoming more diverse and attacks like DDoS (Distributed Denial-of-Service), ransomware and spyware have emerged. Figure 2 depicts DDoS attacks.

WAFs (Web Application Firewalls) [2] are used by many organizations to protect their web-based applications. They analyze web requests and block malicious content. But these products are generic, operate on rules, and protect only commonly

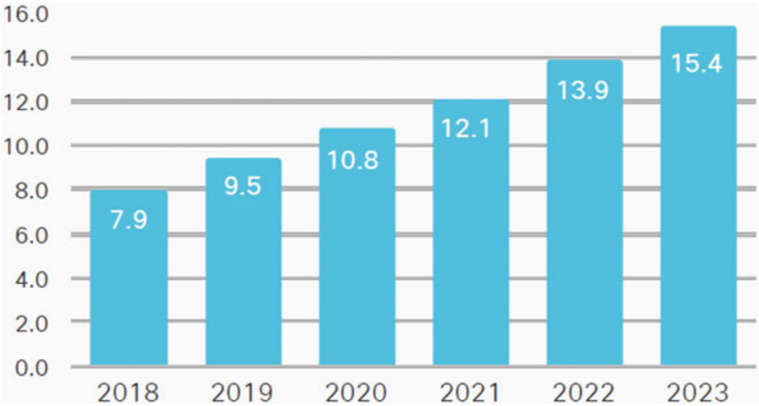


Fig. 2 DDoS cyberattacks in millions [1]

known attack sequences IDSs are tools which detect and report intrusions to administrators [3]. IDSs can be deployed in two forms namely HIDSs (Host-based Intrusion Detection systems) and NIDSs (Network-based Intrusion Detection systems). In order to analyze and monitor the network changes, IDS based on networks is used so that it protects the system from network-based threats. NIDSs collect information from network packets and analyze them with the aim of detecting malicious activity in networks. Target applications and its vulnerabilities can be detected by HIDSs that analyzes the system's events or system calls/logs. This research work proposes an efficient NIDS based on CNNs (Convolution Neural Networks) for quick identification of malicious packets in loaded networks using an optimal feature selection scheme. The scheme called QNNBADT (Quick Neural Network-Based Attack Detection Technique) is evaluated on the CIRA-CIC-DoHBrw-2020 dataset. This introductory section is followed by a related literature review in Sect. 2. Section 3 discusses the proposed.

2 Related Work

Cybersecurity is a significant area of research as Internet exposes global networks to individuals and organizations and in addition cyberthreats. Firewalls and AVs are widely used to preserve the user privacy and data. Intrusion detection system is another important research thrust in cybersecurity and it is used to identify the malicious activities [4]. IDSs can only be implemented only when effective or current datasets are used. KDD98 (Knowledge Discovery in Databases) and UNM (University of New Mexico) datasets were the earliest used for IDSs which are no longer valid for current cybersecurity threats. The study in [5] implemented an efficient kNN (k-Nearest Neighbor) based HIDS for analyzing frequency of system call traces with above 60% detections. IDSs have managed to detect new kinds of attacks when implemented. The earliest implementation of IDSs for network security used MLTs (Machine Learning Techniques) namely DTs (Decision Trees) using Bagged boosting [6] and Kernel Miner [7]. Rising interests in the area of AI has evolved several recognition or anomaly detection mechanisms. NNs have become a common choice for complex computations due to increase of cheaply available computational power and thus driving IDS implementations with NNs. The study in [8] found IDSs created large redundant, false alarms in an online approach that uses DARPA 1999 dataset where their results reduced false alarms by 94%. Web pages were classified as malicious or benign in [9]. The study used a JAVA program to generate static web pages and then processed it for known signatures with Res (Regular Expressions) which were then classified based on a honeypot system. The study in [10] identified APTs (Advanced Persistent Threats) which use different attack methods to access unauthorized systems. Their identification model was built on Search-Patterns, Events, Rules and Hypothesis. Misuse attacks were detected in [11] using their Web Anomaly Misuse Intrusion Detection (WAMID) model to detect SQL injection attacks as a combined detection algorithm. The study created profiles based on legitimate database behavior in training Association Rule Mining (ARM) is processed with SQL queries included XML files and detects the malicious activities. Res and pattern matching were exploited in [12] to detect attacks as a developed IDS model termed as web STAT. It automatically detects the signatures using a high-level language module. Interaction in cyberspace is reported in [13] that proposed a multi-perspective view for IDS interactions in TCP/IP's using four service layers to assess the unusual behavior. In order to detect intrusions, MLTs are used. Literature [14] discusses the feature benefits of combined model that incorporates support vector machine with genetic algorithm. In [15] to detect cyberintrusions, random forest algorithm was used. Deep learning techniques reported in [16] detects malware sequences. The call sequences are classified to classify malware so that accuracy of the deep learning techniques increases. CNN and RNN are widely used techniques. Neural networks are used to obtain the feature vector outputs and the correlation between input and generated feature vectors are obtained by processing the RNN-LSTM cells. Proposed work attains accuracy of 89% and feed-forward network clocked into 80%. A hardware assisted detection method reported in [17] presents an epoch-based malware detection to identify malicious patterns. The authenticated handler used in the detection model automatically detects the malwares and classifier produces better performance metrics for the parameters such as true positive rate. From the analysis, it could be observed that machine learning and deep learning-based intrusion detection models perform better in malware identification process.

3 Proposed Methodology

The proposed QNNBADT model identifies the malicious packets in very quick time as it considers only 5 important parameters for its identification. It follows the main steps of data cleaning, feature extraction, feature selection, selected optimal features cross-validation and classification using the CIRA-CIC-DoHBrw-2020 dataset. The architecture of QNNBADT is depicted in Fig. 3.

3.1 Data Preprocessing

In the data preprocessing of QNNBADT, cleaning the data to removes incomplete and incorrect, corrupted data in the dataset. The process varies for each data type or dataset features. In general irrelevant or incomplete or duplicate data is removed from the dataset. Data cleaning increases the decision accuracy since incomplete or incorrect data might affect the decision process. Input contained non-numeric, infinite or values for converting them into feature vectors for extractions. Unwanted/duplicate columns were dropped.

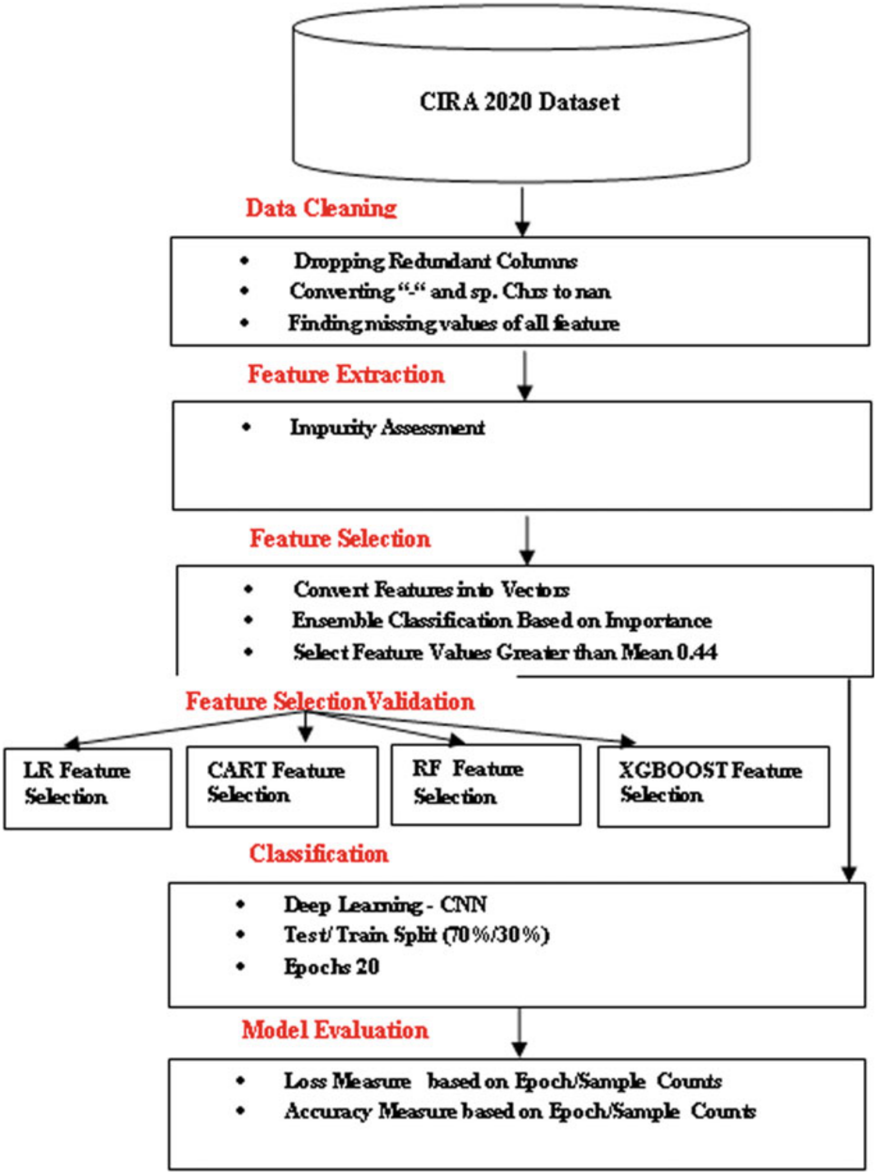

CIRA 2020 Dataset
Data Cleaning
Dropping Redundant Columns
Converting "-" and sp. Chrs to nan
Finding missing values of all feature
Feature Extraction
Impurity Assessment
Feature Selection
Convert Features into Vectors
Ensemble Classification Based on Importance
Select Feature Values Greater than Mean 0.44
Feature SelectionValidation
LR Feature Selection
CART Feature Selection
RF Feature Selection
XGBOOST Feature Selection
Classification
Deep Learning - CNN
Test/ Train Split (70%/30%)
Epochs 20
Model Evaluation
Loss Measure based on Epoch/Sample Counts
Accuracy Measure based on Epoch/Sample Counts

Fig. 3 QNNBADT architecture

3.2 Feature Extraction

In the feature extraction processes, the essential features are extracted by employing various metrics in information theory, uncertainty and correlation factors. All these metrics remove the redundant and irrelevant features. However, proposed model used to optimal features from existing impurity of features.

3.3 Feature Selection

It is a significant step in MLTs and pattern recognitions [18, 19]. Feature selection is a process that selects the feature subsets specified to the model. It is also an effective dimensionality reduction and identifies relevant and irrelevant features [20, 21]. The features are selected based on an ensemble method. The dataset's 31 features are trained for 20 maximum features to identify the exact optimal feature set using 250 estimators in the classification. The results are then aggregated to present a coherent set. Figure 4 depicts the importance of selected features.

QNNBADT uses an ensemble Trees Classification for aggregating multiple de-correlated decision trees of the forest for its classification. The process considers random forest classifier in parallel operation however the progress is different in its tree construction. The best features are identified and selected considering the mean feature values. QNNBAD further verifies this selection by cross verifying them with a feature correlation map and is depicted in Fig. 5.

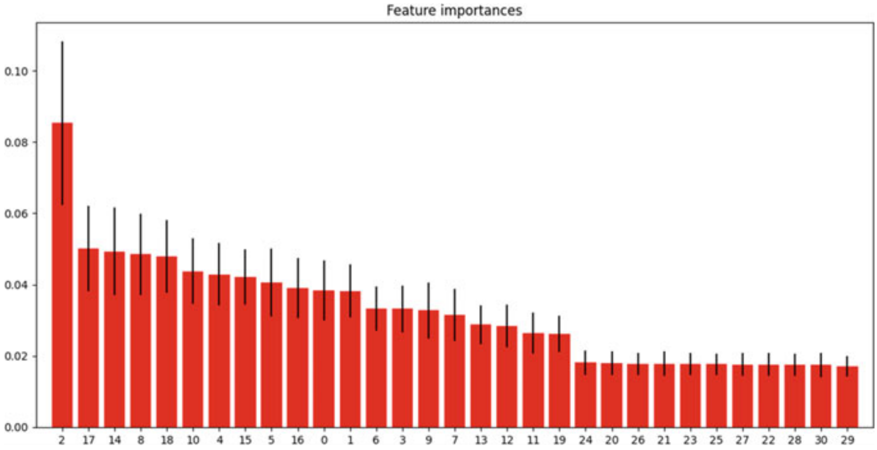


Fig. 4 QNNBADT feature selection

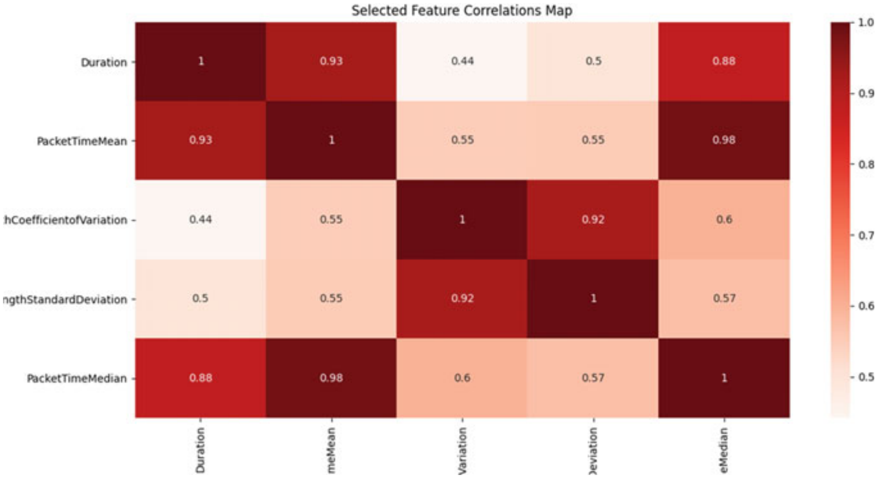


Fig. 5 QNNBAD feature cross verification

3.4 *Classification*

In the proposed QNNBADT model, convolutional neural network is used to classify the intrusions. CNN is a deep learning model that utilizes filter characteristics to learn and classify the data. Compared to other learning methodologies the preprocessing steps required for CNN is less. The neuron connectivity patterns and temporal, spatial features in data are used in the CNN model to classify the given data. The CNN architecture used in the proposed model is illustrated in Fig. 6.

The learning ability of CNN differs from traditional machine learning techniques in terms of pooling and weight sharing mechanisms. The convolution kernel in the CNN generates feature maps and the neurons connects the feature maps into the next layer. The spatial data in the feature maps are shared by the kernel. Finally, all the layers output are given to the fully connected layer which then becomes the base for classifications [22–24]. The weights are used to learn the pattern repetitions

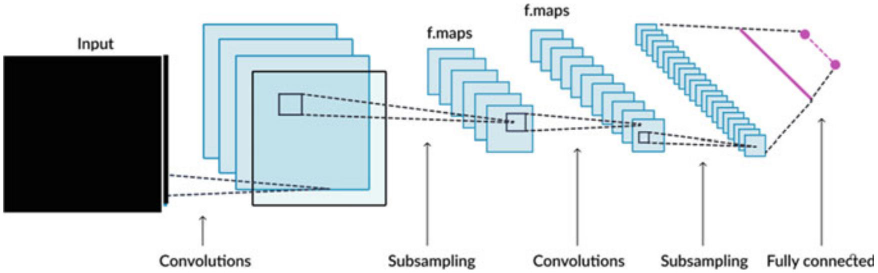


Fig. 6 CNN

without using separate detectors that makes the CNN robust against inputs [25]. The computational complexities are reduced by pooling layers. The connection count is reduced from one convolution to other convolution layer using pooling layers. Simply CNN layer convolutes input data using a set of kernels to produce a feature map. The transformation is mathematically formulated as in Eq. (1).

$$xk^{l} = \sigma\left(wk^{l-1} * x^{l-1} + bk^{l-1}\right) \tag{1}$$

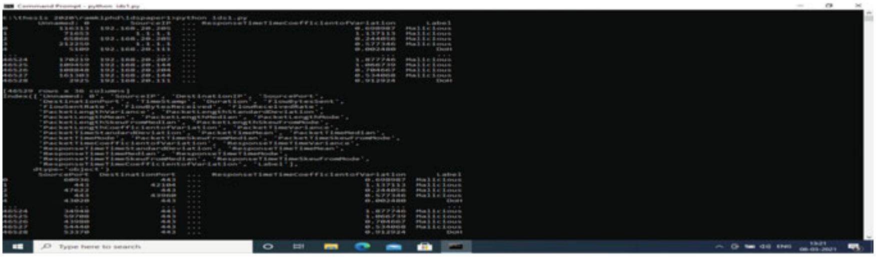
where l is the convolution layer, $W = \{w1, w2, \ldots, wn\}$ are n kernels and $B = \{b1, b2, \ldots, bn\}$are n biases.

4 Results and Discussions

The proposed model was evaluated using python 3 on the CIRA-CIC-DoHBrw-2020 compiled by CIC (Canadian Institute for Cybersecurity) project funded by CIRA (Canadian Internet Registration Authority). The system configuration was AMD Radeon processor with 16 GB RAM. The operating system used is Windows 10, 64 bit version. The software used to experiment the proposed work is Keras and Tensorflow. CNN architecture is framed in an exponential manner GPU enabled TensorFlow in a single Nvidia-GK110BGL-Tesla-k40. Modern, complex and more threat environment data presented dataset is selected for experimentation. The SAMPLES were partitioned into training (181,218) and test (94,854) sets. The sets had four categories ‘Malicious’, ‘Benign’, ‘DoH’ and ‘Non-DoH’. The proposed method’s stage-wise results are detailed below. Figure 7 depicts a snapshot of the dataset.

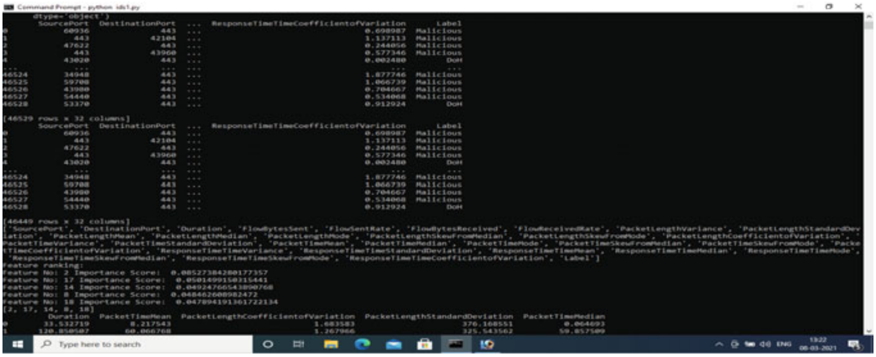
4.1 QNNBADT Data Preprocessing

To improve the classification quality and overall productivity data preprocessing is performed in the proposed work. Through finite difference measure, the missing



```
E:\thesis 2020\ramkiphd\idspaper1>python ids1.py
       Unnamed: 0        SourceIP  ...  ResponseTimeTimeCoefficientofVariation      Label
0          116313  192.168.20.205  ...                                0.698087  Malicious
1           71653         1.1.1.1  ...                                1.137113  Malicious
2           65866  192.168.20.205  ...                                0.244056  Malicious
3          212250         1.1.1.1  ...                                0.577346  Malicious
4            5109  192.168.20.111  ...                                0.002480        DoH
...           ...             ...  ...                                     ...        ...
46524      170219  192.168.20.207  ...                                1.877746  Malicious
46525      189459  192.168.20.144  ...                                1.866739  Malicious
46526      188848  192.168.20.204  ...                                0.784667  Malicious
46527      161303  192.168.20.144  ...                                0.534868  Malicious
46528        2925  192.168.20.111  ...                                0.912924        DoH

[46529 rows x 36 columns]
Index(['Unnamed: 0', 'SourceIP', 'DestinationIP', 'SourcePort',
       'DestinationPort', 'TimeStamp', 'Duration', 'FlowBytesSent',
       'FlowSentRate', 'FlowBytesReceived', 'FlowReceivedRate',
       'PacketLengthVariance', 'PacketLengthStandardDeviation',
       'PacketLengthMean', 'PacketLengthMedian', 'PacketLengthMode',
       'PacketLengthSkewFromMedian', 'PacketLengthSkewFromMode',
       'PacketLengthCoefficientofVariation', 'PacketTimeVariance',
       'PacketTimeStandardDeviation', 'PacketTimeMean', 'PacketTimeMedian',
       'PacketTimeMode', 'PacketTimeSkewFromMedian', 'PacketTimeSkewFromMode',
       'PacketTimeCoefficientofVariation', 'ResponseTimeTimeVariance',
       'ResponseTimeTimeStandardDeviation', 'ResponseTimeTimeMean',
       'ResponseTimeTimeMedian', 'ResponseTimeTimeMode',
       'ResponseTimeTimeSkewFromMedian', 'ResponseTimeTimeSkewFromMode',
       'ResponseTimeTimeCoefficientofVariation', 'Label'],
      dtype='object')
       SourcePort  DestinationPort  ...  ResponseTimeTimeCoefficientofVariation      Label
0           60936              443  ...                                0.698087  Malicious
1             443            42104  ...                                1.137113  Malicious
2           47622              443  ...                                0.244056  Malicious
3             443            43060  ...                                0.577346  Malicious
4           43028              443  ...                                0.002480        DoH
...           ...              ...  ...                                     ...        ...
46524       34948              443  ...                                1.877746  Malicious
46525       59708              443  ...                                1.866739  Malicious
46526       43988              443  ...                                0.784667  Malicious
46527       54440              443  ...                                0.534868  Malicious
46528       53370              443  ...                                0.912924        DoH
```

Fig. 7 Snapshot of the database



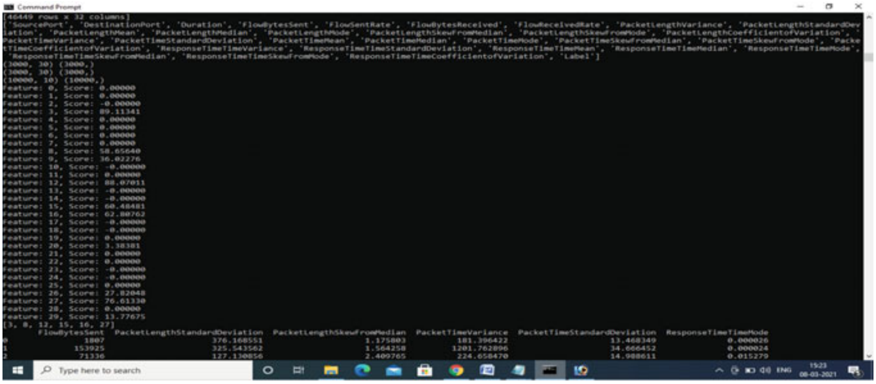
```
      dtype='object')
       SourcePort  DestinationPort  ...  ResponseTimeTimeCoefficientofVariation      Label
0           60936              443  ...                                0.698987  Malicious
1             443            42104  ...                                1.137113  Malicious
2           47622              443  ...                                0.244056  Malicious
3             443            43960  ...                                0.577346  Malicious
4           43020              443  ...                                0.002480        DoH
...           ...              ...  ...                                     ...        ...
46524       34948              443  ...                                1.877746  Malicious
46525       59708              443  ...                                1.066739  Malicious
46526       43980              443  ...                                0.704667  Malicious
46527       54440              443  ...                                0.534068  Malicious
46528       53370              443  ...                                0.912924        DoH

[46529 rows x 32 columns]
       SourcePort  DestinationPort  ...  ResponseTimeTimeCoefficientofVariation      Label
0           60936              443  ...                                0.698987  Malicious
1             443            42104  ...                                1.137113  Malicious
2           47622              443  ...                                0.244056  Malicious
3             443            43960  ...                                0.577346  Malicious
4           43020              443  ...                                0.002480        DoH
...           ...              ...  ...                                     ...        ...
46524       34948              443  ...                                1.877746  Malicious
46525       59708              443  ...                                1.066739  Malicious
46526       43980              443  ...                                0.704667  Malicious
46527       54440              443  ...                                0.534068  Malicious
46528       53370              443  ...                                0.912924        DoH

[46449 rows x 32 columns]
['SourcePort', 'DestinationPort', 'Duration', 'FlowBytesSent', 'FlowSentRate', 'FlowBytesReceived', 'FlowReceivedRate', 'PacketLengthVariance', 'PacketLengthStandardDev
iation', 'PacketLengthMean', 'PacketLengthMedian', 'PacketLengthMode', 'PacketLengthSkewFromMedian', 'PacketLengthSkewFromMode', 'PacketLengthCoefficientofVariation', '
PacketTimeVariance', 'PacketTimeStandardDeviation', 'PacketTimeMean', 'PacketTimeMedian', 'PacketTimeMode', 'PacketTimeSkewFromMedian', 'PacketTimeSkewFromMode', 'Packe
tTimeCoefficientofVariation', 'ResponseTimeTimeVariance', 'ResponseTimeTimeStandardDeviation', 'ResponseTimeTimeMean', 'ResponseTimeTimeMedian', 'ResponseTimeTimeMode',
 'ResponseTimeTimeSkewFromMedian', 'ResponseTimeTimeSkewFromMode', 'ResponseTimeTimeCoefficientofVariation', 'Label']
Feature ranking:
Feature No: 2 Importance Score:  0.08527384280177357
Feature No: 17 Importance Score:  0.05014991503315441
Feature No: 14 Importance Score:  0.04924766543890768
Feature No: 8 Importance Score:  0.0484626008982472
Feature No: 18 Importance Score:  0.047894191361722134
[2, 17, 14, 8, 18]
     Duration  PacketTimeMean  PacketLengthCoefficientofVariation  PacketLengthStandardDeviation  PacketTimeMedian
0   33.532719        8.217543                            1.683583                     376.168551          0.064693
1  120.850507       60.066768                            1.267966                     325.543562         59.857509
```

Fig. 8 QNNBADT preprocessing output

features are identified and filled with correct values. The preprocessing step of proposed model is depicted in Fig. 8.

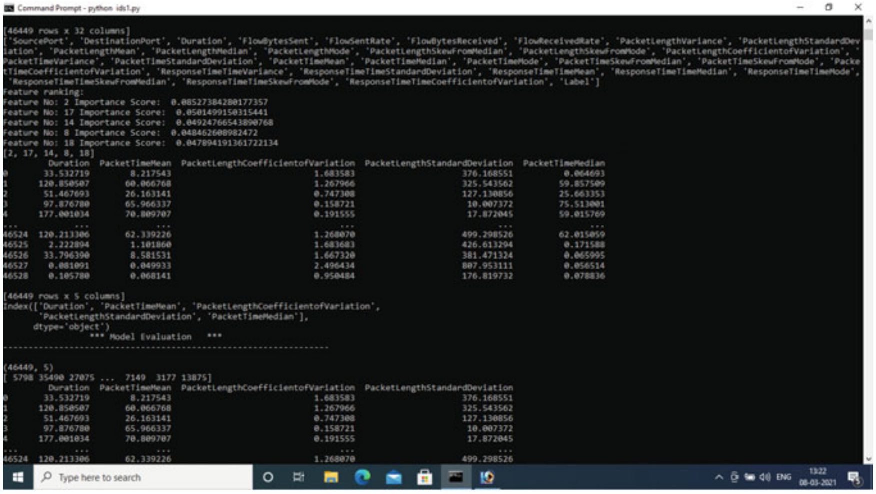
4.2 *QNNBADT Feature Extraction*

The process of identifying important features to improve the classification accuracy is feature extraction. Using suitable filtering techniques, the unnecessary features are removed in few feature extraction process. Figure 9 depicts feature extraction process of QNNBADT.



```
[46449 rows x 32 columns]
['SourcePort', 'DestinationPort', 'Duration', 'FlowBytesSent', 'FlowSentRate', 'FlowBytesReceived', 'FlowReceivedRate', 'PacketLengthVariance', 'PacketLengthStandardDev
iation', 'PacketLengthMean', 'PacketLengthMedian', 'PacketLengthMode', 'PacketLengthSkewFromMedian', 'PacketLengthSkewFromMode', 'PacketLengthCoefficientofVariation', '
PacketTimeVariance', 'PacketTimeStandardDeviation', 'PacketTimeMean', 'PacketTimeMedian', 'PacketTimeMode', 'PacketTimeSkewFromMedian', 'PacketTimeSkewFromMode', 'Packe
tTimeCoefficientofVariation', 'ResponseTimeTimeVariance', 'ResponseTimeTimeStandardDeviation', 'ResponseTimeTimeMean', 'ResponseTimeTimeMedian', 'ResponseTimeTimeMode',
 'ResponseTimeTimeSkewFromMedian', 'ResponseTimeTimeSkewFromMode', 'ResponseTimeTimeCoefficientofVariation', 'Label']
(3000, 30) (3000,)
(3000, 30) (3000,)
(10000, 10) (10000,)
Feature: 0, Score: 0.00000
Feature: 1, Score: 0.00000
Feature: 2, Score: -0.00000
Feature: 3, Score: 89.11341
Feature: 4, Score: 0.00000
Feature: 5, Score: 0.00000
Feature: 6, Score: 0.00000
Feature: 7, Score: 0.00000
Feature: 8, Score: 58.65640
Feature: 9, Score: 36.02276
Feature: 10, Score: -0.00000
Feature: 11, Score: 0.00000
Feature: 12, Score: 88.07011
Feature: 13, Score: -0.00000
Feature: 14, Score: -0.00000
Feature: 15, Score: 60.48481
Feature: 16, Score: 62.80762
Feature: 17, Score: -0.00000
Feature: 18, Score: -0.00000
Feature: 19, Score: 0.00000
Feature: 20, Score: 3.38381
Feature: 21, Score: 0.00000
Feature: 22, Score: 0.00000
Feature: 23, Score: -0.00000
Feature: 24, Score: -0.00000
Feature: 25, Score: 0.00000
Feature: 26, Score: 27.82048
Feature: 27, Score: 76.61330
Feature: 28, Score: 0.00000
Feature: 29, Score: 13.77675
[3, 8, 12, 15, 16, 27]
   FlowBytesSent  PacketLengthStandardDeviation  PacketLengthSkewFromMedian  PacketTimeVariance  PacketTimeStandardDeviation  ResponseTimeTimeMode
0           1807                     376.168551                    1.175803          181.396422                    13.468349              0.000026
1         153925                     325.543562                    1.564258         1201.762896                    34.666452              0.000024
2          71336                     127.130856                    2.409765          224.658470                    14.988611              0.015279
```

Fig. 9 QNNBADT feature extraction



```
[46449 rows x 32 columns]
['SourcePort', 'DestinationPort', 'Duration', 'FlowBytesSent', 'FlowSentRate', 'FlowBytesReceived', 'FlowReceivedRate', 'PacketLengthVariance', 'PacketLengthStandardDev
iation', 'PacketLengthMean', 'PacketLengthMedian', 'PacketLengthMode', 'PacketLengthSkewFromMedian', 'PacketLengthSkewFromMode', 'PacketLengthCoefficientofVariation', '
PacketTimeVariance', 'PacketTimeStandardDeviation', 'PacketTimeMean', 'PacketTimeMedian', 'PacketTimeMode', 'PacketTimeSkewFromMedian', 'PacketTimeSkewFromMode', 'Packe
tTimeCoefficientofVariation', 'ResponseTimeTimeVariance', 'ResponseTimeTimeStandardDeviation', 'ResponseTimeTimeMean', 'ResponseTimeTimeMedian', 'ResponseTimeTimeMode',
 'ResponseTimeTimeSkewFromMedian', 'ResponseTimeTimeSkewFromMode', 'ResponseTimeTimeCoefficientofVariation', 'Label']
Feature ranking:
Feature No: 2 Importance Score:  0.08527384280177357
Feature No: 17 Importance Score:  0.0501499150315441
Feature No: 14 Importance Score:  0.04924766543890768
Feature No: 8 Importance Score:  0.048462608982472
Feature No: 18 Importance Score:  0.047894191361722134
[2, 17, 14, 8, 18]
         Duration  PacketTimeMean  PacketLengthCoefficientofVariation  PacketLengthStandardDeviation  PacketTimeMedian
0       33.532719        8.217543                            1.683583                     376.168551          0.064693
1      120.850507       60.066768                            1.267966                     325.543562         59.857509
2       51.467693       26.163141                            0.747308                     127.130856         25.663353
3       97.876780       65.966337                            0.158721                      10.007372         75.513001
4      177.001034       70.809707                            0.191555                      17.872045         59.015769
...           ...             ...                                 ...                            ...               ...
46524  120.213306       62.339226                            1.268070                     499.298526         62.015059
46525    2.222894        1.101860                            1.683683                     426.613294          0.171588
46526   33.796390        8.581531                            1.667320                     381.471324          0.065995
46527    0.081091        0.049933                            2.496434                     807.953111          0.056514
46528    0.105780        0.068141                            0.950484                     176.819732          0.078836

[46449 rows x 5 columns]
Index(['Duration', 'PacketTimeMean', 'PacketLengthCoefficientofVariation',
       'PacketLengthStandardDeviation', 'PacketTimeMedian'],
      dtype='object')
                 *** Model Evaluation   ***
-----------------------------------------------------------------

(46449, 5)
[ 5798 35490 27075 ...  7149  3177 13875]
         Duration  PacketTimeMean  PacketLengthCoefficientofVariation  PacketLengthStandardDeviation
0       33.532719        8.217543                            1.683583                     376.168551
1      120.850507       60.066768                            1.267966                     325.543562
2       51.467693       26.163141                            0.747308                     127.130856
3       97.876780       65.966337                            0.158721                      10.007372
4      177.001034       70.809707                            0.191555                      17.872045
...           ...             ...                                 ...                            ...
46524  120.213306       62.339226                            1.268070                     499.298526
```

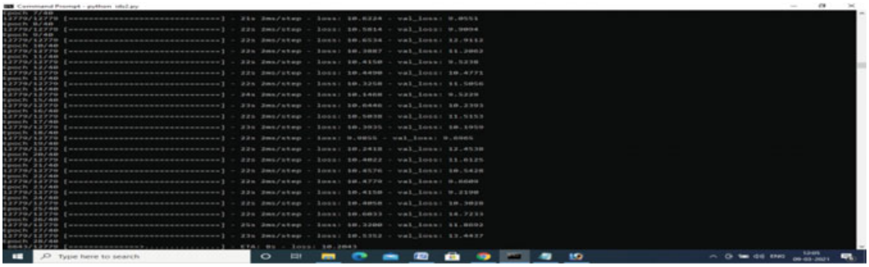
Fig. 10 QNNBADT feature selection

4.3 QNNBADT Feature Selection

The feature selection in the proposed work is mainly used to reduce the data dimensionality so that the classifier accuracy can be improved. Weighted sum of the input values is used in the QNNBADTs as an ensembled model that generates a set of coefficients. QNNBADT selects optimal features by eliminating features whose score is below a threshold value (0.44). Five features were selected and depicted in Fig. 10 along with co-efficient values of features.

4.4 QNNBADT Classification

CNN's dense layer is a regular layer with neurons which receive inputs from previous layers thus are connected. The weight matrix in the CNN module is represented as W, and the bias vector is represented as b. The previous layer activation function is represented as 'a'. This study's configuration used 128 layers with uniform kernel of size 4. Dropout is also handled by CNN. For all the layers the activation function used in the CNN is Rectified Linear Unit (ReLU) whereas for the last layer Softmax function is used. Adam optimizer is used to optimize the functions. Initially, the model is experimented with 40 epochs and it was changed later for different sample sizes. The entire process is split into 70% for training and 30% testing. The test samples and training results are depicted in Fig. 10. Figure 11 depicts the modeled CNN learning.



```
Epoch 7/40
12779/12779 [==============================] - 21s 2ms/step - loss: 10.6224 - val_loss: 9.0551
Epoch 8/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.5814 - val_loss: 9.9804
Epoch 9/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.6534 - val_loss: 12.9112
Epoch 10/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.3887 - val_loss: 11.2862
Epoch 11/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.4158 - val_loss: 9.5238
Epoch 12/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.4490 - val_loss: 10.4771
Epoch 13/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.3258 - val_loss: 11.5056
Epoch 14/40
12779/12779 [==============================] - 24s 2ms/step - loss: 10.1468 - val_loss: 9.5229
Epoch 15/40
12779/12779 [==============================] - 23s 2ms/step - loss: 10.6446 - val_loss: 10.2393
Epoch 16/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.5038 - val_loss: 11.5153
Epoch 17/40
12779/12779 [==============================] - 23s 2ms/step - loss: 10.3935 - val_loss: 10.1959
Epoch 18/40
12779/12779 [==============================] - 22s 2ms/step - loss: 9.0855 - val_loss: 9.6865
Epoch 19/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.2418 - val_loss: 12.4538
Epoch 20/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.4822 - val_loss: 11.0125
Epoch 21/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.4576 - val_loss: 10.5428
Epoch 22/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.4779 - val_loss: 9.8689
Epoch 23/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.4158 - val_loss: 9.2198
Epoch 24/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.4058 - val_loss: 10.3028
Epoch 25/40
12779/12779 [==============================] - 22s 2ms/step - loss: 10.6033 - val_loss: 14.7233
Epoch 26/40
12779/12779 [==============================] - 25s 2ms/step - loss: 10.3200 - val_loss: 11.8892
Epoch 27/40
12779/12779 [==============================] - 23s 2ms/step - loss: 10.5352 - val_loss: 13.4437
Epoch 28/40
 8843/12779 [====================>.........] - ETA: 8s - loss: 10.2043
```

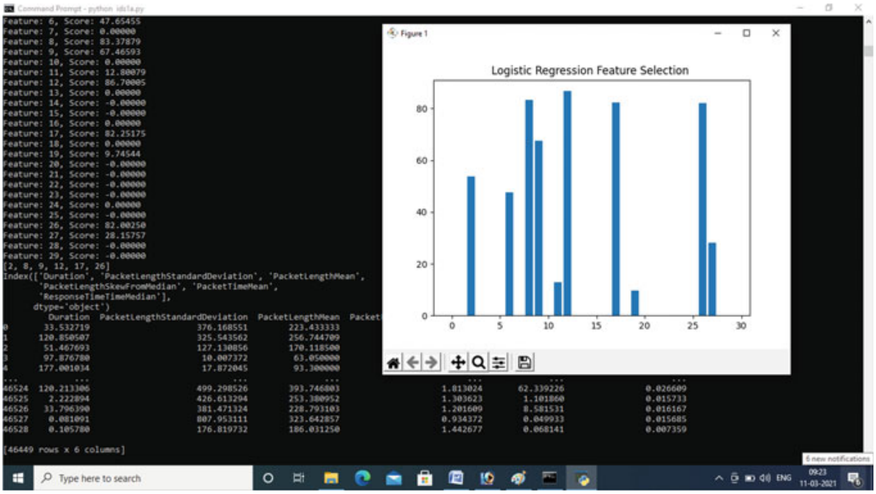
Fig. 11 CNN learning epochs

4.5 QNNBADT Model Evaluation

The feature selections were compared with other MLTs such as Classification And Regression Tree (CART), Logistics Regression (LR), Extreme Gradient Boosting (XGBoost) and Random Forest (RF). LR feature Importance was assessed and coefficients found for each input variable. LR feature selection output was 6 optimal features and depicted in Fig. 12.

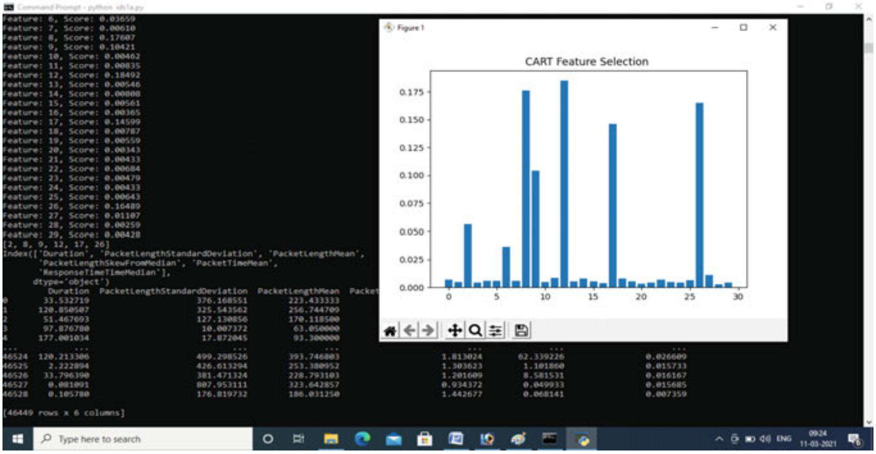
CART Feature Importance is a form of regressor and classifier. Each input feature relative importance score is retrieved using feature importance score once the model is fitted into the process. CART feature selection output has 6 optimal features and depicted in Fig. 13.

RF also uses Regressor and Classifier classes to generate feature importance after fitting a model. RF feature selection output was 6 optimal features and depicted in Fig. 14.



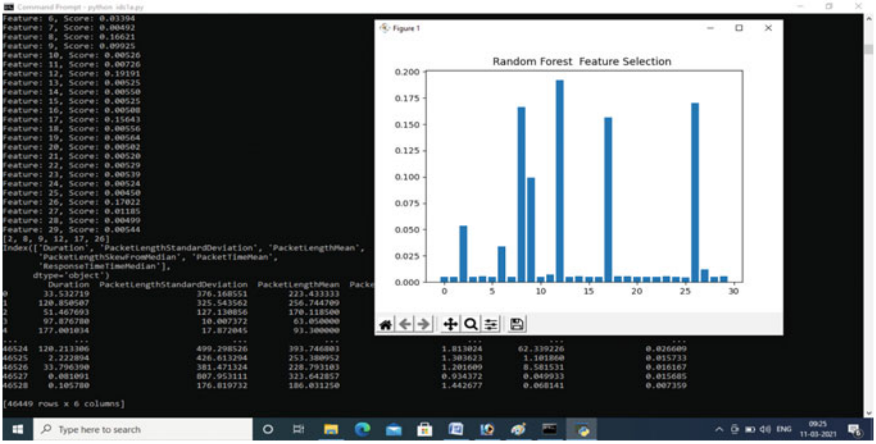
```
Feature: 6, Score: 47.65455
Feature: 7, Score: 0.00000
Feature: 8, Score: 83.37879
Feature: 9, Score: 67.46593
Feature: 10, Score: 0.00000
Feature: 11, Score: 12.80079
Feature: 12, Score: 86.70005
Feature: 13, Score: 0.00000
Feature: 14, Score: -0.00000
Feature: 15, Score: -0.00000
Feature: 16, Score: 0.00000
Feature: 17, Score: 82.25175
Feature: 18, Score: 0.00000
Feature: 19, Score: 9.74544
Feature: 20, Score: -0.00000
Feature: 21, Score: -0.00000
Feature: 22, Score: -0.00000
Feature: 23, Score: -0.00000
Feature: 24, Score: 0.00000
Feature: 25, Score: -0.00000
Feature: 26, Score: 82.00250
Feature: 27, Score: 28.15757
Feature: 28, Score: -0.00000
Feature: 29, Score: -0.00000
[2, 8, 9, 12, 17, 26]
Index(['Duration', 'PacketLengthStandardDeviation', 'PacketLengthMean',
       'PacketLengthSkewFromMedian', 'PacketTimeMean',
       'ResponseTimeTimeMedian'],
      dtype='object')
         Duration  PacketLengthStandardDeviation  PacketLengthMean  Packet
0       33.532719                     376.168551        223.433333
1      120.850507                     325.543562        256.744709
2       51.467693                     127.130856        170.118500
3       97.876780                      10.007372         63.050000
4      177.001034                      17.872045         93.300000
...           ...                            ...               ...   ...         ...              ...
46524  120.213306                     499.298526        393.746803   1.813024   62.339226        0.026609
46525    2.222894                     426.613294        253.380952   1.303623    1.101860        0.015733
46526   33.796390                     381.471324        228.793103   1.201609    8.581531        0.016167
46527    0.081091                     807.953111        323.642857   0.934372    0.049933        0.015685
46528    0.105780                     176.819732        186.031250   1.442677    0.068141        0.007359

[46449 rows x 6 columns]
```

Fig. 12 LR feature selection output



```
Feature: 6, Score: 0.03659
Feature: 7, Score: 0.00610
Feature: 8, Score: 0.17607
Feature: 9, Score: 0.10421
Feature: 10, Score: 0.00462
Feature: 11, Score: 0.00835
Feature: 12, Score: 0.18492
Feature: 13, Score: 0.00546
Feature: 14, Score: 0.00808
Feature: 15, Score: 0.00561
Feature: 16, Score: 0.00365
Feature: 17, Score: 0.14599
Feature: 18, Score: 0.00787
Feature: 19, Score: 0.00559
Feature: 20, Score: 0.00343
Feature: 21, Score: 0.00433
Feature: 22, Score: 0.00684
Feature: 23, Score: 0.00479
Feature: 24, Score: 0.00433
Feature: 25, Score: 0.00643
Feature: 26, Score: 0.16489
Feature: 27, Score: 0.01107
Feature: 28, Score: 0.00259
Feature: 29, Score: 0.00428
[2, 8, 9, 12, 17, 26]
Index(['Duration', 'PacketLengthStandardDeviation', 'PacketLengthMean',
       'PacketLengthSkewFromMedian', 'PacketTimeMean',
       'ResponseTimeTimeMedian'],
      dtype='object')
        Duration  PacketLengthStandardDeviation  PacketLengthMean  Packet
0      33.532719                     376.168551        223.433333
1     120.850507                     325.543562        256.744709
2      51.467693                     127.130856        170.118500
3      97.876780                      10.007372         63.050000
4     177.001034                      17.872045         93.300000
...          ...                            ...               ...           ...         ...                  ...
46524 120.213306                     499.298526        393.746803      1.813024   62.339226            0.026609
46525   2.222894                     426.613294        253.380952      1.303623    1.101860            0.015733
46526  33.796390                     381.471324        228.793103      1.201609    8.581531            0.016167
46527   0.081091                     807.953111        323.642857      0.934372    0.049933            0.015685
46528   0.105780                     176.819732        186.031250      1.442677    0.068141            0.007359

[46449 rows x 6 columns]
```

Fig. 13 CART feature selection output

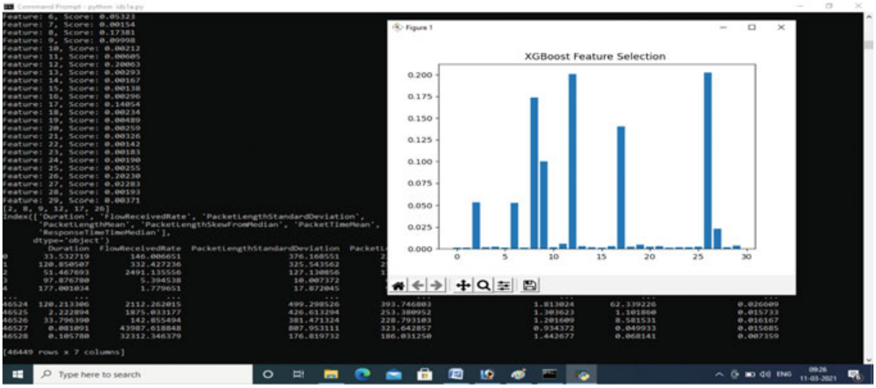


```
Feature: 6, Score: 0.03394
Feature: 7, Score: 0.00492
Feature: 8, Score: 0.16621
Feature: 9, Score: 0.09925
Feature: 10, Score: 0.00526
Feature: 11, Score: 0.00726
Feature: 12, Score: 0.19191
Feature: 13, Score: 0.00525
Feature: 14, Score: 0.00550
Feature: 15, Score: 0.00525
Feature: 16, Score: 0.00508
Feature: 17, Score: 0.15643
Feature: 18, Score: 0.00556
Feature: 19, Score: 0.00564
Feature: 20, Score: 0.00502
Feature: 21, Score: 0.00520
Feature: 22, Score: 0.00529
Feature: 23, Score: 0.00539
Feature: 24, Score: 0.00524
Feature: 25, Score: 0.00450
Feature: 26, Score: 0.17022
Feature: 27, Score: 0.01185
Feature: 28, Score: 0.00499
Feature: 29, Score: 0.00544
[2, 8, 9, 12, 17, 26]
Index(['Duration', 'PacketLengthStandardDeviation', 'PacketLengthMean',
       'PacketLengthSkewFromMedian', 'PacketTimeMean',
       'ResponseTimeTimeMedian'],
      dtype='object')
        Duration  PacketLengthStandardDeviation  PacketLengthMean  Packe
0      33.532719                     376.168551        223.433333
1     120.850507                     325.543562        256.744709
2      51.467693                     127.130856        170.118500
3      97.876780                      10.007372         63.050000
4     177.001034                      17.872045         93.300000
...          ...                            ...               ...           ...         ...                  ...
46524 120.213306                     499.298526        393.746803      1.813024   62.339226            0.026609
46525   2.222894                     426.613294        253.380952      1.303623    1.101860            0.015733
46526  33.796390                     381.471324        228.793103      1.201609    8.581531            0.016167
46527   0.081091                     807.953111        323.642857      0.934372    0.049933            0.015685
46528   0.105780                     176.819732        186.031250      1.442677    0.068141            0.007359

[46449 rows x 6 columns]
```

Fig. 14 RF feature selection output

XGBoost is an effective implementation of stochastic gradient boosting algorithm and uses a regressor and classifier tor retrieving relative importance scores of input features. XGBoost feature selection output was 6 optimal features and depicted in Fig. 15.

Thus in Feature Selection of optimal parameters from a dataset QNNBADT chooses the optimal reduced set of features from a dataset without losing the importance of features.



```
Feature: 6, Score: 0.05323
Feature: 7, Score: 0.00154
Feature: 8, Score: 0.17381
Feature: 9, Score: 0.09998
Feature: 10, Score: 0.00212
Feature: 11, Score: 0.00605
Feature: 12, Score: 0.20063
Feature: 13, Score: 0.00293
Feature: 14, Score: 0.00167
Feature: 15, Score: 0.00138
Feature: 16, Score: 0.00296
Feature: 17, Score: 0.14054
Feature: 18, Score: 0.00234
Feature: 19, Score: 0.00489
Feature: 20, Score: 0.00259
Feature: 21, Score: 0.00326
Feature: 22, Score: 0.00142
Feature: 23, Score: 0.00183
Feature: 24, Score: 0.00190
Feature: 25, Score: 0.00255
Feature: 26, Score: 0.20230
Feature: 27, Score: 0.02283
Feature: 28, Score: 0.00193
Feature: 29, Score: 0.00371
[2, 8, 9, 12, 17, 26]
Index(['Duration', 'FlowReceivedRate', 'PacketLengthStandardDeviation',
       'PacketLengthMean', 'PacketLengthSkewFromMedian', 'PacketTimeMean',
       'ResponseTimeTimeMedian'],
      dtype='object')
         Duration  FlowReceivedRate  PacketLengthStandardDeviation  PacketL
0       33.532719        146.006651                     376.168551
1      120.850507        332.427236                     325.543562
2       51.467693       2491.135556                     127.130856
3       97.876780          5.394538                      10.007372
4      177.001034          1.779651                      17.872045
...           ...               ...                            ...       ...                              ...           ...                    ...
46524  120.213306       2112.262015                     499.298526   393.746803                     1.813024     62.339226              0.026609
46525    2.222894       1875.033177                     426.613294   253.380952                     1.303623      1.101860              0.015733
46526   33.796390        142.855494                     381.471324   228.793103                     1.201609      8.581531              0.016167
46527    0.081091      43987.618848                     807.953111   323.642857                     0.934372      0.049933              0.015685
46528    0.105780      32312.346379                     176.819732   186.031250                     1.442677      0.068141              0.007359

[46449 rows x 7 columns]
```

Fig. 15 XGBoost feature selection output

4.6 QNNBADT Classification Accuracy

Learning curves plot a model's learning performance. These curves diagnose MLTs that learn incrementally in training. This evaluation is held back to be evaluated on the validation dataset after each training update. These values are then plotted creating learning curves which can help in identifying problems like underfits/overfits in the model. A loss function can optimize a MLT. The model performance for each set is used to evaluate the loss function in the training and validation process. Also, loss is estimated using error samples sum in the training or validation sets. Higher loss values depict a model's poor performance or its behavior after iterations. Accuracy is an interpretable measure of algorithmic performances. Accuracy determines model's parameter validity as a percentage. It also measures a model's prediction in real data. The proposed model was cross-validated in terms of samples and iterations which are detailed between Figs. 16, 17, 18, 19, 20 and 21.

It can be seen from Fig. 16 that convergence is fast on lower iterations and the gap widens with increasing samples. It can also be seen as depicted in Fig. 17 that accuracy is better on lesser number of iteration.

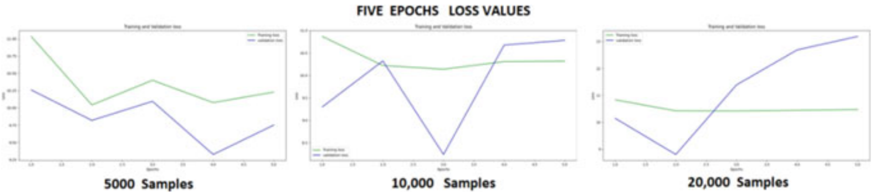


Fig. 16 Loss values against iterations

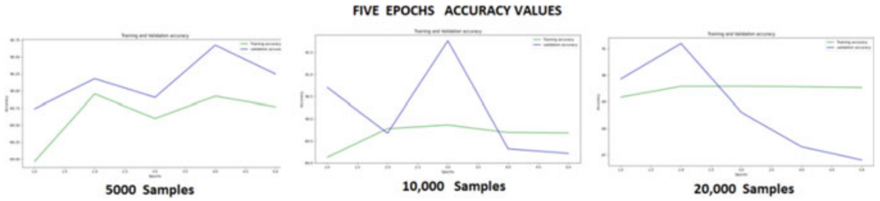


Fig. 17 Accuracy versus iteration

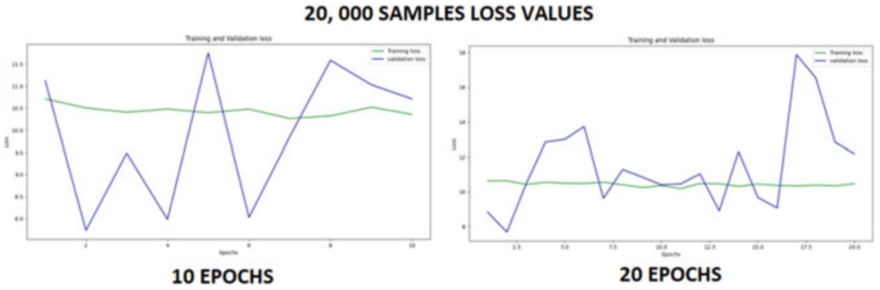


Fig. 18 Samples count against loss

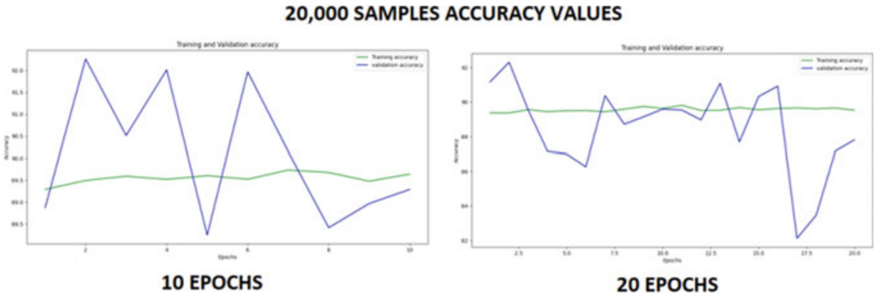


Fig. 19 Samples count versus accuracy

It can be seen from Fig. 18 that the model also converges on higher sample counts and higher iterations. These convergences result in effective accuracy convergence toward the preferred baseline as depicted in Fig. 19.

The proposed model can converge faster even if a complete dataset is checked as depicted in the figures which are for 5 iterations and 45,000 samples. This is particularly useful for loaded networks where innumerous packets flow and need to be asses in quick time as is evident from Figs. 20 and 21.

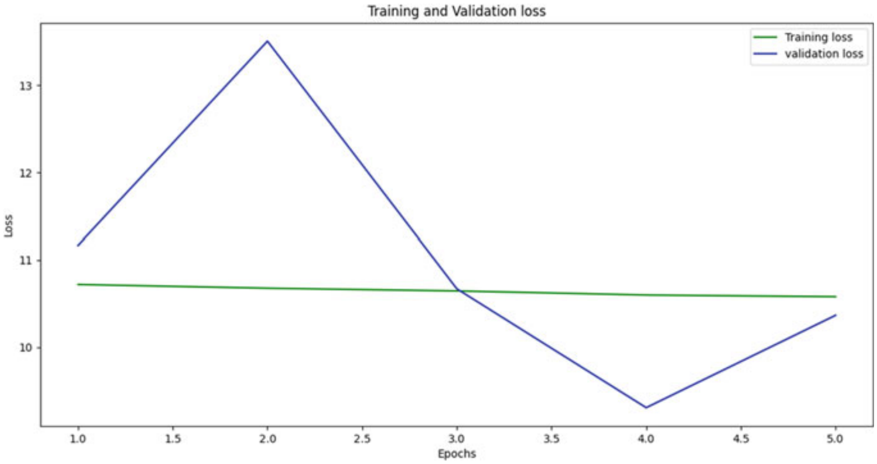


Fig. 20 Higher number of samples versus lower iterations (loss)

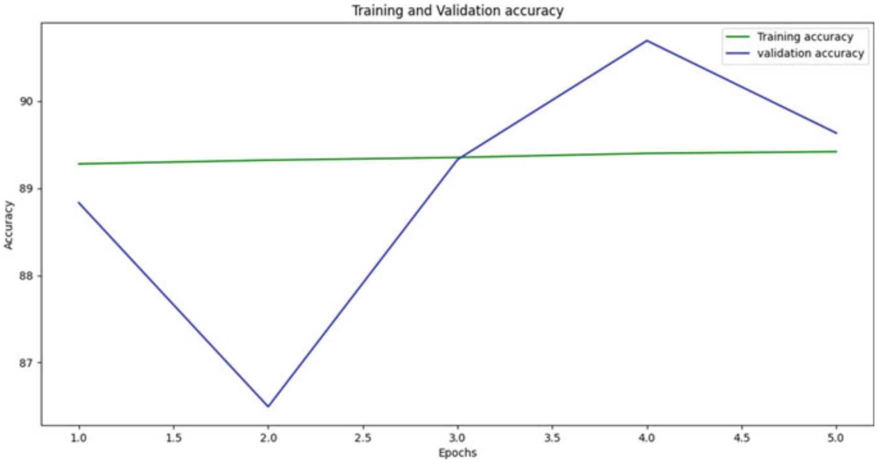


Fig. 21 Higher number of samples versus lower iterations (accuracy)

5 Conclusion

Network security can be enhanced through network intrusion detection system and numerous research works are evolved in the recent years to detect multiclass intrusions. This research work summarizes the issues in network attacks classification and malware identification process in view of improving the network security. Using benchmark dataset, the proposed work classified and identified the respective elements that breaches network security. Correlation and clustering-based tree classifications used in the proposed work identify the important features and reduce the dimensionality by removing the redundant and irrelevant features. In the second phase, using deep learning architecture convolution neural network, the network attacks are identified and classifies as malwares. The overall accuracy of the proposed multiclass detection model is around 90% that indicates the proposed work can be suitable for any networks to identify malwares. Experimental results prove that proposed QNNBADT model efficiently detects the unknown attacks in short duration compared to existing techniques. Further, this research work can be extended to detect attacks on mobiles using artificial intelligence methods.

References

1. Cisco Annual Internet Report (2018–2023). Online at https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html
2. Pałka, D., Zachara, M.: Learning web application firewall-benefits and caveats. In: International Conference on Availability, Reliability, and Security, Aug 2011, pp. 295–308. Springer, Berlin Heidelberg
3. McHugh, J.: Intrusion and intrusion detection. Int. J. Inf. Secur. **1**(1), 14–35 (2001)
4. Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C., Tung, K.-Y.: Intrusion detection system: a comprehensive review. J. Netw. Comput. Appl. **36**, 16–24 (2013). https://doi.org/10.1016/j.jnca.2012.09.004
5. Xie, M., Hu, J.: Evaluating host-based anomaly detection systems: a preliminary analysis of ADFA-LD. In: 6th International Congress on Image and Signal Processing (CISP), Hangzhou, China (2013)
6. Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., Gao, R.X.: Deep learning and its applications to machine health monitoring: a survey. IEEE Trans. Neural Netw. Learn. Syst. (2016)
7. Lee, H., Kim, Y., Kim, C.O.: A deep learning model for robust wafer fault monitoring with sensor measurement noise. IEEE Trans. Semicond. Manuf. **30**(1), 23–31 (2017)
8. Barghi, M.N., Hosseinkhani, J., Keikhaee, S.: An effective web mining-based approach to ımprove the detection of alerts in ıntrusion detection systems. Int. J. Adv. Comput. Sci. Inf. Technol. (IJACSIT) (ELVEDIT) **4**(1), 38–45 (2015)
9. Koo, T.M., Chang, H.C., Hsu, Y.T., Lin, H.Y.: Malicious website detection based on honeypot systems. In: 2nd International Conference on Advances in Computer Science and Engineering (CSE 2013), July 2013
10. Friedberg, I., Skopik, F., Settanni, G., Fiedler, R.: Combating advanced persistent threats: from network event correlation to incident detection. Comput. Secur. **48**, 35–57 (2015)
11. Salama, S.E., Marie, M.I., El-Fangary, L.M., Helmy, Y.K.: Webanomaly misuse intrusion detection framework for SQL injection detection. Int. J. Adv. Comput. Sci. Appl. (IJACSA), Editorial Preface 3.3-Citeseer **3**(3), 123–129 (2012)
12. Vigna, G., Robertson, W., Kher, V., Kemmerer, R.A.: A stateful intrusion detection system for world-wide web servers. In: Proceedings of the Annual Computer Security Applications Conference (ACSAC 2003), Las Vegas, NV, Dec 2003, pp. 34–43
13. Shaikh, S.A., Chivers, H., Nobles, P., Clark, J.A., Chen, H.: A deployment value model for intrusion detection sensors. In: Lecture Notes in Computer Science in 3rd International Conference on Information Security and Assurance, June 2009, vol. 5576, pp. 250–259
14. Davis, J.J., Clark, A.J.: Data preprocessing for anomaly based network intrusion detection: a review. Comput. Secur. **30**(6–7), 353–375 (2011)
15. Mohammadpour, L., Hussain, M., Aryanfar, A., Raee, V.M., Sattar, F.: Evaluating performance of intrusion detection system using support vector machines: review. Int. J. Secur. Appl. **9**(9) (2015)

16. Litjens, G., et al.: A survey on deep learning in medical ımage analysis. CoRR, vol. 1702.05747 (2017)
17. Singh, R., Kumar, H., Singla, R.K.: An intrusion detection system using network traffic profiling and online sequential extreme learning machine. Expert Syst. Appl. **42**(22), 8609–8624 (2015)
18. Iguyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
19. Zhou, P., Hu, X., Li, P., Wu, X.: Online feature selection for high-dimensional class-imbalanced data. Knowl.-Based Syst. **136**, 187–199 (2017)
20. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. **17**(4), 491–502 (2005)
21. Benabdeslem, K., Hindawi, M.: Efficient semi-supervised feature selection: constraint, relevance, and redundancy. IEEE Trans. Knowl. Data Eng. **26**(5), 1131–1143 (2014)
22. Shibahara, T., Yagi, T., Akiyama, M., Chiba, D., Yada, T.: Efficient dynamic malware analysis based on network behavior using deep learning. In: 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1–7. IEEE (2016)
23. Raman, K., et al.: Selecting features to classify malware. InfoSec Southwest, vol. 2012. In: A Survey on Malware Detection from Deep Learning, p. 79 (2012)
24. Firdausi, A.E., Nugroho, A.S., et al.: Analysis of machine learning techniques used in behavior-based malware detection. In: 2010 Second International Conference on Advances in Computing, Control and Telecommunication Technologies (ACT), pp. 201–203. IEEE (2010)
25. Kolosnjaji, B., Zarras, A., Webster, G., Eckert, C.: Deep learning for classification of malware system call sequences. In: Australasian Joint Conference on Artificial Intelligence, pp. 137–149. Springer (2016)

Deep Learning-Based Approach for Satellite Image Reconstruction Using Handcrafted Prior

Jaya Saxena, Anubha Jain, and Pisipati Radha Krishna

Abstract We propose a randomly initialized neural network as handcrafted prior to distorted satellite image for its restoration. The model is applied for cloud removal and proved efficient. Extensive experiments on the satellite datasets demonstrate efficiency of the proposed model both quantitative and qualitative. Further, the proposed approach also removed the dependency on pre-training datasets. In our study, RGB monochromatic satellite images were considered with the obscured area of varying shapes, lying in the range of 14–30%. Reconstructed image with MSE 0.131 and PSNR of 80.937 is obtained. Another inference deduced from the results is structural symmetry index (SSIM) values are better for red and green bands when compared to blue band. Image hash value is also calculated and found satisfactory.

Keywords Convolutional neural networks · Satellite image · Cloud removal · Evaluation metrics

1 Introduction

With the advancement in research and technology, convolutional neural networks (CNN) and its variants have emerged as a popular tool for image generation, classification and restoration [1, 2]. Their excellent performance is attributed to the large training data set with an ability to learn realistic image priors from these. However, this is also a bottleneck in their execution as availability of such training data set is always not possible. This initiates an idea for research where the distorted image,

J. Saxena (✉)
NRSC, ISRO, Hyderabad, India
e-mail: jayasaxena@nrsc.gov.in

A. Jain
Department of CS and IT, IIS University, Jaipur, India
e-mail: anubha.jain@iisuniv.ac.in

P. Radha Krishna
Department of CS, National Institute of Technology, Warangal, Warangal, India
e-mail: prkrishna@nitw.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_44

as a handcrafted prior, itself acts as an input to the model, thereby removing the dependency on the preexisting large training data set.

The existence of clouds is one of the main factors that contribute to missing information in optical remote sensing images, restricting their further applications for Earth observation. Clouds hinder the monitoring of vegetations, land surfaces, water bodies, etc. Removal of cloud cover on the satellite remote sensing images can effectively improve the availability of remote sensing images.

In this work, randomly initialized neural network used as a handcrafted prior was applied to satellite images reconstruction and shown promising result. Further, as cloud is most common and prominent atmospheric interferer in remote sensing imagery, the study is also applied for cloud removal. Encoder-decoder "hourglass" architecture is used for the study with varying hyperparameters. Results are evaluated using MSE, PSNR, SSIM and image hash evaluation metrics and found to be promising.

This also removed the dependency on huge pre-training datasets, in contrast with most of the other deep learning algorithms.

The major advantage of the proposed model over other deep learning models like CNN, RNN, SpaGans, etc., is its non-dependency on pre-training datasets. Sometimes, it becomes really difficult to get near real-time good cloud free images of the interest area, especially in the rainy season to train the model and this becomes a bottleneck. Also, this model efficiently removes thick and thin clouds unlike SpaGans.

The organization of the paper is as follows: Sect. 2 briefs about the related works in this area, Sect. 3 covers the methodology, architecture adopted and operational requirements of the study. Section 4 contains experiments and results. Conclusion and future scope of the work are presented in Sect. 5, followed by references.

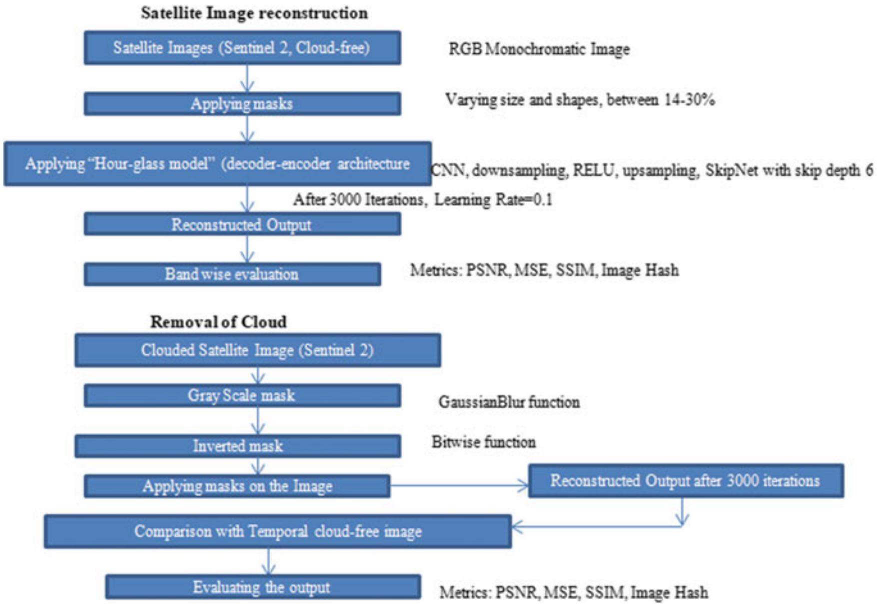
2 Related Works

Extensive research and application work is being carried out in the field of neural networks, including convolutional, fully convolutional, recurrent, residual, generative adversarial networks, etc., for image recognition, classification, segmentation and reconstruction [3, 4]. Many algorithms are worked upon depending upon the requirements. It basically started with the work of Schmidhuber [2] who presented the research work on deep learning in his book "Deep Learning in Neural Networks: An Overview Neural Networks." This marked the beginning of a new era in this field. Liu et al. [5] explored Image Inpainting technique for irregular holes using partial convolutions. Van den Oord et al. [6] used deep recurrent neural networks to improve generative models for natural images. Pixel RNNs with up to 12 LSTM layers were proposed and evaluated. Inglada and Garrigues [7] used simple linear interpolation with temporal images for cloud removal from optical satellite imagery. Similarly, Ren et al. [8] used overlapping region detection, matching point pair's extraction, image rectification for thick cloud removal from MODIS data. Other works are also referenced while doing this study and experiments which include different deep learning models like CNN, RNN, Relu-Net, different SpaGans and different types of clouds. Not only AI and DL algorithms but traditional approaches of cloud removal like image fusion, wavelet transformations, Savitzky-Golay filters, ECDR algorithm, etc., were also applied and found that deep learning algorithms are more efficient and provide better output [9–16].

3 Methodology and Operational Requirements

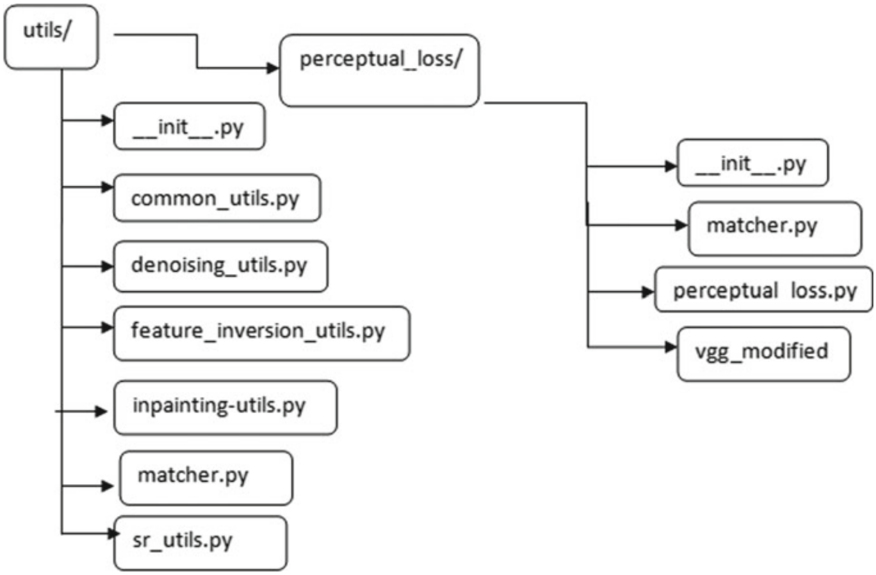
Four monochromatic RGB images from Sentinel 2 satellite with different spectral features in red, green and blue bands are considered for the study. Four masks of different shapes and sizes, covering between 14 and 30% of the image area, are generated. These masks are overlaid on the satellite images to cover the image area. The covered image area due to the masks depicts gaps, which may arise due to cloud presence, line loss, pixel losses, band misregistration, etc. Now, these distorted satellite images are reconstructed using handcrafted prior. The reconstructed image is evaluated in all the three spectral bands.

Further, similar study was applied on the real clouded image and the distorted image was reconstructed to achieve cloud free output image. A temporal image of the same location with same spectral features is used for evaluating the output.


Satellite Image reconstruction
Satellite Images (Sentinel 2, Cloud-free)
RGB Monochromatic Image
Applying masks
Varying size and shapes, between 14-30%
Applying "Hour-glass model" (decoder-encoder architecture
CNN, downsampling, RELU, upsampling, SkipNet with skip depth 6
After 3000 Iterations, Learning Rate=0.1
Reconstructed Output
Band wise evaluation
Metrics: PSNR, MSE, SSIM, Image Hash
Removal of Cloud
Clouded Satellite Image (Sentinel 2)
Gray Scale mask
GaussianBlur function
Inverted mask
Bitwise function
Applying masks on the Image
Reconstructed Output after 3000 iterations
Comparison with Temporal cloud-free image
Evaluating the output
Metrics: PSNR, MSE, SSIM, Image Hash

3.1 Software Used

Python 3.7 is used. The required image reconstruction tools are selected and tested from the available library, and only these tools are kept in the folder for efficient working, as shown below:



3.2 Architecture

The model uses Encoder-decoder "hourglass" architecture (possibly with skip connections) with a number of depending variables and hyperparameters [17–19]. The best results are achieved by carefully tuning the relation of depending variables and hyperparameters, learning rate, batch size, momentum, number of epochs. In our model, we have used 3000 epochs per image, which can be increased or decreased as per the time accuracy trade-off need in particular cases.

The details of depending variables and hyperparameters in the architecture are:
$z \in R32 \times W \times H \sim U(0, 1/10)$ nu = nd = [16, 32, 64, 128, 128, 128]
kd = [3, 3, 3, 3, 3, 3] Satellite Image
ku = [5, 5, 5, 5, 5, 5] ns = [0, 0, 0, 0, 0, 0] $\sigma p = 0$
num iter = 5000 LR = 0.1 upsampling = nearest
The architecture is shown in Fig. 1.

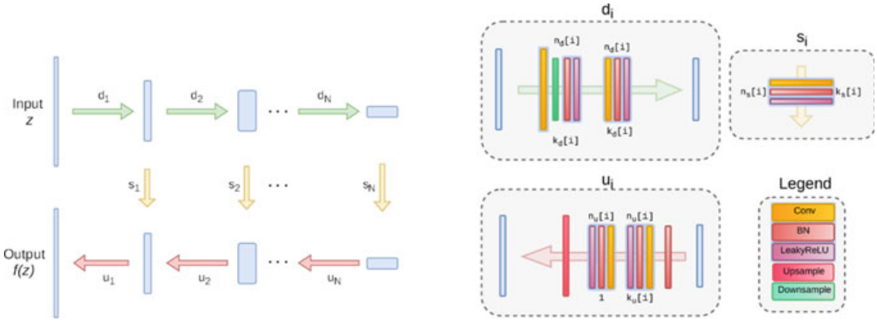


Fig. 1 Encoder-decoder “hourglass” architecture

Skip connections (yellow arrows) nu[i], nd[i], ns[i] correspond to the number of filters at depth i for the upsampling, downsampling and skip connections, respectively. The values ku[i], kd[i], ks[i] correspond to the respective kernel sizes.

3.3 Evaluation Metrics

Mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and image hash [20, 21] are used for the study. The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better. The PSNR ratio is used as a quality measurement between the original and a compressed image. The higher the PSNR, the better is the quality of the compressed or reconstructed image. SSIM is used for measuring the similarity between two images. SSIM is designed to improve on traditional methods such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE).

4 Experiments and Results

Dataset: Four monochromatic RGB images with bounding box coordinates as (79.76, 23.00) (81.00, 23.00) (79.76, 22.24) (81.01, 22.25) from Sentinel 2 satellite with 10 m spatial resolution and spectral resolution between 458 and 899 nm along with four different shapes and sizes masks are considered for the study, as shown in Fig. 2.

Further, observations are recorded on these 4 sample images having shape (3, 256, 256) with different percentages and types of black masks as shown in Table 1. These masks are applied bitwise. Mask and image combinations M-i, M: Mask, i: Image.

We observe from results that structural symmetry index (SSIM) values are better for red and green bands rather than blue band.

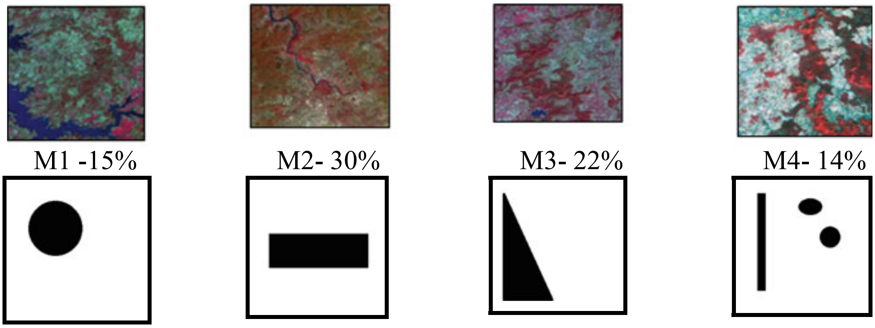


Fig. 2 Mask shapes, as applied on images. From left to right: M1, M2, M3 and M4

Table 1 Loss function (MSE) after 3000 iterations and SSIM of individual bands on applying masks over images

Mask-image	MSE (× 10^{-4})	SSIM		
		Red	Green	Blue
M1-1	3.310	0.893	0.900	0.811
M2-1	5.703	0.806	0.814	0.734
M3-1	3.669	0.852	0.864	0.781
M4-1	3.852	0.900	0.910	0.819
M1-2	2.742	0.904	0.916	0.854
M2-2	3.658	0.847	0.845	0.785
M3-2	3.494	0.863	0.866	0.803
M4-2	3.353	0.922	0.924	0.859
M1-3	3.382	0.918	0.915	0.881
M2-3	3.901	0.843	0.837	0.812
M3-3	2.479	0.871	0.870	0.837
M4-3	2.933	0.917	0.9171	0.882
M1-4	18.406	0.846	0.832	0.830
M2-4	15.588	0.785	0.781	0.774
M3-4	8.527	0.799	0.786	0.787
M4-4	19.170	0.861	0.848	0.850

Analyzing different spectral bands of the images with https://www.geotests.net/couleurs/frequences_en.html, it is observed that, among the compositions of all images, last image has most colors falling in range of blue and with reference to Table 1, rest images perform better than the last, hence, supporting our hypothesis of red and green bands performing better than blue band (Fig. 3).

Now, considering the general trend observed here for each image

M1 < M3 < M4 < M2

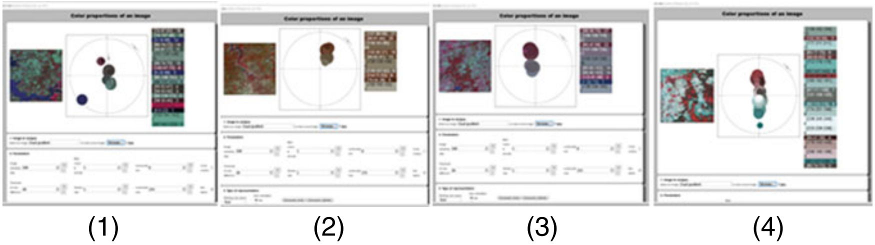


Fig. 3 Composition of the images in RGB spectral bands

M1 < M4 < M3 < M2
M3 < M4 < M1 < M2
M3 < M2 << M1 < M4

According to above information, we see that masks other than M2, which is having a horizontal rectangular mask, perform better than M2.

Now, considering the least and maximum obtained MSE we have following results, as shown in Figs. 4 and 5.

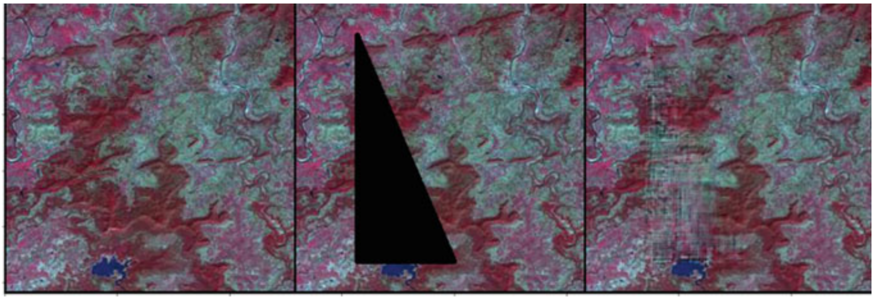


Fig. 4 Least overall MSE→M3-4, 0.0002.479

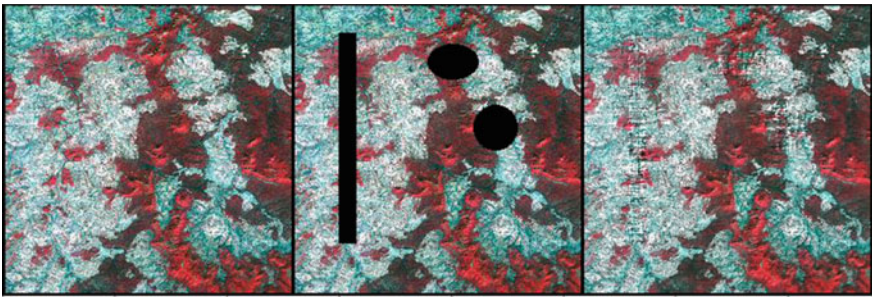


Fig. 5 Max overall MSE→M4-5, 0019.170

4.1 Cloud Removal

The study was also applied to detect presence of cloud in the image, using below algorithm:

```
Step 1: #Read the image with cloud using cv2.imread("filepath")
Step 2: # Calculating mask using threshold values for white color
gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
blurred = cv2.GaussianBlur(gray, (11, 11), 0)
# Calculate threshold the image to reveal light regions in the blurred image
thresh = cv2.threshold(blurred, 110, 255, cv2.THRESH_BINARY)[1]
cv2_imshow(image)
cv2_imshow(gray)
cv2_imshow(thresh)
Step 3: # Inverting the mask
mask = np.invert (thresh)
cv2_imshow(mask) (please refer Fig. 6, shown above)
Step 4
#Applying mask on image using bitwise and
output = cv2.bitwise_and(image, image, mask = mask)
cv2_imshow(output) (please refer Fig. 7)
```

For evaluation of image, we would now have some metric evaluation:

MSE of reconstructed portion—7.155447383411229e−05

MSE of whole image—3.314674387401269

PSNR of whole image—67.00879454985356

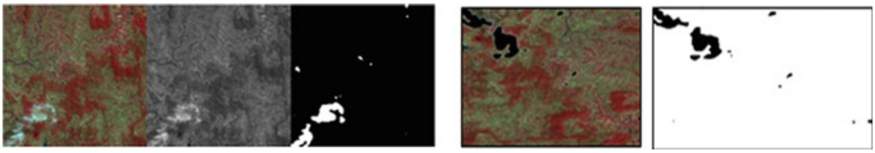


Fig. 6 Left side vertical three images: image, grayscale and masked by threshold values. Right side two images: inverted mask and mask applied on cloudy image

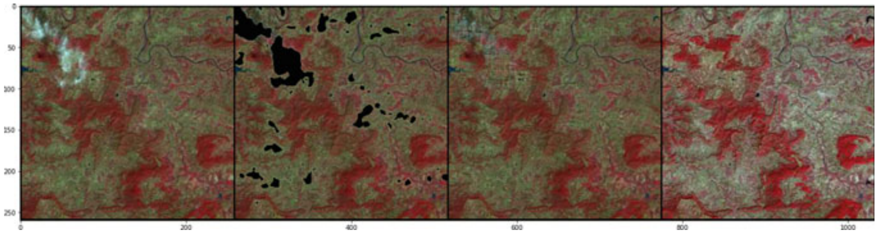


Fig. 7 Left to right; original cloudy image, masked image, reconstructed and temporal image

Image Hash

Calculating Hash Difference values of images is the process which uses the fingerprint technique to store value of an image as a hash value of 16-bits. After calculating hash value of reconstructed image and another image retrieved of the same location is generated, and then, their difference is calculated in order to find differences between two images. We have calculated average, perception and difference hash values.

Average hash 10
Perceptual hash 32
Difference hash 13

Another ways of evaluating the reconstruction area is by using the values of only reconstructed pixels rather than the whole image. Work has started in this direction and can be progressed further. The mask is a rectangular mask thus can be easily cropped using available libraries. We find the coordinates of the mask in whole image, using following code:

```
img = Image.open('/gdrive/My Drive/Reconstruct/data/mask/m7.png').convert('1')
pixels = img.load()
xlist = []
ylist = []
for y in range(img.size[1]):
   for x in range(img.size[0]):
      if pixels[x, y] == 0:
         xlist.append(x)
         ylist.append(y)
# Four corners of the black square
xleft = min (xlist)
xright = max(xlist)
ytop = min(ylist)
ybot = max(ylist)
print(xleft,xright,ytop,ybot)
```

Next, we will crop the original and output image using the bounding box coordinates obtained from the above code. Further applying the evaluation metrics we got:

MSE of whole image 0.1341598667592543
MSE of masked region 0.0186756302884764
PSNR of whole region 80.93697671848453
PSNR of masked region 69.36525511201556
hash1 = ffffd4f5e1031f1e
hash2 = ff1c1e643717203f
Hash1− hash2 = 27 (Please refer Fig. 8)

Here are some constraints that need further study to generate better results. Firstly, we are able to apply only the rectangular mask, and hence, the application area is getting very limited. Secondly, although the MSE value is quite acceptable the other values are not suitable, and hence, some hyperparameter and parameter tuning are required which can be done easily using the provided architectures, like setting

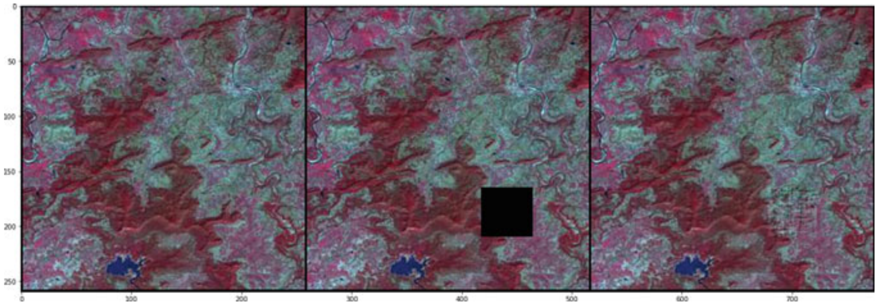


Fig. 8 Original, masked and output images after iteration 05000 loss 0.00014157035911921412

different skip depth, changing the learning rate, number of iterations (More the number of iterations more smooth the output will be), etc.

5 Conclusion and Future Scope

The masks of varying shapes, obscuring an area between 14 to 30%, are applied to the RGB monochromatic satellite images. Reconstructed images are obtained with least overall MSE of 0.000247 with a triangular mask and maximum overall MSE of 0.00191 with rectangular and circular masks. We also observed that structural symmetry index (SSIM) values are better for red and green bands as compared to the blue band. The algorithm was further applied on the image for cloud removal and précised reconstructed cloud free satellite images are obtained. Average hash, perceptual hash and difference hash values came satisfactory. Further, reconstructed satellite image was evaluated using values of only reconstructed pixels rather than the whole image. Best result obtained is MSE—0.000141 and PSNR—80.936.

Considering the fact that not many sensors are using blue band, the architecture can be useful for predicting better structural symmetry and application of deep learning in image reconstruction. In light of above observations, we see the opportunity of further improving this model using the relation of depending variables and hyperparameters. Results may vary on changing the variables, like number of iterations, skip depth, etc. Skip depth used here is 6. Results may change with skip depth of 8. Use of alternative architectures like ResNet, Unet can also be studied.

References

1. Malladi, R.M.V., Nizami, A., Mahakali, M.S., Krishna, B.G.: Cloud masking technique for high-resolution satellite data: an artificial neural network classifier using spectral & textural context. J. Indian Soc. Remote Sens. **47**(4), 661–670 (2019)

2. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
3. Simard, P., Bottou, L., Haffner, P., LeCun, Y.: Boxlets: a fast convolution algorithm for neural networks and signal processing. In: Advances in Neural Information Processing Systems (NIPS 1998), vol. 11. MIT Press (1999)
4. Cheng, J.Y., Chen, F., Alley, M.T., Pauly, J.M., Vasanawala, S.S.: Highly scalable image reconstruction using deep neural networks with band pass filtering. Computing Research Repository. http://arxiv.org/abs/1805.03300 (2018)
5. Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Computer Vision—ECCV 2018—15th European Conference, Munich, Germany, 8–14 Sept 2018, Proceedings, Part XI, pp. 89–105
6. van den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: Proceedings of the 33nd International Conference on Machine Learning, ICML, 2016, New York City, NY, USA, 19–24 June 2016, pp. 1747–1756
7. Inglada, J., Garrigues, S.: Land-cover maps from partially cloudy multi-temporal image series: optimal temporal sampling and cloud removal. In: IEEE International Geoscience & Remote Sensing Symposium, IGARSS 2010, 25–30 July 2010, Honolulu, Hawaii, USA, Proceedings
8. Ren, R., Guo, S., Gu, L., Wang, H.: Automatic thick cloud removal for MODIS remote sensing imagery. In: 2009 International Conference on Information Engineering and Computer Science, Wuhan, pp. 1–4 (2009)
9. Schmidt, U., Roth, S.: Shrinkage fields for effective image restoration. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014
10. Sutskever, I., Martens, J., Hinton, G.: Generating text with recurrent neural networks. In: Proceedings of the 28th International Conference on Machine Learning (2011)
11. Hornik, K., Stinchcombe, M., White, H.: Multilayer feed forward networks are universal approximators. Neural Netw. **2**, 359–366 (1989)
12. Bengio, Y.: Greedy layer-wise training of deep networks, In: Advances in Neural Information Processing Systems, Jan 2007
13. Lowe, D.G.: Distinctive image features from scale-invariant key points. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
14. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 7–12 June 2015. IEEE
15. Sungheetha, A., Sharma, R.: A novel CapsNet based image reconstruction and regression analysis. J. Innov. Image Process. (JIIP) **2**(03), 156–164 (2020)
16. Chawan, A.C., Kakade, V.K., Jadhav, J.K.: Automatic detection of flood using remote sensing images. J. Inf. Technol. Digit. World **2**(01), 11–26 (2020)
17. Li, E.Y.: Human pose estimation with stacked hourglass network and TensorFlow. In: Towards Data Science, Mar 2020
18. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV) (2016)
19. Babu, S.C.: A 2019 guide to human pose estimation with deep learning. Nanonets (2019)
20. Rajkumar, S., Malathi, G.: A comparative analysis on image quality assessment for real time satellite images. Indian J. Sci. Technol. **9**(34) (2016). https://doi.org/10.17485/ijst/2016/v9i34/96766
21. Silva, E.A., Panetta, K., Agaian, S.S.: Quantifying image similarity using measure of enhancement by entropy

CLOP Ransomware Analysis Using Machine Learning Approach

E. S. Aiswarya, Adheena Maria Benny, and Leena Vishnu Namboothiri

Abstract Machine learning seems to be evolving day by day in various technological aspects. It has become a powerful tool that may use to do both harm and good. Data breaches, ransomware threats, Internet of Things (IoT), etc., are some of the threats faced by cybersecurity. Traditional cybersecurity methods could not tackle these attacks. Among malware, ransomware is a particularly diabolical type of malware. Once ransomware gets on your computer usually through an infected email attachment or all-too-common Trojan horse attack it will lock your computer or your data in some way and demand payment in exchange for giving control of your system back to you. Some simple ransomware model will simply try to fool the users and make them to spend more money for fixing it. Clop ransomware is considered as one of the most dangerous malwares. Most of the computers will become victims to Clop ransomware. Nowadays, it is an increasing concern among large companies. So, the main purpose of this study is to provide extreme surveillance, for that a survey has been proposed on Clop detection using machine learning methods. Using machine learning algorithms, the study analyzes clop detection in three different datasets using different machine learning algorithms and measured these conclusions with a similar malware detection study. Based on our evaluation, it is observed that XGBoost outperforms all other machine learning algorithms. The proposed study answers the question regarding the best machine learning algorithm for Clop detection.

Keywords Malware · Machine learning · KNN · XGBoost · Random forest · Logistic regression

E. S. Aiswarya (✉) · A. M. Benny · L. V. Namboothiri
Department of Computer Science and IT, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

L. V. Namboothiri
e-mail: leena@asas.kh.amrita.edu

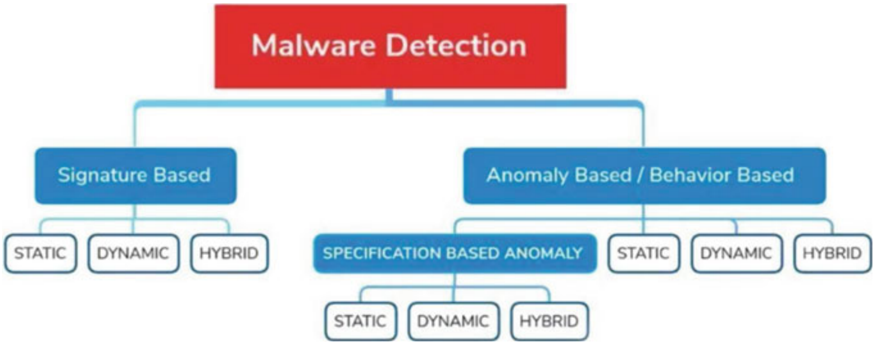
S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_45

1 Introduction

Malware is a software blueprint for causing harm to a computer, computer network, etc. There is a plethora of malware, some of them are

- Viruses
- Worms
- Trojan horse
- Ransomware
- Spyware
- Adware
- Scareware.

These are some kinds of malware that exist today. In this study, we are trying to detect Clop ransomware using machine learning algorithms. Malware detection is the method of inspecting malware samples and benign samples. There are some basic malware detection methods: Signature-based and Anomaly-Based/Behavior-Based. In signature-based, unknown malware remains undetected, because only signatures of known malware are stored, so Clop can't be detected since Clop is newly introduced. Small mutation in code or file can change the entire signature and pattern of the malware. So, nowadays this method is not appropriate. The disadvantage of another one, heuristic-based, is we need to update the malware properties day by day. So, we need more resources in terms of time and space. So, machine learning methods are more accurate. Among malware, Clop or CryptoMix is a ransomware that targets windows. Clop was first introduced in February 2019. Clop tries to stop many of the windows processes and uninstall security software. Encryption of infected files is using RSA 1024 bit public key. After encryption, a ransom note is released. Many algorithms are unsuccessful in the efficiency of malicious code [1]. This study aims to flaunt the best accurate among machine learning algorithms in the detection of Clop malware. In the existing papers, the malware prediction is carried out by using many machine learning algorithm and in proposed system, a comparison between five machine learning algorithms to find out which algorithm is best in predicting Clop malware.



2 Related Work

A few methodologies had proposed on malware family classification. The paper explicates by Gavriluţ et al. [2] suggested malware recognition using distinctive ML calculations such as cascade one-sided perceptron and cascade kernelized one-sided perceptron. Their study results in very few false positives. Another paper by Xu et al. [3] put forward a strategy to recognize malware utilizing ML and profound learning techniques utilizing opcode recurrence as highlight vector. They used kernel rootkits and memory corruption attacks on user programs which results in 99.0% detection rate. Planes et al. [4] present a review of different traditional methods like static, dynamic and hybrid-based detection for malware detection. They reviewed a total of 67 research papers on windows. It also signifies significant glitches faced by researchers. Kim et al. [5] had done an investigation on malware detection using 10 procedures and found that the best among ML is random forest. They also reviewed accuracy neural networks, LSTM and feed-forward network. Sami et al. [6] set forward a thought of extracting application programming interface examples and identifying malwares by utilizing an idea of iterative pattern mining which results in detection rate of 99.7% and accuracy as high as 98.3%.

Kolosnjaji et al. [7] proposed a method detect malware families with deep neural networks, like RNN, to categorize malware into families using API call patterns. From the experimental results, it showed that by combining CNN with LSTM, which a type of RNN 89.4% is recall was obtained.

RNN plays a vital role in malware detection. Ficco [8] compared the malware detection performance using two detectors—CAD and MCD. The results obtained from experiments showed that the MCD is better than CAD, and also, these two detectors are also very robust against hackers. Pascanu et al. [9] proposed a method to find whether the files were malicious or benign using RNN and echo state networks. They found that their combination gave good performance results and 95% accuracy rate.

There are many methods that make use of API information for detecting malwares. One among them is Sundarkumar et al. [10] presented a method that uses text mining and topic modeling to search out malware, based on the categories of API call patterns. They used LDA, a dimension reduction algorithm, as the feature selection model. It provides good performance results when compared with the previous methodology. Ki et al. [11] extracted API call sequence patterns from malwares in numerous classes that specialize in common malware functions and are collected in a signature database. And they compared with the API signatures which was collected. If they are matching, it was classified as a malware (Fig. 1).

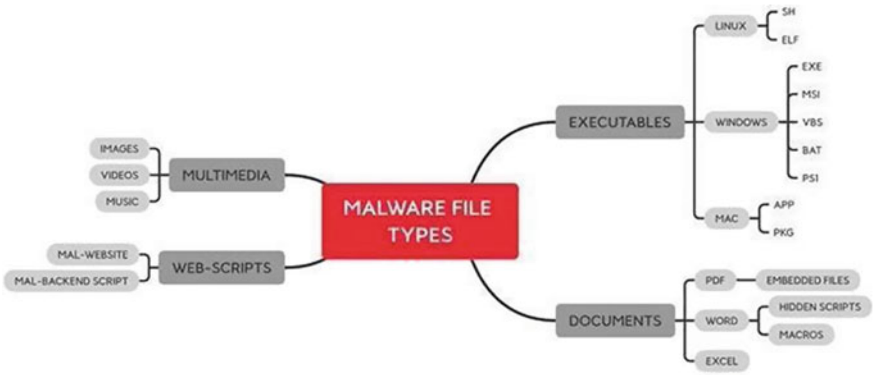


Fig. 1 Malware file types

3 Proposed Work

Nowadays, there are many methods to detect malwares but Clop is not detected using that methods. So, it is vital to find which one is best. In this paper, we aim to identify the most accurate machine learning algorithm for detecting Clop Ransom malware. Several machine learning algorithms used for malware detection are KNN, RNN, XGBoost, logistic regression, etc. These machine learning models learn on themselves and will predict whether the file is Clop or not.

Malware usually is executable. But, it can be a script or can be droppers that drop malware by other methods or can contain in files. There are many types of malware files available on various platforms. The three major platforms are Linux, Windows and Mac. These have extensions that are executable like Windows have EXE, Linux have SH and Mac have APP. Other extensions are documents, web scripts and multimedia. Most of them act as a payload or dropper. Among this, we are focusing on Clop since it is new.

3.1 Algorithms Used

- KNN: KNN is the super simple way to classify data and it is often called as Lazy algorithm. It is called lazy because ii memorize and doesn’t learns itself. It can be easily used when there is little or no prior of data. KNN is widely used in many malware detection research papers because KNN is based on feature similarity. So, our dataset has Clop which goes into our trained model and it predicts it is a Clop. We used this algorithm to detect Clop and made predictions on our test data. For evaluating algorithm, we used confusion matrix, precision, recall and f1 Score.

- Decision Tree: DT is a flowchart in which each node is a test. It can have both numerical and categorical data. The effort of users for data preparation is less. In this study, decision trees made complex trees which made it tough to generalize.
- Random Forest: Random forest is first proposed in 1995. In general, if there are more tress in forest, there are more predictions, and thus, it leads to higher accuracy. For classification using random forest, many decision trees are formed based on attributes and each tree gives a
- XGBoost: XGBoost is widely used because it can handle missing values. It also uses incremental training. This method is used for malware detection because it prevents overfitting and produces best accuracy.
- Logistic Regression: Regression is normally a predictive modeling technique. It takes the relationship of dependent and independent variables into account. When coming to logistic regression, it predicts results in binary format which is very useful. In logistic regression, there is a threshold value that shows the probability of winning and losing. So, it is very efficient in detecting malware.

3.2 Dataset

When we are working on machine learning or deep learning algorithms, the fundamental thing that we need is data. In the primary 1990s [12], there was a marvelous development in World Wide Web (WWW) which was directed to throw the multimedia content communication by a digital network. The most common multipedia is in the form of images, audio, video, numbers, text, etc. We have three datasets that consists of *n* number of entries that contain Clop, other malwares and benign. One of the datasets (Dataset A) has PE (portable executable) parameters which are produced from VirusTotal, VirusBay and kaggle. It is the most common format for any executables on windows. Every parts of executables startup with DOS stop which is an application, and its used to print the message. In the PE features, legitimate is a dependent variable, and all others are independent variables other (Dataset B) with images of malwares and benign. Generating image dataset is pretty simple, but it took a lot of time to run. Dataset C with malware API Information. The API sequences are extracted and with its help, and it will Clop with high accuracy.

3.3 Comparison Result

Using the five machine learning algorithms: KNN, XGBoost, random forest, logistic regression and decision tree classifier features of the dataset are compared and evaluated. For evaluation, a confusion matrix is used for obtaining the correct accuracy (Fig. 2).

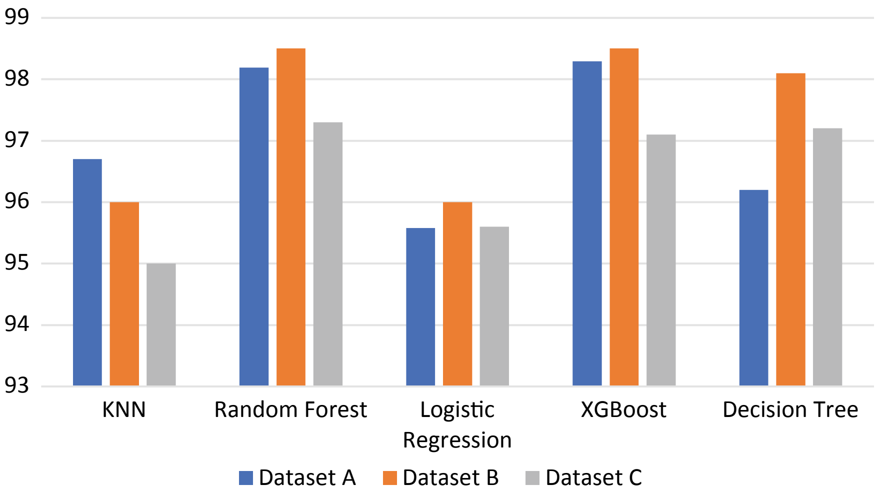


Fig. 2 Accuracy levels of machine learning mode

Accuracy can be improved by increasing the number of steps. The validation loss in XGBoost is much lesser than others. Based on our analysis, XGBoost performs better in predicting CLOT more accurately and efficiently.

4 Performance Evaluation Result and Analysis

- **XGBoost (Accuracy and loss are shown in the graph in Fig. 3)**
- **Others (Accuracy and loss are shown in the graph in Fig. 4).**

The result is followed in two phases training and testing using machine learning algorithms. The data set was randomly divided into two groups 70% of which were training sets and 30% of which were test sets. We achieved an accuracy of 99.80% by using XGBoost algorithm (as given in Fig. 3) and others less than 96 (as given

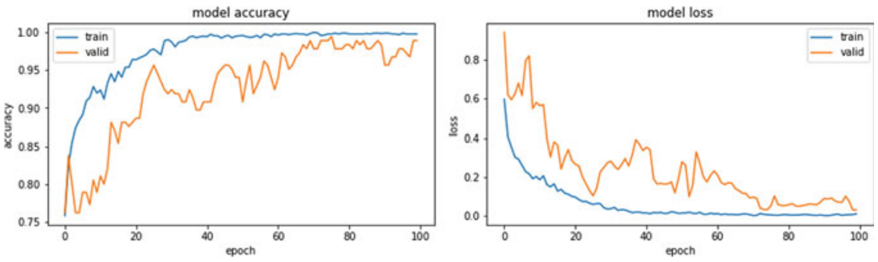


Fig. 3 Model accuracy and loss

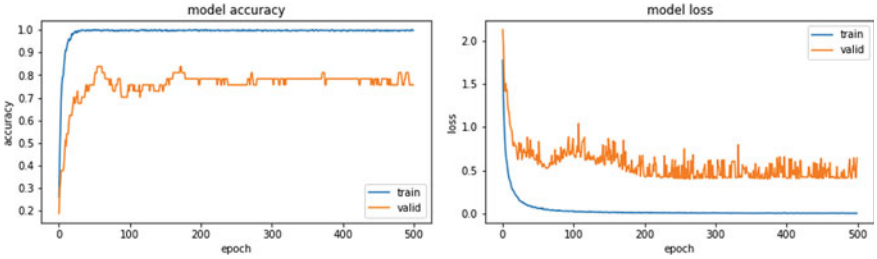


Fig. 4 Model accuracy and loss

Table 1 Results obtained

Algorithm	Dataset	Accuracy	Precision	Recall	F1 score
KNN	A	0.96726	0.977	0.977	0.97
	B	0.95622	0.969	0.980	0.97
	C	0.98100	0.971	0.979	0.97
Random forest	A	0.98197	0.981	0.992	0.98
	B	0.98100	0.987	0.987	0.98
	C	0.97651	0.982	0.986	0.98
Logistic regression	A	0.95588	0.944	0.992	0.96
	B	0.96111	0.955	0.995	0.96
	C	0.95001	0.940	0.991	0.96
XGBoost	A	0.98292	0.983	0.991	0.98
	B	0.98890	0.981	0.997	0.98
	C	0.98776	0.996	0.965	0.98
Decision tree	A	0.97438	0.951	0.997	0.96
	B	0.99110	0.931	0.991	0.96
	C	0.97777	0.965	0.96	

in Fig. 4). Based on the classified forms, matching is displayed on the output with accuracy. Accuracy can be improved by increasing the number of steps. By comparing all the algorithms, it is concluded that XGBoost performs better in predicting the CLOT ransomware (Table 1).

5 Conclusion

In this study, we compared by studying and implementing a script used for data extraction from the PE-files that created a dataset with infected and clean files, images of malwares and benign and API information of malwares, that trained using machine learning algorithms: KNN, XGBoost, random forest, logistic regression. KNN is slow compared to others and also requires high memory. When compared to other algorithms, decision tree is unbalanced and is often inaccurate. Random forest is also slow and ineffective. So, among algorithms, XGBoost is best in all terms. The results show that XGBoost has provided the most effective and attainable classification accuracy rate when compared with others. Future work is feasible, and it can include additional features and different algorithms.

Acknowledgments We would like to extend our deepest gratitude to all those who have directly or indirectly helped us in completing this paper. We want to thank our guide Leena Vishnu Namboothiri for providing adequate information with constant guidance throughout our research, and it helped us a lot to complete our research work successfully.

References

1. Karunakaran, P.: Deep learning approach to DGA classification for effective cyber security (2020)
2. Gavriluţ, D., Cimpoesu, M., Anton, D., Ciortuz, L.: Malware detection using machine learning. In: International Multi Conference on Computer Science and Information Technology. IMCSIT'09 (2009)
3. Xu, Z., Ray, S., Subramanyan, P.: Malware detection using machine learning based analysis of virtual memory access patterns. IEEE Explore (2017)
4. Planes, J., Mateu, C., Gibert, D.: The rise of machine learning for detection and classification of malware: research developments, trends and challenges. J. Netw. Comput. Appl. (2019)
5. Kim, K., et al.: A survey on malware detection using deep learning methods. In: Network Intrusion Detection Using Deep Learning
6. Sami, A., Yadegari, B., Peiravian, N., Hashemi, S., Hamze, A.: Malware detection based on mining API calls (2010)
7. Kolosnjaji, B., Zarras, A., Webster, G., Eckert, C., Bai, Q.: Deep learning for classification of malware system call sequences
8. Ficco, M.: Comparing API call sequence algorithms for malware detection
9. Pascanu, R., Stokes, J.W., Sanossian, H., Marinescu, M., Thomas, A.: Malware classification with recurrent networks
10. Sundarkumar, G.G., Ravi, V., Nwogu, I., Govindaraju, V.: Malware detection via API calls, topic models and machine learning (2015)
11. Ki, Y., Kim, E., Kim, H.K.: A novel approach to detect malware based on API call sequence analysis
12. Chitra, K., Prasanna Venkatesan, V.: An antiquity to the contemporary of secret sharing scheme (2020)

Integration of Wireless Sensors to Detect Fluid Leaks in Industries

N. Santhosh, V. A. Vishanth, Y. Palaniappan, V. Rohith, and M. Ganesan

Abstract The industrial internet of things is a specific domain that deals with industrial machines and their communications. This allows us to bring better reliability and efficiency in the work operations. To increase the reliability in the system, suitable error detection methods should be embedded in the system process. Design implementation and testing for the pipe leak detection are done in this paper. In this work, a prototype is created with small wireless nodes distributed along the pipelines, pressure sensors are installed next to the manually created leak nodes. The leak is detected using the pressure point analysis method. Various pressure data are obtained from the fluid flow networks in the pipes, and its variations are analyzed using Bernoulli's equation, to monitor leaks in pipes.

Keywords Pipeline leak detection · Pressure point analysis · Water distribution systems

1 Introduction

In India, industrial accidents claim over 6300 lives dead and over 51,000 injured between 2014 and 2019. The main cause for this is not identifying the problem in the particular setup much earlier, i.e., when the problem is at starting stage, like when the temperature in a particular tank or furnace rises or a leak in the gas pipeline, and there is a chance of fire and identifying these in earlier stage can avoid the hazard that could happen. Nowadays, modern industries are demanding more sophisticated instruments for monitoring and control of risk parameters in the hazardous area.

A pipeline mostly will be of two types, a straight pipeline and an elbow or a bend. The pipe elbows commonly include 45°, 90° and 180° bends. In most cases, elbows are most susceptible to corrosion or erosion than the straight pipes.

The two main components that are most notable in a flow and are responsible for leaks are the velocity and the pressure developed in the pipes. The velocity of the

N. Santhosh (✉) · V. A. Vishanth · Y. Palaniappan · V. Rohith · M. Ganesan (✉)
Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore 641112, India
e-mail: m_ganesan1@cb.amrita.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_46

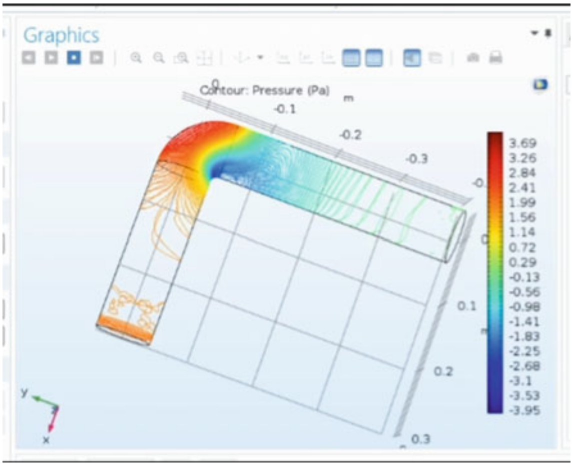


Fig. 1 Pressure developed by the fluid

flow is always greater in a straight pipeline than the velocity on the elbows. When a fluid flows through a pipeline, it is observed that mostly the pressure developed along the elbow is more when compared to the straight pipelines. It is also observed that the pressure on the outer curvature of the pipe is much greater than that on the inner curvature of the pipe. Considering the factors like corrosion, erosion, velocity and pressure developed along the elbows are more vulnerable to leaks than the straight pipelines (Fig. 1).

The common leak detection methods used are the acoustic noise correlation method and the pressure point analysis. In the acoustic noise correlation method, the sound or the noise of the escaping fluid is used to detect the leaks. But, it is also said that the sound greatly attenuates as the distance from the leak increases, especially in the larger pipes. Also, the noise from the surrounding can interfere with the leak noise, claiming false alarm.

In the pressure point analysis method, the pressure of the fluid is continuously monitored. When a leak is present, it produces a sudden change in the pressure. This method is mostly used in gas and oil pipeline systems, and unlike acoustic methods, it is not affected by the surrounding environment.

2 Literature Survey

Dehkordi et al. did a survey on data integration techniques in IoT sensor networks. In that, they observed that there is an increase in interest in wireless sensor technologies in various cases of internet of things. Since there is a huge growth in smart objects and its application, the need for analyzing the data of these objects is a huge challenge these days. The sensor node neglects the needless information in the received data in the neighboring node before transferring the final data to the central station. The rapid growth in RFID systems, NFC and Wi-Fi communications has contributed to the growth of IoT systems. Lots of researchers are working on solving the problems faced in smart cities, smart factories, etc. The challenges faced by international union of communication are issues regarding inconsistency and data security [1].

Murvay and team did a survey on gas leak detection and localization techniques. Gas leakage causes a huge loss both in financial manner and injuries caused to the humans. To prevent these losses, efforts were taken in development of reliable techniques used for gas leakage. In this paper, it speaks about the localization methods because just by detecting the leakage is not enough to take any immediate actions. That is why many techniques for common leakage techniques such as hardware-based methods, non-technical methods and software-based methods are analyzed and compared with one another based on its performance level. The performance level is determined based on some criteria like ability to find the localization of the leak, the speed at which it detects and estimating the size of the leak. From these several techniques and solutions, it is important to go with the technique that is well suited for that particular system [2].

Deshmukh et al. developed a wireless sensor network that automatically detects gas leak in the presence of air. Lab VIEW Software in virtual instrument software architecture (VISA) configuration is used for monitoring the gas sensor wirelessly. A wireless sensor network is developed using the IEEE 802.15.4 standard and a low-power Xbee module. A prototype was designed by employing 4 nodes based on an Atmega 328 microcontroller with an IEEE 802.15.4 standard Zigbee module for wireless communication. A coordinator node is designed to collect data from all the nodes equipped with the MQ-2 gas sensor. Through USB connection the coordinator node is connected to the PC, the data is managed in the Lab VIEW environment. During a gas leak, the sensor system detects and sends SMS to the inmates and activates the alarm. A switch to a solenoid is also activated which disables the gas flow [3].

Lee et al. have discussed the different hazards that happen in an electrical laboratory. The following functions are performed by a wireless sensor network connected with Arduino: Firstly, when hazardous readings are logged by the sensors, safety actions are taken. Secondly, logged data can be viewed by the client through computers and mobile phones. Thirdly, control actions can be done using actuators and controllers in a remote way. In the proposed system, Wi-Fi module is used for transmitting the data from sensor to the base station through routers as Wi-Fi has long-range transmitting capability and the transmitting speed is high when compared to Bluetooth. Moreover, a mobile app was developed to remotely monitor the system. The six sensors used were DS18B20 sensor (for Temperature), BMP280 sensor (for air pressure), HR202 sensor (for humidity), current sensor (using Hall Effect Principle), Smoke sensor and IR flame sensor. The firebase of Google was used as the cloud database in the proposed system [4].

Zope and team designed an IoT sensor and deep neural network-based wildlife prediction system. As wildfires may cause significant damage to animals, human lives and other properties. Wildfires are dangerous as they cause more damage in a short span of time. Wildfires which are caused naturally can be predicted by factors like temperature, humidity, pressure, etc. Here, they've used machine learning by operational monitoring a region, and the changes in climate are detected by the use of sensor. Wiprespy predicts the intensity of the fire in real-time data by monitoring and recording the climatic conditions. In machine learning, a simple logistic regression is used in the problem with the use of dense neural network created by using Keras API where the IoT sensors give the input. The factors like temperature, humidity are measured with sensors which act like a real-time input for prediction. The values are frequently updated and stored in cloud. By the prediction of real-time data, we can be ahead in preserving the environment [5].

Chraim et al. designed a system that detects gas leaks. Gas sensors were placed broadly all over the sensitive part of an industry. Data from the sensors are connected in a mesh format using Zigbee protocol. The sensor data is continuously tabulated in a single storage device. Various localization algorithms and detection algorithms are applied on it. The sensor data is collected in a cyclic manner and analyzed accordingly. If the data after processing exceeds the threshold limit over a period of time, alert messages are sent [3].

Murvay et al. classified various technologies in hazard prevention. Leak detection technologies can be classified into automated detection, semi-automated detection and manual detection. Many techniques for common leakage techniques such as hardware-based methods, non-technical methods and software-based methods are analyzed and compared with one another based on its performance level. Some of the detection techniques discussed were acoustic, visual, flow difference and mass balance. The performance level is determined based on some criteria like ability to find the localization of the leak, the speed at which it detects and estimating the size of the leak. But, it is important to go with the technique based on what that particular system needs. It is important to go with the technique that is well suited for that particular system [2].

Stoainov et al. induced leaks in two locations on a pipe and with two sensors placed in the pipe. The data traces are taken to measure the effectiveness of leak detection. They are divided into non-overlapping segments. Each segment is labeled as leak or no leak, and one of them is selected as base segment and the other one will be compared against it. Decision tree classifier is used to find the best linear separators between the leak and no-leak training data set and used this to predict the leak and no-leak values of the test data. It correctly classifies 87% of the test data. Same data set is used for leak localization, but leak segments are cross-correlated. Error produced is 0.2 m on average [6].

Pasha briefed about ThingSpeak web service which acts as a host for various sensors to monitor the sensed data at cloud level. Arduino UNO board, ESP8266 Wi-Fi Module helps to process and transfer the sensed data to the ThingSpeak cloud. Sensing and monitoring operation is done by loading sensor libraries in Arduino IDE. The program is executed, and the sensed data is visualized. Enter the network credentials in IDE and execute the program one more time and visualize the output in ThingSpeak cloud. The sensed data is uploaded to the MATLAB R2016a using channel ID and the read API key assigned by services [7].

Miry designed water quality monitoring and analytics based on ThingSpeak. Using sensor fusion techniques TDS and Turbidity data is acquired and uploaded to the ThingSpeak platform which monitors and analyzes the data. The sensors are connected to Arduino and the data from it is uploaded to ThingSpeak. The data collected is analyzed using different codes of MATLAB. When the system identifies any abnormal data from the sensors alert messages are triggered. The user will receive a warning through mail [8].

3 Proposed Method

3.1 Methodology

Real-time water leaks in pipes can be detected using real-time data from the sensors. ThingSpeak and MATLAB are the most popular platforms to analyze real-time data. The collected data is then uploaded to the ThingSpeak using a Wi-Fi module (ESP 8266) for visualization and study purposes.

Continuous-time data from sensors are logged and from the uploaded data various calculations and operations can be performed. The pressure difference between the pressure sensors (SKU237545) is calculated and logged in a different field. It basically checks for the pressure difference between the nodes. If a pressure difference is detected, it triggers a mail stating a leak is present at the particular site (Fig. 2).

3.2 Theoretical Calculation

The flow in the pipeline can be calculated theoretically by using the Bernoulli equation. The Bernoulli equation represents the energy conservation principle for steady-state fluid flow systems. Considering the two points before and at the constriction, we have the conservation in pressure energy, kinetic energy and potential energy in a steady-state flow expressed as Eq. (1):

$$p_1 + 1/2\,\rho v_1^2 + \rho g h_1 = p_2 + 1/2\,\rho v_2^2 + \rho g h_2 \tag{1}$$

where p is pressure, ρ is density of fluid, g is gravitational constant, h is elevation from ground and v denotes velocity of flowing fluid.

The pipe is horizontal, so both points are at the same height. Bernoulli's equation can be simplified in this case, i.e., $\mathbf{A_1} = \mathbf{A_2}$, $\mathbf{h_1} = \mathbf{h_2}$. This is expressed in Eq. (2):

$$A_1 v_1 = A_2 v_2 \tag{2}$$

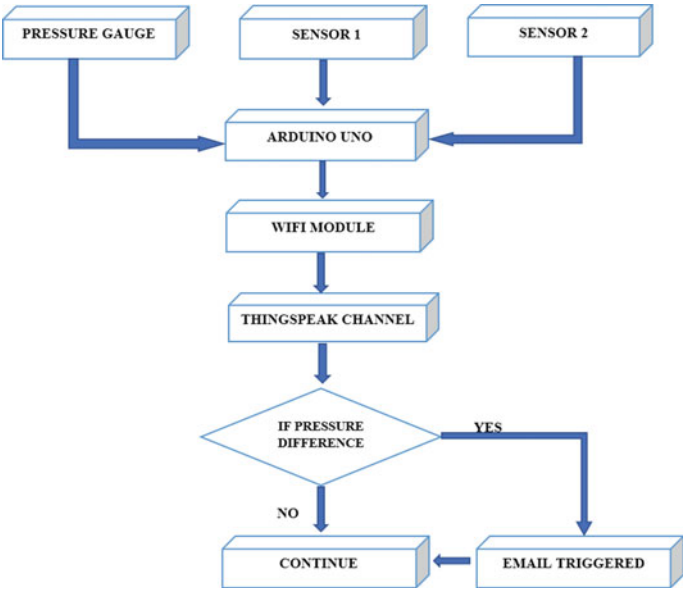


Fig. 2 Workflow diagram

When the fluid starts to flow initially from the pump to the pipe, there would be a difference in height, so the equation for the segment can be simplified as: $p_1 > p_2$ since $h_2 > h_1$.

When there is a leak, total output is the same as total input. So, the equation can be simplified as the summation of products of velocities of fluid and area of the pipe or the leak. This is expressed in Eq. (3):

$$\sum(\Delta v/\Delta T_{\text{in}}) = \sum(\Delta v/\Delta T_{\text{out}}) \tag{3}$$

$$A_1 v_1 = A_2 v_2 + A_3 v_3 \tag{4}$$

where A is the area of pipe, v is the velocity of fluid, h is the height at which liquid is flowing, p_1 and p_2 are pressure at different points. This is expressed in Eq. (4).

3.3 Pipeline Model

A simple pipeline network was built to implement the pressure point analysis technique for detection of leaks. The fluid (water in this case) was allowed to flow through the main pipeline and leaks are created manually using nodes and gate valves. Before and after the leak nodes pressure sensors (SKU237545) are introduced to read the

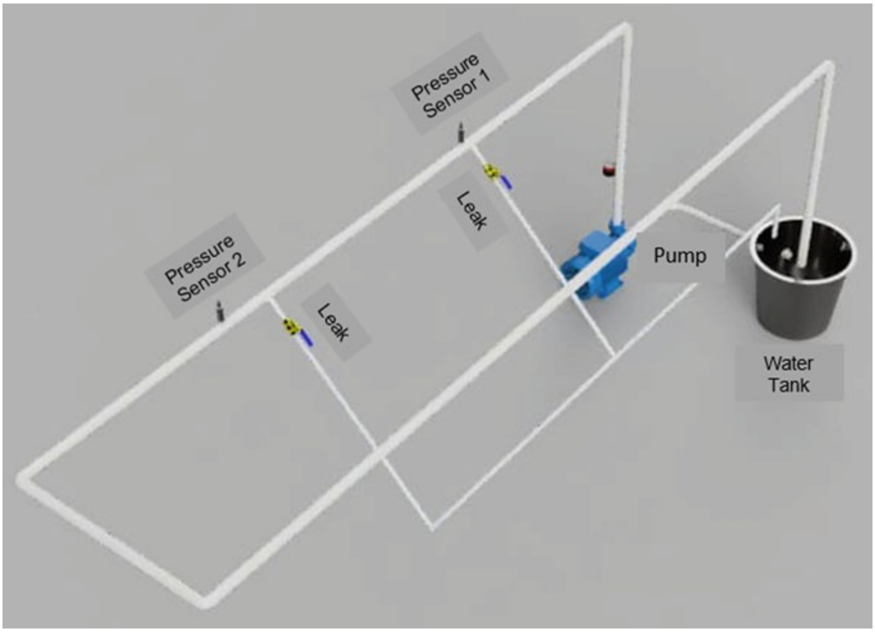


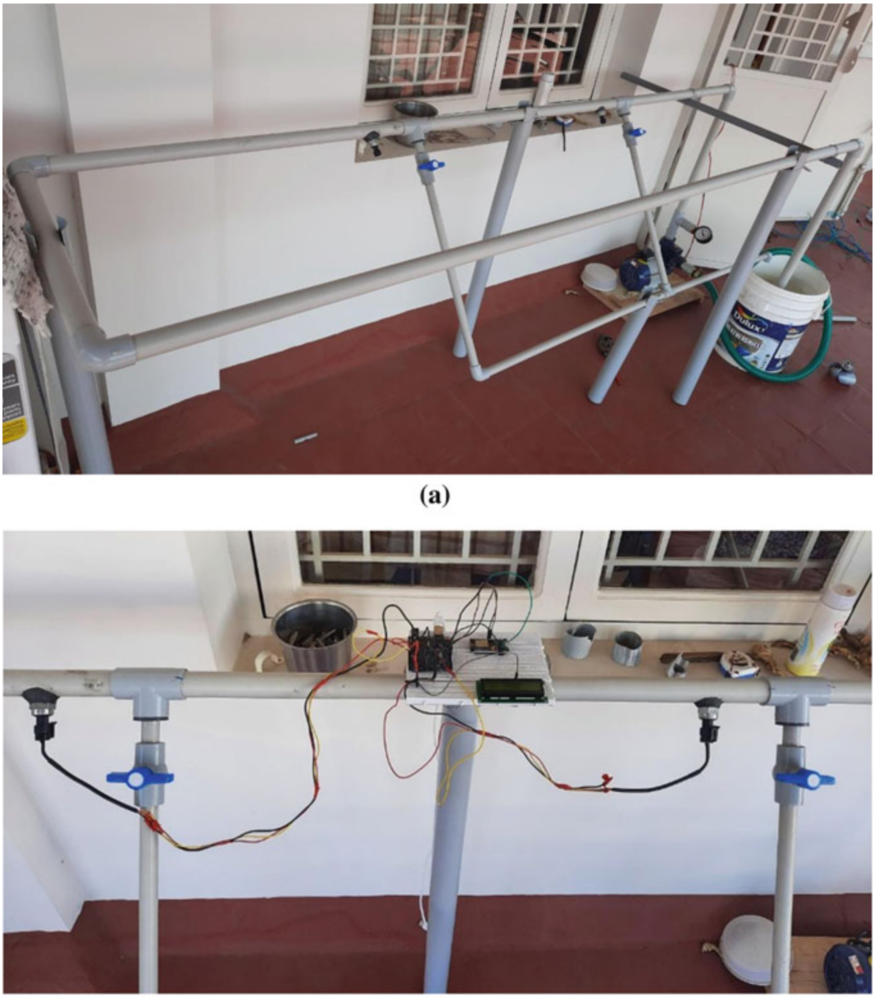
Fig. 3 3D model of the setup

pressure data. Gate valves are attached to the leak pipes on the nodes to control the proportion of the leak. By opening the valve, a sudden change in pressure (pressure difference) is observed by the pressure sensor (Fig. 3).

The sensors give the real-time pressure at a particular point exerted by the fluid. Using Bernoulli's equation, flow of the fluid is calculated at a particular point on the pipe. Separate leak points have also been made at different distances in the pipe. During leakage, there will be a variation in pressure across the pressure sensors. The obtained pressure data is then uploaded to the ThingSpeak channel using the Wi-Fi module (Fig. 4).

3.4 IoT (ThingSpeak)

ThingSpeak platform is a cloud channel. It provides a platform to quickly collect and analyze data collected from sensors. It provides instant visualization in real time. It has apps to visualize, manipulate and trigger actions on the data. Online data analysis and data processing can be done in real time. Automatically, act or react on the real-time data and interface with third-party applications.



(a)

(b)

Fig. 4 **a**, **b** Built pipeline model and circuit

4 Result

Pressure value or pressure data on the sensor is noted for each degree of valve opening and plotted (Fig. 5). Significant amount of pressure drop is observed for each degree of valve opening. When the valve opening is greater, the pressure drop on the sensor would also be greater. When the valve is in fully open condition (90°), the pressure drop goes approximately to 0 psi.

Appreciable amount of pressure drop is observed for each degree of valve opening in the leak node 2. When the valve is in fully open condition (90°), the pressure drop

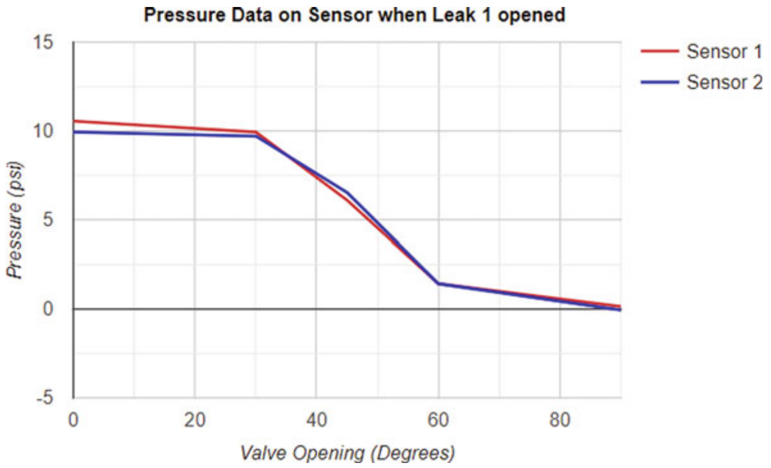


Fig. 5 Pressure data on sensors when leak 1 is opened

goes nearly to 0 psi. A notable amount of pressure difference is seen between the pressure readings on the two sensors (Fig. 6).

Both the valves are opened simultaneously at the same degrees, and the pressure data was plotted (Fig. 7). Considerable pressure drop is observed for each degree of valve opening. It is seen that at approximately 75° of valve opening, and the pressure drop on both sensors becomes 0 psi as all the water flows through the leak itself (Figs. 8 and 9).

React condition allows to trigger a ThingHTTP request or ThingTweet request when the condition is met, and here, the condition used is in numeric form as the

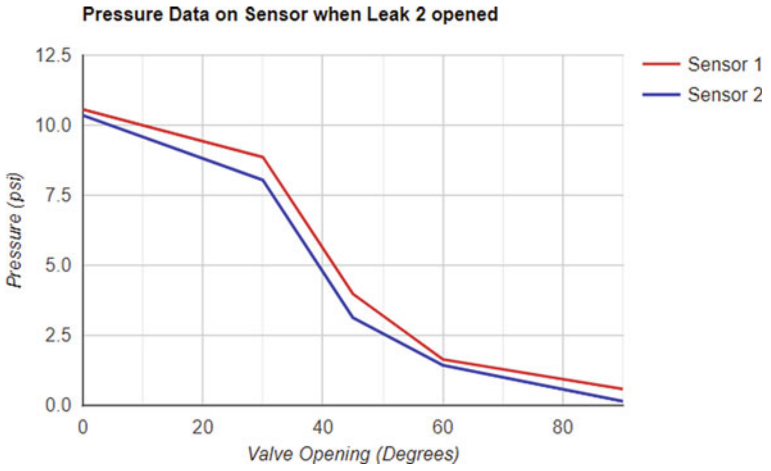


Fig. 6 Pressure data on sensor when leak 2 is opened

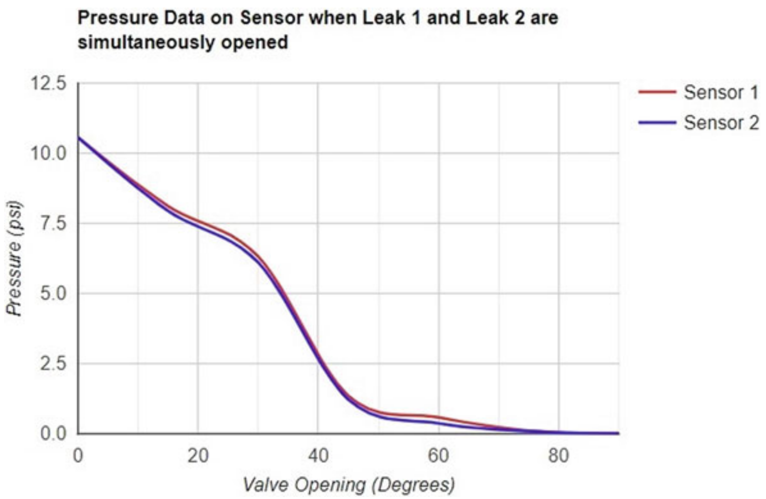


Fig. 7 Pressure data on sensor when leak 1 and leak 2 are simultaneously opened

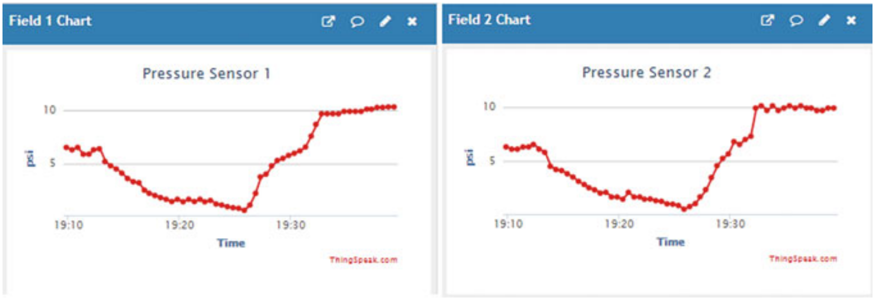
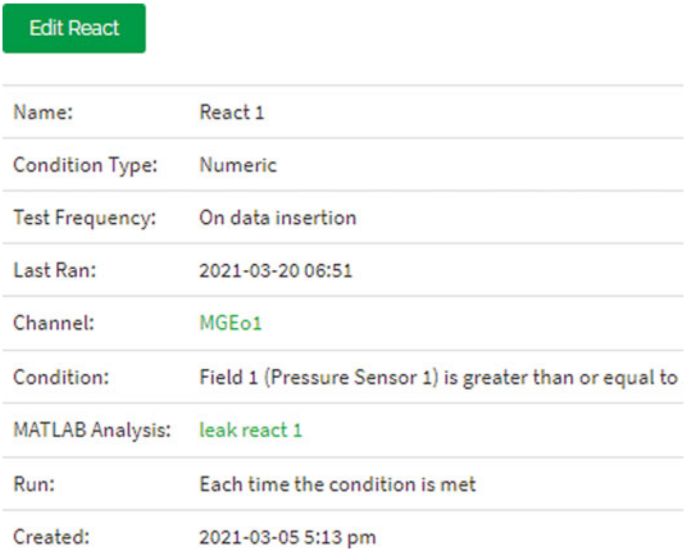


Fig. 8 Pressure data uploaded in ThingSpeak channel

data are values/readings from sensors. The frequency of this test condition is set to check each data after insertion and also made to trigger the request each time the condition is met.

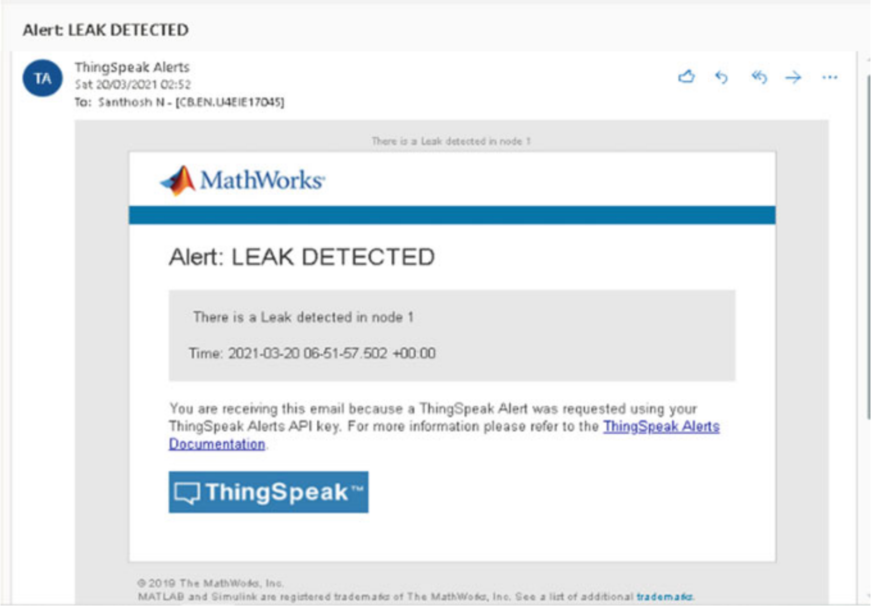
5 Conclusion

The above discussed wireless sensor network can detect leaks (Fig. 10) by monitoring the entire pipeline system continuously. Leaks can be corrected by monitoring the pressure sensor inside the pipe. The pressure sensor data collected is uploaded to ThingSpeak channel. The nodes periodically report the measured pressure signals through a wireless sensor network.



Name:	React 1
Condition Type:	Numeric
Test Frequency:	On data insertion
Last Ran:	2021-03-20 06:51
Channel:	MGEo1
Condition:	Field 1 (Pressure Sensor 1) is greater than or equal to
MATLAB Analysis:	leak react 1
Run:	Each time the condition is met
Created:	2021-03-05 5:13 pm

Fig. 9 Settings for the email to be triggered


Alert: LEAK DETECTED

ThingSpeak Alerts
Sat 20/03/2021 02:52
To: Santhosh N - [CB.EN.U4EIE17045]

There is a Leak detected in node 1

Alert: LEAK DETECTED

There is a Leak detected in node 1

Time: 2021-03-20 06-51-57.502 +00:00

You are receiving this email because a ThingSpeak Alert was requested using your ThingSpeak Alerts API key. For more information please refer to the ThingSpeak Alerts Documentation.

© 2019 The MathWorks, Inc.
MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See a list of additional trademarks.

Fig. 10 Alert message

In future, an array of sensors can be placed in the pipeline and integrated to detect leaks. The exact location of the leak can be found using node localization algorithm and can be transferred efficiently. Machine learning techniques can also be employed to detect leaks even before leak has been created. Also, this method and be implemented and checked for various fluids.

References

1. Dehkordi, S.A., Farajzadeh, K., Rezazadeh, J., Farahbakhsh, R., Sandrasegaran, K., Dehkordi, M.A.: A survey on data aggregation techniques in IoT sensor networks. Wireless Netw. **26**(2), 1243–1263 (2020)
2. Murvay, P.-S., Silea, I.: A survey on gas leak detection and localization techniques. J. Loss Prev. Process Ind. **25**(6), 966–973 (2012)
3. Chraim, F., Erol, Y.B., Pister, K.: Wireless gas leak detection and localization. IEEE Trans. Ind. Inf. **12**(2), 768–779 (2015)
4. Deshmukh, L.P., Mujawar, T.H., Kasbe, M.S., Mule, S.S., Akhtar, J., Maldar, N.N.: A LabVIEW based remote monitoring and controlling of wireless sensor nodes for LPG gas leakage detection. In: 2016 International Symposium on Electronics and Smart Devices (ISESD), pp. 115–120. IEEE (2016)
5. Zope, V., Dadlani, T., Matai, A., Tembhurnikar, P., Kalani, R.: IoT sensor and deep neural network based wildfire prediction system. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 205–208. IEEE (2020)
6. Stoainov, I., Nachman, L., Tokomouline, S.M.T.: A wireless sensor network for pipeline monitoring
7. Pasha, S.: ThingSpeak based sensing and monitoring system for IoT with Matlab analysis. Int. J. New Technol. Res. **2**(6) (2016). ISSN: 2454-4116
8. Miry, A.H., Aramice, G.A.: Water monitoring and analytic based ThingSpeak. Int. J. Electr. Comput. Eng. (2020)

Performance Analysis of Abstract-Based Classification of Medical Journals Using Machine Learning Techniques

A. Deepika and N. Radha

Abstract Researchers face many challenges in finding the opt web-based resources by giving the queries based on keyword search. Due to advent of Internet, there are huge biological literatures that are deposited in the medical database repository in recent years. Nowadays, as many web-based medical researchers evolved in the field of medicine, there is need for an intelligent and efficient extraction technique required to filter appropriate and opt literature from the growing body of biomedical literature repository. In this research work, new combination of model is proposed in order to find the new insights in applying the combination of algorithm on biological data set. The information in the biomedical field is the basic information for healthy living. National Center for Biotechnology Information (NCBI)'s PubMed is the major source of peer-reviewed biomedical documents for researchers and health practitioners in the field of health-related management. In this paper, abstracts available in PubMed database is used for experimentation. In recent years, deep learning-based neural approach models provide an efficient way to create an end-to-end model that can accurately measure classification labels. This research work is a systematic analysis of performance of the supervised learning models such as Naïve Bayes (NB), support vector machine (SVM) and long short-term memory (LSTM) by implementing on textual medical data. The novelty in this work is the process of incorporating certain topic modelling techniques after the pre-processing phase to automatically label the documents. Topic modelling is a useful technique in increasing the efficiency and improves the ability of researchers to interpret biological information. So, the classification algorithms thus proposed are implemented in combination with popular topic modelling algorithms such as latent Dirichlet algorithm (LDA) and non-negative matrix factorization (NMF). The final performance of the combination of algorithms is also analysed and is found that SVM with NMF outperforms the other models.

A. Deepika (✉)
Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India

N. Radha
Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_47

Keywords Text classification · Cancer · Naïve Bayes · Support vector machine · LSTM · Latent Dirichlet algorithm · Non-negative matrix factorization · Topic modelling

1 Introduction

Non-communicable diseases (NCDs) are now responsible for a lot of death globally, and cancer is expected to also be the major cause of death and the single most important obstacle to increase life expectancy around the world [1]. Worldwide, an estimated 19.3 million new cancer cases and 10.0 million cancer deaths occurred in 2020. For both sexes combined, half of all cases and 58.3% of cancer deaths are expected to occur in Asia in 2020, where 59.5% of the world's population lives. Europe accounts for 22.8% of all cancer cases and 19.6% of all cancer deaths, while it accounts for 9.7% of the world's population, followed by 20.9% incidence in America and 14.2% mortality worldwide [2]. Due to the increase of cancer in twentieth century, in 2021, 1,898,160 newer cases and 608,570 cancer deaths are expected to occur in United States (USA). After a long time, increasing cancer mortalities now started decreasing. As per the statistics the total decline by 30% is because of the early finding of decrease in smoking, technology improvement in early detection of cancer and its treatment [3].

Importantly, lifestyle changes are playing a vital role in preventing breast cancer among women. Physical activity, smoking, alcohol consumption and mineral or vitamin usage are among the risk factors that can cause breast cancer for women. If we alter the factors, women can be able to reduce their risk of getting breast cancer [4].

Not only cancer-related information but also many important basic life information are explained in different types of medical documents. Such texts are an extremely rich source of knowledge, but because of their abstract form, extracting insights from them can be difficult and time consuming. So, society is in need of new computing tools to better organize, browse and interpret the large amounts of data.

Text classification using machine learning algorithms can help to automatically, easily, cost effectively to manage the processes and improve data-driven decisions. In many applications, text classification is an important element in the management of unstructured textual content, such as Internet search, retrieval of information and sentiment analysis, emotion analysis. It has therefore gained substantial interest for research experts. This is the process by which the text is classified with specific tags or categories which has the connection with its content. Many models in text analysis follow the bag-of-words model, where all the words in a document are considered and the relationship among word is neglected. Due to this drawback and poor representation of text, they yield unsatisfactory results. Recently, new and efficient methods of statistical models called topic models became popular in the field of text analysis and text classification.

So, on thinking of the two somewhat similar concepts, text classification and topic modelling can yield better efficiency if they are applied together. To check the combination efficiency of them, topic modelling is implemented first in order to label the biological documents and then text classification is employed to classify the labelled text. The efficiency of each and every algorithm in both concepts is specifically evaluated and compared accordingly.

This topic modelling is different from conventional text classification methods and bag-of-words methods that classify the text based on the frequency of a word. It can be classified as an unsupervised approach used for identifying the topic in document and classifying the large texts to one topic [5]. So, in this research, the combination of classification and topic modelling algorithms is then used as hybrid approach. These works are done by using the advanced Python language.

The development in the scientific and technology facilitates researchers to explore new insights in many fields. Due to its growing nature it is hard even for medical researchers to access the deposited precious information as journals. For this research work, cancer data downloaded from PubMed website is chosen for the experimentation as this disease threatening the whole world by its incurable nature.

2 Related Work

The research work has been done in this topic modelling and text classification is analysed, and algorithms and their results also discussed in this section.

Wang Haoxing proposed a research work using recurrent neural network to do emotional analysis on the social media content which often prone with fake news or bogus details. The work is done on emotional perspective, and the fake messages are compared with original messages and false messages [6].

Ayushi Mitra implemented hybrid approach of machine learning and lexicon-based methods for sentiment analysis on movie review data set classification [7].

Harjule et al. implemented multinomial Naive Bayes, logistic regression, support vector machine and RNN on already categorized data sets to perform sentiment analysis. In this paper, seperate topic modelling algorithm is applied to automatically categorize the data set, and further classification algorithms are implemented [8].

Curiskis et al. compared many clustering algorithm with topic modelling algorithms on reddits and twitter data sets. In our work, topic modelling and classification are mainly concentrated on biological data set [9].

Jedrzejowicz et al. applied topic modelling algorithm latent Dirichlet algorithm only for topic modelling and compared with LDA with Word2vec. In our work, it is extended to compare two topic modelling algorithms LDA with non-negative matrix factorizarion [10].

Luo et al. build the three-layer Bayesian model with convolutional neural network on the LDA generated text to formulate new laws. This concept of combining isextended and two algorithms are implemented and new findings are tried in using biological data set [11].

Ti Do et al. employed deep neural networks to achieve classification on protein using biological subwords to detect protein S-sulphenylation. This also motivated to perform this biological text classification [12].

Jang et al. proposed BI − LSTM + CNN to increase the classification accuracy for text. The hybrid BI + CNN + LSTM model produces high accurate results [13].

Venkataraman et al. examined LSTM and RNN on large amount of clinical records on human and veterinary data stores to automatically assign some clinical codes to the records. From that study, LSTM model performed well and provides accuracy slightly higher than the other algorithms proposed [14].

Alaa M. El-Halees implemented NB algorithm on Arabic web data for text classification, and it showed the average accuracy of 68.78% but the topic modelling algorithms combined with text classification algorithms to provide better accuracy [15].

Akbani et al., UCI data set, a non-text data set, believe that the importance given to the text classification on medical data set can benefit all scientists as well as researchers in this field [16].

Chau et al. implemented SVM successfully for text classification, and the supervised learning on the strategies proposed in this paper can improve the classifiers in SVM algorithm [17].

Mccallum et al. used the Naïve Bayes algorithm's probabilistic model, which depends primarily on context data described by the existence of specific terms in a known class text. The algorithm is experimented on Reuter's documents as a data set and contrasted the multinomial model and multivariate Bernoulli model with two kinds of Naïve Bayes approaches found out that the multinomial NB model outperforms than the Bernoulli multivariate model. This multinomial model is implmented in our medical data set [18].

Abdelwadood Mesleh implemented NB, SVM and KNN with chi square on 1445 text documents data set that are collected from Arabic newspapers. Then the documents are classified into nine different news topics, as we have topic modelling algorithms applied in combination with classification algorithm to classify the data set to classify into topics [19].

Luo processed long sentences or texts using RNN model-based LSTM and gated recurrent unit (GRT) was proposed. The model can memorize the relationship of long-distanced dependence and keep main semantice information [20].

Thorsten Joachims implemented support vector machine on Reuters-21578 data set and WebKB collection of WWW pages. And it outperformed other machine learning algorithms like Naïve Bayes, decision trees and KNN with good accuracy [21].

3 Proposed Methodology

The proposed methodology works as the following workflow diagram (Fig. 1).

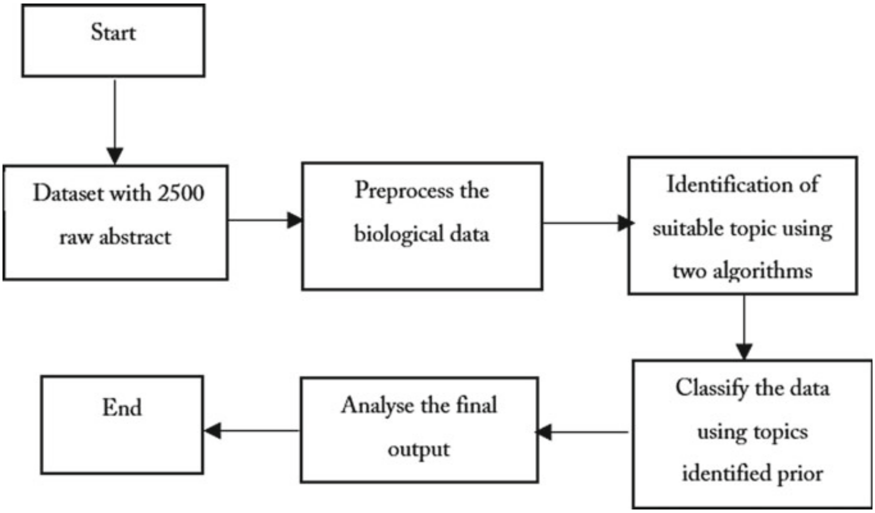
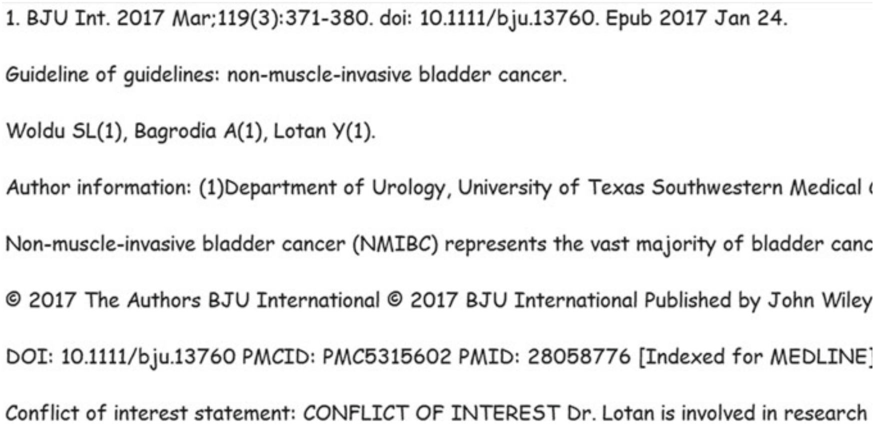

Start
Dataset with 2500 raw abstract
Preprocess the biological data
Identification of suitable topic using two algorithms
Classify the data using topics identified prior
Analyse the final output
End

Fig. 1 Raw data set downloaded from PubMed website

3.1 Data set

Abstracts are downloaded from PubMed website and data set is created. It consists of 2500 medical abstracts which belong to five different categories of cancer. It is shown in Fig. 2. All of the documents somehow semantically related to one another. The downloaded text format data is then converted to comma separated value (csv) format using Pandas library and the fields that are collected from PubMed are shown in Fig. 3. From that abstract field is used in this work.



1. BJU Int. 2017 Mar;119(3):371-380. doi: 10.1111/bju.13760. Epub 2017 Jan 24.

Guideline of guidelines: non-muscle-invasive bladder cancer.

Woldu SL(1), Bagrodia A(1), Lotan Y(1).

Author information: (1)Department of Urology, University of Texas Southwestern Medical

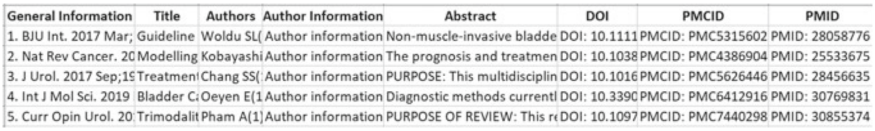
Non-muscle-invasive bladder cancer (NMIBC) represents the vast majority of bladder canc

© 2017 The Authors BJU International © 2017 BJU International Published by John Wiley

DOI: 10.1111/bju.13760 PMCID: PMC5315602 PMID: 28058776 [Indexed for MEDLINE]

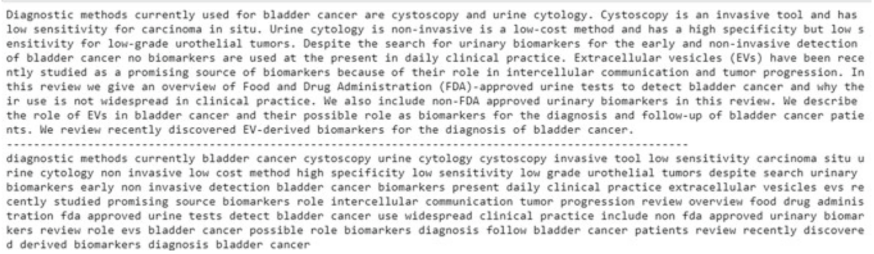
Conflict of interest statement: CONFLICT OF INTEREST Dr. Lotan is involved in research

Fig. 2 Raw data set downloaded from PubMed website



General Information	Title	Authors	Author Information	Abstract	DOI	PMCID	PMID
1. BJU Int. 2017 Mar;	Guideline	Woldu SL(	Author information	Non-muscle-invasive bladde	DOI: 10.1111	PMCID: PMC5315602	PMID: 28058776
2. Nat Rev Cancer. 20	Modelling	Kobayashi	Author information	The prognosis and treatmen	DOI: 10.1038	PMCID: PMC4386904	PMID: 25533675
3. J Urol. 2017 Sep;19	Treatmen	Chang SS(	Author information	PURPOSE: This multidisciplin	DOI: 10.1016	PMCID: PMC5626446	PMID: 28456635
4. Int J Mol Sci. 2019	Bladder Ca	Oeyen E(1	Author information	Diagnostic methods currentl	DOI: 10.3390	PMCID: PMC6412916	PMID: 30769831
5. Curr Opin Urol. 20	Trimodalit	Pham A(1)	Author information	PURPOSE OF REVIEW: This re	DOI: 10.1097	PMCID: PMC7440298	PMID: 30855374

Fig. 3 Modified data set in CSV format



```
Diagnostic methods currently used for bladder cancer are cystoscopy and urine cytology. Cystoscopy is an invasive tool and has
low sensitivity for carcinoma in situ. Urine cytology is non-invasive is a low-cost method and has a high specificity but low s
ensitivity for low-grade urothelial tumors. Despite the search for urinary biomarkers for the early and non-invasive detection
of bladder cancer no biomarkers are used at the present in daily clinical practice. Extracellular vesicles (EVs) have been rece
ntly studied as a promising source of biomarkers because of their role in intercellular communication and tumor progression. In
this review we give an overview of Food and Drug Administration (FDA)-approved urine tests to detect bladder cancer and why the
ir use is not widespread in clinical practice. We also include non-FDA approved urinary biomarkers in this review. We describe
the role of EVs in bladder cancer and their possible role as biomarkers for the diagnosis and follow-up of bladder cancer patie
nts. We review recently discovered EV-derived biomarkers for the diagnosis of bladder cancer.
-----------------------------------------------------------------------------------------------
diagnostic methods currently bladder cancer cystoscopy urine cytology cystoscopy invasive tool low sensitivity carcinoma situ u
rine cytology non invasive low cost method high specificity low sensitivity low grade urothelial tumors despite search urinary
biomarkers early non invasive detection bladder cancer biomarkers present daily clinical practice extracellular vesicles evs re
cently studied promising source biomarkers role intercellular communication tumor progression review overview food drug adminis
tration fda approved urine tests detect bladder cancer use widespread clinical practice include non fda approved urinary biomar
kers review role evs bladder cancer possible role biomarkers diagnosis follow bladder cancer patients review recently discovere
d derived biomarkers diagnosis bladder cancer
```

Fig. 4 Sample data set before and after pre-processing

3.2 Pre-processing

The main and the first process to be performed in text mining is pre-processing. Main objective of pre-processing is to remove the stopwords from the data set as this has no useful meaning in the further process. Using scikit-learn model, the stopwords are removed before calculating TFIDF vectors using TFIDFVectorizer from scikit-learn module.

Some more pre-processing steps are also implemented in the data set to trim the text data using the NLTK library. They are

Tag removal
Lemmatization
Tokenization
Lemmatization (Part of Speech) removal.

The most popular Gensim package is used to remove the punctuations, whitespaces and numeric values from the text. The data set after pre-processing is displayed in below Fig. 4.

3.3 Vectorization

The vectorization is the process used to map the words in the given vocabulary to the corresponding real number or vectors. The real number thus obtained is used to find the word similarities. This process of converting the words into the numbers is called as vectorization.

After the pre-processing step, this vectorization process is performed to calculate the importance of the words present in our data set and that should be transformed into vectors before the data is used in the algorithm. In this work, two vectorization models are used. They are TFIDFVectorizer (term document frequency) and CountVectorizer from Sklearn library.

TFIDF. Term frequency–inverse document frequency (TF-IDF) is the weighing factor for words in the given document collection to find out how important its presence in a document. The TF calculates how often a word comes in the document. The term has more meaning than other words in the document if a term appears more times than the other terms in a document. IDF calculates how much information the word provides by calculating the term frequency for the entire corpus.

Mathematically, TF-IDF for document set represented as *D* for the word referred as t in document represented as d is calculated as

$$\text{tfidf}\,(t,\ d,\ D) = \ \text{tf}\,(t,\ d).\,\text{idf}\,(t,\ D) \tag{1}$$

where

$$\text{tf}\,(t,\ d)\ =\ \log\ (1 + \text{freq}\,(t,\ d)) \tag{2}$$

$$\text{idf}(t, D) = \log(N/(\text{count}(dED : tEd)) \tag{3}$$

CountVectorizer. The CountVectorizer is the easy and effective technique. This model generates the tokenization for the given medical data set and creates a word vocabulary. It returns an integer by counting the words. Scikit-learn library provides CountVectorizer that has fit() and transform() functions. The fit() function learns vocabulary from given documents. The transform() function encodes the given documents as vectors.

4 Experiment and Performance Analysis

In our research work, the following topic modelling algorithms are implemented prior to the classification algorithms and compared the following algorithms followed by their outcomes are explained:

Latent Dirichlet Algorithm
Non-Negative Matrix Factorization.

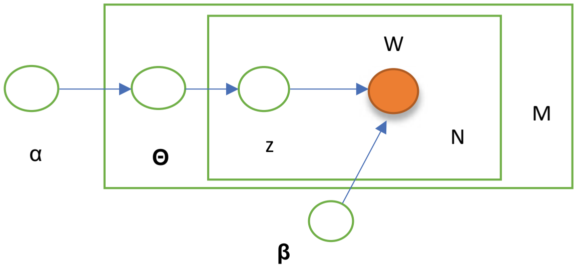


Fig. 5 Graphical representation of LDA

4.1 *Latent Dirichlet Algorithm*

Latent Dirichlet algorithm (LDA) is a statistical model used to do topic modelling by finding the semantic relationship between words in a document. This is an iterative algorithm. LDA has to identify some set of labels or topics for certain documents. In order to do its work, LDA needs number of topics to be find in prior. Then this will go through each and every word in the document and randomly collect the specified number topics with the probability of the word for the specified topic. The probability of the word is calculated for all topics. This process will take many iterations and the list of words and the topics with probabilities. For each of the topic, the highest probability word is selected. Naturally, these words often appear in the same document. High-frequency words will be highlighted more prominently in each topic. The fact that all the documents in a collection share the same set of topics, but each document exhibits certain topics in different proportions, is a distinguishing feature of latent Dirichlet allocation.

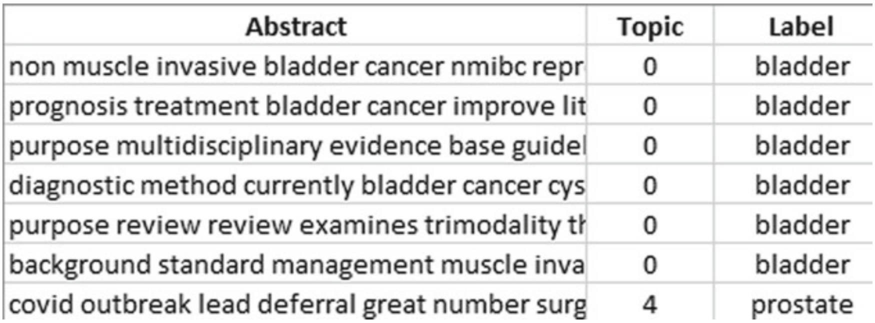
LDA is a generative probabilistic topic model, where each document is represented as a random mixture of latent topics and each topic is distributed across a fixed set of terms [22]. This algorithm works under the principle, three-level Bayesian graphical model, where nodes indicate random variables, the edges represent the dependencies that exists between variables. This is depicted as Fig. 5.

In Fig. 4, α referred to Dirichlet parameter, Θ is the topic variables at the document level, z referred to the assignment of the topic per word and β referred to the topic and w referred to the word that is observed.

This algorithm is used in the medical abstracts to find the hidden topics in the PubMed abstracts data set. On the results of the topic with the LDA, there is a similarity between each topic. In each modelling with the LDA, model that produces one topic can be used as similarity.

4.2 *Non-negative Matrix Factorization*

NMF decomposes multivariate data by producing a number of features that the user determines. The coefficients of these linear combinations are non-negative, and each



Abstract	Topic	Label
non muscle invasive bladder cancer nmibc repr	0	bladder
prognosis treatment bladder cancer improve lit	0	bladder
purpose multidisciplinary evidence base guidel	0	bladder
diagnostic method currently bladder cancer cys	0	bladder
purpose review review examines trimodality tl	0	bladder
background standard management muscle inva	0	bladder
covid outbreak lead deferral great number surg	4	prostate

Fig. 6 Sample data set after labelling

function is a linear combination of the original attribute set. NMF decomposes a data matrix V into the product of two lower rank matrices W and H, yielding a result that is roughly equal to W times H. The initial values of W and H are modified by NMF using an iterative process until the product reaches V. When the approximation error converges or the required number of iterations is reached, the process ends. An NMF model maps the original data into the new set of attributes (features) discovered by the model while the application of the model. Thus, NMF is used to organize the documents by the derived from the non-negative factors.

Given a non-negative matrix $X \in R^{n\times m}$, and an integer $k \ll \min(n, m)$, NMF finds a lower-rank approximation given by

$$X \approx WH \tag{4}$$

where W and H are non-negative factors that belong to $R^{n\times m}$. This method minimizes the distance between the two non-negative matrices X, WH with respect to W and H, based on the constraints $W, H \geq O$ [3].

After the execution of the topic modelling algorithms, the entire data set is labelled, and the labelled data set id is displayed in the following Fig. 6.

4.3 Classification Methods

In machine learning, classification is the process of assigning the items in a collection to some of the specific target classes or specific categories. The success of the specific classifier lies in predicting the target class accurately. To accurately analyse and predict the classes, this uses many machine learning algorithms. Classification algorithms can be implemented on images, text and numeric values. For some of the specialized approaches, different types of algorithms are used. In this work, the specific work is implementing topic modelling with classification method on biological text data.

The algorithms applied in this work are explained below.

Naive Bayes. Naive Bayes (NB) is a kind of classifier that works based on Bayes theorem. To understand the working of NB, the basic knowledge of Bayes theorem is required. Bayes theorem works based on conditional probability. The probability that something will happen by something else has already happened and is known as conditional probability. We can measure the possibility of an occurrence using conditional probability and prior knowledge. It calculates membership probabilities for each class, such as the possibility that a particular record or data point belongs to that class. This method works based on the following equation

$$\frac{P(B|A)P(A)}{P(B)} \tag{5}$$

where *P*(*A*|*B*) is the posterior probability of the Class A given predictor *B*. *P*(*A*) is the prior of the class *A*. *P*(*B*|*A*) is the likelihood which is the probability of predictor of the class. *P*(*B*) is the prior of predictor *B*.

Scikit-learn library is used to implement this algorithm. This algorithm has three types, namely Gaussian NB, Multinomial NB and Bernoulli NB. In this work, multinomial type is implemented. This model will determine frequency of the term that is, how many times the specific term appears in the document.

Support Vector Machine. This is one of the important supervised machine learning algorithms for text classification. Linear, nonlinear classification, regression and outlier identification can be done using SVM. It transforms the data using the kernel and then finds an optimal boundary between the possible outputs based on these transformations. Simply put, it performs some incredibly complex data transformations before determining how to separate the data using the labels or outputs that is specified.

This algorithm is preferable over the other algorithms in classification because this uses less computation and gives significant accuracy. It uses the following equation for classification.

$$f(x) = \text{sign}((w.x) + b) \tag{6}$$

where *w* is a weighted vector in Rn. It separates space Rn into two half spaces to find hyperplane ($y = (w.x) + b$) which has maximum margin. For the nonlinear problems, the linear SVMs is applied by mapping the data into the new H space.

Long Short-Term Memory. Long short-term memory (LSTM) emerged in the 1980s. This belongs to one variant of recurrent neural network (RNN) and this has the capability to remember all long-term past information it has evaluated. This work is handled by creating many activation layers and they are called as gates. With the advent of LSTM, the society began to give special attention to this because this approach performs much better than other classical recursive networks. The important concepts of LSTM are as follows.

- Cell state—this is the memory of LSTM and the cell state is passed to the next steps,
- Forget gate—this gate responsibility is to remove the information that are no longer needed,
- Input gate—this gate will determine what information should be forwarded or written on the internal or next cell state,
- Hidden state—this is the output of the LSTM cell (this means the current output not to be confused with final output).

It has the characteristics of long-distance context-dependent learning and can store context history information.

5 Result Analysis

This proposed work is analysed, and the evaluation metrics are validated. The three hybrid approaches LDA with NB, SVM, LSTM compared with NMF with NB, SVM, LSTM are evaluated using accuracy and F1 score. F1 score has to be calculated using precision and recall, and these two metrics are also included in evaluation metrics. The data set is experimented in Keras, Python. All of the data set are labelled using topic modelling algorithms and classified using classification methods (Figs. 7 and 8).

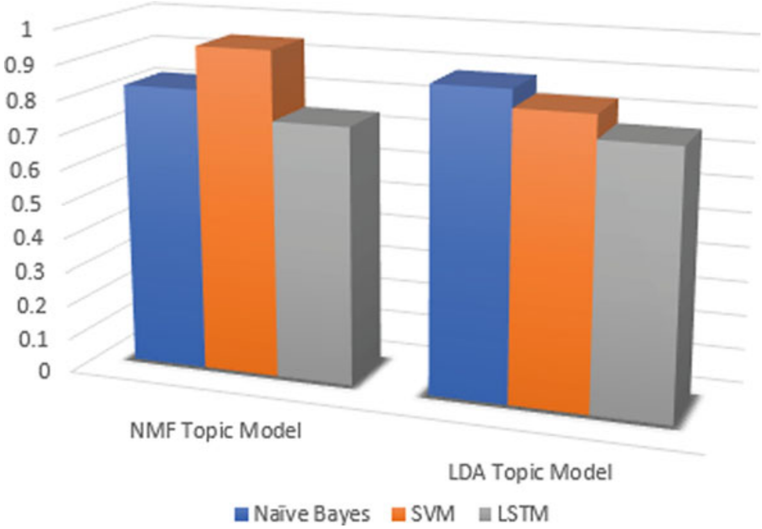


Fig. 7 Comparison of accuracies

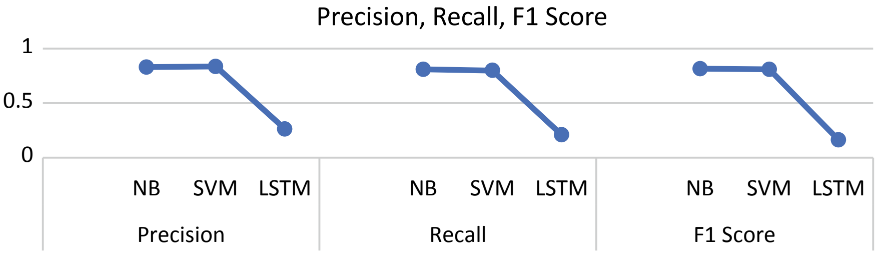


Fig. 8 Precision, recall and F1 score

Table 1 Comparison of accuracies are depicted in the table

Model	Accuracy	Model	Accuracy
LDA + NB	0.8832	NMF + NB	0.8224
LDA + SVM	0.8320	NMF + SVM	0.9472
LDA + LSTM	0.7680	NMF + LSTM	0.7360

Table 2 Precision, recall and F1 score

Algorithms	LDA			NMF		
	NB	SVM	LSTM	NB	SVM	LSTM
Precision	0.8293	0.8357	0.2618	0.8293	0.9535	0.8121
Recall	0.8100	0.7998	0.2079	0.8100	0.9426	0.8121
F1 score	0.8155	0.8093	0.1633	0.8155	0.9470	0.0862

In this paper, the above models are used to experiment on a given medical document data set. The final experimental results are shown in Tables 1 and 2.

6 Conclusion and Future Work

This research work discussed the problem of classifying PubMed abstract data set. In order to identify the exact biological literature, the documents are labelled accordingly by topic modelling algorithms such as latent Dirichlet algorithm (LDA) and non-negative matrix factorization (NMF). Based on the documents that are labelled by the prior algorithms, the classification is done by the classification algorithm such as Naïve Bayes (NB), support vector machine (SVM) and long short-term memory (LSTM). Thus, the combination of topic modelling algorithms with each classification algorithm is analysed. From the proposed model, NMF with SVM attain accuracy of 94% that is little better than other algorithms. In future work, this algorithm can be improved by various combination of new algorithms especially with ensemble learning algorithms on multilabel classification.

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. **68**(6), 394–424 (2018). doi: https://doi.org/10.3322/caac.21492. Epub 2018 Sep 12. Erratum in: CA Cancer J Clin. **70**(4), 313 (2020). PMID: 30207593
2. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **71**, 209–249 (2021). https://doi.org/10.3322/caac.21660. Epub ahead of print. PMID: 33538338
3. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer Statistics, 2021. CA Cancer J Clin. **71**(1), 7–33 (2021). https://doi.org/10.3322/caac.21654. Epub 2021 Jan 12 PMID: 33433946
4. Hashemi, S.H., Karimi, S., Mahboobi, H.: Lifestyle changes for prevention of breast cancer. Electron. Physician. **6**(3), 894–905 (2014). Published 2014 Jul 1. https://doi.org/10.14661/2014.894-905
5. Ramamonjisoa, D.: Topic modeling on users's comments. In: 2014 Third ICT International Student Project Conference (ICT-ISPC), Nakhonpathom, Thailand, pp. 177–180 (2014). https://doi.org/10.1109/ICT-ISPC.2014.6923245
6. Haoxiang, W.: Emotional analysis of bogus statistics in social media. J. Ubiquitous Comput. Commun. Technol. (UCCT) **2**(3), 178–186 (2020)
7. Mitra, A.: Sentiment analysis using machine learning approaches (Lexico based on movie review dataset). J. Ubiquitous Comput. Commun. Technol. (UCCT) **2**(3), 145–152 (2020)
8. Harjule, P., Gurjar, A., Seth, H., Thakur, P.: Text classification on Twitter data. In: 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), Jaipur, India, pp. 160–164 (2020)
9. Curiskis, S.A., Drake, B., Osborn, T.R., Kennedy, P.J.: An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. Inf. Process. Manage. **57**(2), 102034 (2020). ISSN 0306-4573
10. Jedrzejowicz, J., Zakrzewska, M.: Text classification using LDA-W2V hybrid algorithm. In: Czarnowski, I., Howlett, R., Jain, L. (eds.) Intelligent Decision Technologies 2019. Smart Innovation, Systems and Technologies, vol. 142. Springer, Singapore (2020)
11. Luo, W., Gao, J. (2021). Text classification model for public opinion management in colleges and universities based on improved CNN. https://doi.org/10.1007/978-3-030-51431-0_68
12. Thi Do, D., Trang Le, T.Q., Khanh Le, N.Q.: Using deep neural networks and biological subwords to detect protein S-sulfenylation sites. Briefings Bioinform. (2020)
13. Jang, B., Kim, M., Harerimana, G., Kang, S.U., & Kim, J.W.: Bi-LSTM model to increase accuracy in text classification: combining word2vec CNN and attention mechanism. Appl. Sci. (Switzerland), **10**(17), 5841 (2020)
14. Venkataraman, G.R., Pineda, A.L., Bear, O.J., Zehnder, A.M., Ayyar, S., Page, R.L., Bustamante, C.D., Rivas, M.A.: FasTag: automatic text classification of unstructured medical narratives. PLoS One **15**(6), e0234647 (2020). https://doi.org/10.1371/journal.pone.0234647. PMID: 32569327; PMCID: PMC7307763
15. El-Halees, A.: Arabic text classification using maximum entropy. Islamic Univ. J. (Ser. Nat. Stud. Eng.) **15**, 157 (2007)
16. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Boulicaut, J.F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) Machine Learning: ECML 2004. ECML 2004. Lecture Notes in Computer Science, vol. 3201. Springer, Berlin, (2004)
17. Chau, M., Chen, H.: A machine learning approach to web page filtering using content and structure analysis. Decis. Suppor. Syst. **44**(2), 482–494 (2008). ISSN 0167-9236
18. Mccallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. Work Learn Text Categ. **752** (2001)
19. Mesleh, A.: Chi square feature extraction based SVMS Arabic language text categorization system. J. Comput. Sci. (2007). https://doi.org/10.3844/jcssp.2007.430.435

20. Luo, Y.: Recurrent neural networks for classifying relations in clinical notes. J. Biomed. Inf. **72**, 85–95 (2017). ISSN 1532-0464
21. Joachims, T.: Transductive Inference for Text Classification Using Support Vector Machines. ICML (2001)
22. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). https://doi.org/10.1162/jmlr.2003.3.4-5.993

Development of Improved SoC PTS Algorithm for PAPR Reduction in OFDM Underwater Communication

M. Asha and T. P. Surekha

Abstract Orthogonal frequency division multiplexing (OFDM) is a multi-carrier communication technique used in many modern day communication platforms. It was designed with a very high data rate during wireless communication. Underwater communication is an emerging technology in the field of wireless communication. Due to the detrimental effect of time and frequency spreading, achieving good data transfer in underwater wireless communication is challenging. This paper proposes a multi-level optimization model called partial transmission sequence (PTS). This work utilizes PTS algorithm to reduce the peak-to-average power ratio (PAPR) of OFDM systems in underwater communication. This method increases the number of optimization stages and reduces the number of elements in the present phase set for each level of optimization to find the optimal phase rotation. The computational complexity of traditional PTS algorithm increases exponentially with the increase of the number of sub-blocks and the number of phase set elements can be selected. The proposed implementation would reduce the complexity of the system as the sub-blocks are executed in parallel. The development of proposed method is divided into two stages. The proposed PTS model is first developed in MATLAB and then implemented on system on chip (SoC) platform. The proposed method is compared with the existing methods. The experimental results prove that the proposed method greatly reduces the PAPR value, thereby increasing the overall performance.

Keywords Multi-level optimization · Orthogonal frequency division multiplexing · Partial transmission sequence · Peak-to-average power ratio · System on chip · Underwater communication

M. Asha (✉)
Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, Visvesvaraya Technological University, Belagavi, India

T. P. Surekha
Dean (Academic Affairs), Professor, Department of Electronics and Communication Engineering, VidyaVardhaka College of Engineering, Visvesvaraya Technological University, Belagavi, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_48

1 Introduction

Orthogonal frequency division multiplexing (OFDM) consists of a frequency multiplexing of different carriers, where each one carries one information modulated following an M-QAM or M-PSK constellation [1]. The result is a signal that is transmitted in bandpass and that contains its times N transmission sub-bands belonging to a series of carriers orthogonal at a low rate. So, by transmitting all at the same time, it is achieved a much higher rate. Each of these carriers behaves like a channel independent that only suffers attenuation, and there is no dispersion in each subchannel.

The use of carriers orthogonal to each other allows a better use of the transmission band. In conventional FDM, the separation between adjacent subcarriers is at least 2/T [2], while in OFDM the separation is 1/T, which is the minimum for adjacent subcarriers to be orthogonal, thereby improving spectral efficiency, as we can see in the following Fig. 1.

The meaning of orthogonality in OFDM is based on the idea that each subcarrier has an integer number of cycles during the symbol period. This assumes the advantage that each subcarrier has a null in the center of the adjacent subcarrier [3]. The result is that intersymbolic interference (‘Inter Symbol Interference,’ ISI), and ideally, it is zero. We can express the orthogonality according to the following Eq. (1) mathematical development:

$$\frac{1}{T_{\mathrm{TFT}}}\int\limits_{t=0}^{T_{\mathrm{FFT}}} \mathrm{e}^{-i2\pi\Delta f(k-k')^{t}}\,\mathrm{d}t = \begin{Bmatrix} 1 \text{ para } k = k' \\ 0 \text{ para } k \neq k' \end{Bmatrix} \tag{1}$$

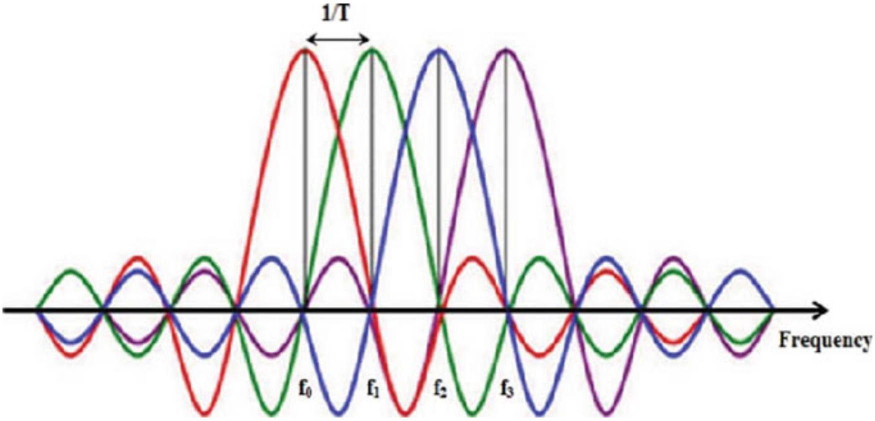


Fig. 1 Spectrum of OFDM signals

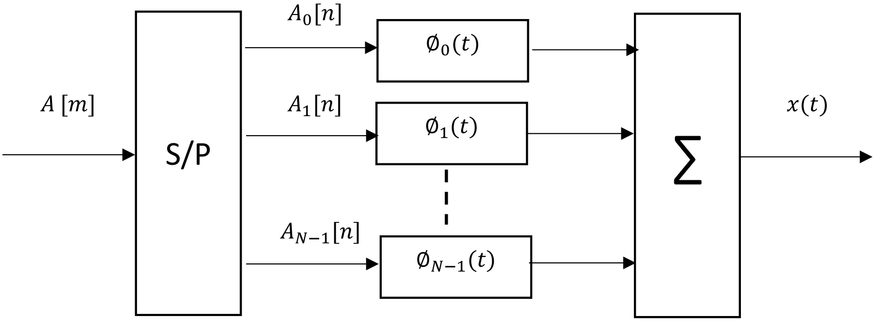


Fig. 2 OFDM modulation process

where T_{FFT} denotes the duration of an OFDM symbol, Δf is the frequency between two adjacent subcarriers, and k and k' are the subcarrier number for values from 1 to the total number of subcarriers.

The process to obtain OFDM modulation is illustrated in the following Fig. 2

A [m] is the set of symbols to be transmitted and the $x(t)$ represented in the figure is obtained from the signal $s(t)$ in the usual way, and the $x(t)$ shows the in Eq. (2).

$$x(t) = \sqrt{2}Re\left\{s(t)\mathrm{e}^{jw_c t}\right\} \tag{2}$$

In this scheme, a set of pulses is used that are generated multiplying a prototype filter by a set of N different carriers. The $S(t)$ expresses in Eq. (3).

$$s(t) = \sum\nolimits_n A^T[n]\varnothing(t - nT) = \sum\nolimits_n \sum\nolimits_{1=0}^{N-1} A_1[n]\varnothing_1(t - nT) \tag{3}$$

where the base functions $\Phi_1(t)$ are of the form Eq. (4)

$$\varnothing_1(t) = \frac{1}{\sqrt{T}}\mathrm{e}^{j\frac{2\pi i t}{T}}.w_T(t) \tag{4}$$

where $w_T\ (t)$ is a rectangular time window of duration T.

These base functions form an orthonormal base. Thus, we observe how the OFDM is actually the superposition of N bandpass modulations that are transmitted simultaneously.

This modulation, described as we have done, has the disadvantage that its practical implementation is difficult because it is necessary to generate N carriers complex (real 2N) perfectly locked in phase. If this condition is not fulfilling, the base functions are no longer orthogonal and the known effect appears as intercarrier interference ('InterCarrier Interference,' ICI) [4]. To avoid it, we sample $s(t)$ with period T/N obtaining: $s[m]$ is shown in Eq. (5)

$$S[m] = \sum_{t=0}^{N=1} A_1[0]\emptyset_1\left(\frac{mT}{N}\right) = \frac{1}{\sqrt{T}}\sum_{t=0}^{N-1} A_1[0]\mathrm{e}^{-j\frac{2\pi lm}{N}}$$
$$; m = 0, \ldots., \; N-1 \tag{5}$$

where we can verify that the term on the right is not more than the Inverse DFT of the Al sequence multiplied by a constant factor in Eq. (6).

$$\mathrm{DFT}_N^{-1} : x[n] = \frac{1}{N}\sum\nolimits_{m=0}^{N-1} X(m)\mathrm{e}^{-j\frac{2\pi nm}{N}} \tag{6}$$

Therefore, taking advantage of the efficiency provided by the algorithms that calculate the Fourier transform in discrete time, like the FFT, we arrive at the implementation of the modulator system.

OFDM Modulator

The information bits that reach the modulator are separated into blocks parallels that will be associated with the different carriers. Each block of bits is associates according to the type of constellation used (QPSK, QAM, etc.). The N symbols resulting are modulated by using the inverse Fourier transform, which it obtains two signals (one real and one imaginary) which is the modulated information [5]. For the transmission of the signals, they are converted to continuous time by means which is the modulated information [5]. For the transmission of the signals, they are converted to continuous time by means of a converter digital-analog (D/A) and both signals in phase and quadrature are modulated in the RF band. Figure 3 block diagram of OFDM modulator.

The reception system follows the reverse steps to each block, as shown in Fig. 4.

The receiver collects the signal $r(t)$, after which it is fed to quadrature mixer and is converted to baseband utilizing the cosine and sine functions. In addition, this generates signals centered on $2f_c$ which low-pass filters reject. After this, the

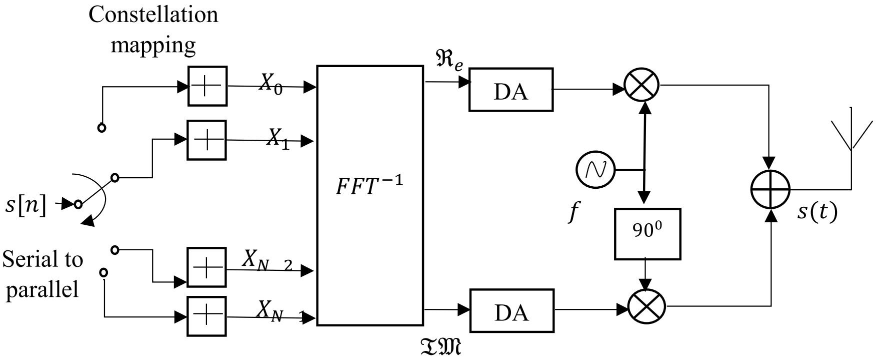


Fig. 3 Block diagram of an OFDM modulator

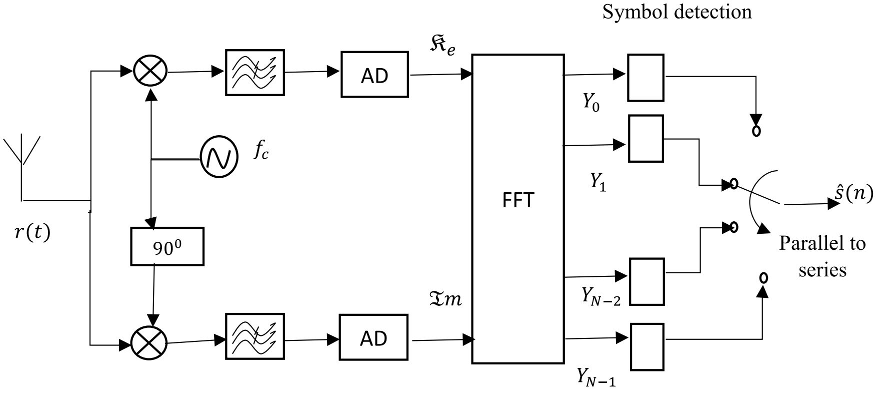


Fig. 4 Block diagram of an OFDM demodulator

baseband signals are sampled and converted to digital form with the help of analog-to-digital converters, after which a forward-FFT transforms them to frequency domain [6]. This gives back N parallel streams, which are all translated into a binary stream using the suitable symbol detector. Then, these streams are merged back into a serial stream $\hat{s}[n]$, which is close in comparison to the actual binary stream at the transmitter end.

OFDM technology [7] can effectively utilize the advantages of frequency bandwidth and anti-multipath interference. In OFDM, the signal is composed of multiple modulated independent subcarrier signals superimposed on each other. If the phase of each subcarrier is the same or close, the superimposed signal will produce a larger instantaneous power peak. Compared with traditional single-carrier transmission method, a linearly dynamic range system has high requirements. If the system is linearly dynamic, the range cannot meet the change in signal, and it will cause signal distortion and making the changes in frequency spectrum of the superimposed signal [8]. Which leads to the destruction of the orthogonality between the sub-channels and resulting in serious mutual interference, which makes the communication system performance has deteriorated seriously.

In view of the problem of reducing the PAPR [9], the exhaustive literature has been conducted and mainly focusing on limiting the technology, probability technology and coding techniques. Among them, the most typical method of probability technology is to choose linear mapping technology and partial transmission sequence (PTS) technology [10]. This technique reduces the probability of peak occurrence through linear interference because this will not distort the signal.

In PTS technology, the impact reduces the PAPR of the OFDM system. The main factors of performance include: data segmentation method, segmentation subsequence number and the number of elements in the phase set to be selected. Segmentation methods include: Adjacent segmentation, interlaced segmentation and random segmentation [11]. Most of the research articles are reported that performance of inter-leaved segmentation method is not up to the mark to reduce the PAPR. However, under a certain segmentation method, the PTS technology should reduce the PAPR [12] if the effect is greatly improved, it is necessary to increase the number of sub-sequences or increase the number of phase set elements to be selected.

In the PTS, number of divided sub-sequences [13] can be denoted as 'V' and the number of phase set elements is represented by 'W.' After calculating the IDFT of sub-sequences V at N points, there will be $W\,V \uparrow$ phases which called as bit rotation vector combination, so that each phase rotation vector corresponds to 'VN' complex number multiplications and $(V - 1)N$ complex number additions. Therefore, When PTS technology is used to reduce the peak-to-average ratio, the performance is improved at the same time and its amount of calculation will increase with 'W' (usually $W \leq 8$, because the value of 'W' increases, the performance of reducing PAPR is not significantly improved, and the amount of calculation will be becoming very large) and 'V' increases exponentially.

Based on the research of PTS technology, this work proposes a multi-level search the scheme of the optimal algorithm. PTS algorithm is utilized to reduce the PAPR of the OFDM systems in underwater communication. The number of optimization stages is increased, and the number of elements in each level is reduced to find the optimal phase rotation. The computational complexity of traditional PTS algorithm increases exponentially with the increase of the number of sub-blocks, and the number of phase set elements can be selected. The proposed implementation would reduce the complexity of the system as the sub-blocks are executed in parallel. Finally, the performance simulation analysis and comparison of peak-to-average ratio reduction are reported. The rest of the paper is organized in the following manner. Section II provides a detailed literature survey explaining the latest techniques proposed by different authors. Section II presents the description of PAPR and PTS technique. Section IV presents the multi-level optimization PTS algorithm proposed in the paper.

2 Literature Survey

The instability of the OFDM envelope can be expressed by PAPR. Therefore, to improve the performance of the system, PAPR reduction has to be done. Most of the research scholars are focusing on PAPR reduction in OFDM system. The main algorithms related to PAPR reduction in OFDM system can be summarized as follows.

2.1 Limiting Method

The introduction of the rectangular window in the limiting method [14, 15] will affect the frequency spectrum of the original signal, this causes new out-of-band noise and reduces spectral efficiency. This kind of non-linear change will produce serious in-band distortion, thereby reducing the bit error rate and resulting in system performance degradation. To overcome the bit error performance caused by limiting the deterioration of the channel, effective channel coding and decoding has been used. A method of combining limiting and error coding is introduced in [16].

2.2 Windowing Method

This method [17, 18] uses a window function with better spectral characteristics than a rectangular window, however, the signal needs to be processed at a higher rate after up sampling, so the realization is difficult and will affect the signal spectrum characteristics.

2.3 Weighted Multicarrier Modulation Method

Weighted multi-carrier modulation method [19] refers to the use of Gaussian before Fourier transform or hamming window function weights of the input signal to reduce PAPR.

2.4 Carrier Suppression Peak Method

The main idea of the carrier suppression peak method [20] is, when the peak value of the OFDM signal power appears, some subcarriers of OFDM are not used to transmit data, however, it sends some designed signals which can suppress peaks. Usually, different frequency bands are recommended to use as the frequency of this carrier. The advantage of this technique is neither it will reduce the SNR (signal to noise ratio) of the system and will not introduce out-of-band interference. The disadvantage is reducing the data rate of the system and increases the complexity of the system.

2.5 Compression and Expansion Method

The main idea of the traditional expansion method is to increase the low amplitude value in the signal to maintain its peak amplitude which increases the average power of the signal, thereby reducing the purpose of PAPR. However, this increases the average transmit power of the system, making the symbol power value of the signal is closer to the non-linear transformation area of the power amplifier, which is an easy cause of signal distortion. Therefore, reference [21] gives an improved compression and expansion transformation method, reference [22] evaluates this method. The method in [21], the high-power transmission signal is compressed, and the small power signal is amplified so that the average power of the transmitted signal can be compared and the pair remains unchanged. This will not only reduce the PAPR of the system but also the anti-interference ability of low-power signals is enhanced.

2.6 Other Algorithms

Agarwal et al. [23] have projected and compared the reduction methods of PAPR for OFDM. The signal compresses for scaling back amplitude distortion by using projected precoding technique. Here, this method is independent of data block avoiding through the companding and optimization theme. Precoding performance is better compared to the companding and PAPR reduces up to 4 dB than the standard OFDM system. In companding, the good performance achieves with the conjointly μ-law companding than the A-law companding. Anoh et al. [24] have presented the process of iterative clipping and filtering that optimizes the OFDM systems through the reduction of designing complexity and utilization of resources. At all clipping ratios, the adaptive technique's PAPR gain shows better outputs. Rateb et al. [25] have presented a reduction technique of PAPR that shouldn't affect the BER performance while maintaining the effectiveness in signal distortion techniques. Even at 99% of spectral efficiency, the PAPR reduces with the proposed method not applying the channel coding at the side data.

Anoh et al. [26] have introduced a uniform distribution technique and Lagrange Multiplier (LM) optimization for solving the PAPR reduction problem of OFDM and minimizing the number of involved iterations, respectively. To enhance the BER, the technique of Minimum Mean Square Error (MSME) has exploited in the proposed method. Compared to the traditional clipping and filtering without LM, the proposed technique has shown improved results in terms of PAPR reduction. Ikram et al. [27] have proposed Peak Insertion (PI) technique to reduce the PAPR of OFDM signals. Compared to the other methods, the proposed technique is faster, simpler, and greater achieving of PAPR reduction. Ce et al. [28] have proposed a PAPR reduction technique based on the integration of PEC and FWFT for OFDM systems. By implementing the proposed technique, the BER improves for OFDM systems when using in multipath fading channels.

Minhoe et al. [29] have proposed a novel reduction technique of PAPR known as PAPR reducing network (PRNet) with deep learning. By comparing with conventional algorithms, the better results like improved BER and reduced PAPR have provided using the proposed scheme. Kim et al. [30] have proposed criteria for CSS scheme to choose good SV sets for ensuring the optimization of PAPR reduction. In this scheme, the ACF of the OFDM is considered for three partition cases including adjacent, inter-leaved, and random. The proposed criteria show the best performance in PAPR reduction.

Wang et al. [31] have proposed a low-complexity TI scheme using distortion signals to reduce the PAPR for OFDM systems. In this method, the reduction of original integer programming problem into a sequence search problem for maintaining a good PAPR. Al-jawhar et al. [32] propose a novel subblock partitioning scheme known as terminals exchanging segmentation (TE-PTS) scheme for enhancing the PAPR reduction performance. The simulation results have shown that the PAPR reduction ratio about 6% with low computational complexity with the proposed scheme compared to the IL-PTS scheme.

Vittal et al. [33] have proposed a novel low-complexity optimized PTS method based on RPSM for achieving the reduction of PAPR in OFDM. Using proposed method, the improved performance results such as reduced PAPR and computational complexity than the traditional PTS. Wang et al. [34] propose a technique based on linearized alternative direction method of multipliers (LADMM) for optimization of OFDM systems through the reduced PAPR. Better BER and larger PAPR reduction have been achieved with the implementation of LADMM algorithm compared to the other existing algorithms based on the simulation results.

Jie et al. [35] propose a novel modified chaos clonal shuffled frog leaping algorithm based on Markov chain theory for OFDM. When compared to the existing algorithms like quantum evolutionary, the selective mapping, genetic algorithm, and the original approach, the proposed technique outperforms in achieving higher convergence speed and reduced PAPR.

Bao et al. [36] propose a perturbation-based approach for MIMO-OFDM systems for achieving the reduction of PAPR while transmission of signals. The proposed technique results improved PAPR reduction and faster convergence rate in the simulation results. Liu et al. [37] have proposed a new ICF method based on time-domain kernel to achieve the different PAPR requirements in OFDM. With the practical testing of proposed novel technique, optimal performance provides in the reduced EVM and PAPR than the iterative clipping and filtering.

Ali et al. [38] have examined the performance of BER, PDPR, and PSD with a clipping technique based on OFDM symbol statistics. Based on the appropriate selection of scaling factor, the smooth control of clipping achieves with the proposed technique than the clipping and mapping methods. Hu et al. [39] have proposed a generalized PLC scheme to decrease the distortion in decompanded signals. The GPLC technique provides an extra freedom degree through the average signal power. It leads to the maintenance of BER enhancement and PAPR reduction.

Zhang et al. [40] have proposed a low complex scheme of PAPR reduction in OFDM and it is a modified version of SLM. For a given number of IFFTs with no side data, the proposed method outperforms than the conventional SLM technique. Mhatre et al. [41] have proposed threshold selective mapping technique to reduce the PAPR in MIMO-OFDM systems. However, the simulation results provide the better performance with proposed technique in terms of reduced PAPR without incurring data loss and increasing the power requirement.

Zhang et al. [42] propose a hybrid PAPR reduction method based on a post-IFFT stage and a modified class-III SLM scheme. The PAPR reduces with low complexity achieves using a proposed method compared to the traditional SLM and class-III

SLM. Matsumine et al. [43] introduce a novel PAPR reduction technique based on polar codes for OFDM systems. By comparing with the conventional SLM, reduction of PAPR is greater than 4.5 dB achieves with the proposed shaping method based on the analyzation of performance results.

3 PAPR and PTS Algorithm in OFDM System

The PAPR is the ratio of maximum power of a sample in an OFDM transmit symbol to the mean power of the same OFDM symbol. It is represented in the units of dB. PAPR comes into existence when distinct subcarriers are not in phase with each other in a multi-carrier system. At every point in time, they are in contrast to each other at different phase values. When all the points accomplish the peak value at the same time, it makes the output envelope to balloon which leads to a ‘peak.’ As there are several separately modulated subcarriers in an OFDM system, the maximum value in the system can be significant in comparison to the mean value of the entire system. This ratio of maximum power to the mean power is known as peak-to-average power ratio.

In an OFDM system with N subcarriers, $X_k(k = 0, 1, \ldots, N - 1)$ represents PSK or QAM modulated for transmission the original frequency domain signal, so the complex baseband signal $s(t)$ is written as in Eq. (1)

$$s(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k \mathrm{e}^{j2\pi f_k t}, 0 \le t \le T_s \tag{7}$$

where T_s is the symbol period of the OFDM signal. For OFDM, signal sampling is performed at an interval of $\Delta t = T_s/JN$, and the sampled OFDM separation the scattered time signal is expressed in Eq. (8)

$$S_n = s(n\Delta t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k \mathrm{e}^{j2\pi k(n/JN)}, 0 \le n \le JN - 1 \tag{8}$$

where J is oversampling factor. Equation (8) can also be expressed in Eq. (9)

$$S_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{JN-1} X'_{kk} \mathrm{e}^{j2\pi k(n/JN)}, 0 \le n \le JN - 1 \tag{9}$$

where $X'_k = \begin{cases} X_k, k < N \\ 0, k \ge N \end{cases}$.

Similarly, the oversampling OFDM signal PAPR is defined Eq. (10)

$$PAPR_{[s_n]} = 10\log \frac{\underset{n}{\max}\left\{|s_n|^2\right\}}{E\left\{|s_n|^2\right\}} \tag{10}$$

where $E\ \{.\}$ represents the average power of the OFDM signal.

The Complementary Cumulative Distribution Function (CCDF) is usually used to measure the performance of peak-to-average ratio reduction, ie., Eq. (11)

$$\mathrm{CCDF} = \Pr\{\mathrm{PAPR} > \mathrm{PAPR_0}\} = 1 - \left(1 - \mathrm{e}^{\mathrm{PAPR_0}}\right)^{\mathrm{JN}} \tag{11}$$

3.1 Basic Principles of PTS Algorithm

The PTS is the one of many signal representation methods, which segments the data block into multiple groups, from which a transmission signal is chosen after scrambling. The idea behind the PTS method is to divide the fed data symbols into discrete subsets and these subsets are rotated with corresponding rotation factors. Then, the adapted divided subsets are merged again to produce a group of candidate signals known as partial transmit sequences. Consequently, one candidate sequence having the least amount of PAPR value has opted for transmission. Also, PTS is considered to be a distortion-less technique since it depends on the scrambling signal technique to decrease the PAPR value. Therefore, PTS employs a probabilistic method to cut down the PAPR of the OFDM signal. As a result, it is not negatively impacted by the bit error rate distortion.

The traditional PTS method combines vector $\boldsymbol{X} = [\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{N-1}]$ and divide into V groups of non-overlapping sub-vector, respectively, composed of $\boldsymbol{X}_v,\ v = 1,\ 2, \ldots, V$. Which means that for each subcarrier of each sub-vector is multiplied by the same phase rotation factor of b_v, and then, V vectors are combined according to formula (12), and the principal block diagram of PTS is shown in Fig. 5.

$$\boldsymbol{X}' \sum_{v=1}^{V} b_v \boldsymbol{X}_v \tag{12}$$

where $b_v = \mathrm{e}^{j\varphi_v}$, $\varphi_v \epsilon [0, 2\pi]$. Perform IDFT on Eq. (12), we get to (13)

$$x' = \mathrm{IDFT}\{\boldsymbol{X}'\} = \sum_{v=1}^{V} b_v.\mathrm{IDFT}\{\boldsymbol{X}_v\} = \sum_{v=1}^{V} b_v x_v \tag{13}$$

The purpose of PTS method is to find an appropriate phase rotation factor and satisfy Eq. (14)

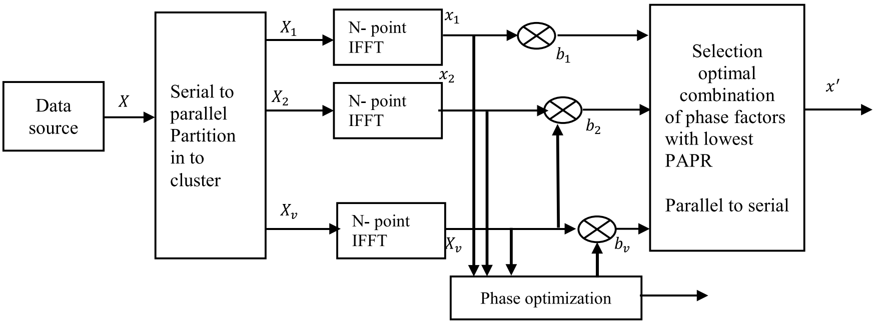


Fig. 5 Basic Principle block diagram of PTS

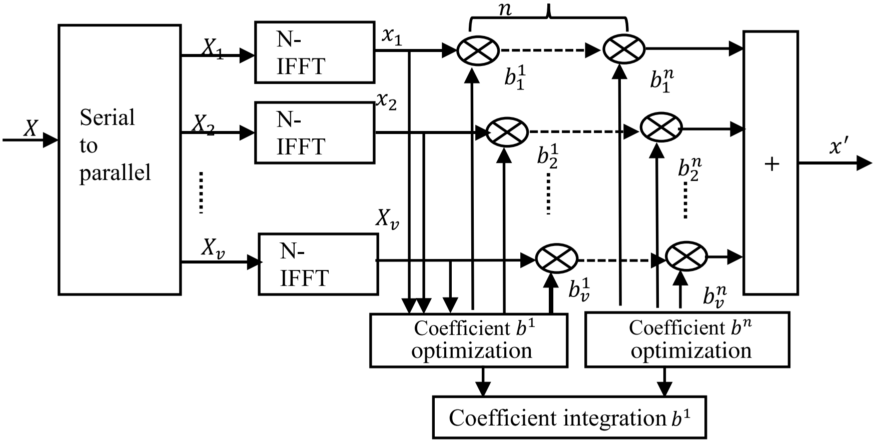


Fig. 6 PTS principle block diagram of multi-level optimization

$$\{b_1,\ b_2,\ \ldots,\ b_v\} \underset{\{b_1,\ b_2,\ \ldots,\ b_v\}}{\arg\min} \left\{ \underset{1 \le n < N}{\max} \left| \sum_{v=1}^{V} b_v x_v \right|^2 \right\} \tag{14}$$

where arg min (·) is used when the function obtains the minimum value condition. In theory, φ_v can take any value in the interval [0, 2π], but in actual operation, b_v is generally in a predetermined discrete phase set in the value.

4 Multi-level Optimization PTS Algorithm

In this section, a multi-level optimization PTS calculation is proposed. The law process of proposed block diagram is shown in Fig. 6. The main contribution of this work is discussed in this section.

4.1 Theoretical Basis of PTS Algorithm

The input data X is divided into sub samples by the serial to parallel converter. Each sub sample is given to the N point inverse FFT. The phase values of the output signals are divided into multiples parts ach each part is optimized separately. This operation will facilitate parallelism and the optimization of the phase value becomes easier and faster. Consider two phase rotation factors $b_v = \mathrm{e}^{j\varphi_v}$ and $b'_v = \mathrm{e}^{j\varphi'_v}$, which in $\varphi_v \epsilon \{\varphi_i, i = 1, 2, \ldots, V\}, \varphi'_v \epsilon \{\varphi'_j, j = 1, 2, \ldots V'\}$.

Here, define in Eq. (15)

$$\begin{aligned}&\{\varphi_1,\ \varphi_2,\ \ldots,\ \varphi_3\} \oplus \left\{\varphi'_1, \varphi'_2, \ldots\ \varphi'_{v'}\right\}\\ &\triangleq \left\{\frac{\varphi_1+\varphi'_1,\ \varphi_1+\varphi'_2,\ \ldots,\ \varphi_1+\varphi'_{v'},\ \ldots\varphi_v+\varphi'_1,\ \varphi_v+\varphi'_2,\ \ldots,\ \varphi_v+\varphi'_{v'}}{vv'}\right\}\end{aligned} \tag{15}$$

By $b_v.b'_v = \mathrm{e}^{j(\varphi_v+\varphi'_v)}$ there $\varphi_v + \varphi'_v \epsilon \{\varphi_1, \varphi_2, \ldots, \varphi_3\} \oplus \left\{\varphi'_1, \varphi'_2, \ldots \varphi_{v''}\right\}$. Assuming a discrete set of phases Φ_w is in Eq. (16)

$$\Phi_w = \{0, \frac{1}{W}2\pi, \frac{2}{W}2\pi, \ldots, \frac{W-1}{W}2\pi\} \tag{16}$$

where the W represents the number of elements in the phase set.

Then another phase set B_N can be expressed in Eq. (17)

$$B_N = \{\theta_i | \theta_i \le \frac{\pi}{2}; i = 1, 2, \ldots, N; N \ge 2\} \tag{17}$$

From the definition of formula (15), we can get Eq. (18)

$$\Phi_{w/2} \oplus B_N = \left\{\underbrace{\theta_i,\ \frac{1}{W/2}2\pi + \theta_i,\ \frac{2}{W/2}2\pi + \theta_i,\ \ldots,\ \frac{W/2-1}{W/2}2\pi + \theta_i, \ldots}_{NW/2}\right\} \tag{18}$$

If we choose $B_N = \{0, \pi/2, \ldots\}$, from Eq. (18), we can get Eq. (19)

$$\Phi_w \subseteq \Phi_{w/2} \oplus B_N \tag{19}$$

4.2 Improved PTS Algorithm

It can be seen from the analysis when the number of elements in the phase rotation is set to be $w = 2$, PTS method requires least number of exhaustive searches. The value of the elements to be selected in the phase set is uniformly divided within in the range [0, 2π]. Similarly, the phase set optimized by the first-level coefficient is $B_2^1 = \{0, \pi\}$ (The number 1 in B_2^1 indicates the first level of optimization, and the value 2 indicates that the number of elements in the phase set is 2). After applying optimized PTS algorithm, the optimal coefficient $b_v^1(v = 1, 2, \ldots, v)$ (where b_v^1 represents the vth segmentation subsequence after level 1 optimization of the corresponding coefficient) is combined with x_v, then the time-domain number of each segmentation subsequence data in Eq. (20).

$$x_v^1 = b_v^1 x_v \tag{20}$$

where $b_v^1 = \mathrm{e}^{j\varphi_v^1}$, $\varphi_v^1 \epsilon B_2^1$, $v = 1, 2, \ldots, V$ At this point,

After the first-level optimization coefficient vector is stored, the optimized subsequence data is continuing with the second level of coefficient optimization. The group of phase sets to be selected at this time for $B_2^2 = \{0, \pi/2\}$, then the optimized subsequence time-domain data in Eq. (21)

$$x_v^2 = b_v^2 x_v^1 \tag{21}$$

where $b_v^2 = \mathrm{e}^{j\varphi_v^2}$, $\varphi_v^2 \epsilon B_2^2$, $v = 1, 2, \ldots, V$. Similar in order, Let the $k(1 \leq k \leq n)$ stage optimized phase set $B_2^k = \{0, \pi/2^{k-1}\}$, then the time-domain subsequence data after k$^{\text{th}}$ optimization is in Eq. (22)

$$x_v^k = b_v^k x_v^{k-1} \tag{22}$$

where $b_v^k = \mathrm{e}^{j\varphi_v^k}$, $\varphi_v^k \epsilon B_2^k$, $v = 1, 2, \ldots, V$. Finally after the optimization of the level coefficient, the output time-domain data is Eq. (23)

$$x' = \sum_{v=1}^{V} b_v^n x_v^{n-1} = \sum_{v=1}^{V} b_v^n (b_v^{n-1}(\ldots (b_v^1 x_v) \ldots)) \tag{23}$$

where $b_v^n = \mathrm{e}^{j\varphi_v^n}$, $\varphi_v^n \epsilon B_2^n$. Let $b_v' = b_v^n b_v^{n-1} \ldots b_v^1 = \mathrm{e}^{j\varphi_v'}$, i.e., $\varphi_v' = \varphi_v^1 + \varphi_v^2 + \cdots + \varphi_v^n$. Then, Eq. (23) can be written in Eq. (24)

$$x' = \sum_{v=1}^{v} b'_v x_v \tag{24}$$

where $\varphi'_v \epsilon B_2^1 \oplus B_2^2 \cdots \oplus B_2^n$. From Eq. (18), we can see and we written in Eq. (25)

$$B^1_{(2)} \oplus B^2_{(2)} \cdots \oplus B^n_{(2)} = \left\{\pi k/2^{(n-1)}, k = 0, 1, 2, \ldots, 2^n - 1\right\} \tag{25}$$

The phase rotation factor after multi-level optimization b'_v and the phase is $\varphi'_v \epsilon \{\pi k/2^{n-1}, k = 0, 1, 2, \ldots, 2^n - 1\}$. Similarly, consider multiple phase set elements (W $\geq$ 4), and it is transformed into a multi-stage optimization process with only two-phase elements per stage.

The algorithm is as follows:

1. Divide N number of subcarriers into V set of sub-sequences.
2. Apply Inverse Desecrate Fourier Transform (IDFT) on N sub-sequences
3. Assume $b_v^n = 1, v = 1, 2, \ldots, V$.
4. Measure Peak-to-average ratio $\text{PAPR}_0 = 10\log(\max|x'_n| E\{|x'_n|^2\}$, where$x'_n = \sum_{v=1}^{v} b_v^n.\text{IDFT}\{\boldsymbol{X}_v\}$, and set index $=$ 1.
5. Assume $b_{\text{index}} = \pi/2^{n-1}$, and calculate PAPR
6. If PAPR > PAPR$_0$ then $b_{\text{index}} = 1$; otherwise PAPR$_0$ = PAPR, and index = index + 1.
7. If index < V + 1, return to step (4); otherwise go to Step (7)
8. Get the nth phase rotation coefficient $b_v^n(1, 2, \ldots, V)$, and $n = n + 1$.
9. If n is less than the present optimization times, return to step (3); otherwise go to step (9).
10. Calculate the coefficients after n optimizations.

5 Experimental Results

In SoC implementation, the proposed algorithm is developed in Simulink environment. In Simulink, each blocks are custom functions which are written in fixed point code. The simulation model is as shown in Fig. 7. In initialization block, the input configuration parameters are assigned such as IFFT length, carrier count, no of subcarriers and SNR values, etc., is indicated in Table 1.

In the test_PAPR_for_SoC block, signal enters the OFDM transmitter, IFFT is performed before transmission. A channel model is then applied to the transmitted signal. The model identifies the inputs like signal to noise ratio, multipath and peak power clipping, selective mapping and partial transmit sequence to be controlled. Output of the test_PAPR_for_SoC is given to clipping model and the OFDM symbol model. The other input of the OFDM symbol model is taken from the initial symbols_per_carrier model. At the OFDM receiver, the data is received and

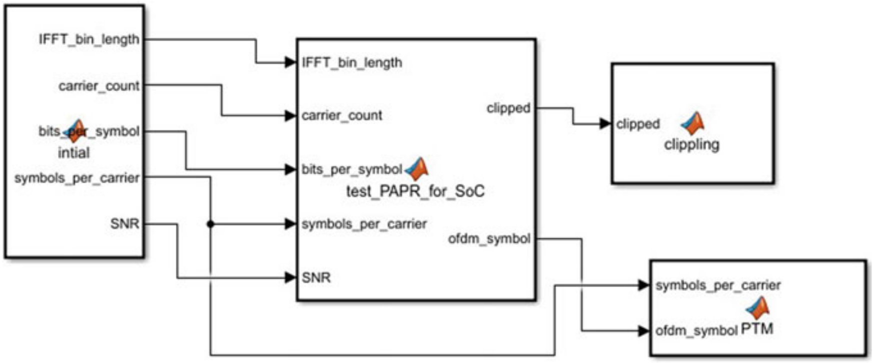


Fig. 7 Simulation model of SoC

Table 1 Simulation parameters

S. No.	Parameter	Value
1	IFFT Length	1024
2	Carrier count	128
3	Bits per symbol	3
4	Symbols per carrier	100
5	SNR	20
6	Carrier frequency	24 MHz
7	Bandwidth	8 MHz

then FFT is applied to convert the signal back to the frequency domain and demodulation, decoding procedures are performed. Depending on the channel factors, code rate and modulation scheme are observed. Based on the state of the received signal, PAPR calculations are displayed.

The simulation parameters as consider for the proposed method in underwater communication scenario are reported in Table 1.

6 Results and Discussion

In this section, the proposed PTS algorithm simulations are discussed and the CCDF distribution curve of PAPR is analyzed. The proposed method is implemented on MATLAB and Simulink environments. In the simulation process, QPSK is used, because of the higher data rate compared with other modulation techniques along with higher PAPR. The number of subcarriers $N = 128$ modulated on the signal and pseudo-random division with oversampling rate is considered as $J = 4$. For every simulation, 10,000 iterations are required to find the CCDF curve. The one symbol

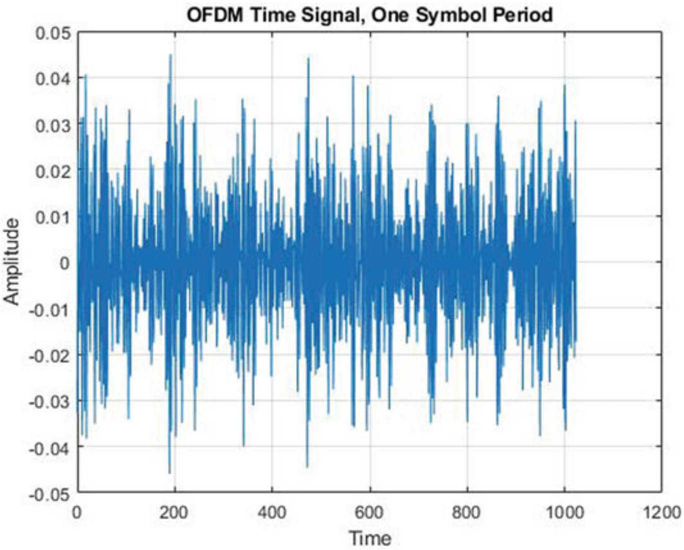


Fig. 8 One symbol period of OFDM signal

period of OFDM signal is depicted in Fig. 4 and the spectrum of OFDM signal is depicted in Fig. 8.

Based on the simulation parameters which reported in Table 1, the traditional PTS method requires large number of computations. Figure 9 shows the computational complexity of proposed method by using the same configurations with different SNR values. At each SNR, the proposed method requires less computational time for completion of algorithm. Figure 10 shows the Computational time comparison.

It can be seen that every time the improved method is optimized, the PAPR is reduced and the performance has been improved. Through simulation, it is found that as the proposed method exhibits better results compared with the traditional approach. The proposed method is compared with conventional PTS method, and corresponding results are shown in Fig. 11. From Fig. 12, it is clearly indicating that PAPR is reduced compared with other methods.

In order to validate proposed method is compared with SLM, original and conventional PTS method. In Fig. 12, all three methods are compared with the original CCDF results which is no PAPR reduction algorithm is applied. The PAPR results of proposed method have been reduced when compared with other methods.

In OFDM system, at the receiver side the bit error rate (BER) metric calculation is important. After calculating the PAPR at transmitter side, the BER values are measured at receiver side in underwater communication scenario and corresponding BER performance is depicted in Fig. 13. Similarly, the PAPR values of each methods at random locations are mentioned in Table 2, and also, the PAPR reduction performance comparison results are reported in Table 3.

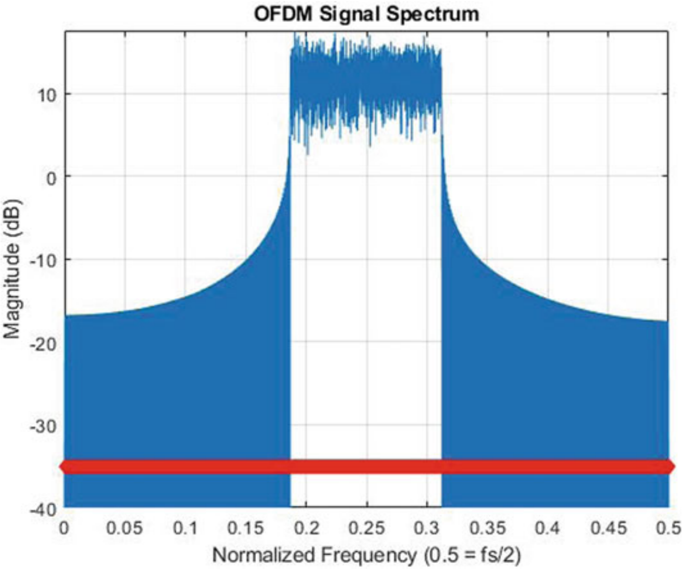


Fig. 9 Spectrum of OFDM signal

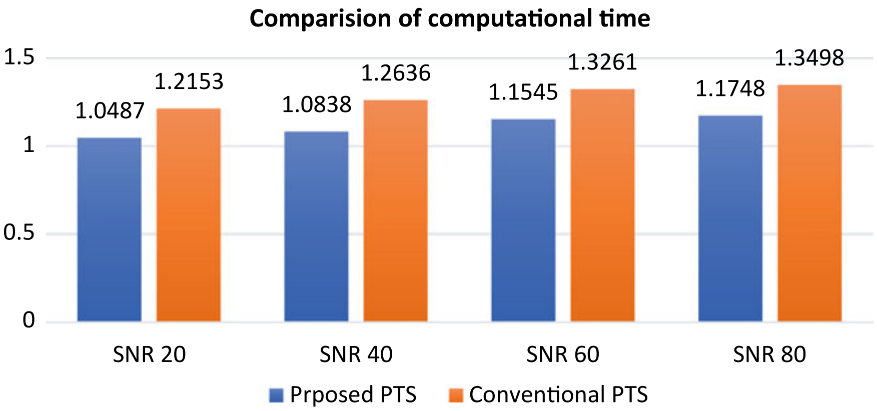


Fig. 10 Computational time comparison

Table 3 shows the proposed PTS method's PAPR reduction. The proposed PTS method is 44.44% more efficient than the conventional PTS technique.

In Fig. 14 shows that SoC-based PAPR reduction implementation results nearly matched with Matlab results obtained in Fig. 11. It is clearly indicating that, 64.28% PAPR reduction is achieved compared with conventional methods.

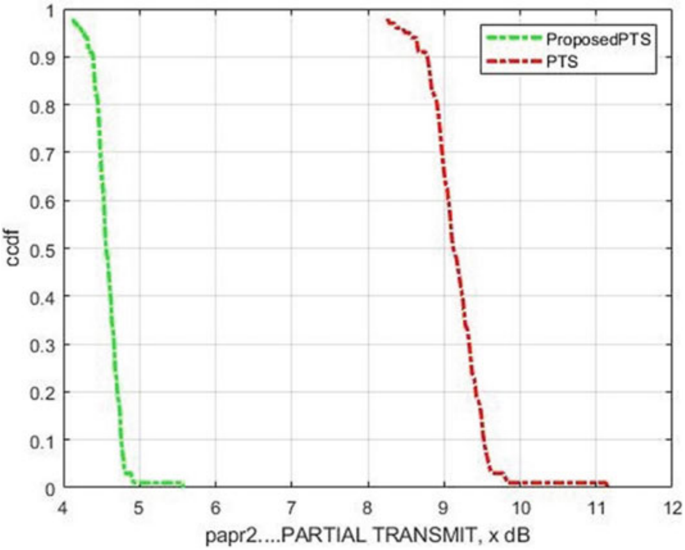


Fig. 11 PAPR comparison results in underwater communication

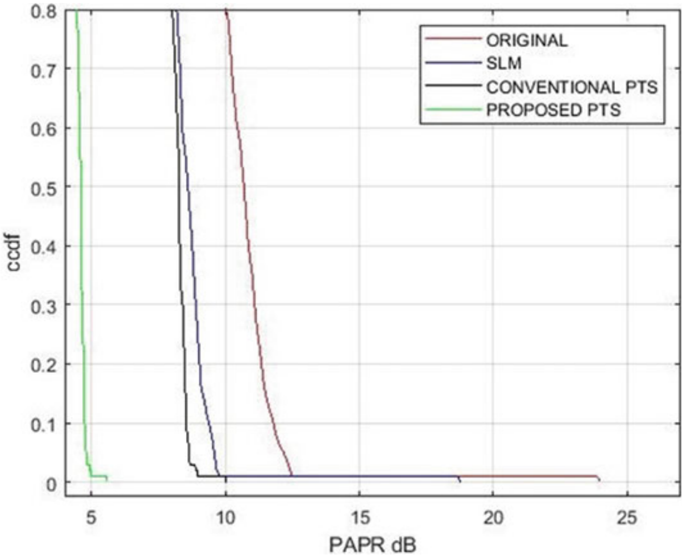


Fig. 12 Comparison between proposed PTS with other method in underwater communication

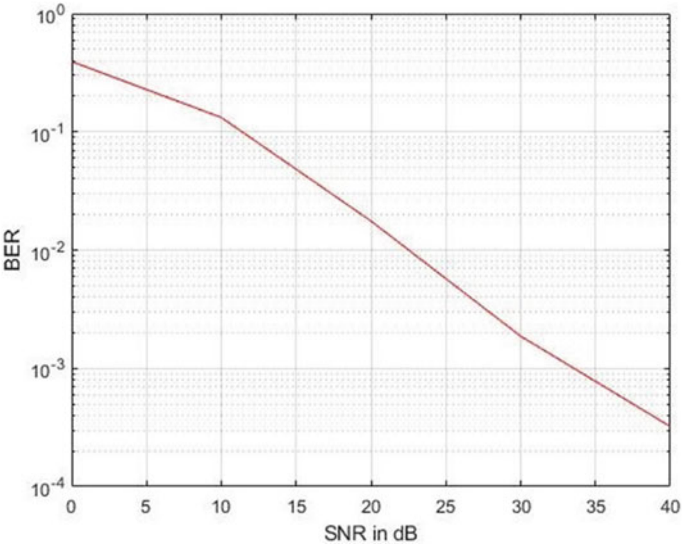


Fig. 13 Results of BER versus SNR OFDM for underwater communication

Table 2 PAPR values of proposed PTS, conventional-PTS, SLM and original

Proposed PTS	Conventional PTS	SLM	Original
4.2707	7.6872	8.3479	9.7161
4.3120	7.7615	8.6775	10.3178
4.3670	7.8606	9.1169	10.7691
4.4496	8.0092	9.7760	11.6717
4.5321	8.1826	10.5449	12.7247

Table 3 PAPR reduction comparison results

Proposed PTS versus original	Proposed PTS versus SLM	Proposed PTS versus conventional PTS
64.28%	56.89%	44.44%

7 Conclusion

In general, PAPR reduction plays important role in OFDM communication. Due to the characteristics of underwater channels, especially the limited bandwidth, OFDM is widely used due to its high spectral efficiency and ability to resist multipath fading. However, OFDM also has its shortcomings, one of which is the relatively high PAPR. This problem leads to saturation of the power amplifier and therefore distortion of signals that are not allowed in underwater communication. The number of optimization stages is increased to find the optimal phase in multiple small intervals. The phase angle of 0 to π can be divided into multiple parts to increase the performance.

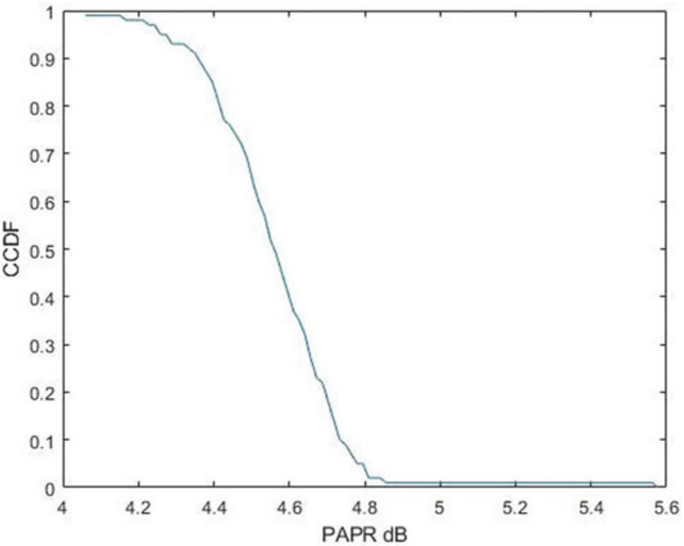


Fig. 14 Proposed PTS PAPR result in Simulink

The computational complexity of the proposed PTS algorithm is much lower than the conventional algorithm. As per the results and discussion, this work proposes the multi-level optimization PTS algorithm which reduces the phase element of each step optimization. The number of optimization stages is increased to improve the performance of the system. By simulating the proposed multi-level optimization algorithm, 64.28% PAPR is reduced compared with other conventional methods.

References

1. Wen, M., Ye, B., Basar, E., Li, Q., Ji, F.: Enhanced orthogonal frequency division multiplexing with index modulation. IEEE Trans. Wireless Commun. **16**(7), 4786–4801 (2017)
2. Dang, S., Ma, G., Shihada, B., Alouini, M.-S.: Enhanced orthogonal frequency-division multiplexing with subcarrier number modulation. IEEE Internet Things J. **6**(5), 7907–7920 (2019)
3. Chin, W.-L., Kao, C.-W., Qian, Y.: Spectrum sensing of OFDM signals over multipath fading channels and practical considerations for cognitive radios. IEEE Sens. J. **16**(8), 2349–2360 (2016)
4. Zheng, J., Chen, R.: Linear processing for intercarrier interference in OFDM index modulation based on capacity maximization. IEEE Signal Process. Lett. **24**(5), 683–687 (2017)
5. Mao, T., Wang, Z., Wang, Q., Chen, S., Hanzo, L.: Dual-mode index modulation aided OFDM. IEEE Access **5**, 50–60 (2016)
6. Rashich, A., Kislitsyn, A., Gorbunov, S.: Trellis demodulator for pulse shaped OFDM. In: IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), pp. 1–5 (2018)

7. Ma, X., Wang, T., Lin, Y., Jin, S.: Parallel iterative inter-carrier interference cancellation in underwater acoustic orthogonal frequency division multiplexing. Wireless Personal Commun. **2**, 1603–1616 (2018)
8. Galton, I.: Digital cancellation of D/A converter noise in pipelined A/D converters. IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process. **47**(3), 185–196 (2000)
9. Adebisi, B., Anoh, K., Rabie, K.M.: Enhanced nonlinear companding scheme for reducing PAPR of OFDM systems. IEEE Syst. J. **13**(1), 65–75 (2018)
10. Jawhar, Y.A., Audah, L., Taher, M.A., Ramli, K.N., Mohd Shah, N.S., Musa, M., Ahmed, M.S.: A review of partial transmit sequence for PAPR reduction in the OFDM systems. IEEE Access **7**, 18021–18041 (2019)
11. Fulai, Z., Luokun, L., Jinjin, Y.: DFT-spread combined with PTS method to reduce the PAPR in VLC-OFDM system. In: IEEE 5th International Conference on Software Engineering and Service Science, pp. 629–632. IEEE (2014)
12. Wang, L., Liu, J.: Cooperative PTS for PAPR reduction in MIMO-OFDM. Electr. Lett. **47**(5), 351–352 (2011)
13. Zhang, X., Tao, G.: The research of improved PTS method for peak-to-average power ratio reduction. In: IET 3rd International Conference on Wireless, Mobile and Multimedia Networks, pp. 104–107 (2010)
14. Li, X., Cimini, Jr L.J.: Effects of clipping and filtering on the performance of OFDM. IEEE Commun. Lett. **5**, 131–133 (1998)
15. Ochiai, H., Imai, H.: Performance analysis of deliberately clipped OFDM signals. IEEE Trans. Commun. (2002)
16. Wulich, D., Goldfeld, L.: Reduction of peak factor in orthogonal multicarrier modulation by amplitude limiting and coding. IEEE Trans. Commun. **50** (2002)
17. Pauli, M., Duchenbecker, H.P.: Minimization of the intermodulation distortion of a nonlinearly amplified OFDM signal. Wireless Pers. Commun. **4**(1), 93–101 (1996)
18. Van Nee, R.J., De Wild, A.: Reducing peak-to-average power ratio of OFDM. In: IEEE Vehicular Technology Conference (VTC'99-fall), Amsterdam, Netherlands (1999)
19. Nikookar, H., Prasad, R.: Weighted multicarrier modulation for peak-to-average power reduction. In: IEEE ICT'99, Cheju, South Korea (1999)
20. Gatherer, A., Polley, M.: Controlling clipping probability in DMT transmission. In: Proceedings of the 31st Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA (1997)
21. Wang, X., Tjhung, T.T.: Reduction of peak-to-average power ratio of OFDM system using a companding technique. IEEE Trans. Broadcast. **45**(3) (1999)
22. Huang, X., Lu, J., Zheng, J., et al.: Reduction of peak-to-average power ratio of OFDM signals with companding transform. IEEE Electron. Lett. **37**(8), 506–507 (2001)
23. Agarwal, D., Sharan, N., Ponmani Raja, M., Agarwal, A.: PAPR reduction using precoding and companding techniques for OFDM systems. In: International Conference on Advances in Computer Engineering and Applications, pp. 18–23. IEEE (2015)
24. Anoh, K., Tanriover, C., Adebisi, B.: On the optimization of iterative clipping and filtering for PAPR reduction in OFDM systems. IEEE Access **5**, 12004–12013 (2017)
25. Rateb, M., Labana, M.: An optimal low complexity PAPR reduction technique for next generation OFDM systems. IEEE Access **7**, 16406–16420 (2019)
26. Anoh, K., Tanriover, C., Adebisi, B., Hammoudeh, M.: A new approach to iterative clipping and filtering PAPR reduction scheme for OFDM systems. IEEE Access **6**, 17533–17544 (2017)
27. Ikram, S.A.: PAPR reduction in OFDM systems using peak insertion. AEU-Int. J. Electron. Commun. **69**(2), 573–578 (2015)
28. Ce, K., Liu, Y., Hu, M., Zhang, H.: A low complexity PAPR reduction method based on FWFT and PEC for OFDM systems. IEEE Trans. Broadcast. **63**(2), 416–425 (2017)
29. Minhoe, K., Lee, W., Cho, D.-H.: A novel PAPR reduction scheme for OFDM system based on deep learning. IEEE Commun. Lett. **2**(3), 510–513 (2017)
30. Kim, K.-H.: On the shift value set of cyclic shifted sequences for PAPR reduction in OFDM systems. IEEE Trans. Broadcast. **62**(2), 496–500 (2016)

31. Wei, W., Hu, M., Li, Y., Zhang, H.: A low-complexity tone injection scheme based on distortion signals for PAPR reduction in OFDM systems. IEEE Trans. Broadcast. **62**(4), 948–956 (2016)
32. Al-Jawhar, Y.A., Ramli, K.N., Ahmed, M.S., Abdulhasan, R.A., Farhood, H.M., Alwan, M.H.: A new partitioning scheme for PTS technique to improve the PAPR performance in OFDM systems. Int. J. Eng. Technol. Innov. **8**(3), 217 (2018)
33. Vittal, M.V.R., Rama Naidu, K.: A novel reduced complexity optimized PTS technique for PAPR reduction in wireless OFDM systems. Egyptian Inf. J. **18**(2), 123–131 (2017)
34. Yajun, W., Wang, M., Xie, Z.: A PAPR reduction method with EVM constraints for OFDM systems. IEEE Access **7**, 171830–171839 (2019)
35. Jie, Z., Dutkiewicz, E., Liu, R.P., Huang, X., Fang, G., Liu, Y.: A modified shuffled frog leaping algorithm for PAPR reduction in OFDM systems. IEEE Trans. Broadcast. 61(4), 698–709 (2015)
36. Bao, H., Fang, J., Wan, Q., Chen, Z., Jiang, T.: An ADMM approach for PAPR reduction for large-scale MIMO-OFDM systems. IEEE Trans. Veh. Technol. **67**(8), 7407–7418 (2018)
37. Liu, X., Zhang, X., Xiong, J., Gu, F., Wei, J. An enhanced iterative clipping and filtering method using time-domain kernel matrix for PAPR reduction in OFDM systems. IEEE Access **7**, 59466–59476 (2019)
38. Ali, M., Rao, R.K., Parsa, V.: PAPR reduction in OFDM systems using clipping based on symbol statistics. In: 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1–5. IEEE (2017)
39. Hu, M., Wang, W., Cheng, W., Zhang, H.: A generalized piecewise linear companding transform for PAPR reduction in OFDM systems. IEEE Trans. Broadcast. **66**(1), 170–176 (2019)
40. Zhang, S.-Y., Shahrrava, B.: A selected mapping technique using interleavers for PAPR reduction in OFDM systems. Wireless Pers. Commun. **99**(1), 329–338 (2018)
41. Mhatre, K., Pandit Khot, U.: Efficient selective mapping PAPR reduction technique. Procedia Comput. Sci. **45**, 620–627 (2015)
42. Zhang, S.-Y., Shahrrava, B.: A hybrid PAPR reduction scheme for OFDM systems using perfect sequences. IEEE Trans. Broadcast. **66**(1), 177–186 (2019)
43. Matsumine, T., Ochiai, H.: A novel PAPR reduction scheme for polar-coded OFDM systems. IEEE Commun. Lett. **23**(12), 2372–2375 (2019)

Analysis of Twitter Data for Identifying Trending Domains in Blockchain Technology

Sahithya Mareddy and Deepa Gupta

Abstract Opinion data collection is one of the most important forms of data analysis to understand and gain more insight about the trending information related to any domain or technology. The need for opinion data mining on the Twitter data is the demand of the indeed titled historical big data era. Blockchain is the technology which was introduced for cryptocurrency and later claimed to be embraced in most of the technologies because of its efficiency in ensuring privacy, security, and data management. With the increase in the popularity of blockchain technology, the opinion data collection for the blockchain technology is becoming compulsive to identify its substance in the practical application of different sectors. The utmost intention of this research is twitter data analysis centred on the domains that are believed to be domains that apply blockchain technology and hence ascertain that they are active and trending domains. The data analysis is performed on the tweets downloaded using tweepy API. This research engages different data visualizations, and Domain Identification by extracting features from tweets. The trend analysis for the opinion mining is accomplished by considering the re-tweets of the considered tweets. The proposed analysis is carried out on tweets which are carefully streamed using filter words related to the domain which claim to be active applicants of the blockchain technology. We will focus on techniques for the extraction of the Twitter data related to blockchain, processing, segregation, pattern visualization and trend identification by considering natural language processing paradigm.

Keywords Blockchain · Twitter · Social media analytics · Opinion mining · Trending · Domain · Re-tweet · Application

S. Mareddy (✉) · D. Gupta
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India

D. Gupta
e-mail: g_deepa@blr.amrita.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_49

1 Introduction

The freedom of expression becomes a reality in the real word only after the immense rise in the spread of the internet and because of the social networking platforms. Social networking platforms are reasoned to be a medium for people to express their views, share their thoughts, etc. Such openness made social network sites favourable among the current universe. Opinion data collection is the key in the data manipulated world to understand the people's needs which requires data for the analysis. Social Media Analytics is the one which provides the data for opinion data analysis. The data available on social media platforms such as Twitter, Facebook, LinkedIn, Pinterest, and Instagram are gathered in social media analytics from which valuable information is collected for taking the right decision in any business or domain front. Among the aforementioned social media platforms, Twitter is considered more useful because of its widespread use and tweet-based information sharing. Tweets are nothing but data aggregated in Twitter terms. The reason for conceiving Twitter as the data source is because of its increase in active users over each successive month. As per the statistics, Twitter produces roughly 500 million which is enormously large enough for any data analysis application.

Twitter is started in 2006 as a micro-blog platform with some post format called tweets [1]. Tweets allow the user to express user information as a post which is no longer than 140 characters which made it more reliable. The privacy of the posted tweets is controlled by the individual user where it can private amongst the user followers or public for all Twitter users. To make the analysis easier, Twitter introduced a new feature called hashtags which helps in segregating or grouping the tweets. Such availability made the Twitter user explore the tweets on a particular hashtag. When a hashtag is tweeted more frequently, it may be classified as trending by Twitter's trend detection algorithm [2]. This principle is adopted in this research for analyzing the blockchain trend. There are three varieties of tweets on Twitter namely Original Tweets, replies, re-tweets. Original Tweet is generally written by an individual user which appears in their profile page and such original tweet can be re-tweeted by any other users. re-tweets are useful for sharing attractive posts and links through the Twitter platform. Twitter has a valuable data source related to different emerging technologies like artificial intelligence, cloud computing, blockchain technology, and many more. The first-string intention of this research work is to analyze the Twitter data for the trending domains in which the blockchain technology is utilized upon. The reason for the consideration of the blockchain as the key factor is because of its breakthrough in the change in the world as we know of starting from the banking sectors to education to healthcare and from embedded system to standalone applications and even to security [3]. The enormous motivation of blockchain in several sectors increased its proposal in numerous application in recent years.

1.1 Blockchain

Blockchain is a technology which is linked to the digital currency Bitcoin. It came to the surface in 2008 with the research paper "Bitcoin: A Peer-To-Peer (P2P) Electronic Cash System" [4]. Bitcoin means money made electronically without any central authority to prevent the falsification of the data, duplication in payment, intrusion detection, etc. Blockchain technology is reasoned out to be essential because of underlying technologies such as hashing-based encryption, cryptography, digital signatures, etc. Another reason for the Blockchain uprising is the simplicity of record chain which makes it unchangeable. Blockchain is defined as a distributed ledger with the capability of maintaining the integrity of transactions by decentralizing the ledger among participating users [5]. Blockchain seems more secured private because of its dynamic chain of records encrypted using hash. Once the data is stored in the blockchain with hash, it is very difficult to change the stored data. Some of the quality which describes the importance of the blockchain are Consensus, faster settlement, distributed computation, enhanced security, Information storage, decentralized technology, Provenance, Immutability, and Access control. The major advantages of using Blockchain technology are Transparency, Reduced transaction costs, Decentralization, Faster transactions, User-controlled networks, and Fraud prevention. There are mainly three types of blockchains that have emerged namely public Blockchain, private Blockchain, and Consortium or Federated Blockchain. Blockchain technology is getting growing attention from various fields like IoT, cloud computing, big data and many more. An Enormous number of articles have been published focusing on blockchain technology explaining that it can be used in different applications like healthcare, business, education, security, governance, IoT, and data management. By this, we can understand how the trend is moving towards blockchain in research. Social Media is the best platform to analyze real-world insights.

1.2 Need of Blockchain Trend Identification

The increase in the attention over blockchain is evident because of its application-specific possibilities but it's not well adjusted among many of the blockchain subtopics. The target market accelerating the blockchain usage still needs the fluctuation to be rectified. The predictability factor of any new technology should be well known among the users for its wide acceptance. The future of blockchain as a ubiquitous technology is losing its trendiness in recent years because of its questionable application emergence over different domains. This is the void which tend the need of identification of the blockchain trending domains/technologies.

1.3 Subjective Facts

In this research, the objectives are centred on blockchain opinion mining. It is an open research field area with its application in many of the fields. With this subjective trend identification, the usage of the blockchain will be made well among for the researchers to continue their research using blockchain within their application-specific environment. The oblique aspects in the proposed work are deliberated as perspective below.

- Data collector perspective
- Data Pre-processor perspective
- Hashtag Analysis perspective
- Feature Extraction perspective
- Vector SparseMatrix perspective
- Domain Trend Identification perspective

In this paper, we proposed an opinion mining model centred on blockchain which will give away which domain talk about the blockchain more. In this study, Twitter data will be used to identify whether the blockchain technology is really implemented for the practical applications or is just a buzz word which is used only in research.

The tweets are extracted from the Twitter platform using tweepy [6]. Tweepy is an open-source API on python which enables the user to communicate with the Twitter platform on mutual consent using an authentication method called OAuth. This is a new form of authentication method which is different from nominal username and password-based authentication. OAuth authentication requires consumer key, access key token provided by the Twitter development account of the Twitter user which makes it more secured. Tweepy is trusted over many other existing API because of its possibility to get any object available on Twitter such as user-based tweets, hashtag-based tweets, trend-based tweets, etc. In this research, the real-time Twitter data related to blockchain technology is streamed using the StreamListener object of the tweepy.

The proposed work applies different data visualizations and Domain trend identification by extracting features. The different pattern visualizations focused are hashtag analysis, region visualization, and hashtags with maximum re-tweet count.

2 Literature Review

This section provides an overview of different approaches used for analyzing the Twitter data and type of inferences they have made in the literature. To our knowledge, there has been no related work on analyzing blockchain technology using Twitter data. However, there has been extensive work in other areas related to this research, which include analysis involving Twitter data for analyzing different areas like stocks, chemotherapy, smartphone brands, health domain, and others.

The reason for considering the information from Twitter as the data is because of the highlights in the paper [7], which made the fact clear that the social networking sites such as Twitter, Facebook, etc. had revolutionized the productivity of the information in the form of sharing, viewing, etc. Bordino in [8], demonstrated the stock trading correlation by analyzing the Twitter data using sentiment analysis. The conclusion from their research found query volumes as a viable option. The disadvantage found in this research is because of limitless data collection without any conditional threshold. The tweet based analysis to find the box-office prediction was developed by Asure et al. [9]. The decision in the analysis was found to be influenced by emotions. This research opened a new gateway that the social media can impact decision making also it can accurately guide us in the sentiment analysis by the direct reflection of the larger population opinions. Analysis on minute-by-minute by stock price centred on Twitter data was developed by Pieter de Jon [10]. The result concluded in an evident fact that stock returns were mostly influenced by the Twitter data.

Venugopalan and Gupta in 2015 developed a classifier model for sentiment analysis where Support Vector Machine (SVM) and J48 classifiers are used to perform binary sentiment classification of the subjective tweets [11] in which they analyzed popular smart phone brands. Ling Zhang et al. in 2018 focused on sentiment analysis to know the feelings of patients about the results of chemotherapy and Sentiment analysis in TextBlob was used to find out the overall attitude of patients on chemotherapy cross cancer types and also used to discover the side effects of chemotherapy [12]. A paper on analyzing health technology-related discussions using tweets in which they presented the top technologies in health domain through hashtag analysis and top diseases through word analysis and their association through co-occurrence of words within the tweets [13]. They chose the Twitter as the data source over Facebook because of its tweet principle which made the data easy to handle. Manju et al. focused on exploring the sentiment analysis on Hindi tweets and proposed a model for dealing with challenges in extracting sentiment from Hindi tweets [14].

To overcome the problem of the symbols inclusive with words in tweets for humans, which are not formally defined with any model for the English lexicon, many research works had been developed over the recent years. The effort depicted by the authors [15] in such a words list for sentiment classification using machine learning algorithms were applauded among the research community. Moreover, the research of same type with alternative form of sentiment classification was developed by authors to analyze sentences of different lengths and words [16]. An analysis on different digital payment methods is carried out using Twitter data by Prabhsimran Singh et al. in which they have performed various techniques like hashtag analysis, sentiment analysis, network analysis, geo-location analysis [17]. As presented in [18], the authors depicted the usage of the machine learning algorithm for the classification of the sentiment in the Twitter messages. They also used emoticons as the key to the positive and negative polarity for the sentimental analysis. They adopted many of the unsupervised machine learning algorithms to attain the objective combining basic text mining skill set. Fran Casino et al. [19] in 2019 provided a systematic literature review of applications related to blockchain technology across the multiple industries in which they have demonstrated how blockchain-based applications develop over time in the research. Many of the recent research works studied over the Twitter data analysis framed using logarithmic probability rate concept. The authors in [20] postulated a model which takes into sentiment of all terms in each sentence to formulate an overall sentiment for a sentence that is under consideration. Usage of unsupervised techniques for the Twitter data analysis was pre-compiled as a model by authors in [21]. They depicted the classification of the sentiment through comparison of word lexicons, and sentiment values.

After going through several approaches used for analyzing the Twitter data, this study initially focused on data visualizations like hashtag analysis, region visualization and hashtags with maximum re-tweet count. Then the main focus is to identify the different domains in which blockchain technology is being used for which re-tweets count is considered as selection criteria, as re-tweeting does indicate a level of endorsement of the tweet. Trending Domains are identified by extracting features from the Tweets. To the best of our knowledge, this is the first study which focuses on exploring the Blockchain technology related discussions on Twitter.

3 Proposed Approach

This section briefly describes the methodologies adopted in the proposed approach. The proposed research is to analyze the user tweets related to blockchain technology. The system workflow and basic architecture of the proposed approach is depicted in Figs. 1 and 2 respectively.

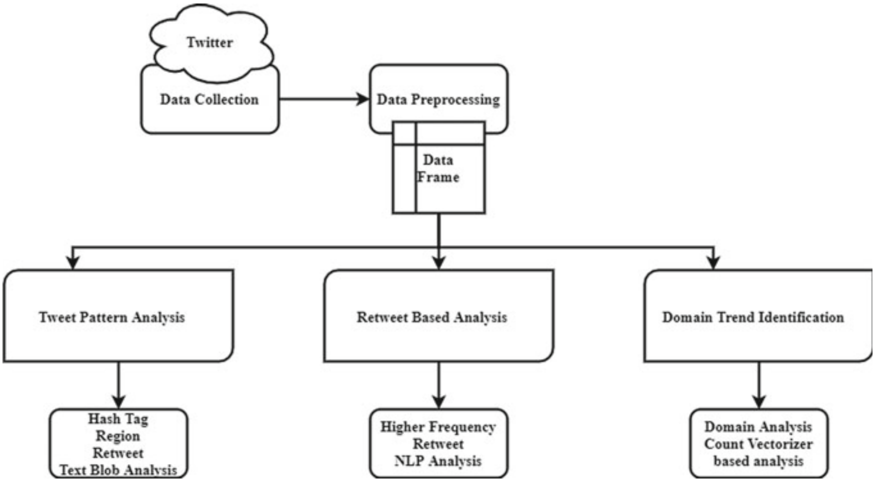
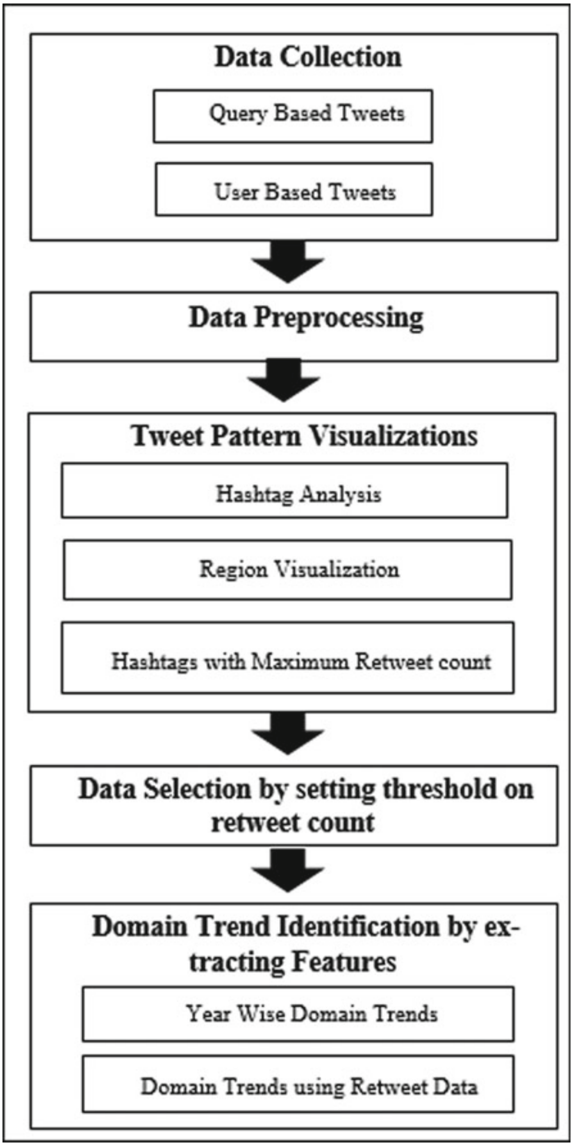


Fig. 1 System workflow of the proposed approach

Fig. 2 Schematic diagram for the proposed approach


Data Collection
Query Based Tweets
User Based Tweets
Data Preprocessing
Tweet Pattern Visualizations
Hashtag Analysis
Region Visualization
Hashtags with Maximum Retweet count
Data Selection by setting threshold on retweet count
Domain Trend Identification by extracting Features
Year Wise Domain Trends
Domain Trends using Retweet Data

As pictured in the aforementioned illustration, the system workflow starts from the data collection and is fed to the preprocessor where the natural language processing tool kit is utilized for the removal of special symbols and unwanted words from the Twitter data. The cleaned data is read as data frame using pandas library of python and are conferred to three work flows severally. They are tweet pattern analysis, re-tweet-based analysis and domain trend identification analysis. In the tweet pattern analysis, text blob is used for the hash tag extraction, region extraction and re-tweet extraction from the Twitter data frames. In the re-tweet based analysis, higher frequency re-tweets are separated from the Twitter cleaned data and are evaluated with different threshold of the re-tweet among the Twitter users. The third flow is in relation to the domain trend identification. In the domain analysis, count vectorizer from the scikit learn library is used for the conversion of the text blob to vectors. The extracted words vectors are considered as the features. The extracted features are then subjected to sparse matrix creation and popular domain segregation using nominal mathematical operations. In the proposed model, source file generation is performed using python as the programming language, and the software which is used for the development is anaconda. The reason for consideration of the python is because of its community reach in data analysis and text mining applications.

3.1 Data Collection

Data collection is the process which tends to acquire the available Twitter data related to the blockchain keyword query from the Twitter account. The data collected are subjected to availability as open source by Twitter. As previously explained, in proposed research tweepy is used as the API for extracting the tweets. The extracted Twitter data is by default in a non-structured JSON format. The JSON data is formatted to data frame and is stored in spreadsheets. The data collection block as specified in Fig. 1 is the building block for the proposed approach.

Two datasets were collected in this study related to blockchain. The first dataset is query based where the specific hashtags (#Blockchain OR #BlockchainConnect OR #BlockchainTechnology) were used to fetch the data. In total 8,802,838 tweets were extracted over a period of five months (August 2019 to December 2019). The second dataset is collected based on users for identifying the year wise domain trends. Here the official blockchain-related Twitter accounts are considered as users. As the blockchain trend has started around 2016, we have collected tweets from 2016 to 2019. In total 8576 tweets were extracted in the second dataset.

3.2 Data Preprocessing

Processing of textual data have an impact on the analysis over it. In most of the data analysis research works, the data processing steps are performed to make use of full handiness of the data collected. Even though the collected data is structured and saved in as a data frame. The nature of the data type is non-structured. There could be typos, slang usage, grammar mistakes, etc. The English lexical typos from the extracted Twitter data are processed using text blob analysis and natural language processing tool kit. The first process is the duplication removal where the tweet column which are duplicated by various users will be skipped. The next process is lower case conversion to make up the analysis process more feasible and easy to handle by the machine followed by stop word removal where all the words with no analytical values will be removed. The final step in the data pre-processing is stemming where the words of the same meaning and different forms are eliminated.

3.3 Tweet Pattern Visualizations

Tweet Pattern visualization is the introductory part of the complex data analysis process in the proposed work. The tweet pattern visualization is the process in which the Twitter data are visualized using three different aspects. The patterns are called upon the preprocessed data based on hashtag, region, and hashtag with maximum re-tweet count. In the hashtag analysis, the top hashtags are visualized in which blockchain is more popular based on the number of tweets. The location-based analysis is attempted in the proposed approach, to know in which regions people are discussing more about blockchain technology. As maximum re-tweet count shows the interest of people on any particular tweet, the final data visualization applied is hashtags with maximum re-tweet count. The extraction of the informants of the three specific factors utilized for the pattern visualization are derived using the nominal data analysis operation where the columns of the Twitter data are separated based on the region, re-tweet, and hashtag which are the part of the data collected through tweepy API. Upon collection, and after processing, the data frames are aggregated into columns for visualization using pandas and matplotlib libraries.

3.4 Data Selection by Setting Threshold on Re-Tweet Count

This is the second of the essence fact considered in the proposed work. Most of the Twitter trend analyzer uses the limitless data from the tweets as input for the opinion mining. In this research, an attempt is made to do a statistical analysis on Twitter data based on re-tweet count. The consideration of re-tweet count as the selection criteria increase the trust factor. As the higher the frequency of re-tweets, the higher the trust. To find the domain trend of blockchain technology on different re-tweet count, we have set the threshold count on re-tweets like 100, 200, 300, and so on. In regard to the tweet pattern visualization, re-tweet based analysis is also done using the matplotlib libraries.

3.5 Domain Trend Identification by Extracting Features

The domain trend identification is the final process in the proposed system workflow. The domain trends are identified by extracting numerical features from the preprocessed data in order to find the relevant information. The training dataset is generated with possible domains where the blockchain technology can excel and is compared with the original feature extracted from the tweet input. The output of the domain trend is the sectors where the blockchain technology is well adapted based on the tweets. The process in which the domain trend identification is made possible is backed up by the word to count vectorizer feature extraction method and sparse matrix generation algorithm. In the word to count vectorizer, the tweet data are subjected to conversion in matrix vectors. The matrix vectors are then exempted to sparse matrix generation where they are combined upon the training data set. The resultant matrix is then compared with original training data set for the extraction of the trending domains. The vectorizer API is used from the feature extraction module of scikit learn library of python.

4 Result Visualization and Trend Analysis

The entire implementation of this study has been done in a python programming language using some standard libraries like pandas, NumPy and scikit-learn. As mentioned earlier in methodology, result analysis is carried out in two levels. The Initial one is pattern visualizations based on hashtag, region and hashtag with maximum re-tweet count. And the second one is analyzing the domain trends. The quantifying parameters are employed for the result analysis because trend identification is the subjective fact rather than the sentimental analysis of the blockchain among the public.

4.1 Tweet Pattern Visualizations

Figures 3, 4 and 5 depicts the hashtag analysis, location-based analysis, and hashtags with maximum re-tweets analysis respectively. It is evident that, even with higher number of the training data given as input for the feature comparison and sparse matrix generation only the subtopics which are more related to the blockchain technology is projected over the plot as output which intercommunicate the advantage of the proposed domain identification model. As highlighted in [22], the most important event in XRP calendar is Swell by Ripple is not surprising, because it is the most used hashtag in Twitter where the ripple is a payment protocol which uses blockchain technology and swell is a conference for avid Ripple fans and holders [23]. It is also found that Swell 2019 was conducted on November 7, 2019, and as mentioned earlier the data is collected from August 2019 to December 2019. So this is one of the reasons that swellbyripple is the most used hashtag at that time period.

From the location-based analysis, it is found that blockchain-related tweets are more observed in Eastern Europe, Central Asia and South-Eastern Europe countries. As presented in [24], there are around 70 Europe Blockchain companies, so it is found that more blockchain-related tweets are from Europe countries.

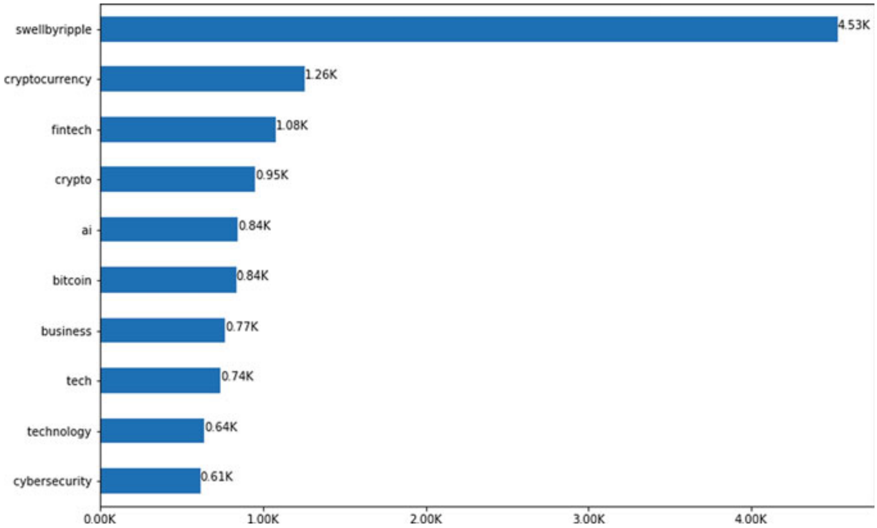


Fig. 3 Top hashtags in which blockchain is popular based on number of tweets

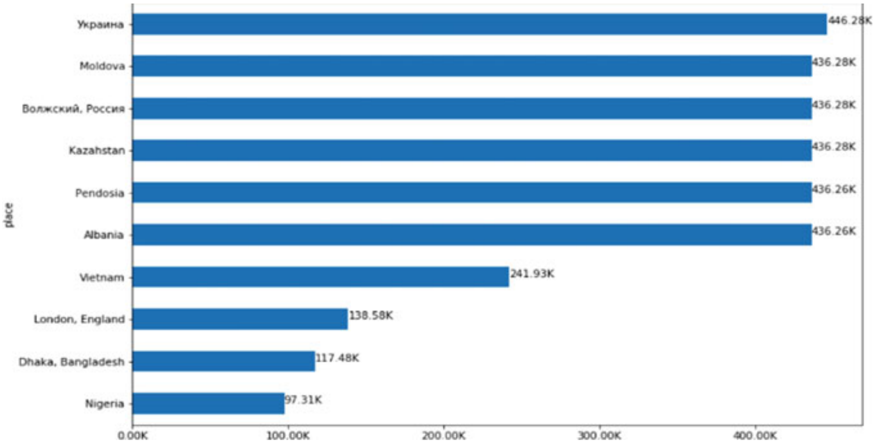


Fig. 4 Top regions with highest discussions about blockchain

From Fig. 4, it is observed that BitGreen is the most used hashtag with maximum re-tweet count because it is a sustainable cryptocurrency which is built on the energy-efficient green blockchain as mentioned in [25]. Sustainability is the next most used hashtag is not surprising, as presented in [26] blockchain can be a vital tool to boost sustainability. So we can observe that people who re-tweeted also felt that blockchain as a key to sustainability.

The results related to the tweet pattern visualization of three aspects convey the importance of the blockchain technology distribution over the different region,

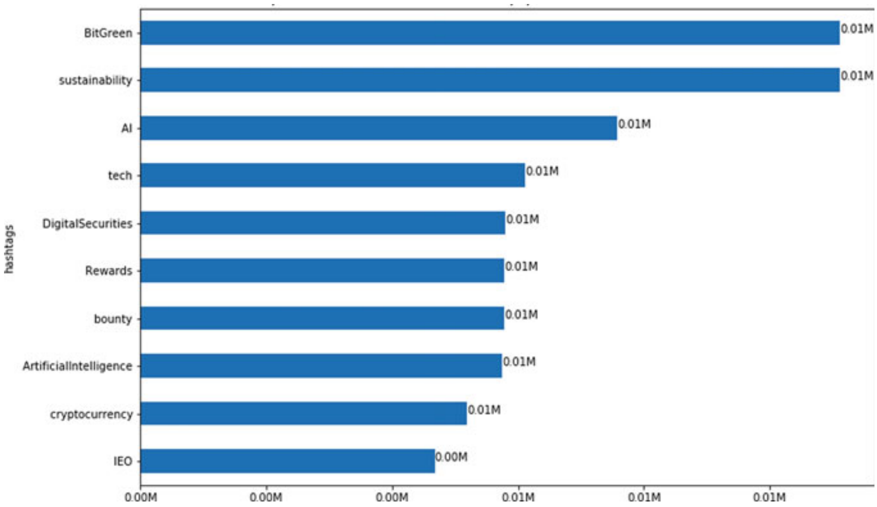


Fig. 5 Distribution of different hashtags according to maximum re-tweets

hashtag, and re-tweet which overcome the findings of the literature related to the blockchain war waged by environmentalist tagging blockchain as power consumption despoiler, carbon emission, and ecological damage. The proposed findings even with the conflict states that, blockchain is widely acknowledged irrespective of the region.

4.2 *Year Wise Domain Trends*

The rational motive for the consideration of the result analysis centred on the year-wise domain trends is to connect the findings of the blockchain growth over the period of year. As discussed earlier, the second dataset which is collected based on users is used for this analysis. The comparative analysis for the trending domain over the year of 2016, 2017, 2018, and 2019 is depicted from Figs. 6, 7, 8 and 9.

From the year-wise domain trend analysis, we observed that finance has huge trend during 2017 time period because IT was the top domain in 2016 but then in 2017 finance domain was equal with IT domain given that finance does not have much discussions in 2016. However there was again downfall of finance domain in 2018. From this analysis, it is clear that IT domain is constantly developing where as finance domain always has its good and bad times. From Fig. 9 it is found that blockchain is involved almost in 15 domains in 2019. The future of the blockchain technology seems omnipresent from the year wise results between 2016 to 2019 as it finds its application emerging thereby increasing its trend. Even though the domains are not fixed to constant stature over the application of the blockchain, they are evenly

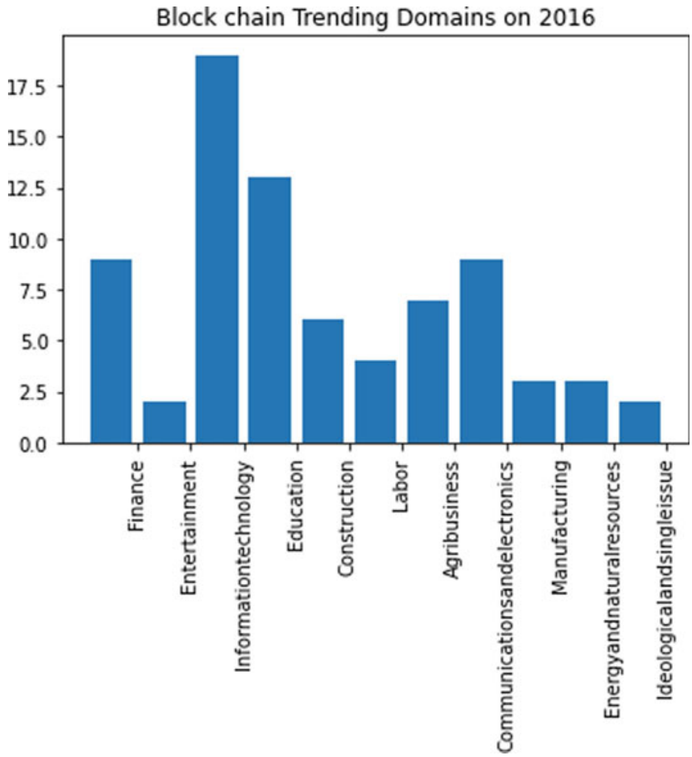


Fig. 6 Blockchain trending domain on 2016

distributed among different domains which shows the usage of the blockchain in the current era and its uniqueness in the future.

4.3 Domain Trend Analysis Using Re-tweet Data

This section briefs the result analysis of the blockchain trends over different frequencies of the re-tweet count. The statistical analysis is carried out in the proposed research work to analyze the results over many re-tweet count conditions headed for the blockchain trend analysis. After 1000 re-tweet count, we found the flat behaviour in the results, so we have stopped at 1000 re-tweet count. Figures 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19 illustrates trending domains of each re-tweet count respectively.

Domain Trends. From the domain trend results, we can observe that even with the people who are discussing about Finance domain, but there is not much trend when re-tweets are considered as we can see it was in top till 300 re-tweet count but then there is a downfall as re-tweet count is increasing. On further analysis., it is found

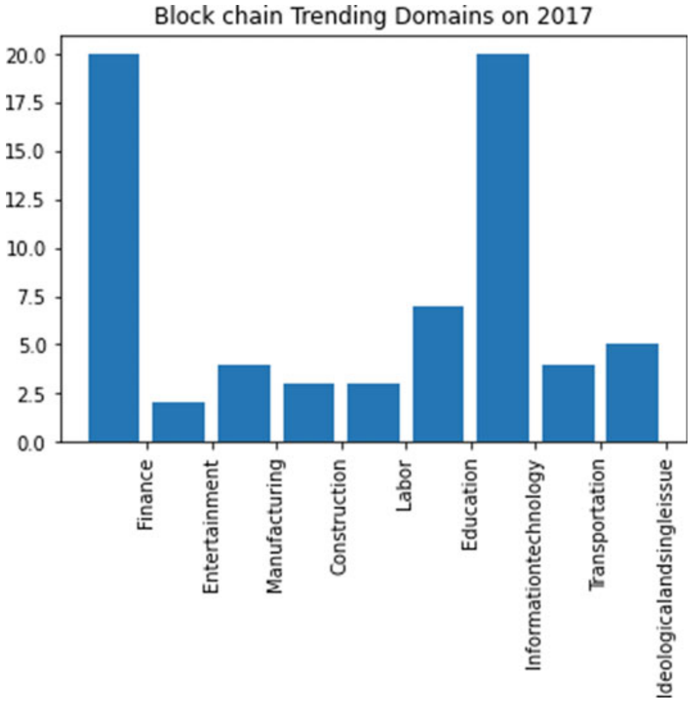


Fig. 7 Blockchain trending domain on 2017

that IT (Information Technology) domain remains constant even when re-tweet count is increasing as it is at the top position even when re-tweet count is increasing. In addition to statistical acceptance, it is evident from the results that with the increase in the re-tweet count, there is decrease in the number of domains. The results inclining nature does not indicate that blockchain is not well used over different domains but it states the fact that even with lower lexicons the blockchain availability is still of need.

From both the domain trend analysis i.e. year wise and re-tweets data, it is obvious that IT domain is constantly advancing where as finance domain always has its ups and downs.

5 Conclusion and Future Work

In today's words, where everything is automated and because of the spacious amount of the data generated, the opinion mining seems relevant and easy to understand the betterment of any technologies in public credence. Social media data has a lot of

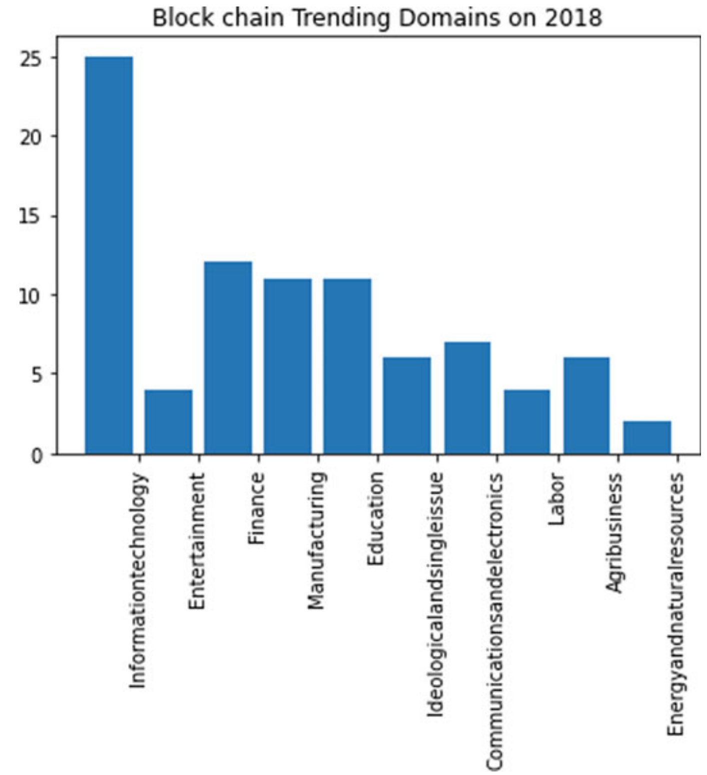


Fig. 8 Blockchain trending domain on 2018

knowledge hidden in its huge volumes of noisy, unstructured and dynamic collection of data. One of the main divisions of this research was finding the blockchain hashtags information relative to the sectors where blockchain technology is used based on the Twitter data. Blockchain technology is changing at a fast pace in different domains. As the trend is moving towards blockchain in research, this project focused on presenting the trend of blockchain technology using Twitter data. Statistical analysis was also deliberated viewing the influence of the limited data collection processed based on the re-tweet count. Basically, our goal was to find the trending domains in the blockchain thereby finding the public opinion of the blockchain in the application-specific environment. Normal survey-based mining are tiresome with better accuracy, but our model was able to project the desired output in automation fashion. Future scope of the proposed work includes endeavour on data collection, stacking sentimental analysis with trend identification, and including a web framework with the ability to find the trending domain for any topics or technologies which makes it feasible for everyone. In the future, this could be expanded to user interest based on age, and work experience which would give us an even clearer picture of the extent of blockchain in technology application-oriented environment. But as of

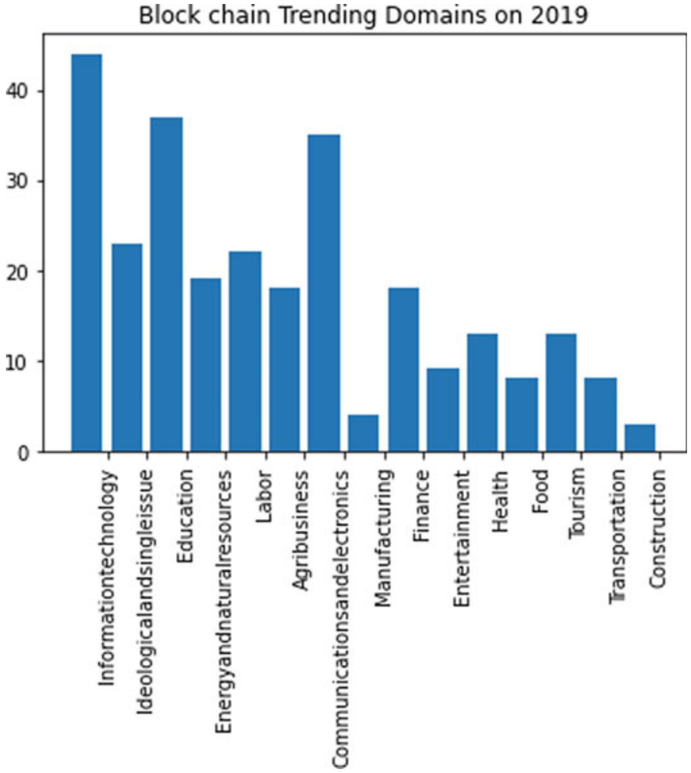


Fig. 9 Blockchain trending domain on 2019

now age-based information, Twitter APIs do not work on these parameters. However, in the long run, algorithms could be derived from users' tweets, photos, and "Bio" fields. Data could also be aggregated for years to see when the frequency of tweets related to the blockchain was being posted and to check if any spikes in data corresponded to any new product launch. Furthermore, we can add some unsupervised learning algorithm for trend pattern identification as of clustering problem.

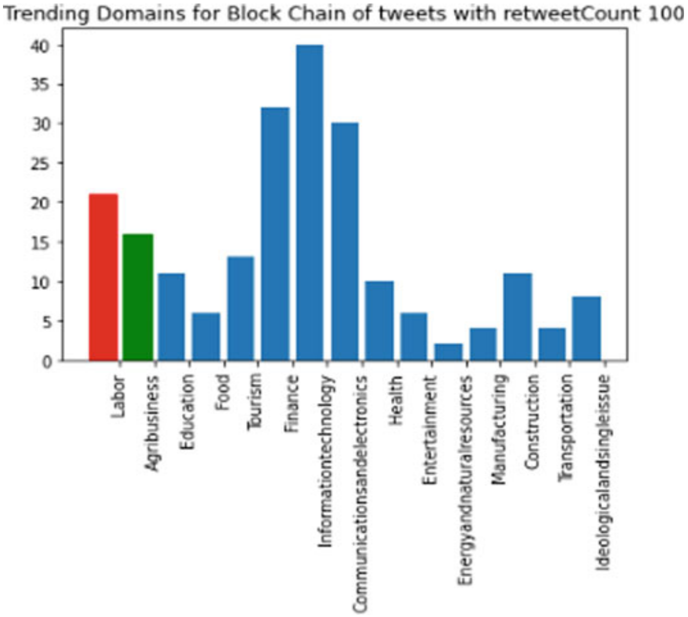


Fig. 10 Re-tweet count 100

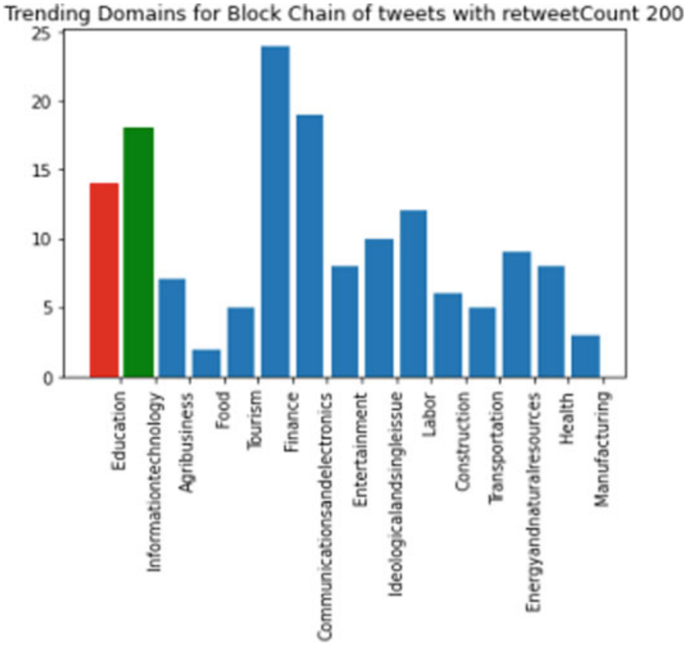


Fig. 11 Re-tweet count 200

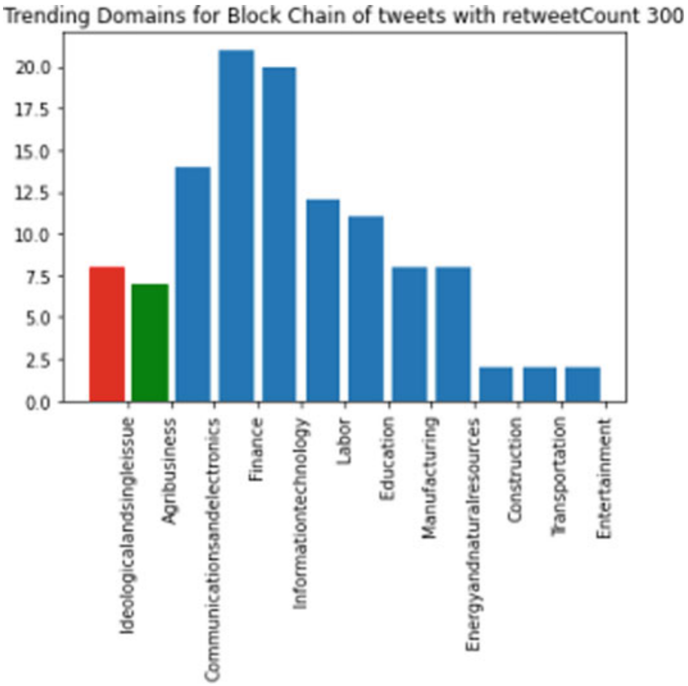


Fig. 12 Re-tweet count 300

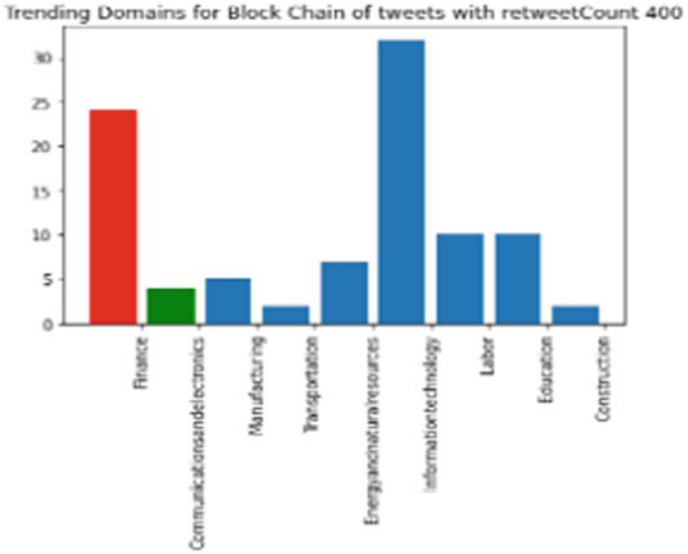


Fig. 13 Re-tweet count 400

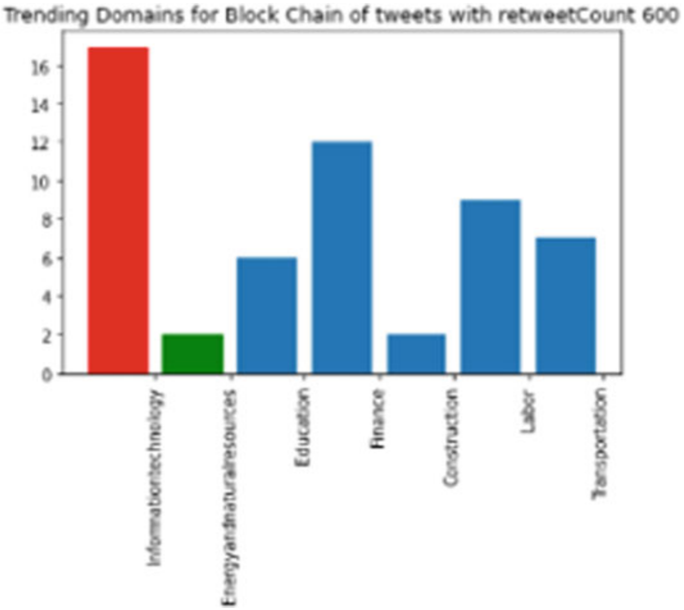


Fig. 14 Re-tweet count 500

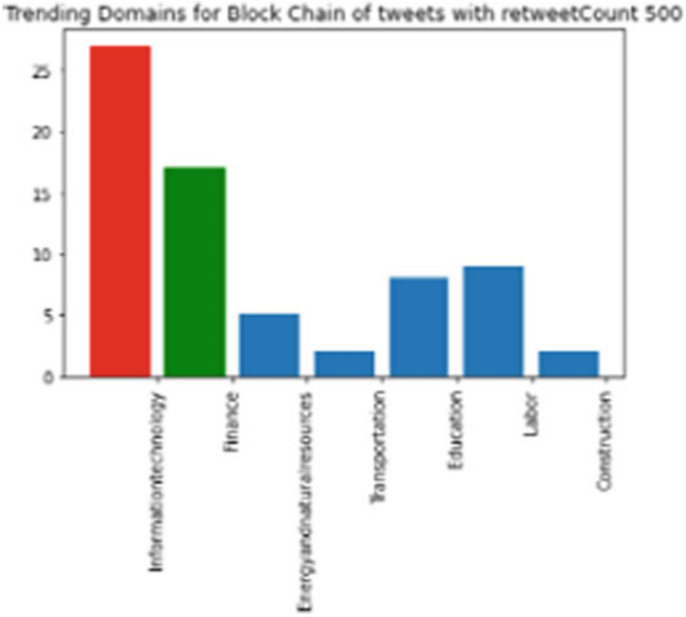


Fig. 15 Re-tweet count 600

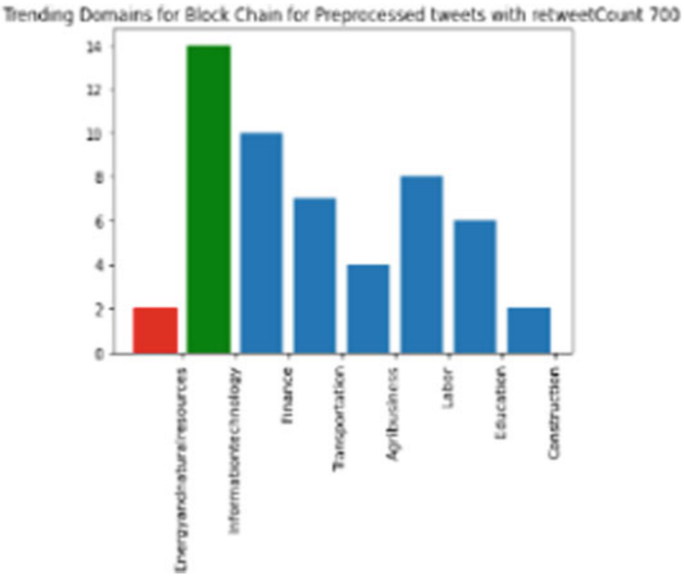


Fig. 16 Re-tweet count 700

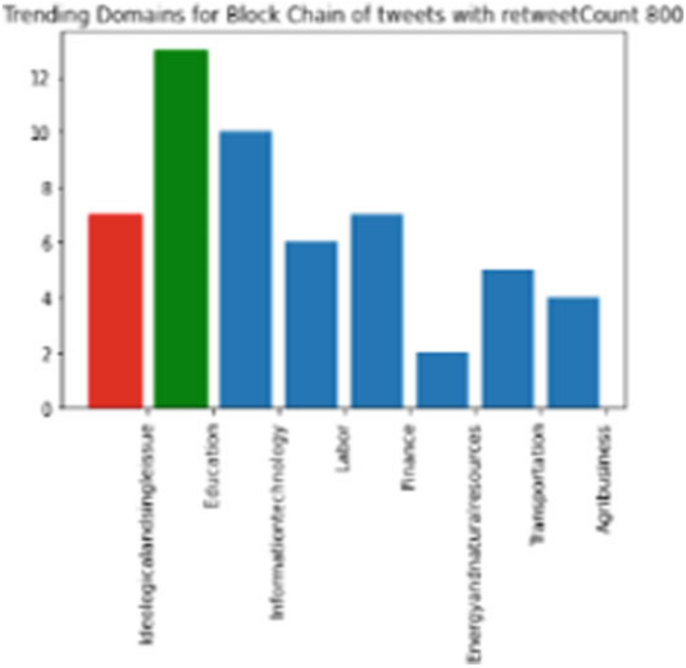


Fig. 17 Re-tweet count 800

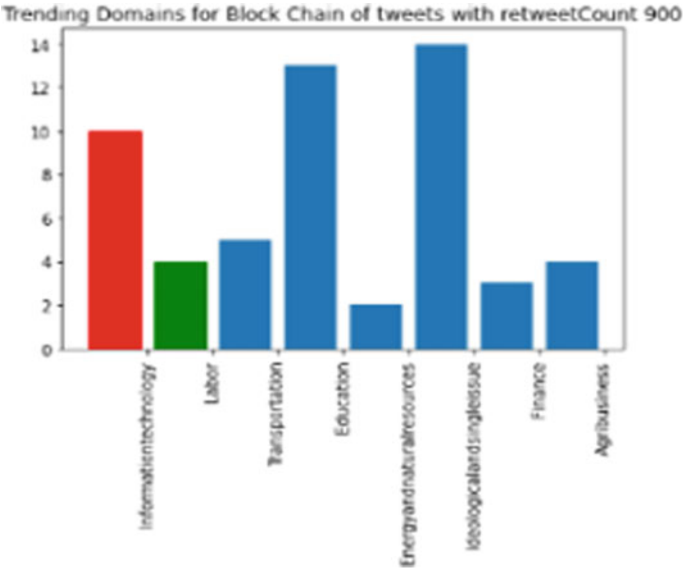


Fig. 18 Re-tweet count 900

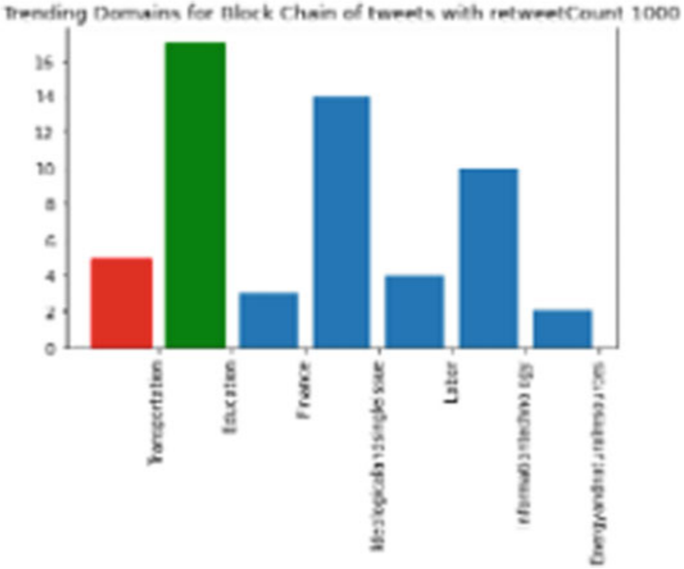


Fig. 19 Re-tweet count 1000

References

1. Jansen, J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: tweets as electronic word of mouth. JASIST **60**, 2169–2188 (2009). https://doi.org/10.1002/asi.21149
2. Inc. Twitter: FAQs about trends on Twitter (2016). https://support.Twitter.com/articles/101125
3. Adithya, M., Scholar, P.G., Shanthini, B.: Security analysis and preserving block-level data DE-duplication in cloud storage services. J. Trends Comput. Sci. Smart Technol. (TCSST) **2**(02), 120–126 (2020)
4. Nakamoto, S.: Bitcoin: a peer-to-peer electronic cash system (2008)
5. https://developer.ibm.com/technologies/blockchain/tutorials/cl-blockchain-basics-intro-bluemix-trs/
6. Twitter API. http://docs.tweepy.org/en/latest/
7. Nisbet, R., Elder, Miner, G.: Handbook of Statistical Analysis & Data Mining Applications (2009)
8. Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., Weber, I.: Web search queries can predict stock market volumes. PLoS ONE **7**(7), 1–17 (2012)
9. Asur, S., Huberman, B.A.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)
10. Pieter de Jong, P., Elfayoumy, S., Schnusenberg, O.: From returns to tweets and back: an investigation of the stocks in the Dow Jones Industrial Average. J. Behav. Finan. **18**(1), 54–64 (2017)
11. Venugopalan, M., Gupta, D.: Exploring sentiment analysis on Twitter data. In: 2015 Eighth International Conference on Contemporary Computing (IC3), Noida, pp. 241–247 (2015)
12. Zhang, L., Hall, M., Bastola, D.: Utilizing Twitter data for analysis of chemotherapy. Int. J. Med. Inf. **120**, 92–100 (2018)
13. Grover, P., Kar, A.K., Davies, G.: Technology enabled health—insights from Twitter analytics with a socio-technical perspective. Int. J. Inf. Manage. **43**(2018), 85–97 (2018)
14. Venugopalan, M., Gupta, D.: Sentiment classification for Hindi Tweets in a constrained environment augmented using tweet specific features. In: Mining Intelligence and Knowledge Exploration, pp. 664–670 (2015)
15. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1181
16. Irsoy, O., Cardie, C.: Opinion mining with deep recurrent neural networks. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 720–728 (2014)
17. Singh, P., Dwivedi, Y.K., Kahlon, K.S., Rana, N.P., Patil, P.P., Sawhney, R.S.: Digital Payment Adoption in India: Insights from Twitter Analytics. Springer International Publishing, pp. 425–436 (2019)
18. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision. http://www.stanford.edu/~alecmgo/cs224n/sigproc-sp.pdf
19. Casino, F., Dasaklis, T.K., Patsakis, C.: A systematic literature review of blockchain-based applications: current status, classification and open issues. Telematics Inform. **36**, 55–81 (2019)
20. Bose, R., Dey, R.K., Roy, S., Sarddar, D.: Analyzing political sentiment using Twitter data. Inf. Commun. Technol. Intell. Syst. 427–436 (2019)
21. Bose, R., Dey, R.K., Roy, S., Sarddar, D.: Topic modeling as tool to gauge political sentiments from Twitter feeds. Int. J. Nat. Comput. Res. **9**, 427–436 (2020)
22. Swell By Ripple, https://timeforcrypto.com/swell-by-ripple/
23. Ripple, https://www.investopedia.com/tech/what-ripple-swell/
24. Europe Blockchain Companies, https://cryptoslate.com/companies/region/europe/
25. BitGreen, https://www.mycointainer.com/assets/bitgreen/
26. Blockchain as Key to sustainability, https://www.sustainability-times.com/sustainable-business/blockchain-can-be-a-vital-tool-to-boost-sustainability/

Enhancing the Security for Smart Card-Based Embedded Systems

G. Kalyana Abenanth, K. Harish, V. Sachin, A. Rushyendra, and N. Mohankumar

Abstract Nowadays, security has become an uncertain phenomenon in the modern world. User data can be stolen and other privacy risks are posing a major threat to the society. With the increase in design complexity and cost of setting up a foundry, it has led to globalization of integrated circuit supply chain which poses many security threats like piracy and hardware Trojans, which leads to data leakage to the outside world. In this paper, we are proposing a security method for a smart card-based embedded system. Remote user authentication and key agreement scheme for smart cards are the way to go. It is a very practical solution to validate the eligibility of a remote user and provide secure operation of the system, hence securing the user data from untrusted sources.

Keywords Design for security · Hardware security · Embedded systems · ARM7 · GSM

1 Introduction

In today's world, technology is being developed at a rapid rate which leads to growth in all the fields. As a result, cost of materials is increasing with the increase in cost of living. Integrated circuits (IC) are essential component of electronic systems ranging from home appliances to military equipment. They are the building blocks of any computation system. Due to the increase in design complexities and cost of setting up a foundry, in current day semiconductor manufacturing, these IC are being designed in a globalized multivendor environment, thus leading to issues like IP piracy, data theft, and hardware Trojans, resulting in serious economic losses to IC designing companies [1]. Many defense techniques like water marking, logic locking, and many more were proposed to counter such issues. In this paper, our aim is to build a system that can withstand one of the powerful attacks called as "sensitization attack" [1, 2] by adding a suitable countermeasure for the attack.

G. Kalyana Abenanth · K. Harish · V. Sachin · A. Rushyendra · N. Mohankumar (✉)
Department of Electronics and Communication Engineering, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: n_mohankumar@cb.amrita.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_50

Nowadays, the use of smart cards has grown dramatically, increasing several security issues. With password-based authentication being the most widely used method, attacks such as brute force and reverse engineering technique, with technology advancement, have become very common and easier to execute. The risk of stealing the password is therefore high and the system is compromised once stolen. Even the owners will often be unaware of the attack on their system (smart card). So, a more efficient method in terms of security is need of the hour. What if the user enters the system even after entering an incorrect password, but deprived from the real application and functionalities?

In this paper, we have worked on this idea for enhancing the security for smart card-based embedded system. In the standard password-based authentication method, if the given input is wrong, then the user or hacker cannot go further into the system. Here in our paper what we are proposing is, in the event of an wrong input entered fifth time the user/hacker will be allowed inside the system, but it will be directed to an incorrect output which is deviated from the original application. This process makes it difficult for the anonymous user to find the correct password. If a person enters an incorrect password for five times continuously, then the entire system along with the processor will go in a locked state and an alert message will be sent to the owner's mobile phone via GSM module [3–6] and the data in the system's memory will be erased. In our system, the user can decide whether the data needs to be corrupted or erased.

Existing Model Versus Our Model

The existing model is only password-based; it cannot detect hacking and alert the user. In our model, we can detect hacking and alert the user and we are also protecting the data of the system.

Live Senario When the System Gets Hacked

As discussed earlier, the hacker will be allowed inside the system (blank screen), and the data in the SD card will be corrupted or erased (users' wish). And the entire system will go in a locked state. The entire system needs to be restarted to make the processor come out of the locked state. Now after the processor goes into locked state the restart of the system will take 3 min, and when the system agains functions and when it asks for the passsword the hacker will get only one chance if the hacker again enters the wrong password the system will again restart. This process will continue and the hacker will be put in an infinite loop which has a time complexity of 3^N min where N is the number of attempts the hacker is attempting. Another major advantage using our system is that it is not connected to the Internet so remote hacking is not possible. The hacker needs to access the mainframe in person to proceed with the hacking [7]. Further, the user will get the message alert when the system restarts for first time. So hacker can be caught and the data can be protected [8].

With this idea to develop a countermeasure for ARM-based devices using embedded systems, various other components with limited facilities and resources were attempted.

Directions

Section two deals with the literature survey and its outcomes. Section three deals with the materials and methods. Section four deals with the system design of the proposed model. Section five deals with the results and discussion obtained from the proposed model. Section six deals with conclusion.

2 Literature Survey and Its Outcomes

A basic understanding of SAT attack is attacking processors and designing of ANTI-SAT block as a countermeasure [2]. An idea on how hardware security modules and its implementation can be done, and how attacks can happen on the system and its prevention [1]. The concept of detecting hardware Trojans and the techniques used to prevent from hacking is explored. This model proposes a home security alarm system by applying WSN and GSM technology. This technique is quite useful since it covers various security issues like theft, gas leakage, and fire prevention [9]. When it detects any of those, it sends alarm message. This method was proved to be so reliable and uses less power. This model is based on GPS, and GSM for automobile anti-theft system's location of the vehicle can be obtained with GPS and can send information to the owner via GSM module [3]. One main advantage is this system is open for future developments like IoT. This paper gave us more information about the GSM modules and its applications. A remote user authentication scheme is using smart cards. This method has many advantages compared to the previous one. Mainly the password length is 64 bits only here.The efficiency of this method is so much better [4].Various authentication schemes and to get more insight on smart card technology and its vast applications are also provided. Further, it also states about its authenticity and security. They used advanced algorithms to provide security [5, 6]. Data sheet of the GSM module gives us the information on how to configure and program the module [10]. Data sheet of the ARM LPC2148 can be used for programming and connections purposes [11]. Security concerns on the integrated circuits (ICs) are one of the key points discussed in this paper. This helped us realize that security is, in general, important to any system. Corruption of data, leakage or theft of information, and obstructing the expected functionality of the system/design are some of the implications of improper security [12]. With the advancement of technology, it has become easier for hackers to breach old security measures. The problem of information security and security in defense and about asymmetric shift-based ciphering are discussed and understood; the required changes to be carried out in our system for increasing the security, the response time, and the other parameters should not be changed [13]. The importance of security of a system and learned how the security logic must be developed for a system such that the introduction of the encrypted circuit does not occupy much area, aimed at maximum efficiency is discussed [14]. The idea about the testing of the unit is obtained from [15]. Proper testing ensures reliability, security, and high performance which further results in time saving, cost effectiveness, and customer satisfaction. An approach is followed for this and carried out unit testing on the USB module for washing machine. For the purpose of data transfer between the washing machine and the USB module, a microcontroller was used for this embedded device. Microcontrollers are intended for embedded systems, and most of them are integral part of our day-to-day life. The use of microcontrollers for multiple applications was one of the driving factors for us to settle on our specific framework.

3 Materials and Methods

3.1 To Execute the Steps and Programs

There are various kinds of processors available in the market. They can be classified as CISC or RISC. Generally, RISC processors are very fast and reliable and less flexible. And as the name suggests the number of instructions for a RISC processor is less and all the instructions can be fetched in a single clock cycle. For this purpose, an ARM7 processor was taken, and it has been checked and verified for its basic functionalities before proceeding further. Before implementing the hardware part, we simulated each step to resolve any issues. The model which we chose for our implementation is LPC214–ARM7TDMI. It is designed by Philips. The reason we chose this particular board is based on our appication which has several inbuilt features, peripherals, and the number of input and output pins. This has 512-kB on-chip FLASH memory as well as 32-kB on-chip SRAM also. It will be more reliable as well as the efficient option for our application.

3.2 To Secure the Smartcard-based Systems

We need a smart card-based system to provide the security for and to safeguard the data it has. So, we decided to go with sim cards. To do that we used a GSM module [12]. There are various models available. We choose the model which will be supported by the ARM7 processor. So, we went with SIM900a GSM module and we integrated the SIM900a with the ARM7 processor.

3.3 To Store the Data the System Collects

We need a drive to store all the necessary data. We integrated a SD card module with the ARM7 processor by which all the data can be stored in the SD card. It is a 16-GB SD card.

3.4 To View the Outputs and Inputs

Liquid-crystal display (LCD) is used for displaying status or parameters in the system. It is an electronic display module. We used LCD 16 * 2. It is a 16-pin device which has 8 data pins and 3 control pins and the remaining 5 pins are for supply and backlight for the LCD. It means it can display 16 characters per line and there are two such lines. We integrated this LCD display to the ARM processor to view the outputs and to give inputs to the system.

3.5 Dot Matrix Keypad

Keypad is used as an input device to read the key given by the user and to process it. The matrix keypads are used in systems where human–machine interface is required in order to change the operating parameters of a system. In our project, we used a 4 * 4 dot matrix keypad. It consists of 4 rows and 4 columns. All these keys in the keypad are configured via programming. We followed the general order which is followed in the mobile phone's keypad.

3.6 Sim Card

To be able to use the GSM module, we need a sim card in it. It supports all the services available like 2G/3G/4G. Now choosing the sim card service providers is important. Because if we choose JIO service, it only provides 4G service meaning if you are not getting a 4G signal then it will show no signal available, whereas it is not the case with Airtel or Vodafone-Idea sim cards. Hence, we choose Vodafone-Idea sim card for this purpose.

3.7 Proteus 8 Professional

It is a simulation software which we have used to simulate all the scenarios and to check if the code and connections are correct.

3.8 Other Materials Include

Wires, breadboard, adapters (power source for ARM board [11] and GSM module) and RS232 cable (Fig. 1).

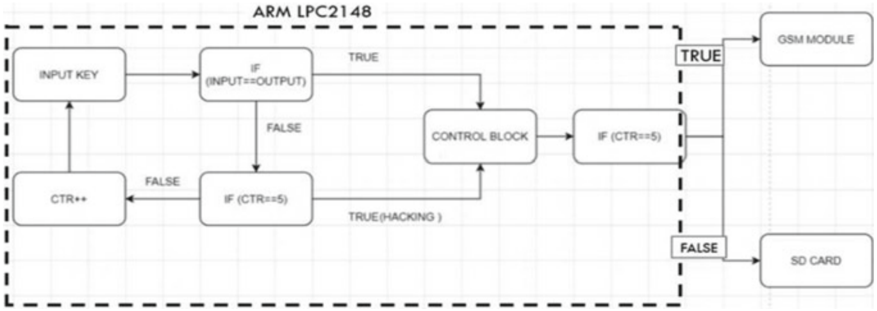


Fig. 1 Block diagram of the overall process

If the password is correct then TRUE will be executed and if the system gets hacked FALSE case will be executed.

4 System Design

The system needs a proper processor to execute all steps. For this purpose, an ARM7 processor was taken, and it has been checked and verified for its basic functionalities before proceeding further [11]. Before implementing the hardware part, we simulated each step to resolve any issues.

The first step in the designing is that the user should be able to enter the key or password once he switches on the system. For this purpose, we need to integrate the dot matrix keypad with the ARM processor. It is a 4 × 4 matrix keypad, and the button switches in it can be configured via programming. We need to write a proper code for the 4 × 4 matrix keypad to function properly. For this, we used the keil u4 vision software. After writing the code, we need to create the hex file so that we can burn it into the ARM processor. We need to connect the respective pins between keypad and ARM processor before uploading the code. We used PORT 0 in the ARM processor to configure the keypad. We used flash magic to burn the code from the PC to the ARM processor. We also used RS-232 cable to connect the PC and ARM processor.

The second step involves the integration of an LCD display to view outputs and to provide inputs for the system. An 16 × 2 LCD was used for this purpose, and respective connections between the ARM processor and the LCD was made. A proper code was also written using the keil software.

The third step is to do the integrating the GSM module [4, 5] with the ARM processor. Separate coding and drivers are needed for the GSM module since it also has a processor in it. After installing the drivers, we were able to sync both the GSM and ARM processor [12–15]. The connection between them is done with the help of male to male RS-232 cable. The sim card should also be inserted into the GSM module before uploading the code. Another important aspect is the antenna of the

GSM module though there are various options we thought it would be better if we go with the covered coil antenna and the height of the antenna was 8 cm. Any antenna can be used but we need to ensure the height of the antenna is approximately half the wave length so that it does not affect the transmission of waves.

In fourth step, we need to integrate an SD card with the ARM processor. Since the default ROM of the ARM is very less hardly, we can use it. Hence, an external data storage device is required so that all the data can be stored in it. The SD card needs to be programmed in such a way that when the system is being hacked only then the data should be erased.

Fifth step involves the simulation part. We verified and compared our simulation outcomes by running it multiple times before proceeding with the hardware connections and integration. At the last we need to combine all the program entities into a single hex file and burn it into the ARM processor so that all works simultaneously and in sync. We used flash magic to burn the code from the PC to the ARM processor. We also used RS-232 cable to connect the PC and ARM processor to upload the code (Figs. 2, 3, 4, 5, and 6).

Figure 7 shows the complete system with all its peripherals integrated with GSM module [10], LCD display, SD card, and 4 × 4 keypad. It also indicates the connections between the ports of the ARM processor and to other components which are connected via its peripherals and all the ports need to be configured separately using embedded C programming. The total cost involved in making this system is around Rs.10000, and this system was entirely built on our own. And the respective coding part for it was also done by us using keil u4 vision embedded C software [3]

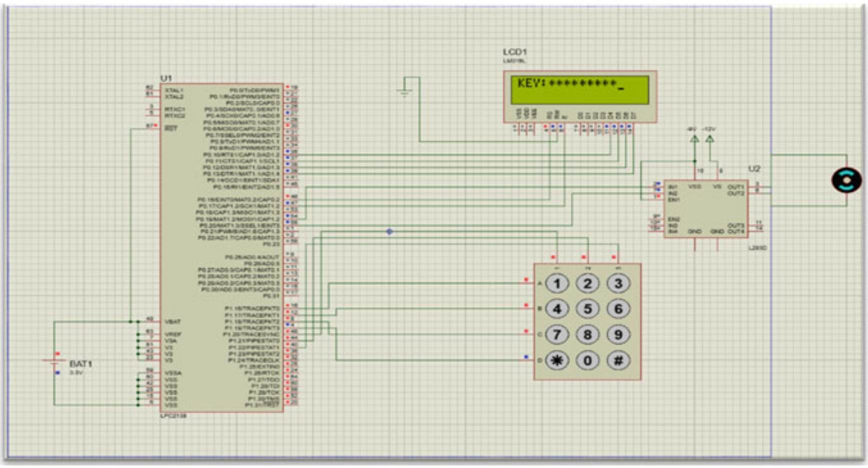


Fig. 2 Simulation: input key

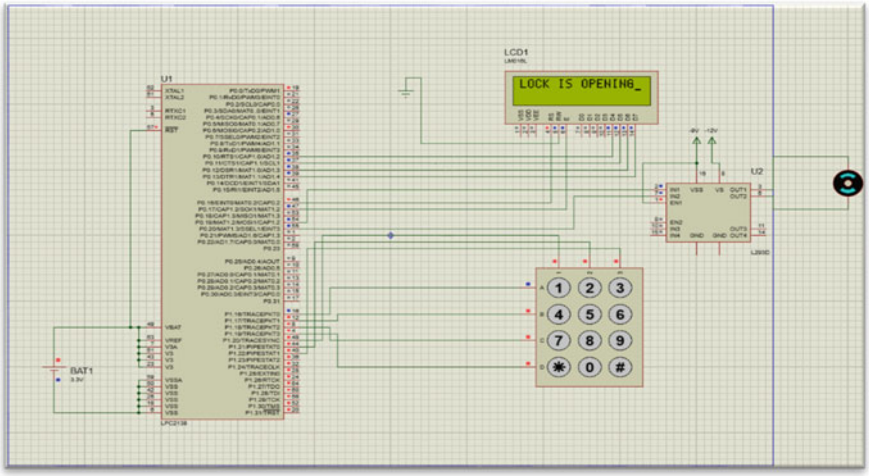


Fig. 3 Simulation: ıf correct key then user can acess the system

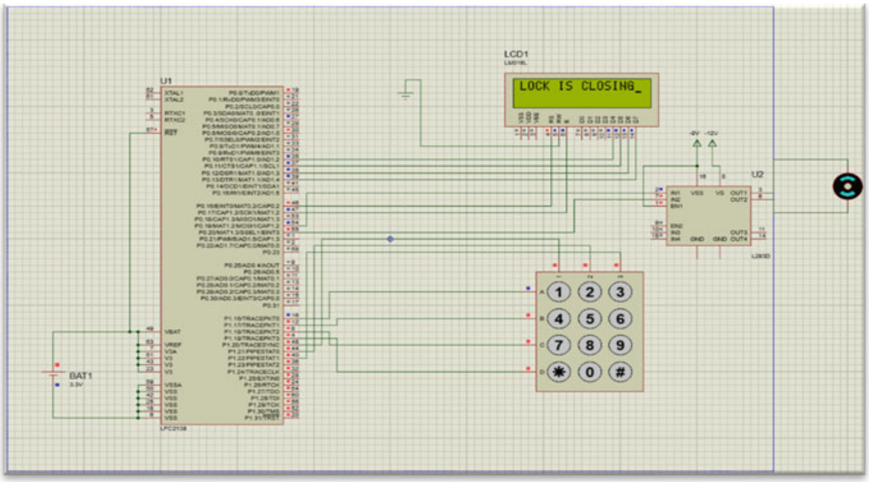


Fig. 4 Simulation: after user completes the tasks, system access will be closed

5 Results and Discussion

The model was successful when tested. It was able to perform the security algorithm perfectly and the data was protected against hacking.

The operations it performed are as follows:

1. When Hacked:

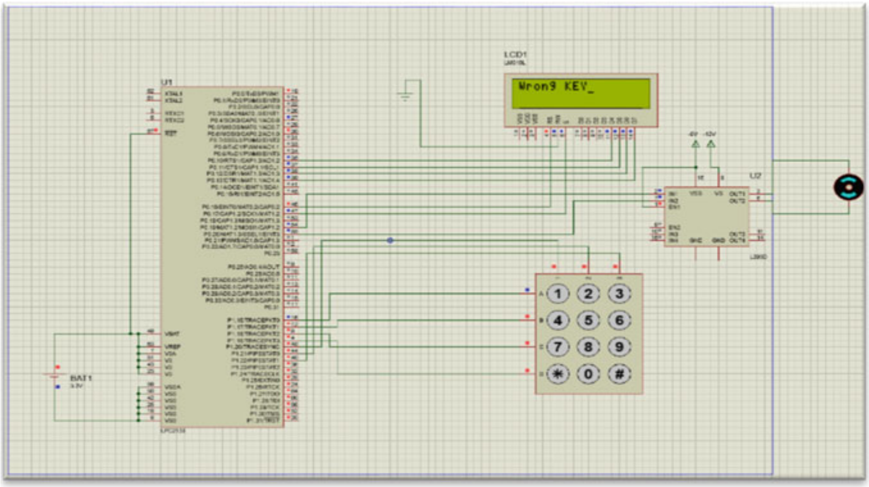


Fig. 5 Simulation: when wrong key is given

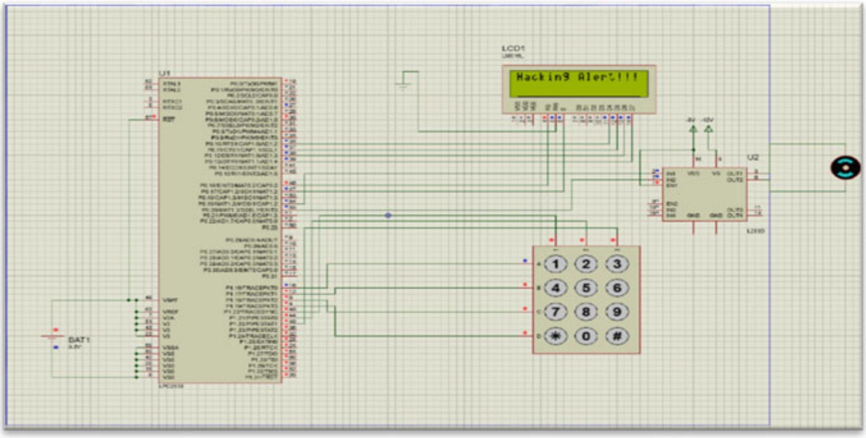


Fig. 6 Simulation: if hacked then hacking alert message

- When the hacker is entering the incorrect key fifth time the entire system and the processor will go in a locked state. And the "**Hacking Alert**" message will be displayed in the LCD display. And the data in the SD card will be erased or corrupted [11–13]. The user authentication or registration data will be inside the ARM inbuilt memory which is of 512 KB. When the system gets hacked the data which is present in the SD card will only be erased. Hence the user authentication credentials will not be affected if the system gets hacked.

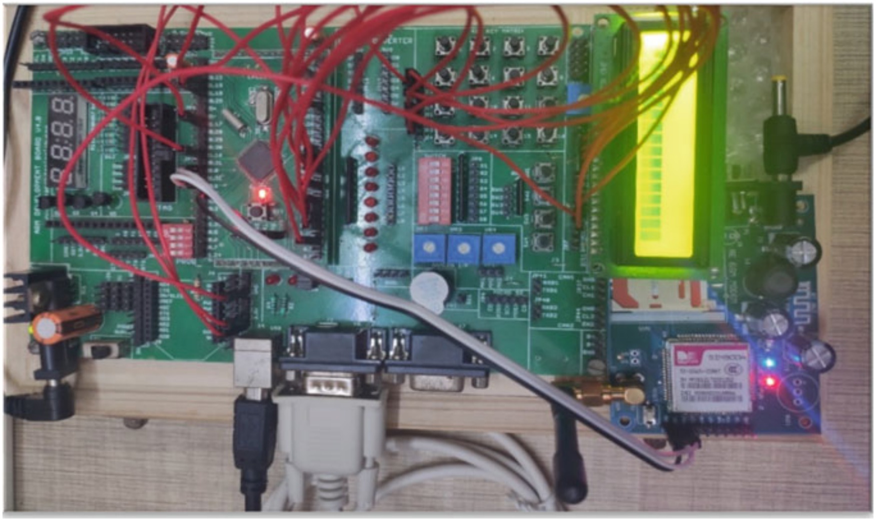


Fig. 7 ARM LPC2148 system

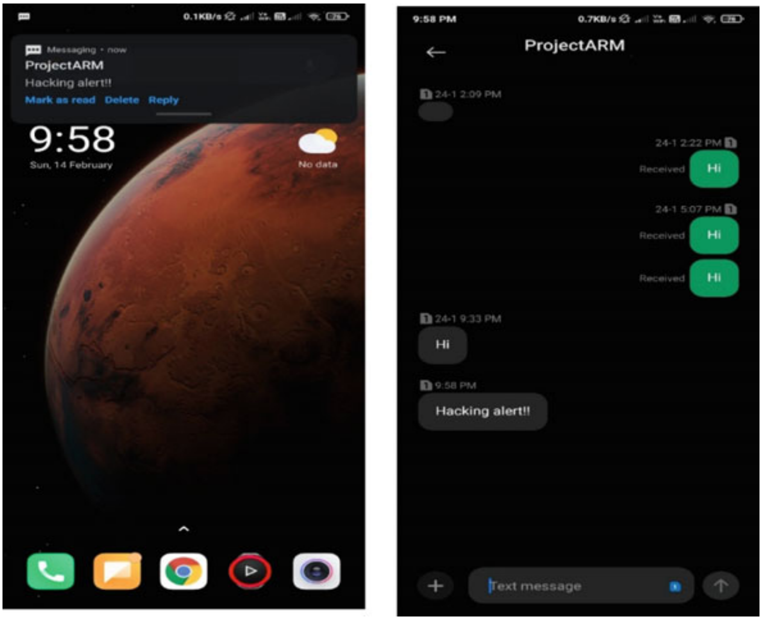
It will take 3 min for the system to restart once the processor went to the locked state, and as discussed in the above sections, the hacker will be put in a loop and the time complexity of the loop is 3^N. Here, N is the number of times the system got restarted. This locking state of the processor and the restart procedure will continue until the correct password is entered.

- Fig. 8: User will also get message alert to his phone when the system gets hacked. Here, we use a 4G sim card for our purpose and the message came to user phone immediately after the hacker entered the password incorrectly for 5 times. We have also included call alert feature in our design but one major drawback if we use call feature is the traffic. The traffic which is present while we make calls are significantly higher when compared with the messages traffic. Crosstalks and handoffs can also happen when we use calls. Because of this, the time delay may increase a lot which in turn increases the alert time to alert the user about the hacking issue in his system. Hence, we proceeded with the message alert notification method

2. Not Hacked

- If the user is accessing the system, then if he enters the correct key, he will be able to proceed to the further steps and he can access the functionality of the system which is being protected. In our case, it is the smart card by accessing it he can dial or message to anyone.

Another important aspect which we came up is with the length of the key which the user will be using to configure the system. In our project, there is no constraint on the length of the key, but one can think if the length of the key increases then the



Messaging • now
ProjectARM
Hacking alert!!
Mark as read Delete Reply

9:58
Sun, 14 February
No data

9:58 PM
ProjectARM

24-1 2:09 PM

24-1 2:22 PM
Received Hi

24-1 5:07 PM
Received Hi

Received Hi

24-1 9:33 PM
Hi

9:58 PM
Hacking alert!!

Text message

Fig. 8 Alert message on user phone

number of bits used in the key increases which will result in more combinations and it will be difficult to decode or hack the key. But what we found out is if we increase the key size the power consumed, the memory used also increases proportionally (Fig. 9).

Fig. 10: CPU operating voltage range of 3.0–3.6 V (3.3 V ± 10%) with 5 V tolerant I/O. The peak current is limited to 25 times the corresponding maximum current. The average power is 0.09 uW × 100

- If we increase the number of bits, the power consumed will also increase. Here, we kept the upper limit as 64-bit.

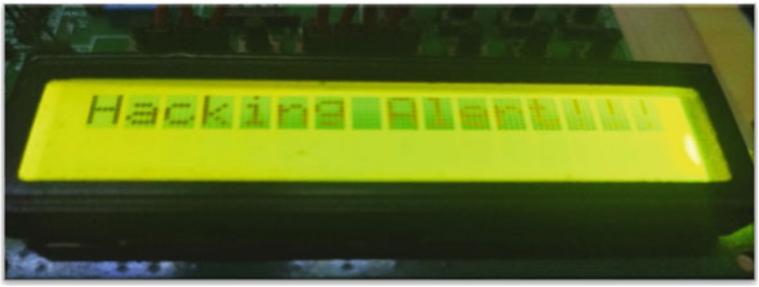


Fig. 9 Hacking alert is displayed on LCD

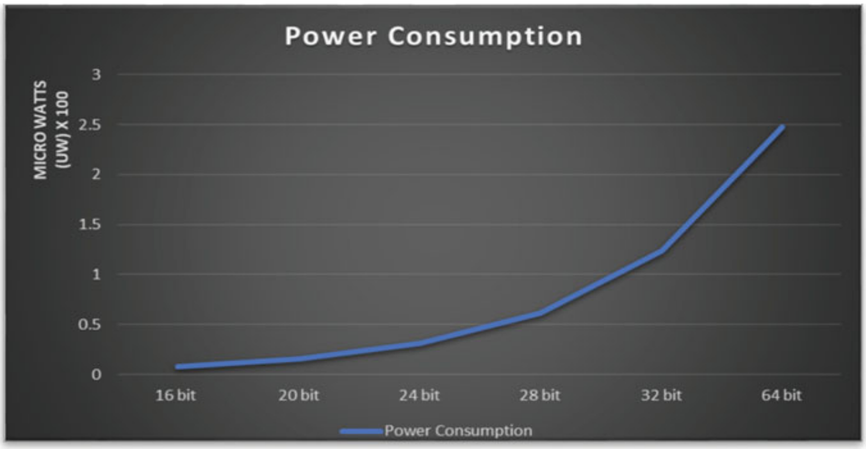


Fig. 10 Power consumption curve

$$\text{Power} = \text{Voltage} \times \text{Current} \tag{1}$$

- We calculated the power with the help of the data sheet for LPC2148. We did a total of 10 trials to get the approximate power consumed by the system.

The future scope of this system is if we can use a Raspberry Pi board and its components then many extra features can be added to the existing system. Furthermore, we can implement IoT in it to make it more connected with the user.

From Table 1 and Fig. 11, we can observe that there is 0 hamming distance for all the correct passwords and when the user enters a correct password the lock will open. If the user enters a wrong password the hamming distance will be calculated with respect to the correct password and it is displayed. And 10 will be displayed in

Table 1 Distance-based analysis

Input	O/P for correct key	O/P for wrong key	Hamming distance
0123456789ABCDEF	Pls enter valid key	Wrong password	38
ABCDEF9876543210	Pls enter valid key	Wrong password	36
54321ABCDEF07896	Correct password	-NIL-	0
3333	Correct password	-NIL-	0
4,894,651,314,865,222	Correct password	-NIL-	0
0000	Pls enter valid key	Wrong password	8
1,981,891,913,547,818	Pls enter valid key	Wrong password	27
54321ABCD0789678	Pls enter valid key	Wrong password	35
7,777,777	Correct password	-NIL-	0
8,486,465	Pls enter valid key	Wrong password	15

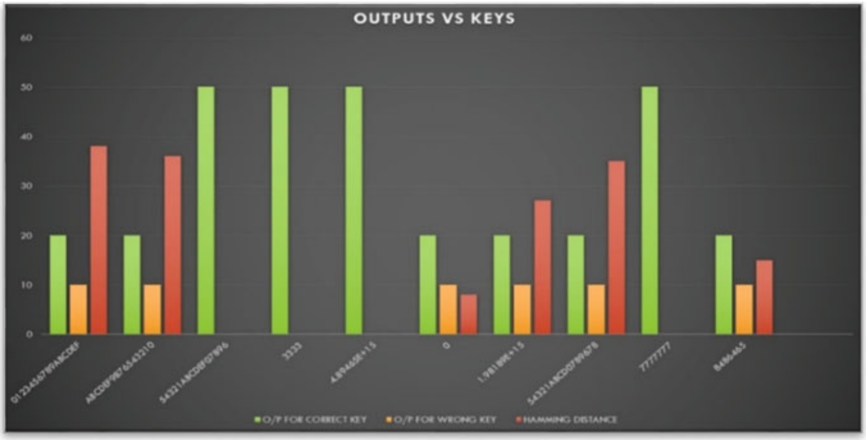


Fig. 11 Outputs versus keys bar graph direction for below graph: 10 ⇒ Wrong password, 20 ⇒ Pls enter a valid key, 50 ⇒ Correct password

the place of o/p for wrong key and 20 in the place of pls enter a valid key. If the user enters the correct key, then 50 is displayed with 0 hamming distance.

6 Conclusion

We succeeded in making the proposed system, and it was functional in all aspects. The experience we gained would be an inspiration for us to work more and build newer designs and algorithms on top of the existing system. With the help of our system, we will be able to protect any ARM-based devices in the real world. The major difficulty we faced while building the system is the programming port of the ARM LPC2148 since we were simultaneously integrating GSM module, SD card and we also need to connect to the computer to upload the code hence some clock synchronization was missing. Further, the programming part was very challenging and difficult to implement but we overcame these obstacles.

References

1. Yasin, M., Mazumdar, B., Rajendran, J., Sinanoglu, O.: Hardware Security and Trust: Logic Locking as a Design-for-Trust Solution: Design and Implementation (2019). https://doi.org/10.1007/978-3-319-93100-5_20
2. Xie, Y., Srivastava, A.: Anti-SAT: mitigating SAT attack on logic locking. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **38**(2), 199–207 (2019). https://doi.org/10.1109/TCAD.2018.2801220

3. Hu, J., Li, J., Li, G.: Automobile anti-theft system based on GSM and GPS module. In: 2012 Fifth International Conference on Intelligent Networks and Intelligent Systems, Tianjin, pp. 199–201 (2012). https://doi.org/10.1109/ICINIS.2012.86
4. Sun, H.-M.: An efficient remote use authentication scheme using smart cards. IEEE Trans. Consum. Electron. **46**(4), 958–961 (2000)
5. Abadi, M.B., Kaufman, C., Lampson, B.: Authentication and delegation with smart-cards. Technical Report 67, DEC Systems Research Center (1990)
6. Mohammed, L.A., Ramli, A.R., Prakash, V., Daud, M.B.: Smart Card Technology: Past, Present, and Future, vol. 12#1, pp 12–22 (2004)
7. Manoharan, S.: Embedded imaging system based behavior analysis of dairy cow. J. Electron. **2**(02), 148–154 (2020)
8. Karunakaran, P.: Deep learning approach to DGA classification for effective cyber security. J. Ubiquitous Comput. Commun. Technol. (UCCT) **2**(04), 203–213 (2020)
9. Huang, H., Xiao, S., Meng, X., Xiong, Y.: A remote home security system based on wireless sensor network and GSM technology. In: 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing, Wuhan, Hubei, pp. 535–538 (2010). https://doi.org/10.1109/NSWCTC.2010.132
10. LinkSprite Technologies, Inc., September 2008, User Manual: GSM/GPRS Module
11. Datasheet LPC2141/42/44/46/48, Philips semiconductor, October 2005, pp. 1–2
12. Manoj Reddy, D., Akshay, K.P., Giridhar, R., Karan, S.D., Mohankumar, N.: BHARKS: built-in hardware authentication using random key sequence. In: 2017 4th International Conference on Signal Processing Computing and Control (ISPCC), pp 200–204 (2017)
13. Mekaladevi, V., Kumar Reddy, D.R., Mohankumar, N.: Embedded device security and access control of quadcopter with a taser through encryption. In: 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 549–553 (2018). https://doi.org/10.1109/CESYS.2018.8723964
14. Saravanan, K., Mohankumar, N.: Design of logically obfuscated n-bit ALU for enhanced security. In: 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 301–305. https://doi.org/10.1109/ICECA.2019.8822129
15. Bhaskar, L., Natak, R.B., Ranjith, R.: Unit testing for USB module using google test framework. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1–3 (2020)

Implementation Mobile App for Foreign Language Acquisition Based on Structural Visual Method

Imad Tahini and Alex Dadykin

Abstract Among all modern communication devices, mobile phones are the most powerful means of communication even richer than e-mail or chat, as it can act as a teaching device despite its technical limitations. Therefore, we need to create a fast and efficient automated system that allow increased interest of students to learn foreign language, which provides the cognitive activity of students, stimulates and develops cognitive processes: thinking, perception, and memory. With such a system, the learner controls the learning process and progress in his own space based on his cognitive state, and learners can speak the target language without effort and psychological barriers. The purpose of the research is to create interactive speech trainer system based on a structural and visual approach and to ensure the formation of stable foreign language skills of trainees on the background of the active use of visual representation of language and interactive speech technology, and this system uses a technique for applying the visual approach and structural visual method in the educational environment by transforming grammatical information from verbal to graphic form and replacing complex text rules with appropriate visual structures in the form of pictures, schemes, and diagrams. We present our steps to implement our proposed architecture based on a visual model as a platform in mobile application with the establishment of content management system to provide the process of controlling the formation of speech skills and allowing the transition from foreign language learning to its improvement and acceleration. We also describe the ideas that will guide the design of this system.

Keywords Structural visual method · Learning management system · Learning content management system

1 Introduction

The modern world of rapidly developing and instantly gaining mass distribution of mobile network technologies is fundamentally changing the whole paradigm of

I. Tahini (✉) · A. Dadykin
Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_51

education. Today, the school, university, and the system of additional education cannot be limited to transferring to students a certain set of knowledge and skills that very quickly become obsolete and lose their practical value. To be successful, a modern person of any age needs to constantly develop in a professional and personal plan (which determines the essence of the concept of lifelong education) [1], using the most affordable and effective educational and information technologies.

As demonstrated by studying existing prototypes, at present there are no programs that allow you to prevent thinking in your own language and ensure that you quickly acquire direct thinking skills in another language. The use of structural visual method (SVM) in teaching helps to implement new generation of learning management system (LMS) to provide individualization and differentiation acquisition language which allows to understand the meaning of perceived or produced constructions of foreign language without reliance on the native language. Implementing this LMS for learning foreign languages with the help of mobile programs aroused great interest among students.

Thus, using SVM, as well as other visual tools developed within the framework of the visual approach, it is possible to significantly simplify the process of obtaining skills in using basic grammatical constructions, as well as to be able to measure and control the learning process at each particular point in the learning or retraining curve.

This visual approach is used in the development of learning management system (LMS) in mobile application, and it will help to provide both management of the process of obtaining skills and the possibility of gamification, socialization, and cooperation for translating the educational process into a modern intensive and effective format using the latest developments in the IT sphere and various areas of pedagogy and psychology. This LMS reduces psychological barriers and increases motivation for classes and independent training, as well as reduces the burden on teachers and allows you to automate routine processes.

This article is organized as follows: the next section discusses the related work, where the visual model and learning content management system (LCMS) were selected to be integrated into the proposed system in order to profile learners; the structure of the proposed adaptive system is discussed in Sect. 3; Sect. 4 describes the method of working the LCMS; Sect. 5 describes the method of working the mobile application; Sect. 6 describes the method of processing data from LCMS to mobile application; and the conclusion is discussed in the last section.

2 Related Work

A new generation of systems is to create language skills for mastering a foreign language for adults using the proposed SVM [2], which is based on scientific theories from different branches of knowledge to facilitate learning in a more efficient way and to provide personalized learning.

The research work [2] presents a visual approach to general scientific theories, and pedagogical principles aimed to improve the structure of activity, ensuring a uniform terminology in the development of complex scientific and educational projects. Brief examples clearly demonstrate features and benefits of the visual approach. This approach has been applied to a variety of scientific theories. The greatest practical importance is a set of models of the structure of English language and its implications for foreign language learning by adults.

The work [3] is aimed to create a formal model of a limited part of the language and the meanings described by it, in order to use it in a subsystem for measuring the rate of change of a skill and managing this process.

In the work [4], two new very promising methods were analyzed and possible ways of their further research, improvement, integration, and automation were identified and directed to the association of two these methods, the methods are the visual-auditory shadowing (VAS) method [5] and the structural visual method (SVM).

The work [6] has developed the generalized structure of a new generation learning management system with the display of the development of the main components of LMS that carried out in interrelated steps and show information—contact form and management of the educational process to set the concept of building the visual models as an interface using modern technology.

The work [7] has developed a prototype system and displayed the implementation of visual approach as application-based technologies into the educational environment; this prototype is the result of the detailed analysis and result models for visual learning, and the construction of this prototype is based on Web-based development programming with the implementation of the first step to build the content management system to store the data for lesson in database.

The learning algorithm for the principles of constructing English sentences for different times [2] as shown in Fig. 1, persons and for different types of sentences is as follows:

First, the logic of describing the various stages of the same action (event) is analyzed in the native language, and these meanings are linked to the color code and move the meanings into the intermediate sign system in the form of the visual objects.

Then the students get acquainted with the forms of the verb of the language being studied and associate visual coding elements and a color code with them.

At the next step, they learn to describe the situation in the target language, using not the terms, rules, or tables as the basis for their speech activity, but the SVM.

This solution allows you to transfer the planning of the utterance and control over its correctness from the speech fields of the left hemisphere of the brain into the visual fields of the right hemisphere, thus freeing up the speech zone for free speaking. Perception has its internal structure, according to the works of [8, 9], with memory and with the limbic system responsible for motivation and emotion. The perception cycle is completed by returning the signal to the primary projection zone, which creates the image of the perceived object.

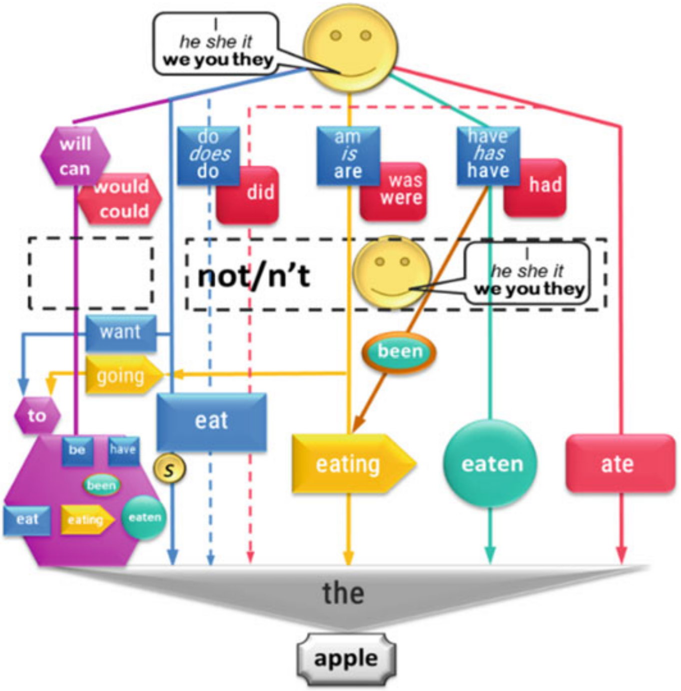


Fig. 1 Full model for all active levels

3 Proposed Approach

Several models have been worked on to implement this project [7], and among these models is the model for creating a mobile application. In this paper, we will show how to implement and build this project and the construction of the components will be discussed and explained.

Building an effective simulation model on the mobile phone required creativity skills in terms of technology and construction. Therefore, work was done on the technical side of the application in a technical manner and experience to achieve.

The working is done on developing this system, as shown in Fig. 2, which contains educational content, storage techniques, content management process and using mobile to display the lesson in different forms (lesson map, pictures, sound and voice recognition). Special language mobile applications can speed up and improve the process of learning English. They also help to develop stable language patterns, communication skills, and the rules of English grammar.

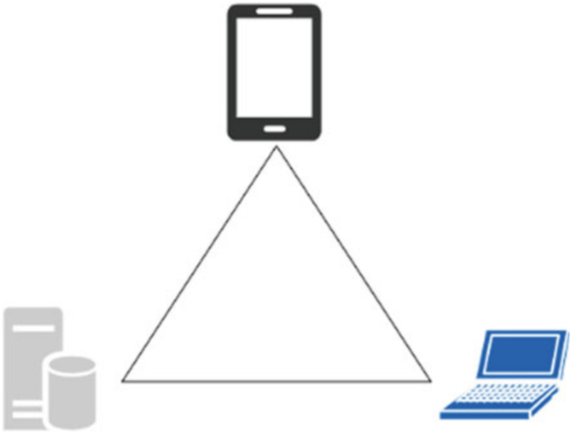


Fig. 2 Architecture of whole system

4 Learning Content Management System (LCMS)

Recently, a new class of systems has been developed [10] that implements the management of educational content (learning content management system, LCMS). LCMS is designed for managing data content on mobile application (adding, deleting, editing content), in our developments that provide functions for administering existing data resources (changing the structure of a Lesson, parameters), and can also be a tool for developing new lesson resources. With the help of this system, the owner can independently manage the content of his resource without resorting to the services of developers.

4.1 From Big Data to Simplified Parameters

The current programs for learning languages available in the market depend on traditional methods of teaching, to manage educational content on the current method of teaching the English language, and methods of storing it require a lot of time to coordinate and divide lessons and increase infinite sentences, and these lessons differ from one program to another so that you do not find limited content from the educational material. The current sentences of any many language require a lot of time to be memorized and displayed in the correct teaching method, and this requires a large space to be stored and displayed in the desired manner. The traditional designs of language programs increased information and efforts to produce a database containing the required sentences for each lesson. Currently, though our research and continuous work to change the method of teaching show a new style of education based on colors and linguistic maps, the teaching effort has been reduced from expanding unintended lessons into limited sentences and dividing them into a scientific way will accelerate education and transform the educational content that depends on fully stored sentences to a hashed word and saved in the database so that it will be compiled as sentences while the program is running.

Any change of any word in a sentence in the current educational programs that exist in market leads to the creation of a new sentence and an increase in the information in the database, wasting time by entering the sentences and corrected and revised for each lesson individually. So, our supposed design leads to the reduction of the method of memorizing the information in the database for a particular lesson as parameters to be formulated in front end use of system.

4.2 Discrete Parameters

Any process in reality is infinitely complex and has an infinite number of parameters and connections with other processes [3], and it is impossible to control a system with an infinite number of parameters.

In addition, the use of scientific terminology greatly complicates the process and reduces motivation. And the use of the verbal description of the language (in the form of rules) to control the planning, execution, and control of language activity prevents this activity, since it uses the same physiological mechanisms and mental processes. This has been repeatedly emphasized by leading psychologists and linguists [11–13].

We propose to change the direction to natural and fix not the formal side of the language, but the events, phenomena, and facts that it describes.

As we need to hold the grammatical and set the forms 8 (subjects) × 4 (grammatical change of model) × 7 (change of forms) = 224 variants of constructions for describing a single-type event.

To extend the many types of events to different verbs into number of N verbs, the grammatical set forms as N × 224 for describing a N-type events where the number of verbs will depend on describing the field of lesson.

So, we will describe the parameters of the grammatical sentences as discrete parameters words to be used to present the complete event.

The method of processing the lesson in the proposed approach depends on the withdrawal of words from the database to reconfigure the sentence during the usage of the lesson, this means we need only to store the words and types and link these words for a particular lesson and when requesting the presentation of the lesson by the user, the program uses a function to formulate the sentence as required programming for educational maps as seen in Fig. 3 and the mathematical formula will be discussed in Sect. 6.

4.3 Block Diagram

The working of LCMS will be explained by describing the block diagram and algorithmic description as seen in Fig. 4:

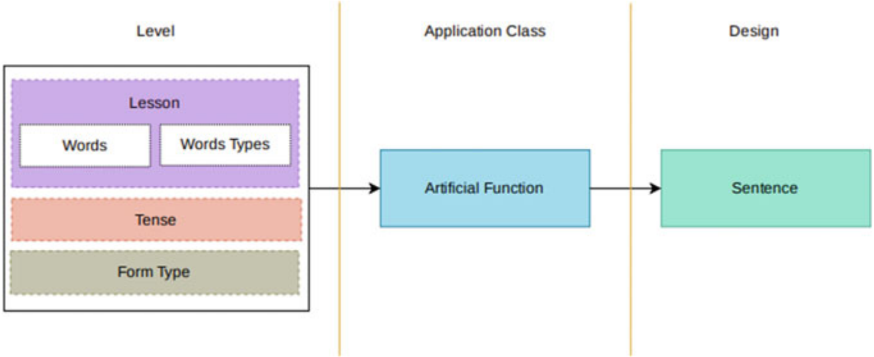


Fig. 3 Method to compose the sentence from parameters

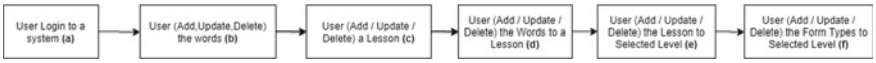

User Login to a system (a)
User (Add,Update,Delete) the words (b)
User (Add / Update / Delete) a Lesson (c)
User (Add / Update / Delete) the Words to a Lesson (d)
User (Add / Update / Delete) the Lesson to Selected Level (e)
User (Add / Update / Delete) the Form Types to Selected Level (f)

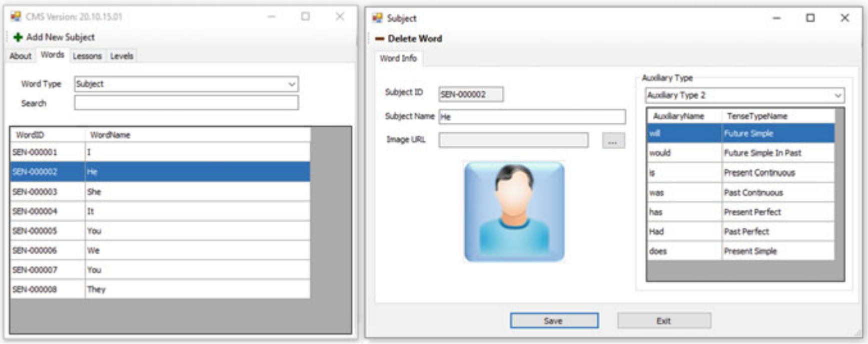
Fig. 4 Block diagram of LCMS

a. User login to a system: The system will begin from user login; the system checks if the login does not exist then the user creates one and then login;
b. User (add, update, delete) the words: User has authority to enter information about subject, verb, and objects and upload their represented images and user can modify and delete the information;
c. User (add, update, delete) a lesson: User has authority to enter information about lesson and upload the represented images and user can modify and delete the lesson;
d. User (add, update, delete) the words to a lesson: The lesson can be completed by adding the words to it, and user can add/delete these words to a lesson;
e. User (add, update, delete) the lesson to a selected level: The level is a set of lessons, user can add/delete these lessons to a level;
f. User (add, update, delete) the form types to a selected level: The role of the parameter of form type is to change the form of sentence (+, −, +?, −?, Do, What, Who), and this can be added to level by user to set the change of model of the map for target sentence.

4.4 Design Overall

The program is designed using windows base techniques as seen in Fig. 5, and the program is divided into screens to enter the information about levels and lessons to be displayed on the mobile application.

The entry of information was previously reviewed on block diagram, and the split screens will be displayed in the figure below.



Word Type: Subject

Search:

WordID	WordName
SEN-000001	I
SEN-000002	He
SEN-000003	She
SEN-000004	It
SEN-000005	You
SEN-000006	We
SEN-000007	You
SEN-000008	They

Subject ID: SEN-000002

Subject Name: He

Image URL:

Auxiliary Type: Auxiliary Type 2

AuxiliaryName	TenseTypeName
will	Future Simple
would	Future Simple In Past
is	Present Continuous
was	Past Continuous
has	Present Perfect
Had	Past Perfect
does	Present Simple

Fig. 5 Interfaces of LCMS

It should include the following key components:

(1) Repository of educational objects. The learning objects repository is the central database that stores and manages learning content for the ways for the efficient retrieval of information from multimedia data storage [14, 15].

(2) Display interface: To present educational objects in accordance with the training profile, for preliminary testing or in accordance for lesson contents, an interface for displaying materials is required. This component also provides the tools to control the information and various options for editing and creation from users. In addition, controls and design elements can be localized for the required SVM.

(3) Administration tools. This application is used to manage student accounts, launch courses from the catalog, track progress, report learning progress, and other simple administrative functions. This information can be passed to an LMS designed to provide more advanced administrative functionality.

5 Mobil Application

5.1 *Block Diagram*

The working of program will be explained by describing the block diagram as seen in Fig. 6 and algorithmic description:

a. User login to a system: First of all, the start use of the system will begin from user login, and the system checks if the login does not exist then the user creates one and then login.

b. System displays a set of levels: After that the system gets the levels from database and display them on screen.

c. User select a level: User can choose any level to go to new next step of training.

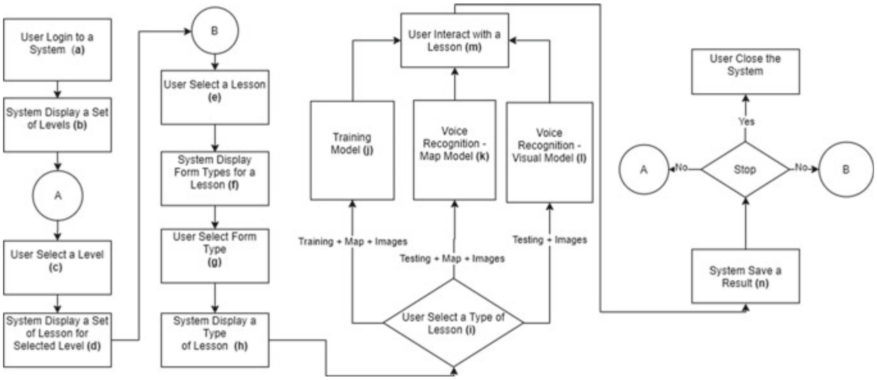

User Login to a System (a)
System Display a Set of Levels (b)
A
User Select a Level (c)
System Display a Set of Lesson for Selected Level (d)
B
User Select a Lesson (e)
System Display Form Types for a Lesson (f)
User Select Form Type (g)
System Display a Type of Lesson (h)
User Interact with a Lesson (m)
Training Model (j)
Voice Recognition - Map Model (k)
Voice Recognition - Visual Model (l)
Training + Map + Images
Testing + Map + Images
Testing + Images
User Select a Type of Lesson (i)
User Close the System
Yes
A
No
Stop
No
B
System Save a Result (n)

Fig. 6 Block diagram of working the system

d. System displays a set of lessons for selected Level: The system gets the lessons from database and displays them on screen and displays form types for lessons.

e. User select a lesson: User can scroll the screen and choose any lesson.

f. System display form types for a lesson: System gets set of form types (+, −, +?, −?, Do, who, what) for a lesson, each form type displays the form of construction for sentence.

g. User select form type: User can choose the form type from a set of options {+, −, +?, −?, Do, who, what).

h. System display a type of lesson: System sets a popup and displays a group of options: training model, voice recognition map model, voice recognition visual model.

i. User select a type of lesson: User chooses an option from: training model, voice recognition map model, voice recognition visual model, to enter the lesson content.

j. Training model: This model contains a map to construct a sentence and option to control this map to change the construction of a sentence, go to next sentence, return to previous sentence, change type of tense, change form type, text to speech, option to change (subject, verb, object).

k. Voice recognition—map model: This model contains a map that represents the construction of a sentence and voice recognition button to change speech to text and option to text to speech, and tools to control the change of form of a sentence.

l. Voice recognition—visual model: This model contains the images that represent the sentence and voice recognition button to change speech to text and option to text to speech.

m. User interact with a lesson: User uses the chosen model under the displayed controls.

n. System save a result: System saves a result in a database after interaction with lesson.

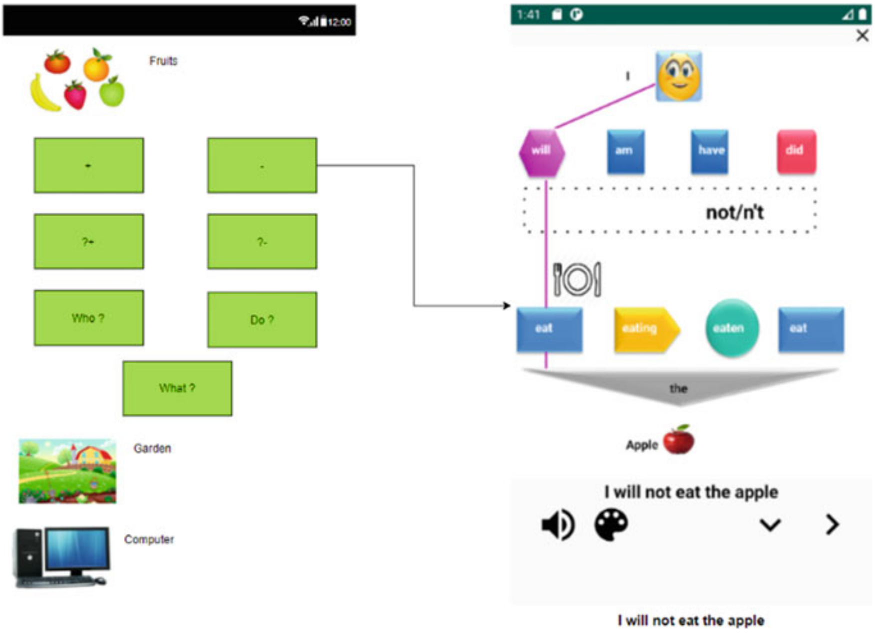
o. Action to end the program or continue: if the user wants to stop the system, then he can go to (p) else user can go to (A) or (B).

5.2 Design Overall

This program contains multiple levels and lessons, for each level has several lessons. Here, we will review how to use the program and divide the program into several interlinked screens so that it facilitates the user's access to the screen and the level to do a training for formation language skills.

Among the screens reviewed, there is a screen that has a group of form types which represent a set of patterns, and each pattern has a role to determine the method of build sentence that belongs to selected lesson, and when choosing one from the set of patterns, the system will switch to a specific screen and focus on maps that represent sentence and then the user has ability to easily understand how to build the sentence. The screens as seen in Fig. 7 are divided into several maps and each map will represent the sentence for particular tense as follows:

- From Type (+)—It will display the screen for constructed sentence as he will eat the apple.
- From Type (−)—It will display the screen for constructed sentence as he will not eat the apple.



I will not eat the apple

I will not eat the apple

Fig. 7 Interfaces the method of different forms for construction of sentences

- From Type (+?)—It will display the screen for constructed sentence as will he eat the apple? That has a postive answer.
- From Type (−?)—It will display the screen for constructed sentence as will he eat the apple? That has a negative answer.
- From Type (who)—It will display the screen for constructed sentence as who will eat the apple?.
- From Type (what)—It will display the screen for constructed sentence as what will he eat?.
- From Type (do)—It will display the screen for constructed sentence as what will he do?.

The screen in Fig. 8 is used to change the lexical words and the grammatical form of sentence. When entering to training screen, a number of options will appear that allow the user to build sentences by practicing in tools of building sentences and change the (subject, verb, object, tense), when pressing the subject and object the screen will be replaced by another new screen, and this screen has a set of images about (subjects and objects) that play a role to change the parameters of subject and object during the building of a new target sentence, same as when user do a touch on verb then system will get set of images about verbs to select the verb to build a new sentence, and any pressing on the button that represents the left arrow the new

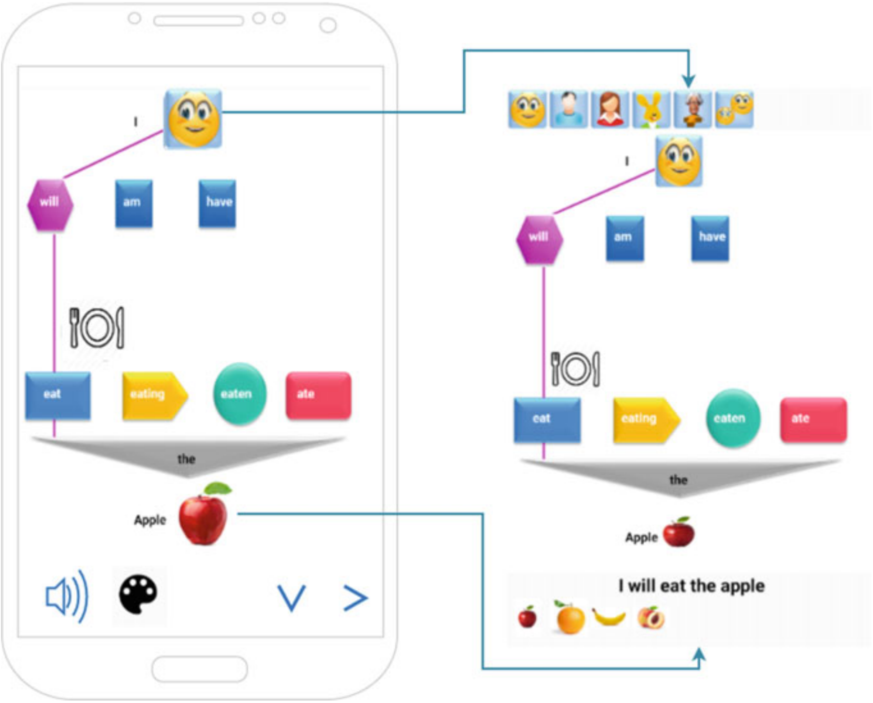


Fig. 8 Lesson interface and key controls for interaction with contents

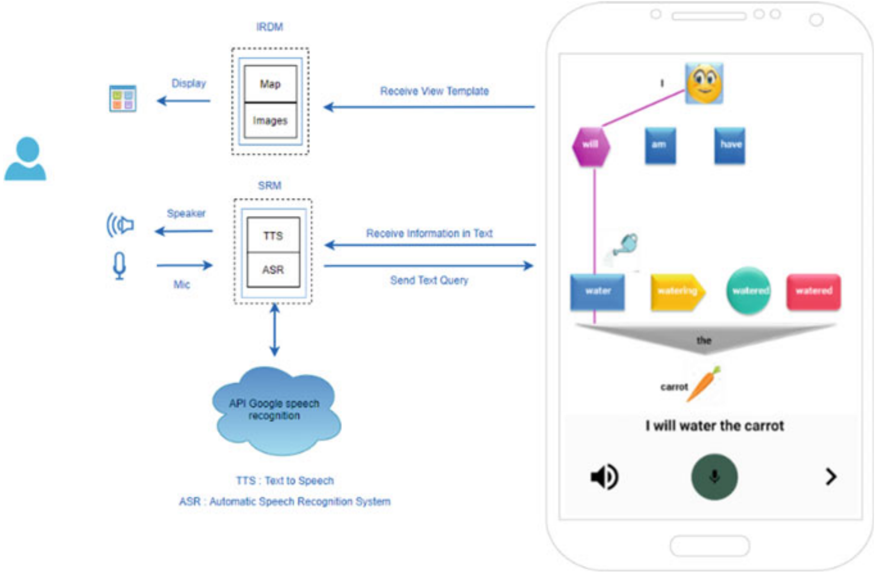
sentence will move to a next sentence with different tense and when pressing the button that represents the down arrow, the new sentence will move to a new sentence with different subject. When pressing a black button we get a pop-up of group tenses to change to a new sentence that has a different tense and image of voice to allow the system to speak the sentence.

5.3 The Method of Implementing the Platform with Voice and Speech Recognition System

To adapt the speech recognition system in Android system as seen in Fig. 9, we suggest a mechanism to facilitate interaction between the user and the platform through an interface that uses voice recognition.

Information Request and Display Module (IRDM): In this module, the images that represent the action of sentence are displayed as action image and map, and this module allows the user to control by the options of program through the mobile screen. Once information is processed, the requested are shown to the user as data lesson.

Speech Recognition Module (SRM): This module allows to the learner to interact in our platform by using voice. For instance, the system returns the information to the SRM in order to be converted to speech using text to speech (TTS). The user speaks for information in the microphone and such consult is converted to text in the



I will water the carrot

Fig. 9 Adapt the speech recognition and voice with Android system

automatic speech recognition system (ASR) [16]. In Android application, the Java functions check from this input data if it is correct to the displayed text at IRDM. We will focus specifically in API Google for (ASR) service because it is a cloud computing system and does not compromise the performance of the mobile.

6 Principles of Formulating Sentences for the Lesson

Artificial function in the programming of mobile application uses the mathematical formula to construct the sentence as follows:

The one dimensions of arrays for subject and object and multidimensional arrays for auxiliary and verb are the discrete values to formulate data lesson that belong to selected level (l) that contains tenses (t) and form types (f). Let $S[i]$, $i \in \{0, 1, \ldots, s\}$ is the array for subjects in lesson, $O[k]$, $k \in \{0, 1, \ldots, o\}$ is the array for objects in lesson, and let $A[s]$ [4] is the multidimensional arrays as matrix for auxiliary with row subjects and column length for tenses.

$$\begin{bmatrix} A_{01} & A_{02} & A_{03} & A_{04} \\ A_{11} & A_{12} & A_{13} & A_{14} \\ . & . & . & . \\ . & . & . & . \\ A_{s1} & A_{s2} & A_{s3} & A_{s4} \end{bmatrix} = (A_{s4})$$

And $V[b]$ [4] is the multidimensional arrays as matrix for verbs with row verbs and column length for tenses.

$$\begin{bmatrix} V_{01} & V_{02} & V_{03} & V_{04} \\ V_{11} & V_{12} & V_{13} & V_{14} \\ . & . & . & . \\ . & . & . & . \\ V_{b1} & V_{b2} & V_{b3} & V_{b4} \end{bmatrix} = (V_{b4})$$

Set the mathematical model in the form.

$$r(i, j, k, t, f) = \begin{cases} S_i + A_{it} + V_{jt} + L + O_k, \ f = 1 \\ S_i + N + A_{it} + V_{jt} + L + O_k, \ f = 2 \\ A_{it} + S_i + V_{jt} + L + O_k, \ f = 3, f = 4 \\ Q_{f-5} + A_{it} + V_{jt} + L + O_k, \ f = 5 \\ Q_{f-5} + A_{it} + S_i + V_{jt}, \ f = 6, f = 7 \end{cases}$$

$$w(i, j, k, t, f) = \begin{cases} Y + S_i + A_{it}, f = 3 \\ E + S_i + A_{it} + N, f = 4 \\ S_i + A_{it} + V_{jt} + L + O_k, f = 5 \\ L + O_k, f = 6, f = 7 \end{cases}$$

where $0 \leq i < s, 0 \leq j < b, 0 \leq k < o, 0 \leq t < 4$, $s =$ *length of array subjects,* $b =$ *length of row dimension of multi array verbs,* $o =$ *length of array objects, t the index of change for the grammatical change.*

The main task is to create the sentence from a set of variable parameters, and it is interesting the guidance of the coordinate parameters $r(i, j, k, t, f)$ to formulate a set of sentences or questions and $w(i, j, k, t, f)$ to formulate the answer which are referring to input variables conditions of form type that represented by (f) where f is the index of changes of array of different forms {+, −, +?, −?, Who, Do, what}, for example the index of '−' is $f = 2$, and the variable L belongs to set of values of articles, L ϵ {'a', 'an', 'the'}, and the array variable of interrogative $Q[f - 5]$ has set of array values ϵ {+?, −?, Who, Do, what} depends to condition of variation of index f.

Example

Set a lesson of two subjects = {'I', 'He'}, two verbs = {eat, Buy}, two objects = {'Apple', 'Banana'}, and this lesson belongs to level that has form types = {+, −, +?, Who},

So, the matrix values for auxiliary

$$]\begin{bmatrix} \overline{\text{will am have } //} \\ \text{will is has } // \end{bmatrix} = (A_{s4})$$

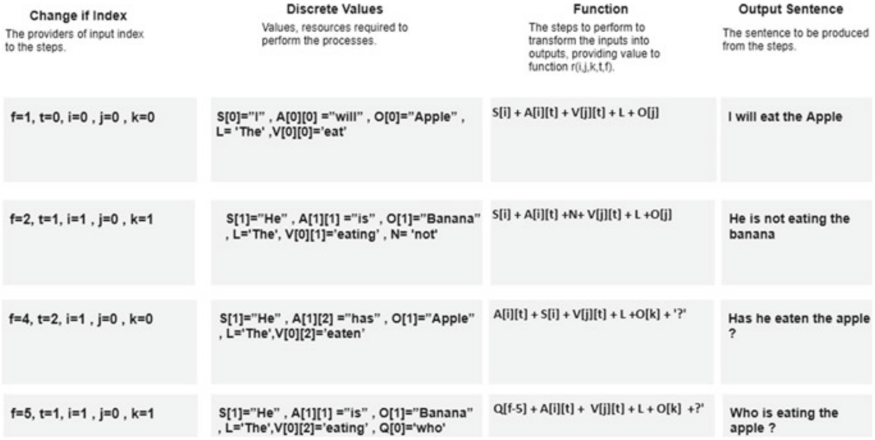
And matrix values for verbs

$$\begin{bmatrix} \text{eat} & \text{eating} & \text{eaten} & \text{ate} \\ \text{buy} & \text{buying} & \text{bought} & \text{bought} \end{bmatrix} = \left(V_{j4}\right)$$

Then, the results of formulated sentence due to the model function where the artificial function at mobile application API provides the input index are listed in Fig. 10.

7 Conclusion

In this paper, we presented an ongoing work to create the first type of method learning language toward use SVM to represent visual language which helps to produce the structure to compose the sentence and give each element from the sentence the visual design with color encode. We showed our steps to implement the proposed



Change if Index The providers of input index to the steps.	Discrete Values Values, resources required to perform the processes.	Function The steps to perform to transform the inputs into outputs, providing value to function r(i,j,k,t,f).	Output Sentence The sentence to be produced from the steps.
f=1, t=0, i=0 , j=0 , k=0	S[0]="I" , A[0][0] ="will" , O[0]="Apple" , L= 'The' ,V[0][0]='eat'	S[i] + A[i][t] + V[j][t] + L + O[j]	I will eat the Apple
f=2, t=1, i=1 , j=0 , k=1	S[1]="He" , A[1][1] ="is" , O[1]="Banana" , L='The', V[0][1]='eating' , N= 'not'	S[i] + A[i][t] +N+ V[j][t] + L +O[j]	He is not eating the banana
f=4, t=2, i=1 , j=0 , k=0	S[1]="He" , A[1][2] ="has" , O[1]="Apple" , L='The',V[0][2]='eaten'	A[i][t] + S[i] + V[j][t] + L +O[k] + '?'	Has he eaten the apple ?
f=5, t=1, i=1 , j=0 , k=1	S[1]="He" , A[1][1] ="is" , O[1]="Banana" , L='The',V[0][2]='eating' , Q[0]='who'	Q[f-5] + A[i][t] + V[j][t] + L + O[k] +?'	Who is eating the apple ?

Fig. 10 Applied values to formulate the sentences

architecture with describing the ideas about the design of this system as mobile platform. We also described the step of working system.

The novelties in this program are considering to display the method of creating LCMS to get simplified storage and accelerate the creation of data lessons that depends on parameters then using algorithm by mobile platform to compose the sentences of lessons.

The proposed architecture has the advantage of allowing the integration of new methods to improve system management with minimal effort. Pronunciation training using the technique helps not only to bring your pronunciation closer to the level of native speakers, but to perceive and understand the language by ear. Thanks to the methodology, almost all the disadvantages of traditional language teaching are eliminated—the time of language acquisition is reduced several times, while the results are significantly improved. This will help you can gain the ability to communicate freely with native speakers and eliminate at the same time for disfigured pronunciation and for a slow response.

Currently, the mobile app has already been created and successfully operating, and interactive simulators for primary and secondary levels of education have been developed. Encouraging results were obtained regarding the accuracy of speech recognition in English language increased to 95%.

Experimental testing of teaching materials, trial simulators and elements of the proposed approach on a limited group of students showed interesting results, close to those obtained in other developments of rapid learning methods according to [17]. There was a decrease in the time spent on learning to perform a specific action without errors by 3–30 times and an increase in the success of training from 10–25% to 80–95%.

The future incorporation is to work on updating and developing this program while conducting experiments to reach an integrated program for teaching all foreign languages.

References

1. Wilson, E.O.: On Human Nature. Harvard University Press (1978)
2. Dadykin, A.K., Dibrova, V.A., Tahini, I.H.: The visual approach in educational projects. Int. J. Soc. Sci. Hum. **7**(6), 373–377 (2017)
3. Tahini, I.H., Dadykin, A.K, Dibrova, V.A.:The model of change as the basis of the knowledge structure in the next generation E-LMS. In: 10th annual International Conference of Education. Research and Innovation, Seville, Spain, pp. 5022–5032 (2017). ISBN: 978-84-697-6957-7. ISSN: 2340-1095. https://doi.org/10.21125/10.21125/iceri.2017.1325
4. Tahini, I., Nakayama, T., Dibrova, V., Dadykin, A.: Cognitive psychology models and approaches to develop language skills. In: 5th International Conference on Education and Psychological Sciences (ICEPS 2018), Seoul, South Korea, January 27–29, 2018, International Journal of Information and Education Technology (IJIET). ISSN: 2010–3689. https://doi.org/10.18178/IJIET
5. Nakayama, T.: Efficacy of Visual-Auditory Shadowing Method in SLA Based on Language Processing Models in Cognitive Psychology. Kaitakusha, Tokyo (2017), 110p
6. Tahini, I.H., Dadykin, A.K.:A study of new techniques for learning management system to accelerate language acquisition using structural visual models. In: 2018 Sixth International Conference on Digital Information, Networking, and Wireless Communications. IEEE, Beirut, pp. 92–97 (2018). https://doi.org/10.1109/DINWC.2018.8357002
7. Tahini, I.H., Dadykin, A.K.: Proposed system of new generation LMS using visual models to accelerate language acquisition. Adv. Sci. Technol. Eng. Syst. J. **3**(5), 277–287 (2018)
8. Ivanitsky, A.M.: Brain science on the way to solving the problem of consciousness. Her. Russ. Acad. Sci. **80**(3), 229–236 (2010)
9. Edelman, G.M.: Wider Than the Sky: The Phenomenal Gift of Consciousness. Yale University Press (2004). ISBN 0-300-10229-1
10. Yakushin, A.: Analysis of Technologies and Management Systems for e-Learning. M. Dialectics, p. 78 (2008)
11. Zhinkin, N.: About code transitions in internal speech. Questions Linguist. **6**, 26–38 (1964)
12. Krashen, S.D.: Principles and Practice in Second Language Acquisition. University of Southern California, p. 202
13. Leontiev, N.: Activities and Consciousness. Personality, Moscow (1975)
14. Vijayakumar, T., Vinothkanna, R.: Retrieval of complex images using visual saliency guided cognitive classification. J. Innov. Image Process (JIIP) **2**(02), 102–109 (2020)
15. Sayantan, D., Ayan Banerjee, A.: Highly precise modified blue whale method framed by blending bat and local search algorithm for the optimality of image fusion algorithm. J. Soft Comput. Paradigm (JSCP) **2**(04), 195–208 (2020)
16. Anggraini, N., Kurniawan, A., Wardhani L., Hakiem, N.: Speech recognition application for the speech ımpaired using the android-based google cloud speech API. Telkomnika 16(6), 2733–2739 (2018). ISSN: 1693-6930
17. Galperin, P.Y.: Psychology of Thinking and Teaching About the Gradual Formation of Mental Actions. Research in the Thinking of Soviet Psychology, Moscow (1966)

RETRACTED CHAPTER: Assessing Deep Neural Network and Shallow for Network Intrusion Detection Systems in Cyber Security

Deena Babu Mandru, M. Aruna Safali, N. Raghavendra Sai, and G. Sai Chaitanya Kumar

Abstract Intrusion detection system [IDS] has become a central layer that unites everything inside the most recent ICT structure on account of the consideration for advanced prosperity inside the ordinary world. Motivations to recall the weakness to search out the sorts of assaults and grow the intricacy of bleeding edge computerized assaults; IDS requires the need to hitch deep neural networks (DNN). During this report, DNNs will not foresee assaults on the N-IDS. A DNN with a learning pace of 0.1 is applied and runs for the assortment of 1000 years, and subsequently, the informational index KDDCup-'99' was utilized for readiness and site meaning association. For assessment purpose, the arrangement is finished on the comparable dataset with another obsolete AI figuring and DNN of levels begin from 1 to 5. The outcomes were broke down, and it had been accepted that a DNN of three levels would be advised for execution.

Keywords Intrusion identification · Deep neural organization · Deep learning · Machine learning

1 Introduction

At present in the bleeding edge world, quick mechanical improvements have driven each relationship to accept information joining and advancement by correspondence

The original version of this chapter was retracted: The retraction note to this chapter is available at https://doi.org/10.1007/978-981-16-3728-5_67

D. B. Mandru
Department of CSE, Malla Reddy Engineering College, Hyderabad, India

M. Aruna Safali
Department of CSE, Dhanekula Institute of Engineering and Technology, Gangur, AP, India

N. Raghavendra Sai (✉)
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

G. Sai Chaitanya Kumar
Department of CSE, MIC College of Technology, Kanchikacherla, AP, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*, Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_52

(ICT). From this point, to forward build up an environment during which all action is coordinated through that structure by leaving the affiliation weak if the well-being of the ICT office is subverted. Subsequently, this necessitates the multi-layered acknowledgment and a protection convenience, which will influence genuinely new attacks on the picture, additionally as self-sufficiently adjust to new data.

There are few systems, which will not shield, and they are ICT structures from shortcomings, particularly the circumstance of idiosyncrasies and IDS. A frightful indication of conflicting and recognizable proof designs is that the intricacy identified with the connection of the definition rules. Every show under scrutiny should be exactly attempted, characterized, and executed. Another plot that is related to the situation of weirdness is that the destructive activity that is a piece of the customary task of utilization has not been visualized. Thus, the requirement for an IDS, which will be adjusted to new attacks, will frequently set up as a delivery by utilizing the flighty scattering informational indexes, which gets fundamental.

Disturbance detection systems (IDS) are a security-based online advancement field that was at first made to spot shaky areas and stresses against a host objective. The sole utilization of the intrusion detection is to recognize the perils. In this way, it is out of band at the lower part of the association, and it is not present inside the genuine consistent correspondence between the sender and beneficiary of the information. Taking everything into consideration, the course of action will regularly utilize TAP or SPAN ports for investigating the copy online traffic stream and plan to foresee the attack that upheld a formerly arranged gage by making the need for human mediation immaterial.

In the field of organization security, AI counts have had a key impact. Especially, because of the splendid presentation and cutoff of significant learning networks as of late in various subjects from a fair sort of fields as of late saw as insoluble, the relentless nature of their application for (AI) and in this way the difficulties without help have extended. Profound learning is basically a section of AI that duplicates components of human cerebrum also subsequently the name counterfeit neural association. The possibility of profound learning includes making bewildering and reformist pictures that incorporate making essential construction squares to handle issues of an important level. Recently, the use of profound learning systems is utilized for different organization security use cases.

As a result, it becomes clear that deep neural associations and IDS, once assembled, can function on a superhuman level. Also, because IDSs are out-of-band within the base, normal attacks like DoS, which are basically intended to stifle the association's data transfer ability to recognize input, cannot hamper their display, for what this level of security cannot adjust the force. Moving up the parts are coordinated as follows: In Sect. 2, it shows IDS-relevant action, several deep neural associations, and a couple of results of the KDDCup-'99' dataset that was performed. Section 3 discusses about the DNN and applications of the ReLU activation function from above. Section 4 analyzes the dataset used in this document, explains its shortcomings, and evaluates the final result, and finally, Sect. 5 concludes and establishes a plausible future workflow for this research paper.

2 Related Work

ID tests in network security have been utilized since the presentation of PC projects. The usage of LD methodologies and responses for all IDSs they understand has gotten customary recently, regardless, the arranging of closing information is confined and is generally used exclusively for posting journal entries. DARPA datasets [1] are maybe the biggest transparently available datasets. The Tepdump information offered by the 1998 DARPA ID appraisal association was tidied up and utilized for the 1999 KDDCup challenge at the fifth International Conference on Knowledge Discovery and Data Processing. The work comprised of planning affiliation records which are presently pre-handled in normal rush hour gridlock, or in one of the going with attack classes: 'DoS,' 'Analyzing,' 'R2L,' and 'U2R.' In the competition of KDDCup-'99,' data preprocessing was finished utilizing the MADAMID system [2]. Entries where the decisions tree varieties utilized showed irrelevant differences in execution incorporated the initial three focuses [3–5]. The initial 17 resistance enlistments were completely saved for great execution and are summed up [6]. The majority of the sent outcomes have been tried and prepared with just 10% of the introduce set seeing the thing decline in the KDDCup-'99' datasets [7–9]. Scarcely any experts have utilized uncommonly collected datasets from the ten planning set KDDCup-'99' [10–12].

There are a few interesting dispersions where results are found during a differentiation inclination because of the utilization of different test and arrangement informational indexes. During an article [13], determination trees and genetic figuring for the age of the modified guidelines were utilized for a clever structure to improve the limit of a current intrusion detection system. The planned utilization of neural associations in IDS has been suggested by [14–16] proposed a utilization of discontinuous neural associations (RNN) and [17] considered the introduction of neural association structures for the limitation of quantifiable anomalies in datasets from four unique circumstances.

Albeit the KDDCup-'99' datasets present various issues [18] contends that so far they are a vigorous diary section check dataset that can be openly gotten to have faith in different limitation strategies for interruptions. ML-based techniques might be the aftereffect of their capacity to assault the perplexing and different dangers that persistently advance to accomplish a recognizable bogus positive rate at sensible computational cost. In the beginning phases, the PNrule strategy acquired from the P rules and N rules was utilized to independently decide the presence and non-presence of the class. This has a benefit because of the advancement of the notoriety rate in different sorts of rounds with the exception of the U2R positioning.

One positive manner of thinking for this article is to manhandle the chance of an approaching computerized hostility inconsistency that is not dubious to natural eyes however which is regularly sifted through by adding a counterfeit layer of information to the association. Along these lines, by setting up the neural association with current computerized attack data, a way can be found to successfully envision an approaching assault and caution the edge or start a pre-customized response that can keep the attack from proceeding with further. In this way, a considerable number of dollars, post-quake payoffs, and over the top data centers are often evaded simply by adding another layer to the security framework. The benchmarking dataset used to set up the affiliations is old, and for a predominant and consistent power of the check, resulting information should be used for retraining before being conveyed to the field. The responsibility of this record is to present the core of fake neural relationship in the field of online security which is advancing rapidly.

3 Background

The DNNs are ANNs with a structure of multi-layered between the data and yield levels. They can show complex nonlinear associations and can make computational models in which the thing is conveyed similar to layered design of locals. Under, we cover commonly direct DNNs and ReLU applications, and how it is efficient over other incitation features.

3.1 DNN

Standard AI tallies are immediate and significant neural affiliations put away in a reformist making of multifaceted plan and reflection. Each layer applies a non-direct change to your information and makes an undeniable model dependent on what it gets it. Essentially, the information layer is taken from the data layer and passed to the previously covered up layer. These puzzling layers are numerically reliant on our wellsprings of data. One of the issues with making neural affiliations is picking the quantity of covered up layers and the quantity of neurons for each layer. Every neuron has an underlying position that is utilized to normalize the neuron's exhibition. The 'critical' in important learning indicates the presence of more than one degree of secret. The exhibition level reestablishes the presentation information. Until the exhibition has been appeared to an admirable degree of accuracy, the hundreds of years go on.

3.2 Use of Straight Ground Units (ReLU)

ReLu has been shown to be more capable and can accelerate the entire planning measure taking everything into account. Neural associations typically use sigmoid activation work or tanh order limits (excessive deviation). Be that as it may, these limitations will ultimately make the problem go away. The evaporation slope occurs when the lower layers in the DNN have points close to zero in light of the fact that the units of the upper layers are almost flooded in the working asymptotes tanh. This offers an option in place of sigmoid nonlinearity that pays attention to the issues mentioned so far.

4 Methodology

This research work considers Keras as a compartment in Tensor-Flow for the construction of items. To fundamentally change information arranging ability into important learning models, a GPU-empowered tensor stream was utilized on a special Nvidia-GK110BGL-Tesla-k40.

4.1 Representation of Educational Records

DARPA's 1998 Identity Assessment Program was administered and organized by MIT's Lincoln Laboratories. The standard point of convergence for this is disassembling and conducting DI research. A standardized dataset was designed and disseminated, which included various types of interference that reproduced a military environment. The 1999 KDD Intrusion Detection Contest dataset was a particularly refined variation on this.

4.2 Insufficiencies of the KDDCup-'99' Informational Index

ReLu winds up being more productive and has an arranged relationship An and the fundamental weaknesses of the made dataset, like KDDCup-'98' and KDDCup-'99,' were talked about by. This fundamental conviction was that they could not keep up their dataset. In spite of every one of these reactions, a few scientists have utilized the KDDCup-'99' informational index as a proficient informational collection to inspect IDS gages after some time rather than the dataset creation reactions, uncovered the point—Spot assessment of substance, seen non-cognizance, and old pieces imitated in the manmade network traffic information.

It was examined why AI classifiers have a limited ability to recognize attacks that have a place with the R2L, U2R content dispositions in the KDDCup-'99' datasets. They assumed that it was beyond the domain of the creative mind to expect a good identification rate using conventional AA calculations. Also, it was found that a high area rate can be refined when in doubt by transmitting an enhanced and extended dataset by merging the train and test sets. In any case, a critical approach has not been verified.

The DARPA/KDDCup-'88' could not assess the traditional IDS, which required a great deal of examination. To annihilate this, it follows that tcpdump utilized the Snort recognizable proof system in DARPA/KDDCup-'98.' The picture was fizzling bringing about inept precision and denying bogus positive rates. He neglected to perceive two and the arrangement of the investigation, in any case, the blended execution was better than the distinguishing proof of R2L and U2R. In spite of the fierce examination, the KDDCup-‘99’ set up the excess parts in maybe the most dependable and generally utilized benchmarking informational indexes, unreservedly open to distinguished people with the IDS appraisal and other related exercises. With an end goal to direct the center issues with the KDDCup-‘99’ set, he proposed a refined type of knowledge gathering called NSL-KDD. Dispose records of over-relationship between the train data and the test data. Aside from that, invalid records were likewise separated from the test data. These estimates hold the classifier back from being uneven toward the most normal records. Even after refinement, this did not deal with the issues uncovered by and another dataset called UNSW-NB15 was proposed,

4.3 DARPA/KDDCup-‘99’ Informational Index

The DARPA ID Assessment Group, amassed information dependent on the IDS network by recreating an Air Force base LAN from in excess of 1000 UNIX hubs and for 9 successive weeks, many clients at any one time at Lincoln Labs which was then isolated into 7 and 14 days of preparing and testing separately to remove crude TCP dump information. The MIT laboratory, with broad monetary help from DARPA and AFRL, utilized Windows and UNIX hubs for virtually all approaching interruptions from an estranged LAN not at all like other working framework hubs. For the reasons for the dataset, 7 unique situations and 32 distinct assaults were reenacted, for an aggregate of 300 assaults.

Since the time of distribution of the KDD-‘99’ dataset, these are the most utilized information to assess different IDSs. This dataset is gathered for almost 4,900,000 individual associations, including a component check of 41. The mimicked assaults were for the most part delegated follows:

- Denial of Service (DoS) Attack: Intrusion in which an individual means to make a host difficult to reach for its genuine reason by incidentally or sometimes intruding on administrations by flooding the objective machine with gigantic measures of applications and accordingly over-burdening the host.
- User Root Attack (U2R): A class of move regularly utilized by the creator who starts by attempting to access a client’s prior access and misusing openings to acquire root control.
- Local Remote Attack (R2L): Intrusion in which the aggressor can send information bundles to the objective yet does not have a client account on that machine, attempts to misuse a weakness to acquire neighborhood access by taking on the appearance of a current client from the objective machine.

- Probing-Attack: The sort where the creator attempts to assemble data about PCs on the organization and a definitive objective for doing so is to overcome the firewall and gain root access.
  The KDDCup-‘99’ set is characterized into the accompanying three gatherings: Basic qualities: the credits got from a TCP/IP association come from this gathering. The vast majority of these qualities bring about a certain postponement in recognition. Traffic qualities: determined attributes with respect to a period window is ordered in this gathering. This can be partitioned into two gatherings:
- Characteristics of ‘a similar host’: This class incorporates associations that have a last host indistinguishable from the association being referred to during the persistent 2 s and are utilized to figure the conduct measurements of the convention, and so on.
- ‘Same Service’ highlight: Only associations that have administrations indistinguishable from the current association over the most recent two seconds are remembered for this classification.
- Content Characteristics: Typically, test assaults and DoS assaults have probably some kind of regular consecutive interruption designs, dissimilar to R2L and U2R assaults. This is on the grounds that they include various associations with a solitary arrangement of hosts in a brief timeframe, while the other 2 interruptions are incorporated into information segment bundles where generally just a single association is included. For the location of these sorts of assaults, we need some interesting attributes through which we can search for unpredictable conduct. They are called content capacities.

4.4 Recognizable Proof of Organization Boundaries

Hyper-tuning of boundaries to decide the ideal arrangement of boundaries to accomplish the ideal outcome is itself a different field with a lot of space for future exploration. In this article, the learning stays consistent at 0.01 while different boundaries are streamlined. They include neurons in a layer which was tried different things with by transforming it in the reach from 2 to 1024. From there on, the check was additionally expanded to 1280 however created no obvious expansion in precision. Thus, the neuron tally was changed in accordance with 1024.

4.5 ID of Organization Structures

Ordinarily, expanding the quantity of layers creates preferable outcomes over expanding the quantity of neurons in a layer. Hence, the accompanying organization geographies were utilized to break down and close the ideal organization structure for our info information.

Table 1 Information of structured network

Layer (type)	Output shape	Param
Dense-1	(NIL, 1024)	43,008
Dropout-1	(NIL, 1024)	0
Dense-2	(NIL, 768)	787,200
Dropout-2	(NIL, 768)	0
Dense-3	(NIL, 512)	393,728
Dropout-3	(NIL, 512)	0
Dense-4	(NIL, 256)	131,328
Dropout-4	(NIL, 256)	0
Dense-5	(NIL, 128)	32,896
Dropout-5	(NIL, 128)	0
Dense-6	(NIL, 1)	129
Activation-1	(NIL, 1)	0

- DNN with 1,2,3,4,5 layers.

For all the above network geographies, 100 pages were run, and the outcomes were noticed. At last, the DNN three layer showed the best exhibition over all others. To widen the quest for the best outcomes, all normal exemplary AI calculations were utilized and the outcomes contrasted with DNN layer 3, which actually beats every exemplary calculation. Point-by-point factual outcomes for various organization structures are appeared in Table 1.

4.6 *Proposed Engineering*

An outline of the proposed DNN engineering for all utilization cases is appeared in Fig. 1. This incorporates a five secret layer tally and a yield layer. The info layer comprises of 41 neurons. Neurons in the information layer for the secret layer and those covered up in the yield layer are completely associated. The back-propagation component is utilized to prepare DNN organizations. The proposed network is made out of completely associated layers, polarization layers and prohibition layers to make the organization more vigorous.

Secret Layers and Inputs: This layer comprises of 41 neurons. They are then embedded into the secret layers. Secret layers use ReLU as a non-direct trigger capacity. At that point, the loads are added to take care of them to the following secret level. The quantity of neurons in each secret layer is continually being diminished from essential to yield to make yields more precise while lessening computational expenses.

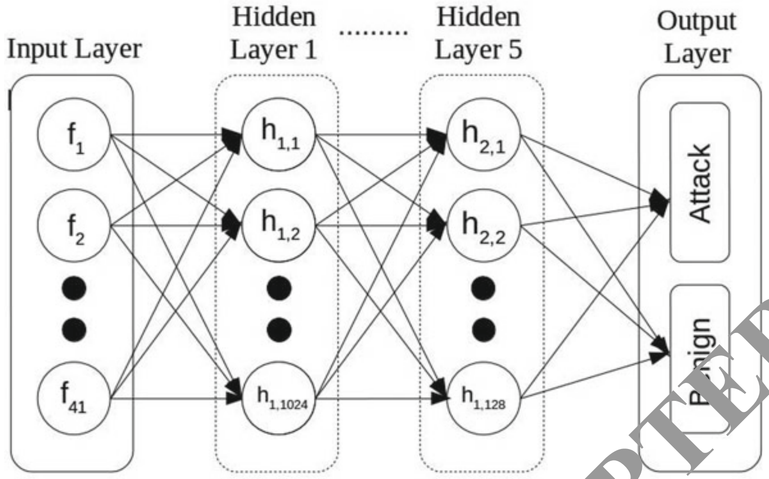


Fig. 1 Proposed architecture

Regularization: The entire cycle is effective and saves time, Abandonment (0.01). The capacity of deserting is to haphazardly disengage neurons, making the model more vigorous and in this way keeping it from being larger than average in the preparation set.

Yield layer and characterization: The yield layer comprises of just two attack and benign neurons. Since the 1024 neurons in the foremost layer need to turn out to be just 2 neurons, a sigmoid actuation work is utilized. Because of the idea of the sigmoid capacity, it just returns two yields, so it favors the parallel characterization given in this article.

5 Results

From the motivations behind this report, the KDDCup-'99' informational index has been consolidated into the exemplary AA and covered up multilayer DNN calculations.

Toward the finish of preparing, the models were looked at for f1-score, exactness, review, and accuracy with the test informational index. Their scores were looked at in detail in Table 2. The three-level DNN has overcome any remaining exemplary AI calculations. This is because of the capacity of DNNs to separate information and usefulness with more noteworthy reflection, and the nonlinearity of the organizations adds a benefit for different calculations.

Table 2 Results

Algorithm	Accuracy	Precision	Recall	f1-score
DNN-1	0.928	0.997	0.914	0.953
DNN-2	0.928	0.997	0.913	0.953
DNN-3	0.929	0.996	0.914	0.953
DNN-4	0.929	0.998	0.912	0.955
DNN-5	0.926	0.997	0.910	0.952
Ada boost	0.924	0.994	0.910	0.950
Decision tree	0.927	0.998	0.911	0.952
K-nearest neighbor	0.928	0.997	0.912	0.953
Linear regression	0.847	0.988	0.820	0.896
Navie Bayes	0.928	0.987	0.922	0.954
Random forest	0.926	0.998	0.911	0.952
SVM*-Linear	0.810	0.993	0.771	0.867
SVM*-rbf	0.810	0.991	0.772	0.867

6 Conclusion

This archive has exhaustively summed up the handiness of DNNs in IDS. For reference purposes, other traditional ML calculations were checked and contrasted the DNN results. The openly accessible informational collection KDDCup-'99' was basically utilized as a benchmarking apparatus for the investigation through which the predominance of DNN over the other looked at calculations was unmistakably recorded. For additional refinement of the calculation, this work considers the DNNs with various secret layer checks, and it was presumed that a DNN with three layers demonstrated effective and precise all through

As neurons are prepared on a past benchmarking dataset, as discussed about a few times in this paper, this addresses a hindrance to this approach. Luckily, it tends to be overwhelmed by utilizing another dataset with the embodiments of the most recent assault methodologies before really conveying this layer of AI in existing organization frameworks to guarantee the spryness of the calculations' certifiable capacities.

From the observational after-effects of this paper, it can be assumed that profound learning strategies are a promising bearing toward online protection assignments; however, notwithstanding the exhibition on the fake dataset is extraordinary, the use of the equivalent continuously network traffic contains more data. Mind boggling and ongoing sorts of assaults are required. Besides, concentrates on the adaptability of these DNNs in conflicting settings are required. The ascent of huge varieties of profound learning calculations requires an overall assessment of these calculations concerning their adequacy against IDS. This will be one of the bearings in which IDS examination can travel, and along these lines, it will stay as a forthcoming task of things.

References

1. Lippmann, R., Haines, J., Fried, D., Korba, J., Das, K.: The 1999 DARPA off-line intrusion detection evaluation. Comput. Netw. **34**(4), 579–595 (2000). https://doi.org/10.1016/S1389-1286(00)00139-0
2. Lee, W., Stolfo, S.: A framework for constructing features and models for intrusion detection systems. ACM Trans. Inf. Syst. Secur. **3**(4), 227261 (2000). https://doi.org/10.1145/382912.382914
3. Pfahringer, B.: Winning the KDD99 classification cup: Bagged boosting. SIGKDD Explor. Newsl. **1**, 6566 (2000). https://doi.org/10.1145/846183.846200
4. Vladimir, M., Alexei, V., Ivan, S.: The MP13 approach to the KDD'99 classifier learning contest. SIGKDD Explor. Newsl. **1**, 76–77 (2000). https://doi.org/10.1145/846183.846202
5. Agarwal, R., Joshi, M.: PNrule: A new framework for learning classier models in data mining. Tech. Rep. 00-015. Department of Computer Science, University of Minnesota (2000)
6. Elkan, C.: Results of the KDD'99 classifier learning. SIGKDD Explor. Newsl. **1**, 63–64 (2000). https://doi.org/10.1145/846183.846199
7. Sung, S., Mukkamala, A.H.: Identifying important features for intrusion detection using support vector machines and neural networks. In: Proceedings of the Symposium on Applications and the Internet (SAINT), pp. 209216. IEEE Computer Society (2003). https://doi.org/10.1109/saint.2003.1183050
8. Kayacik, H., Zincir-Heywood, A., Heywood, M.: Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets. In: Proceedings of the Third Annual Conference on Privacy, Security and Trust (PST) (2005)
9. Krishna Katakam, M.S., Devineni, K., Kanagala, P., Raghavendra Sai, N.: Analysis of artificial neural networks based intrusion detection system. Int. J. Adv. Sci. Technol. **29**(5s), 928–935 (2020). Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/7832
10. Raghavendra, S.N., Jogendra, K.M., Smitha, C.Ch.: A secured and effective load monitoring and scheduling migration VM in cloud computing. In: IOP Conference Series: Materials Science and Engineering ISSN-1757-899X, Vol. 981 (2020)
11. Chebrolu, S., Abraham, A., Thomas, J.: Feature deduction and ensemble design of intrusion detection systems. Comput. Secur. **24**(4), 295307 (2005). https://doi.org/10.1016/j.Cose.2004.09.008
12. Raghavendra Sai, N., Satya Rajesh, K.: A novel based approach for Liaison analysis in data summarization and deep web interface data extraction. Int. J. Control Theor. Appl. (IJCTA) 9(4) (2016). ISSN: 0974-5572
13. Smys, S., Abul, B., Haoxiang, W.: Hybrid intrusion detection system for internet of things (IoT). J. ISMAC **2**(04), 190–199 (2020)
14. Karunakaran, H.: Deep learning approach to DGA classification for effective cyber security. J. Ubiquit. Comput. Commun. Technol. (UCCT) **2**(04), 203–213 (2020)
15. Cannady, J.: Artificial neural networks for misuse detection. In: Proceedings of the 1998 National Information Systems Security Conference (NISSC), pp. 443456. Citeseer (1998)
16. Debar, H., Dorizzi, B.: An application of a recurrent network to an intrusion detection system. In: International Joint Conference on Neural Networks. IJCNN, vol. 2, pp. 478 483 (1992). https://doi.org/10.1109/ijcnn.1992.226942
17. Raghavendra Sai, N., Jogendra Kumar, M., Hussain Basha, P., Sai Chaitanya Kumar, G.: Effective intrusion detection system by using LOS classifier. Int. J. Innov. Technol. Explor. Eng. (IJITEE) 9(2) (2019). ISSN: 2278-3075
18. Raghavendra Sai, N., Satya Rajesh, K.: An efficient los scheme for network data analysis. J. Adv. Res. Dyn. Control Syst. (JARDCS). **10**(9) (2018). ISSN: 1943-023X

Leveraging Association Rules in Feature Selection to Classify Text

Zaher Al Aghbari and Mozamel M. Saeed

Abstract Appropriate feature selection is an important aspect in the fields of data mining and machine learning. Feature selection reduces data dimensionality and produces simpler classification models that have lower variance. In this paper, we propose a robust feature selection method that produces smaller and yet more effective set of features. The proposed feature selection method leverages association rules to select the effective features for text classification. Our experiment shows that the proposed method outperforms its pears in terms of execution time and classification accuracy.

Keywords Feature selection · Association rules · Text classification

1 Introduction

The increase in the number of online data necessitates the automatic classification of text. Text classification methods map textual data into predefined classes that are useful to some application [1], such as sensor networks [2, 3], smart cities [4], and traffic analysis [5]. By mining these classes of document, interesting patterns that define each class can be discovered. However, due the large size of these online documents, efficient feature selection methods that can be used in text classification are required.

Recently, machine learning tools were utilized to classify documents based on their textual content. Such tools are commonly used by information retrieval systems (IRSs) to answer user queries. These machine learning tools require large number

Z. A. Aghbari (✉)
Department of Computer Science, University of Sharjah, Sharjah, UAE
e-mail: zaher@sharjah.ac.ae

M. M. Saeed
Department of Computer Science, Prince Sattam Bin Abdulaziz University,
Riyadh, Kingdom of Saudi Arabia
e-mail: m.musa@psau.edu.sd

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_53

of labeled documents to learn the classification patterns. IRS systems represent each textual document by a set of features. Typically, these feature vectors are highly dimensional, which causes to several issues. The main issues are lower time performance and lower classification accuracy.

To improve the time performance and classification accuracy of the IRS systems, a feature selection method is carefully designed to filter out the irrelevant features from the feature vectors. After feature selection process, the feature vectors that represent the textual documents will be lower in dimensionality, contain the most effective features, require less memory, and processed more efficiently by a classifier. Data mining algorithms, such as association rules, can be utilized in feature selection. Typically, association rules are used to find hidden associations or patterns between items in a large database. They are made up of an antecedent (head) and consequent (body), e.g., bread → milk means that customers who buy bread are likely to buy milk as well. Although this is an example of market basket analysis, the applications of association rules, extend beyond that of the scope of market analysis.

The method in [6] proposes a text classification algorithm which uses association rules in its feature selection phase. This algorithm represents textual documents in a binary structure. Furthermore, it uses binary operations to find the association rules. The main aim of this method is reducing the number of features used and thus, reducing computation time. However, these feature vectors are still produce relatively large features that represent the text of a document.

In this paper, we propose a robust feature selection method that filters out irrelevant features and retain the effective features. The produced feature vectors are smaller and yet more effective set of features. The proposed feature selection method leverages association rules to select the effective features for text classification.

The paper is divided into the following sections: In Sect. 2, we present the related work. In Sect. 3, we present the proposed feature selection method. Section 4, discusses the experiments and comparisons to peer methods. The conclusion is presented in Sect. 5.

2 Related Work

Feature selection methods for text classification can be classified into three categories [7], wrapper, embedded, and filtered. Wrapper methods employ a greedy search approach to compare all possible combinations of features in terms of classification accuracy [8]. Therefore, their computation complexity is high. On the other hand, the embedded methods are developed as part of the classifier [9], e.g., decision tree algorithm. A filtered method is typically a separate component of classification model. These methods are faster than wrapper methods since they do not have to undergo training. Therefore, our proposed feature selection method is catered to the filtered approach to be able to process large number of text documents efficiently.

The work in [10] presented a feature selection method that uses the relative document frequencies. It selects the features according to true positive rate and false positive rate. In [11], presented a feature selection method that can effective memetic features. The work in [7] proposed feature selection methods to employing new parameters to the relevance frequency methods. A feature selection method that utilized relevancy and redundancy of features is proposed by [12].

Many classification algorithms for text documents are available; however, due to the high-dimensionality of these documents, it may take time and high computation power to train the classification models. For that reason, many researchers have used association rules for feature selection to filter out the irrelevant features and extract meaningful ones. The method in [6] proposes a text classification algorithm, which uses association rules in its feature selection phase. Their algorithm called Bit-priori Association Classification Algorithm (BACA) represents textual documents in a binary structure. Furthermore, it uses binary operations to find the association rules.

The approach in [13] is based on the heuristic that in a dataset containing textual documents that belong to one domain, relevant terms are likely to be associated with other relevant terms, while irrelevant terms are distributed randomly among documents. Firstly, they select the support and relative confidence as constraints used to generate the association rules. Then they use the Apriori algorithm to mine for rules that satisfy the support and confidence constraint (higher than a specified threshold). Finally, they score the terms based on the rules they generated; terms that have a high score are deemed relevant, while terms with a low score are irrelevant.

In [14], the paper identified a number of problems in IRS that may result in the wrong information being presented to the user. To address this, they offer SemanQE; a new semantic query expansion algorithm. It is made up of three components: A component for association rule-based query expansion, a feature selection component, and an ontology-based expansion component. The system works by creating a set of sample documents retrieved from a search engine based on a user's query. Each document divided into sentences and processed to select relevant terms and phrases. Finally, a hybrid querying expansion algorithm is applied that uses association rules and ontologies to add further fine-tuned queries that are used to retrieve the final set of documents that are relevant to the user.

The method in [15] reduced the dimensions of feature vectors by mining association rules. The association rules help to select the optimal set of features that influence the class category. Their feature selection algorithm consists of four phases. The first phase is to mine the association rules from the training data using the Apriori algorithm. The second phase is to prune the resulting rules to keep only those that contain predictive features for the class attribute. Phase three counts the frequency of the features that appear in the remaining ARs. The final phase, which is feature selection, selects only the subset of features that are higher than a frequency threshold, which are highly relevant to class. Similar work was carried out by [16], whose experimental results also showed that association rules mining for feature selection dramatically reduces the cost for classification.

3 Feature Selection Based on Association Rules

Raw data usually comes with several problems that need to be resolved before the data can be used for classification. The main problems that affects text classification is the large number of terms (features), the existence of irrelevant terms, and the dependency between terms. Therefore, raw data are usually pre-processed to resolve these problems. In our experiments, we used a dataset consisting of 2226 documents that belong to 5 classes. The proposed method contains five phases: text pre-processing, converting text to structured form, identifying frequent words, creating association rules, and classification.

3.1 Text Pre-processing

In the pre-processing phase, punctuation and stop words are removed. In addition, we applied stemming to the dataset to transform each word to its root and thus unifying the the words for easier processing.

3.2 Converting Text to Structured Form

In this phase, the documents in the dataset are converted to a binary form. This phase includes the following steps: First, we count the occurrences of every word in each document. Next, we discard any word that occurs less than the specified threshold. The threshold in this case is set to 2, so any word that occurs twice or more in an article is kept. Otherwise, it is discarded. The words that are kept are represented as 1, meanwhile discarded words are represented as 0. At the end of this phase, all the documents will be represented in binary form. These are saved in a binary table that has the documents as rows and the frequent words and class labels as columns.

3.3 Identifying Frequent Words

This phase will identify which words are most effective for predicting each class label. To find these words, we perform a binary AND operation between each class label and the frequent words. The resulting "1" s are summed and any word whose totals meet a specified threshold (set to 2) are kept, while those below the threshold are discarded. Similar to the previous phase, we represent the words that are kept with "1," and the discarded words as "0." The results will highlight which words are frequent, and thus effective, for each class label.

3.4 Creating Association Rules

For each class label, we generate subsets from the class label's frequent words. The Apriori algorithm will be applied at this stage to determine if a produced subset is frequent. If it is found to be frequent, it is saved as a rule that can be used to predict that class label. This is done by performing a binary AND operation between the generated subsets and the frequent words of each document. If there are enough matching results (= a specified threshold; 2, in this case), then the subset is saved as a rule. At the end of this phase, we will have generated a number of association rules for every class label.

3.5 Classifying Unseen Documents

Pre-processing is applied to unseen documents, and then they are converted to a structured format. To convert these unseen documents, a new binary table is created that contains the unseen documents as rows and the keywords (the words found in the rules) as columns. If the keyword is found in the document, it is marked as "1," otherwise it is marked as "0." This will give us the binary representation of each unseen document.

Finally, to classify the documents, we perform an AND operation between each unseen document and all the rules. We observe and count which rules match with the results. The class label that has the highest amount of matched rules is the class label of the unseen document.

4 Experimental Results

We used Turi's Graphlab and MATLAB. We experimented with the dataset consisting of 2226 articles that belong to five different classes: business, entertainment, politics, sport, and technology.

4.1 Association Rules Creation

For each class label, we generate subsets from the class label's frequent words. For our dataset, there are 2840 possible frequent words. Among these, several hundred are marked as frequent for each class. Because of the large number of words in this dataset, two or three-word subsets are hundreds of millions of possibilities that are too many to compute. Therefore, we have decided to only generate single-word

Table 1 Possible subsets for "Tech" frequent words

Advert	Advertis	Advic	Advis	aArospac	Affair	Affect
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	0	1

subsets for every class. Table 1 shows the possible subsets for the "Tech" row of the sample documents.

Next, we perform a binary AND between the generated subsets and the frequent words of each document. If the result is similar to the original subset, it is a match. If the matching results are greater than or equal to the threshold (which we have set to 2), then the subset is saved as a rule. The number of saved rules for each class in our dataset were as follows:

{'Business': 27, 'Entertainment': 22, 'Politics': 32, 'Sport': 50, 'Tech': 39}

4.2 Classifying Unseen Documents

Of the 2226 articles in our dataset, we saved 30% of them (684 documents) for testing. These were pre-processed and saved in a binary format. To classify these unseen documents, we perform a binary AND between each document and all the rules. We then count how many results match the rules of each class. The class label with the highest number of votes is chosen as the class label of the unseen document. The output is 684 class labels, one for each unseen document. The following is an example of an unseen document text classified correctly as 'Business'.

Sample unseen document
Peugeot deal boosts Mitsubishi Struggling Japanese car maker Mitsubishi Motors has struck a deal to supply French car maker Peugeot with 30,000 sports utility vehicles (SUV). The two firms signed a Memorandum of Understanding, and say they expect to seal a final agreement by Spring 2005. The alliance comes as a badly needed boost for loss-making Mitsubishi, after several profit warnings and poor sales. The SUVs will be built in Japan using Peugeot's diesel engines and sold mainly in the European market.

4.3 Classification Accuracy

The accuracy of our classifier was 82.2%. Meanwhile, the average precision was 86.9%, recall was 80.6%, and the F-measure was 83.6%. Table 2 shows the confusion matrix of classifying the test documents to the five classes (business, entertainment,

Table 2 Confusion matrix

	Business	Entertainment	Politcs	Sports	Tech
Business	152	1	4	1	2
Entertainment	42	68	0	2	4
Politcs	21	1	88	1	1
Sports	15	2	2	149	0
Tech	17	2	3	1	105

Table 3 Precision, recall, F-measure, and accuracy for each class

	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
Business	61	94	76	84
Entertainment	92	59	73	93
Politcs	91	79	83	94
Sports	96	88	91	97
Tech	93	83	87	96

politcs, sports, and Tech). As can be seen from Table 2, the test documents were classified with high true positives. Some false positives occurred due to the reason that some documents can be classified into more than one class. This is expected, since documents usually discuss more than one topic.

Table 3 shows the precision P, recall R, F-measure $F1$, and accuracy ACC, which are shown in Eq. 1–4, respectively, for each class.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$F1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{4}$$

where TP represents the true positives, FP represents false positives, TN represents true negatives, and FN represents false negatives.

5 Conclusion

In conclusion, we find that the proposed algorithm is effective at predicting the class label of textual documents. It uses association rules to produce high-quality rules that in turn select high-quality features. As a result, it reduces computation time and results in more accurate predictions. Therefore, the propose method is a robust in selecting features that are smaller that peer systems and yet effective.

References

1. Cai, J., Luo, J., Wang, S., Yang, S.: Feature selection in machine learning: a new perspective. Neurocomputing **300**, 70–79 (2018)
2. Al Aghbari, Z., Kamel, I., Elbaroni, W.: Energy-efficient distributed wireless sensor network scheme for cluster detection. Int. J. Parallel Emerg. Distrib. Syst. **28**(1), 1–28 (2013)
3. Al Aghbari, Z., Kamel, I., Awad, T.: On clustering large number of data streams. Intell. Data Anal. **16**(1), 69–91 (2012)
4. Hanif, S., Khedr, A.M., Al Aghbari, Z., Agrawal, D.P.: Opportunistically exploiting internet of things for wireless sensor network routing in smart cities. J. Sensor Actuator Netw. **7**(4), 46 (2018)
5. Alkouz, B., Al Aghbari, Z.: SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks. Inf. Process. Manage. **57**(1), 102–139 (2020)
6. Sheydaei, N., Saraee, M., Shahgholian, A.: A novel feature selection method for text classification using association rules and clustering. J. Inf. Sci. **41**(1), 3–15 (2015)
7. Şahin, D.O., Kılıç, E.: Two new feature selection metrics for text classification. Automatika **60**(2), 162–171 (2019)
8. Al Aghbari, Z., Junejo, I.N.: DisCoSet: discovery of contrast sets to reduce dimensionality and improve classification. Int. J. Comput. Intell. Syst. **8**(6), 1178–1191 (2015)
9. Uysal, A.K., Gunal, S.: Text classification using genetic algorithm oriented latent semantic features. Expert Syst. Appl. **41**(13), 5938–5947 (2014)
10. Kim, K., Zang, S.Y.: Trigonometric comparison measure: a feature selection method for text categorization. Data Knowl. Eng. **119**, 1–21 (2019)
11. Lee, J., Yu, I., Park, J., et al.: Memetic feature selection for multilabel text categorization label frequency difference. Inf. Sci. **485**, 263–280 (2019)
12. Labani, M., Moradi, P., Ahmadizar, F., et al.: A novel multivariate filter method for feature selection in text classification problems. Eng. Appl. Artif. Intell. **70**, 25–37 (2018)
13. Webb, G.I.: Discovering significant patterns. J. Mach. Learn. **68**, 1–33 (2007)
14. Song, M., Song, I.Y., Hu, X., Allen, R.B.: Integration of association rules and ontologies for semantic query expansion. Data Knowl. Eng. **63**, 63–75 (2007)
15. Kaoungku, N., Suksut, K., Chanklan, R., Kerdprasop, K., Kerdprasop, N.: Data classification based on feature selection with association rule mining. In: International MultiConference of Engineers and Computer Scientists, Hong Kong (2017)
16. Xie, J., Wu, J., Qian, Q.: Feature selection algorithm based on association rules mining method (2009)

Protected Admittance E-Health Record System Using Blockchain Technology

Sharyu Kadam and Dilip Motwani

Abstract E-health record is considered as an individual's health document, which gets shared among several amenities. The EHR system is becoming a popular protagonist as it has the potential to transform the paper-based industry into the digital system for maintaining the patient's health records. Nevertheless, the E-health record system needs renovation in terms of confidentiality and access to records since the hospital authority is treated as a central holder for the records. The proposed system comes up with a novel solution to modernize the traditional centralized system by the erection of a decentralized framework, whereas the patient is the sole owner of the health document. Implementation of blockchain technology with a decentralized patient-centered structure promotes a secure healthcare system. Blockchain's open admittance would permit changes to an individual's her, which needs to be restructured in real-time and makes it instantly available to parties involved for the same. The use of shrewd agreements and dispersed stockpiling improves the therapeutic administrations, including clinical records comparably persistent related proof. This high-level model gives high security and ease of utilizing the highlights through disseminated record where information can't be held for payment, where every client has a refreshed duplicate of the blockchain.

Keywords Block chain · Distributed ledger · Electronic health record (EHR) · IPFS · Security

1 Introduction

The present digital era is moving away from the long-standing system and believes in modern practices. This modernization has facilitated the convenience to different sectors so that the health sector also remains as one of the consumers, which discover novel ways for improvement. An electronic health record provides a great feature, where patient acquires clinical records in purely digital form in spite of paper [1].

S. Kadam · D. Motwani (✉)
Computer Engineering Department, Vidyalankar Institute of Technology, Mumbai, India
e-mail: dilip.motwani@vit.edu.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_54

The move towards patient control collaboration has the potential to introduce a new platform for data sharing in healthcare and hence brings new challenges and requirements related to privacy and technology. E-Health record systems are real-time and patient-focus records that offer records immediately and securely to authorized users. This structure facilitates the ownership of the records to patients in order to achieve confidentiality and ease of access. This structure remains helpful to deliver secure data access and exchange with valid approval. This protected admittance of EHR structure will offer cavernous shared faith between each associated with the help of blockchain technology [2]. Furthermore, this system provides the distributed storage to maintain EHRs in a secure and steady form, hence the patient become a sole owner and creates a patient-centred environment to provide interoperability among the specialist [3].

2 Literature Review

[1]**"Securing Blockchain based Electronic Health Record using Multilevel Authentication Radhakrishnan, A Sam Joseph, S. Sudhakar, International Conference on Advanced Computing & Communication Systems (ICACCS), 2019**.

E-medical system has various shades of patient's medical information and their medical antiquity. Damage to electronic health record stimulates a confusing prescription. Societal insurance frameworks serve fewer safety attempts to guarantee about fitness histories. Blockchain is a flowed and distributed data that acknowledges an essential action in guaranteeing the information and exchange. Presentation of blockchain with respect to the social organization framework guards the fitness information against the provokers. This paper suggests an astounding insistence and grounded course of action to guard the blockchain against the assaults referred currently.

[2]**"Using Blockchain for Electronic Health Records", Ayesha Shahnaz1, Usman Qamar1, (Member, IEEE), Ayesha Khalid, (Member, IEEE), IEEE, 2019**.

In time, this effort will revolutionize the world of technology. This change has made it much easier for the patient to handle data in digital form but there are also some problems with its convenience. Safety and confidentiality are the two most important apprehensions for patients and it is equally important to verify the unauthorized access to such vital information. Of course, adaptability and interoperability are also remaining critical issues that require significant research focus. Proposed work addresses the special issues and spotlights the profits of the blockchain development for the association of an ensured and flexible response for clinical data exchange in order to have the best introduction.

[3]**G. Jetley and H. Zhang, "Electronic health records in IS research: Quality issues, essential thresholds and remedial actions," Decis. Support Syst., vol. 126, pp. 113–137, Nov. 2019**.

E-clinical records are serious, extremely delicate isolated data in the medical sector, which is must commonly share with other parties. Blockchain delivers a communal, absolute and secure structure to established accountability and transparency. This research offers an exclusive prospect to develop a confidential and distributed EMR system using blockchain. The perspectives of this paper are to provide a decentralized E-clinical system, for data sharing among the healthcare providers. In association with Stony Brook University Hospital, the executed structure ensures security, availability, and accessibility over E-medical records. Offered effort can expressively shrink the turnaround time for EHR distribution and progress judgment ability for health care.

[4]**"Health Record Management through Blockchain Technology", Harshini V M, Shreevani Danai, Usha H R, Manjunath R Kounte, IEEE, 2019**.

The exchange of E-records on the clinical background has unimaginable progressive importance for the exploration of infirmity and experts' decision. Lately, cloud-based E-clinical data exchange arrangement has transported enormous measures of convenience, yet the monopolization for cloud system opens risks unavoidably to data safety. Blockchain advancement is oftentimes seen as a promising response to deal with these issues by ideals of its amazing assets of decentralization, mystery, and proof. Here, the blockchain-based structure offers safe and shielded exchange of E-health records. Data demander demands for specific e-health record which is retrieved from the blockchain-based cloud server in the encrypted form after the owner's permission. The use of blockchain methodology achieves security and confidentiality aspects for the system.

[5]**"BlocHIE: a BLOCkchain-based platform for Healthcare Information Exchange", Jiang, S., Cao, J., Wu, H., Yang, Y., Ma, M., & He, J. 2018 IEEE International Conference on Smart Computing (SMARTCOMP)**.

HIE submitted remarkable achievements for the clinical sector. Uploading and sharing copious clinical information is a chief requirement as well as a great challenge. In this paper, BlocHIE is a platform that provides interoperability for healthcare information. The initial step as per this paper was to perform an investigation of the different requirements which need to be shared and their different sources. Based on this investigation, the system will hire Blockchain to operate on miscellaneous e-health records. Furthermore, to achieve confidentiality and authenticity combination of on-chain and off-chain verification was used.

3 Problem Statement

Traditional framework believes in centralized storage method and as per this traditional methodology clinical centers are responsible to hold the entire information of a patient into to the central database as well as the access of this data is also done through a central authority [3, 4]. According to the investigation of the current system, one of the problems has been noticed that various types of vital as well as

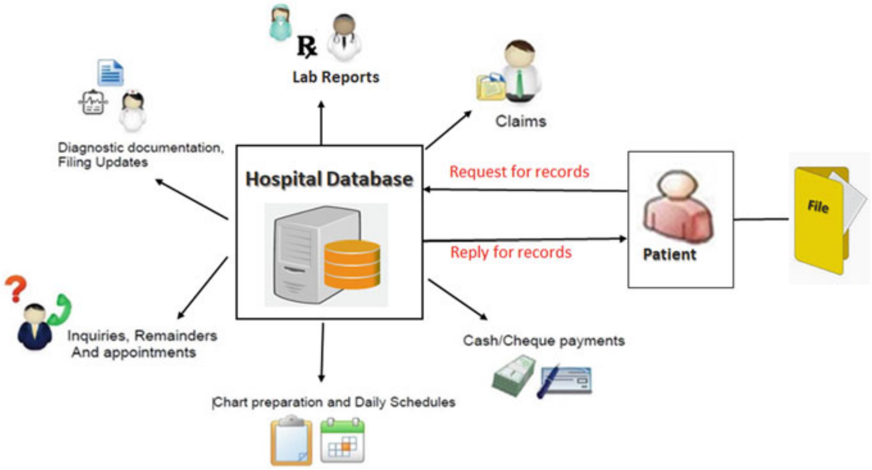

Lab Reports
Claims
Diagnostic documentation, Filing Updates
Hospital Database
Request for records
Reply for records
Patient
File
Inquiries, Remainders And appointments
Cash/Cheque payments
Chart preparation and Daily Schedules

Fig. 1 Existing centralized EHR system

personal information take place into the EHR system like appointment documents, different format of reports and personal information which is in steady form and not available all the time with the patient [5]. Access to past clinical records for a patient is a struggling task as they need to approach hospital authority every time. Provision of secure medical data is another toughest task as this data handle by various clinicians/doctors.

Figure 1 illustrate the traditional framework of data where hospital management is the owner of the data process and data exchange. In this centralized structure, After the patient is examined or treated for his illness, the hospital office authority is responsible to take care of all the relevant documents. Data access is one of the irritating and waiting processes from the patient site as they require the permission of hospital authority every time. By reason of the consolidated storage system faces information damage.

3.1 Objectives

The approach behind the proposed system is to come up with a polished framework to whelm the issues related to the existing system and Provides secured and classified records by exploring “Blockchain Technology” [4]. The main objective of this system is to eliminate centralized authority and avail easy access facility to the patients for their steady form records anytime. Utilization of blockchain technology helps in order to provide secured and unchanged data which reduce the data loss or data modification [6]. Our examination contributes innovative information for the interconnection between the performances of an E-Health Record structure and as needs be the idea of social administrations passed on inside the inpatient setting. The goal of the system is to divert patient from paper-based industry to digitalization hence no need to transport records in the form of papers by patient. Digital gadgets becomes one of the finest media for storing the data in order to improve the usability of the system.

4 Related Work

Until now, most of the hospitals were operating according to the existing system which is based on centralization and due to such a central system records handling is not an easy task for patients [7]. The necessary situation according to that patient always handles records in the paper arrangement. Storage of large volume records and to access those records is another challenge for the existing system [8]. Adoption of the decentralized method provides handy use of records to the patients. Another major concern of this historical structure is security provision for vital information of the patient. According to the centralized approach, all the records handled by hospital management [9]. Hospital is blamed for carelessly supervision the clinical data, changing it or breaking the privacy of the histories. Blockchain-Based EHR along with decentralization delivers a locking system to the patients due to which patient has the facility to restrict the admittance of records for health specialist [10, 11].

5 Proposed System

With most advanced stage of development in computers, an existing structure provides approach for storage of patient's health records. This system provides an excellent type of technique that allows the patient's record, reports and specialist's solution to be handled very well. Due to centralized mechanism there is chances of data leakage and unauthorized access of patient's records. In centralized framework patients are unable to hide and secure their own data. There is always a lobby and patient need to wait outside the lobby when they want to access their data. By considering these issues, the system is come up with a new concept of decentralization by using Blockchain technology. System is able to gain security of records by implementing unique ID and hashing concept of blockchain. Unauthorized access restricted by adoption of decentralized method.

Figure 2 overviewed the decentralized concept for secured information exchange by using latest technology i.e. Blockchain where patient is a decision maker for his/her records by granting and denying the permissions. Uploading records with the unique id is the main task of the patient. Further, these records transferred on Blockchain with secured hash value.

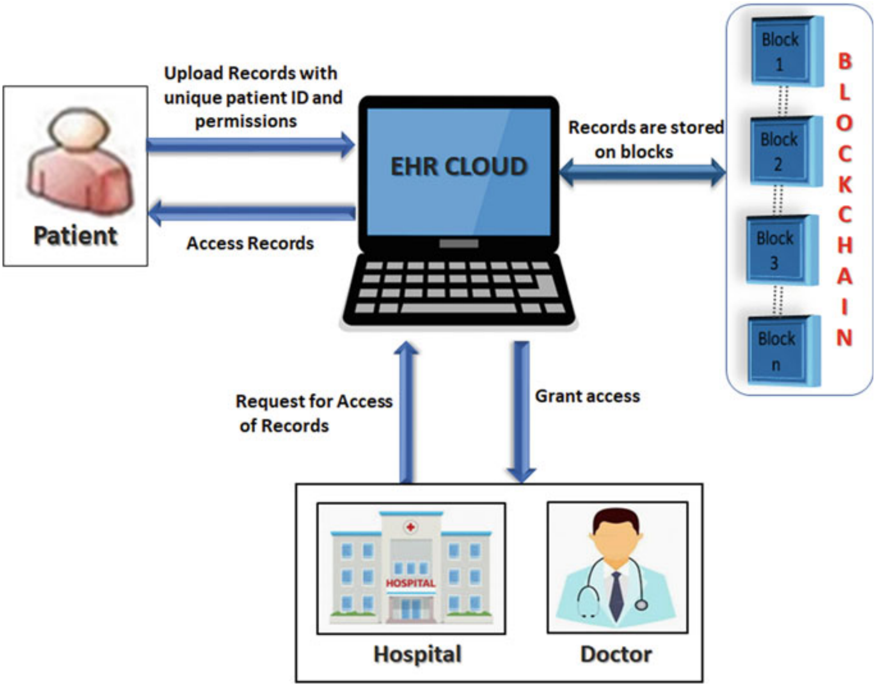

Patient
Upload Records with unique patient ID and permissions
Access Records
EHR CLOUD
Records are stored on blocks
Block 1
Block 2
Block 3
Block n
BLOCKCHAIN
Request for Access of Records
Grant access
HOSPITAL
Hospital
Doctor

Fig. 2 Blockchain-based decentralized EHR system

5.1 *Blockchain Background*

The inauguration of Digital currency was introduced by Satoshi Nakamoto in 2007 [12]. Blockchain technology has a great contribution to every sector as well as many aspects of our lives [13]. This technology provides a large platform for constantly growing ledger which keeps a permanent record of all the transaction in a secure, consecutive and static [13]. This juncture uses a distributed system that allows the information to be spread and that every cycle of particular data or usually identified data have shared belonging. Blockchain innovation breaks the twofold go through the issue with the assistance of public-key cryptography, where every client has relegated a private key and a public key is imparted to any remaining clients [14]. This benefit provides a reasonable substitute for the patient's clinical storage subsequently, on the grounds that the advanced expansion of social administration's commerce focused the safety for the patient's clinical data on a prior basis.

Figure 3 clarifies the universal blockchain measure for exchange in which the sender can make an exchange of a square. This square of exchange is approved through cryptographic hashing. Additional, this conveyed diced exchange is submitted and remunerated by excavators and moved to the recipient.

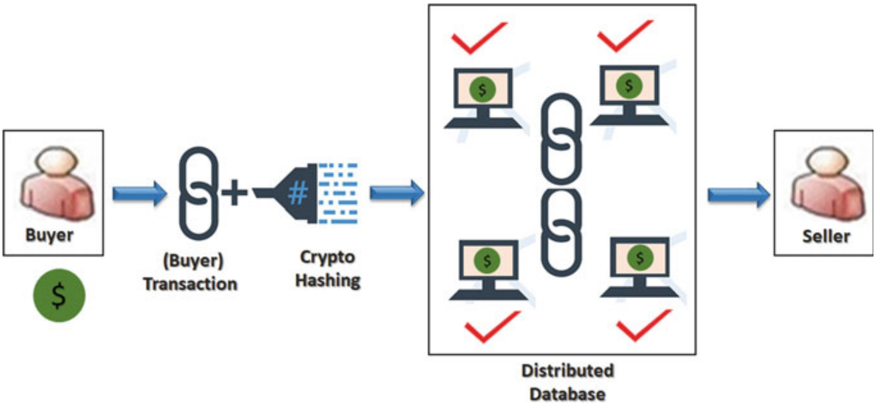


Fig. 3 Blockchain process

5.2 *Flow of the System*

Figure 4 characterizes the progression of the framework which has three principal sections are Patient, Admin, Doctor and Hospital. The operational progression of the proposed framework alongside these sections is as per the following:

1. Creating a clinical specialist's profile is the first task of the admin and for that, the admin needs to login into the system himself.
2. In the beginning, patient must register to the system and then proceed to login.
3. After entering the system, the patient generates his / her identification through a unique medical ID.
4. At this moment data uploading takes place and the patient is ready to upload its clinical and personal information.
5. A special security facility is available to the patient and according to that patient restricts the record access.
6. Hospital staff and doctor can search patient by using their unique medical id and for this first, they have to enter into the system through login.
7. Patient's records can be accessed by the doctor and hospital once it is approved by the patient.
8. Once permission is granted by the patient for accessing records, then only documents can be viewed by a doctor and hospital.

5.3 *System Architecture*

A protected admittance EHR system is a storage platform where vital information of the patient is uploaded to the cloud. Data Uploading and data sharing are important processes executed by this system. E-medical care records may consolidate singular information along with therapeutic data of patients provided by them. A patient able

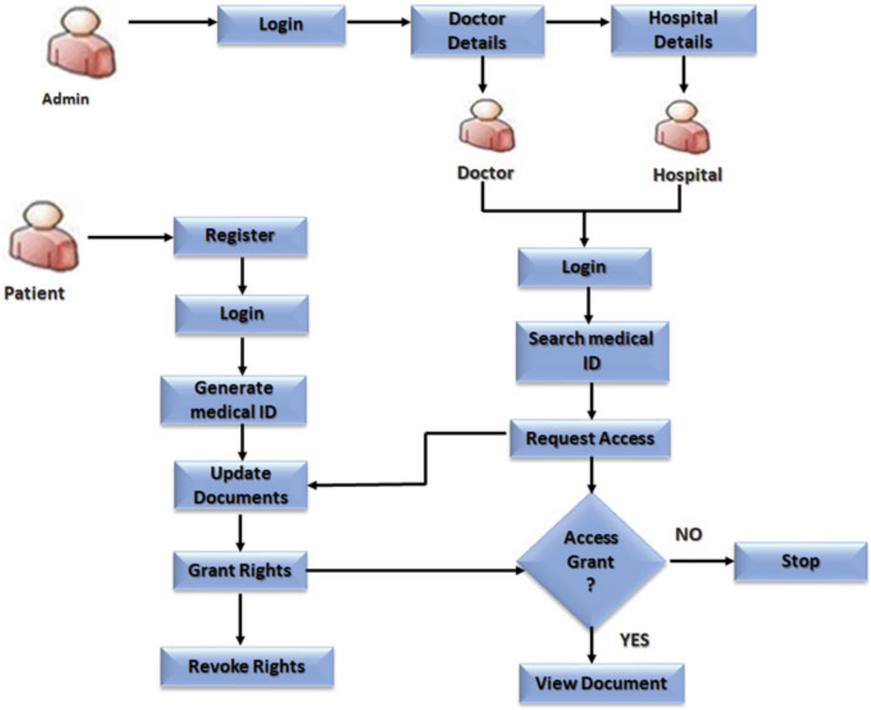


Fig. 4 System Flow

to create their own remarkable clinical ID also they have a special facility to restrict admittance of their histories for health sectors.

As shown in Fig. 5. The entire system is divided into two major parts are data uploading and data sharing.

Data Uploading

In this blockchain-based decentralized system, patients play the chief role in uploading their own vital information on the cloud as well as they can generate a unique medical ID as their identity [15]. Separate blocks are created for an individual patient and generated unique id is linked with the blockchain [16]. Each block of blockchain enclosed with its own Block hash value, previous block hash value, patient ID, and time stamp.

Data Sharing

Data exchange is a crucial part of this system, which facilitates easy access for patients to their own data. Data sharing is done by the following important modules:

Admin: Admin is a mastermind behind exchanges and procedures on the cloud. An administrator is careful to send adroit agreements and the primary component with the ability to invigorate or change game plans in sharp agreements [16].

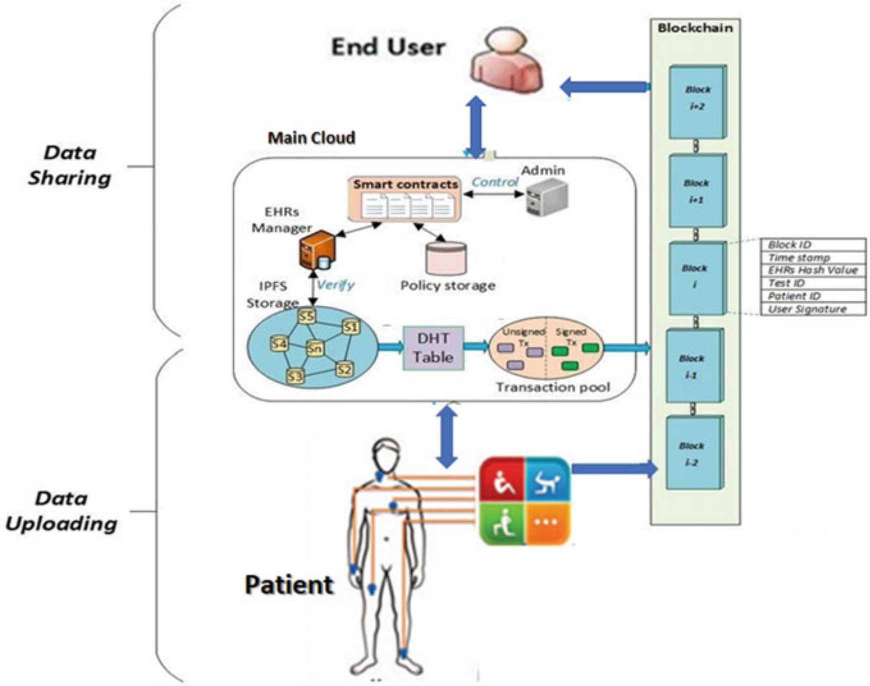


Fig. 5 System structure

EHRs manager: In this system information sharing is a chief constituent. The EHR manager contributes major efforts to control all customer trades on the blockchain [16]. The organization of EHRs overseer is engaged by quick agreements through demanding customer systems.

Smart contracts: They indicate all exercises acceptable in the passage regulator structure. Customers can interface with sharp arrangements by understanding location [17]. A canny arrangement can perceive, affirm sales and grant get to approvals for medical customers by enacting trades [17]. The quick arrangement and its exercises are available to all blockchain components (Fig. 6).

SHA 256 Algorithm

To avail secure hash bits, National Institute of Standards & Technology invented uniques algorithm called as SHA—Secure Hash Algorithm, and now a days hashing done with their latest version SHA-256 The key procedure is given in the name of this algorithm, and accordingly this algorithm calculates the hash length of 256 bits which is very remarkable [18]. The principle behind this is the input value is at the end added to the output and that it transforms an encryption algorithm into a "hashing" algorithm, a building piece of a standard hash function. The underlying block cipher has 64 rounds and thus a 2048-bit expanded internal key (64 × 32 bits).

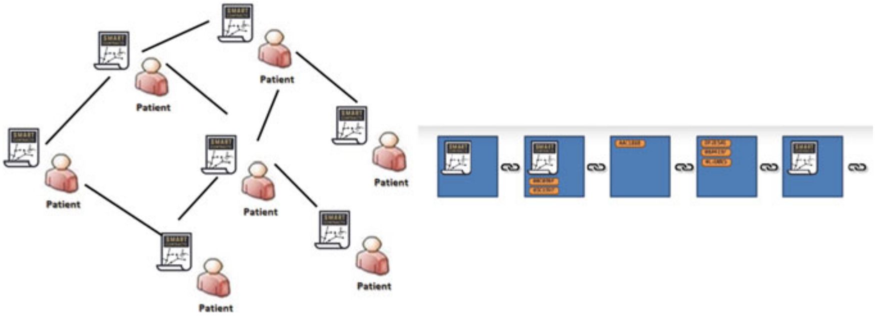


Fig. 6 Smart contracts

This key is obtained from the message block to be compressed, which has 512 bits at the input and is expanded four times to form this 2048-bit internal key for our block cipher. The SHA-256 algorithm is divided into two stages: pre-processing and hash computation [18]. Pre-processing involves padding a message and parsing the padded message into m-blocks. Initialization values are set which is to be used in the hash computation. Hash computation produces a message plan from the cushioned message. The yield hash esteem produced by hash computation is used to decide the message digest. Hash computation includes message plan, functions, capacities, and word activities that are created iteratively to get hash esteem. Security of SHA-256 hash calculation based on the amount of the hash value.

Interplanetary File System

An interplanetary file system is a show which usages conveyed framework for data amassing. It gives secure data storing as data set aside on this file system is shielded from every variation [19]. It uses a cryptographic identifier that protects the data from the change as any undertaking to make a change on the data set aside on IPFS should be done by altering the identifier.

Figure 7 represents the part of the interplanetary file system stockpiling for the blockchain-based system. As this file system is a distributed convention any place each data stores a bunch of diced documents checked by EHR director through shrewd agreement strategy [19]. IPFS is a non-centralized convention that monitors the substance-based tending. When hash esteem produced for the explicit data, that confused data alongside novel key is moved to the conveyed hash table (Fig. 8).

System Screenshots

See Figs. 9, 10, 11, 12 and 13.

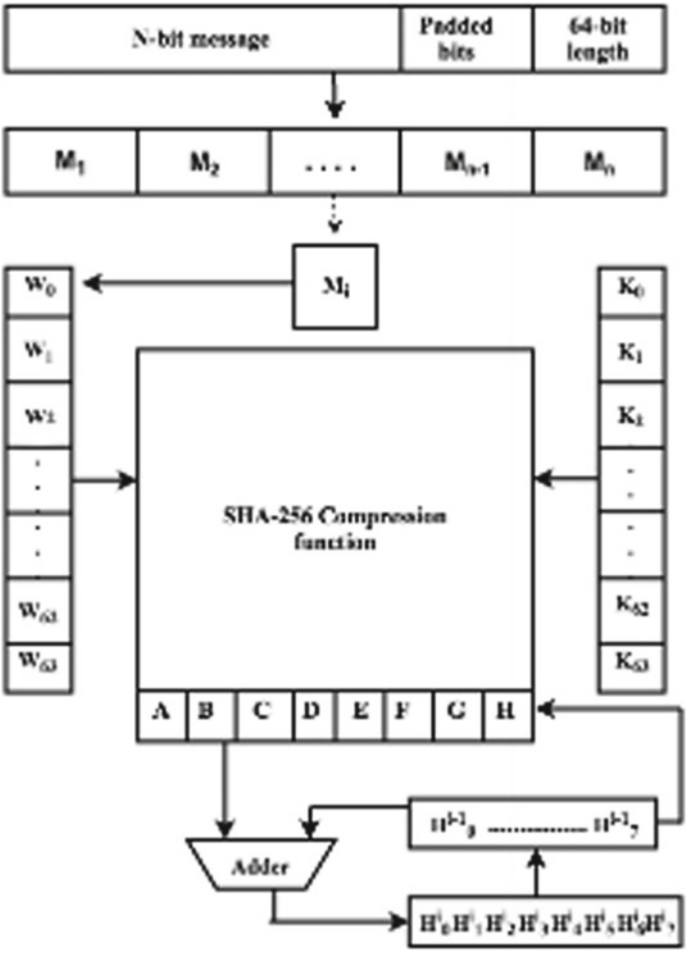


Fig. 7 Processing of SHA 256

6 Result Analysis

Due to its idiosyncratic properties, it has been proven that a decentralized EHR system gives much better performance than a centralized system. According to decentralization, patients have become free birds. They can handle their records anytime and most importantly easily. The security of a patient’s vital and personal information with the help of blockchain technology is inevitable but this ease has created a puzzle and that is scalability. These systems got successes to solve generate code by using the off-chain database with the help of IPFS. According to this methodology, all the documents related to the patient’s medical history are stored in off-chain databases

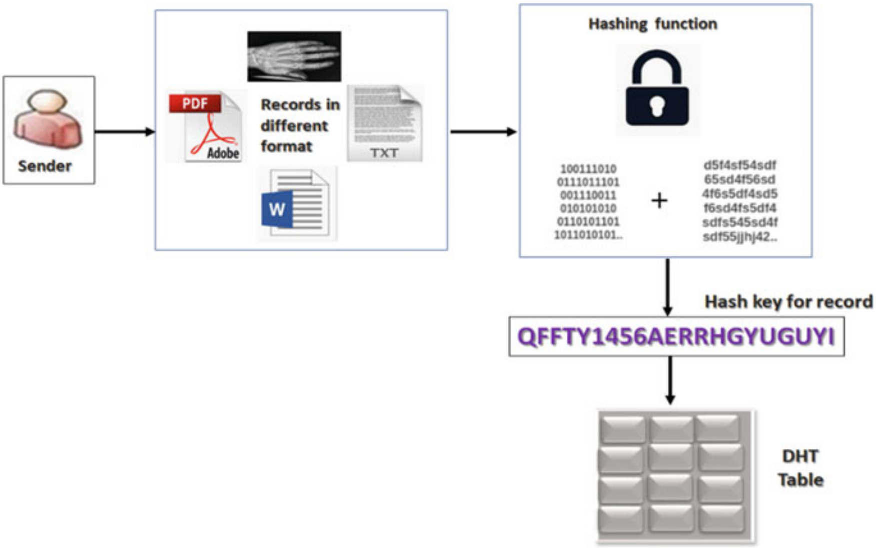
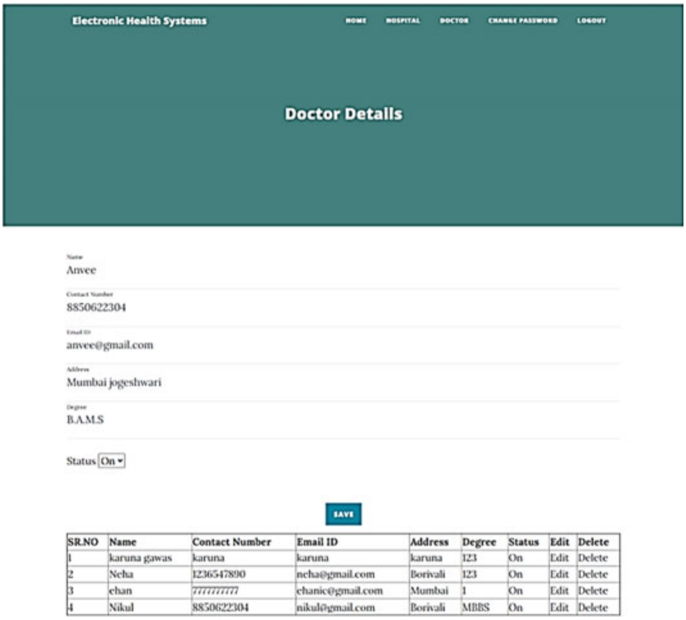


Fig. 8 IPFS process



Name
Anvee

Contact Number
8850622304

Email ID
anvee@gmail.com

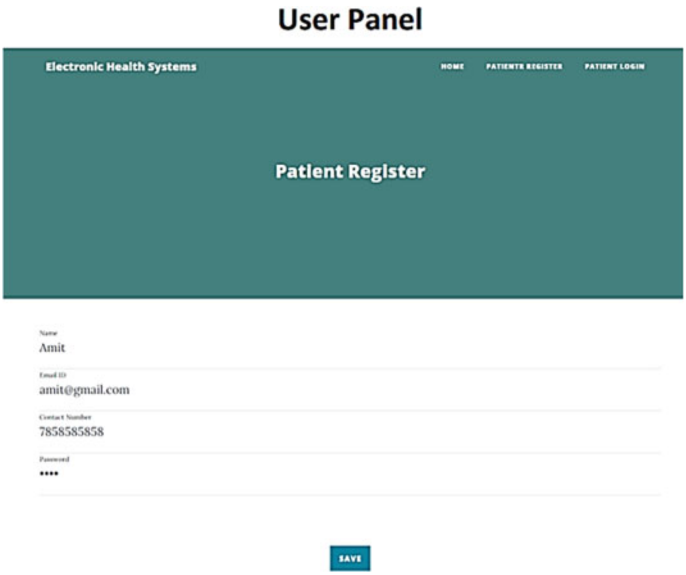
Address
Mumbai jogeshwari

Degree
B.A.M.S

Status On

SR.NO	Name	Contact Number	Email ID	Address	Degree	Status	Edit	Delete
1	karuna gawas	karuna	karuna	karuna	123	On	Edit	Delete
2	Neha	1236547890	neha@gmail.com	Borivali	123	On	Edit	Delete
3	chan	7777777777	chanic@gmail.com	Mumbai	1	On	Edit	Delete
4	Nikul	8850622304	nikul@gmail.com	Borivali	MBBS	On	Edit	Delete

Fig. 9 Doctor/Hospital registration panel



Name
Amit

Email ID
amit@gmail.com

Contact Number
7858585858

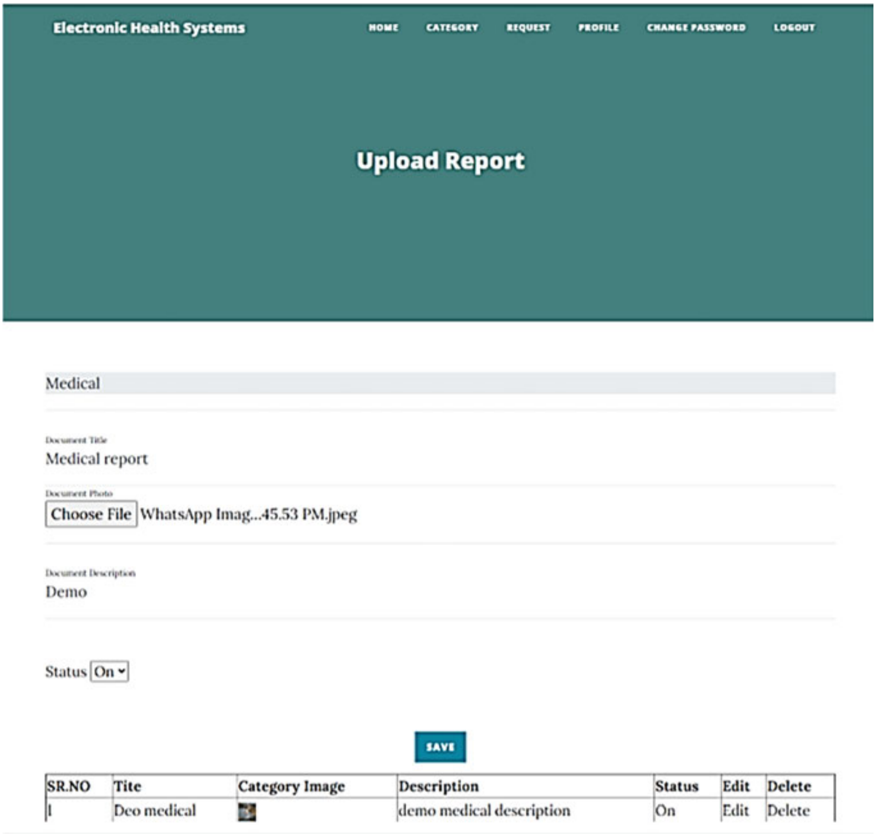
Password
••••

Fig. 10 Patient registration panel

instead of blockchain so that the scalability of the system is enlarged by the attractive number of Percentage. Figure 14 clarifies the comparison between centralized health record system and decentralized health record system on the basis of various characteristics (Table 1).

7 Conclusion

This age in the field of technology is evolving day by day so much that a new invention is becoming very popular where patients became landlord of their record process. For given system, the patient holds decision-maker degree who took all decisions related to data sharing and processing. Provision of high-level security for patients records and avert data leakage is the highest attainment of the system. This milestone achieved by using blockchain technology which fulfils the security aspects. Storing all the vital information of patients in distributed ledger is a very crucial function that blockchain methodology contributes completely to the performance. The system provides fully protected data for the authorized user. In any event, when the patient visits any clinical foundation like an emergency clinic, the patient should throw the clinical id, entire evidence that hospital on the patient's clinical id which can be gotten to anyplace. SHA 256 has an important contribution to this system in order to supply a secure hash value. A 256-bit hash value is the key point of this system. Furthermore, the construction offers procedures to ensure the system can take care



Medical

Document Title
Medical report

Document Photo
Choose File WhatsApp Imag...45.53 PM.jpeg

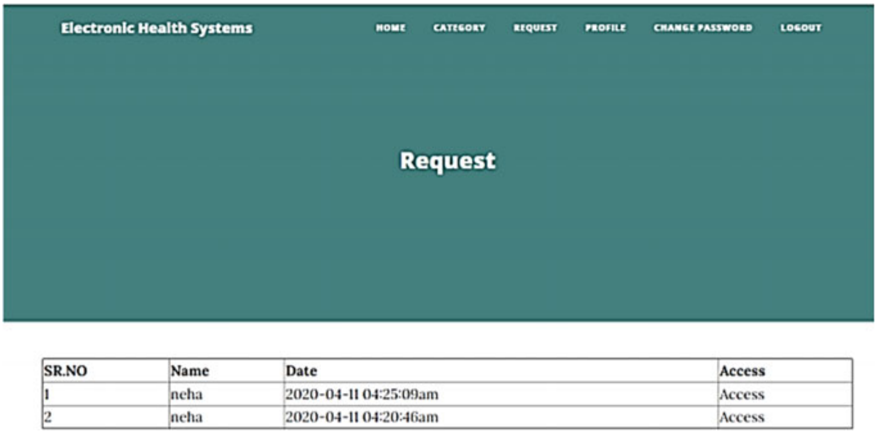
Document Description
Demo

Status On

SAVE

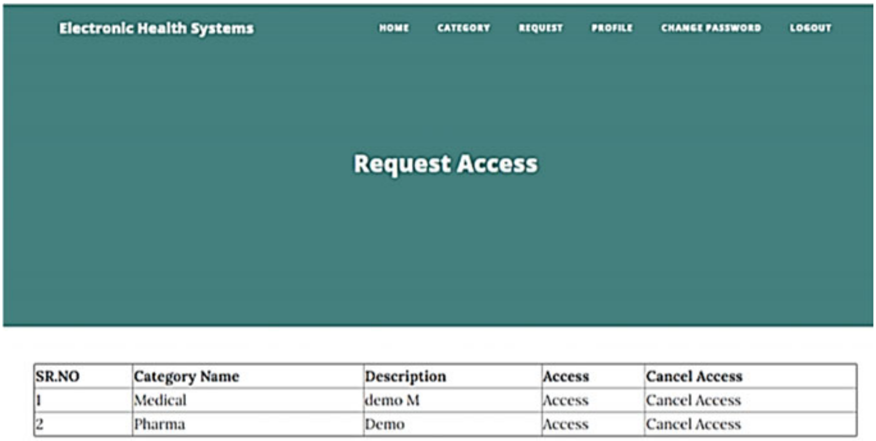
SR.NO	Tite	Category Image	Description	Status	Edit	Delete
1	Deo medical		demo medical description	On	Edit	Delete

Fig. 11 Data uploading



SR.NO	Name	Date	Access
1	neha	2020-04-11 04:25:09am	Access
2	neha	2020-04-11 04:20:46am	Access

Fig. 12 Data sharing (Grant Access)



Request Access

SR.NO	Category Name	Description	Access	Cancel Access
1	Medical	demo M	Access	Cancel Access
2	Pharma	Demo	Access	Cancel Access

Fig. 13 Data sharing (Cancle Access)

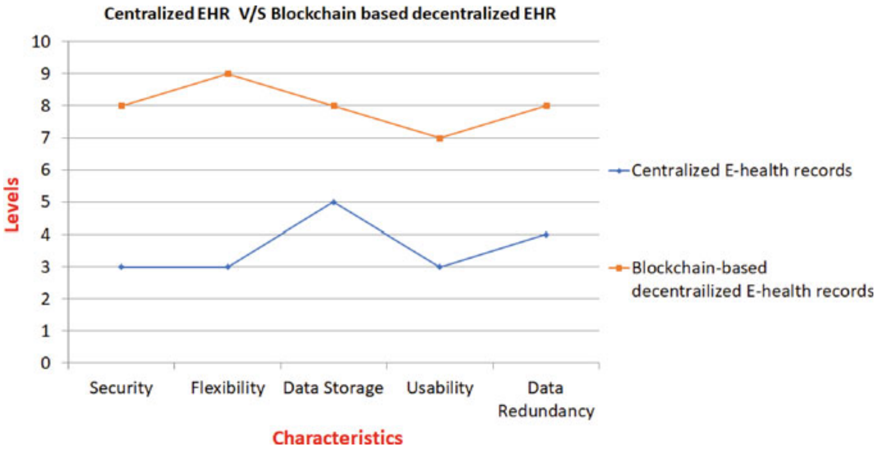


Fig. 14 Centralized EHRs versus blockchain-based decentralized EHRs

Table 1 Performance characteristics of centralized EHRs and decentralized EHRs

Characteristics	Centralized E-health records	Blockchain-based decentralized E-health records
Security	3	8
Flexibility	3	9
Data storage	5	8
Usability	3	7
Data redundancy	4	8

of the concern of data storing as it utilizes the off-chain accumulating part of IPFS. This outlook also passes on a clever arrangement, like a puzzle, which executes on its own when both the social events agree on the course of action of shows. Patients are able to upload insurance-related documents along with medical documents but the insurance claim settlement process is still pending. Given structure is a purely web-oriented system and due to this sometimes causes a delay in permission granting process. A mobile-based application would be a great way to avoid this delay.

Acknowledgements I need to loosen up my real appreciation to all who caused me for the endeavour work. I need to sincerely offer thanks towards Dr. Dilip Motwani for their info and consistent bearing for giving basic information regarding the commission in like manner, for their assistance in finishing this task. I want to express my thanks to people and everyone from the Vidyalankar Institute of Technology for their smart co-action and backing.

References

1. Radhakrishnan, B.L., Joseph, A.S., Sudhakar, S.: Securing blockchain based electronic health record using multilevel authentication. In: International Conference on Advanced Computing & Communication Systems (ICACCS), 2019
2. Shahnaz, A., Qamar, U., Member, IEEE, Khalid, A., Member, IEEE: Using Blockchain for Electronic Health Records. IEEE, 2019
3. Jetley, G., Zhang, H.: Electronic health records in IS research: quality issues, essential thresholds and remedial actions. Decis. Support Syst. **126**, 113–137 (2019)
4. Harshini, V.M., Danai, S., Usha, H.R., Kounte, M.R.: Health Record Management through Blockchain Technology. IEEE (2019)
5. Jiang, S., Cao, J., Wu, H., Yang, Y., Ma, M., He, J.: BlocHIE: a BLOCkchain-based platform for Healthcare Information Exchange. In: 2018 IEEE International Conference on Smart Computing (SMARTCOMP)
6. Wisner, K., Lyndon, A., Chesla, C.A.: The electronic health record's impact on nurses' cognitive work: an integrative review. Int. J. Nursing Stud. **94**, 74–84 (2019)
7. Akbari, E., Zhao, W.: The impact of block parameters on the throughput and security of blockchains. In: ICBCT'20: Proceedings of the 2020 the 2nd International Conference on Blockchain Technology, Mar 2020
8. Bardhan, I.R., Thouin, M.F.: Health information technology and its impact on the quality and cost of healthcare delivery. Decis. Support Syst. **55**(2), 438–449 (2013)
9. Mikula, T., Jacobsen, R.H.: Identity and Access Management with Blockchain in Electronic Healthcare Records. IEEE (2018)
10. Cirstea, A., Enescu, F.M., Bizon, N., Stirbu, C., Ionescu, V.M.: Blockchain technology applied in health. In: ECAI 2018—International Conference—10th Edition Electronics, Computers and Artificial Intelligence, 2018
11. Pramod, P., Tripathy, P.K., Bajpai, H., Kounte, M.R.: Role of natural language processing and deep learning in intelligent machines. In: IEEE International Conference on Electrical, Communication, Electronics, Instrumentation and Computing (ICECEIC), Kanchipuram, India, 30–31 Jan 2019
12. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electrnic Cash System (2008)
13. Vivekanadam, B.: Analysis of recent trend and applications in block chain technology. J. ISMAC **2**(04), 200–206 (2020)
14. Koczkodaj, W.W., Mazurek, M., Strzałka, D., Wolny-Dominiak, A., Woodbury-Smith, M.: Electronic health record breaches as social indicators. Soc. Ind. Res. **141**(2), 861–871 (2019)

15. Spatar, D., Kok, O., Basoglu, N., Daim, T.: Adoption factors of electronic health record systems. Technol. Soc. **58** (2019)
16. Nguyen, D.C., Pathirana, P.N., Senior Member, IEEE, Ding, M., Senior Member, IEEE, Seneviratne, A., Senior Member, IEEE: Blockchain for Secure EHRs Sharing of Mobile Cloud based E-health Systems (2019)
17. Novikov, S.P., Kazakov, O.D., Kulagina, N.A., Azarenko, N.Y.: Blockchain and smart contracts in a decentralized health infrastructure. In: 2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies" (ITQMIS), pp. 697–703. IEEE (2018)
18. Gilbert, H., Handschuh, H.: Security Analysis of SHA-256 and Sisters. France Telecom R&D, FTRD/DTL/SSR 38–40 Rue du General Leclerc, F-92131
19. Zheng, Q., Li, Y., Chen, P., Dong, X.: An innovative IPFS-based storage model for blockchain. In: An Innovative IPFS-Based Storage Model for Blockchain. 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (2018). https://doi.org/10.1109/wi.2018.000-8

Using Hierarchical Transformers for Document Classification in Tamil Language

M. Riyaz Ahmed, Bhuvan Raghuraman, and J. Briskilal

Abstract Document classification is used for various applications from spam detection in email to article classification. Recently document classification in Tamil has been gaining momentum due to increased data available in said language. One of the major advances in document classification is due to bidirectional encoder representations from transformers (also known as BERT), which uses transformer architecture and has been used effectively in many natural language processing problems like sentiment analysis and document classification for Tamil language. One of the main disadvantages of pre-trained BERT model is the number of tokens cannot be higher than 512; otherwise, it has to be retrained. Our implementation mitigates this issue by using hierarchical transformer architecture and is especially useful for resource poor languages like Tamil. We compare hierarchical transformer model and compared with classical machine learning algorithms and found recurrence over BERT shows substantial improvement over SVM, logistic regression and random forest, with a weighted average $F1$ score of 0.88 for news article classification.

Keywords Tamil · Natural language processing · BERT · Large document classification

M. Riyaz Ahmed (✉) · B. Raghuraman · J. Briskilal
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chengalpattu, Tamil Nadu, India
e-mail: mv6796@srmist.edu.in

B. Raghuraman
e-mail: br2367@srmist.edu.in

J. Briskilal
e-mail: briskilj@srmist.edu.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_55

1 Introduction

Document classification is a task which is used to classify a document into either one of the categories. These documents could be either a blog post, mail, news article, book, etc. Document classification could be either supervised, unsupervised, or semi-supervised depending on the problem. Document classification has several applications from classification of texts in books to spam filtering to sentiment analysis. The input tokens are quite large in document classification, when compared to standard natural language processing, thus requiring different approach than standard text classification. Either pooling (either minimum pooling, maximum pooling or average pooling) has to be done or some part of the text has to be removed. An example sentence is given below.

> We choose to go to the moon. We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills, because that challenge is one that we are willing to accept, one we are unwilling to postpone, and one which we intend to win, and the others, too.

This excerpt contains 76 words; but if a model has been trained for 50 words, we might need to remove last few words to fit into model.

> We choose to go to the moon. We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills, because that-

Notice how some words are lost and the model now has less information to start with.

This is especially the case in Tamil language where there are very few pre-trained models available. Our approach uses hierarchical architecture on multilingual BERT (a multilingual model which was trained on 104 languages) instead of base BERT model for Tamil document classification.

1.1 *Tamil Language*

Tamil (தமிழ்) is a Language which is primarily spoken by the Tamil community and is part of the Dravidian language tree. The Dravidian language tree consists of four main branches: Northern, Central, South-Central, and Southern Dravidian languages. Tamil is majorly spoken in many parts of Southern India, Sri Lanka, and Singapore. Unlike many of other Indian languages, Tamil does not have aspirated tones. Tamil is considered to be agglutinative language where affixes are added to a given root word, and these affixes indicate number, noun class, tense, politeness, etc. Most of the affixes in Tamil are suffixes. The order is usually of subject-object-verb (SOV). Tamil also combines both adjectives and adverbs together to form uriccol [1].

Table 1 Sample text in each category for Tamil document classification

Category in Tamil	Category in English	Sample text
இந்தியா	India	… தேரில் பொருத்தப்பட்டிருந்த ரூ.1 …
உலகம்	World	லண்டன்: திருமணம் செய்துகொள்ளும் …
குற்றம்	Crime	… சாலையோரம் உள்ள டிபன் கடையில் …
தமிழகம்	Tamil Nadu	சென்னை:சட்டசபை தேர்தல் பணிக்காக …
தலையங்கம்	Editorial	கணவனைத் தவிர யாரிடம் சிரித்துப் …
விளையாட்டு	Sports	… இந்தியா அணிகள் நேற்று 4வது …

The writing system used to write Tamil is called Tamil script, which was derived from Brahmi script which got popularized by Emperor Asoka (ca. 250 BCE), who used Brahmi script in inscriptions publicizing his edicts [2]. Thus, many of the writing systems derived from it (including Tamil) follow similar rules. Almost all of scripts descended from Brahmi script is classified under Abugida writing system, where each letter represents either a vowel on its own or a consonant followed by vowel either to the left, right, top, or bottom. If there is no vowel nearby, it is assumed to have "default" vowel, /a/.

A small sample of each category is given in Tamil document classification is given in Table 1.

2 Background Work

2.1 BERT Language Model

Bidirectional encoder representations from transformers [3] is a relatively new language representation model using transformer architecture which contains sets of feed forward and self-attention layers [4]. Unlike other models, BERT is designed to be pre-trained using both left and right context in all layers. This makes BERT more language independent as it could understand both left and right context as some languages depend more on later than the former. There are two main parts in training BERT—pre-training and fine tuning. Pre-training is done using Wikipedia (2500 million words) and BooksCorpus (8000 million words). BERT pre-training is done in two parts - masked language modeling (MLM) and next sentence prediction (NSP). MLM hides some words from the given sentence, and the model is trained to predict the hidden word. Note that this is different from classical language modeling (CLM) where the goal is to find the next word from a sequence of words. An example of MLM is discussed below:

I think, therefore [MASK] am.

The model tries to fill the [MASK] using both right and left context. The model is tuned based on how far is the result when compared to the actual value.

I think, therefore I am.

NSP is used to improve the understanding of relation between sentences by using the model to figure the next sentence given previous sentences. This pre-trained model can be used per application for further fine-tuning. The result of this is over 7.7-point absolute improvement in GLUE score to 80.5, 4.6% absolute improvement to 86.7% in MultiNLI and 5.1-point absolute improvement in $F1$ score of SQuAD v2.0 question and answering test.

Multilingual BERT (mBERT) is a variation of BERT but was trained to be multilingual (to be able to understand multiple languages). While there is no difference between BERT and mBERT in terms of model architecture, there is some difference in tokenization and training phase. The entire Wikipedia for 104 languages is the corpus for pre-training multilingual BERT. But using this causes many of low resource languages to be under-represented. So to mitigate this issue, exponentially smoothed weighting of data is performed during both tokenization and pre-training. This allows resource rich language like English to be under sampled and very low resource languages like Irish, Ido, and Swahili are over-sampled. During tokenization, whitespaces are added to CKJ (Chinese, Japanese, Korean although Korean Hangul and Japanese Hiragana and Katakana are not included in the list since they are syllabic) characters before using WordPiece tokenizer. It also removed accent marks and converts everything in lowercase. In Tamil specifically, both "◌்" and "◌ீ" are removed due to how Unicode works.

2.2 RoBERT

Recurrence over BERT or RoBERT is a hierarchical model which uses LSTM along with BERT language model [5]. The document is first split into several chunks with overlap. Overlap ensures some context for previous chunk is passed to the next chunk. Then each chunk is independently fed into the BERT language model. The pooled outputs of each chunk could be considered to be a chunk vector. The pooled outputs are then stacked next to each other and is then passed into a relatively small LSTM layer. The output of the LSTM layer is considered to be the vector of the entire document. Then the document vector is passed to a very small linear layer which is then used for predicting the class of the document.

3 Related Work

This section focuses on two dimensions—one being document classification for English language and other being document classification for Tamil language.

3.1 Document Classification in English Language

Adhikari et al. [6] use knowledge distillation on BERT for document classification task, thus significantly reducing the number of parameters, but the average document length is less than 512 tokens (pre-trained BERT's Limit). Yang et al. [7] use hierarchical attention networks, in which each set of attention layers focuses either on word level or sentence level attention. It was trained and evaluated with average number of words being around 150 words. Rao et al. [8] use word level and sentence level LSTM for document classification, but its metrics could be improved. Pappagari et al. [5] use BERT along with stacking and LSTM or transformer layers to mitigate BERT's 512 word limit, but it was not experimented with other, resource limited languages.

3.2 Document Classification in Tamil Language

Research on document classification for Tamil language is comparatively low due to its resource limited nature. Rajan et al. [9] use vector space modeling and artificial neural networks, but more documents were given to testing than training, thus hindering the model's ability to learn as it has less number of training examples per class. Reshma et al. [10] use SVM and several other machine learning techniques to classify large documents, but the number of classes is very small (three classes), and metrics could change if the number of classes is increased. Sanjanasri et al. [11] use random kitchen sink instead of SVM kernels, but the average number of words per class is around 150 words.

4 Proposed System

For RoBERT, instead of using BERT base model, we have used multilinugal BERT (mBERT) model which allows us to train for Tamil language. Each document is split into chunks of 200 words with an overlap of 50 words. Then WordPiece tokenization is used on each chunk of the document. We then used it as an input for our model.

Table 2 Summary of the processed dataset

Category in Tamil	Category in English	Documents in training	Words in training	Documents in testing	Words in testing
இந்தியா	India	3542	1,043,366	900	262,447
உலகம்	World	790	213,849	188	52,657
குற்றம்	Crime	2341	615,253	561	150,589
தமிழகம்	Tamil Nadu	3972	1,131,599	1027	294,945
தலையங்கம்	Editorial	1210	275,335	297	67,353
விளையாட்டு	Sports	1509	430,511	368	106,043
	Total	13,364	3,709,913	3341	934,064

4.1 Dataset

We used a dataset which contains articles published on Tamil news website tamilmurasu (tamilmurasu.org) from January 6, 2011, upto January 6, 2020. The data is available online [12].

We removed any articles which contains less than 200 words and picked six of the most common classes which are as follows: தமிழகம் (Tamil Nadu), இந்தியா (India), உலகம் (World), குற்றம் (Crime), தலையங்கம் (Editorial) and விளையாட்டு (Sports) Since தமிழகம் (Tamil Nadu) has over three times the number of articles when compared to the next most common category (13,633 articles), and we decided to remove some random amount of articles pertaining to that category. The summary is given in Table 2. Twenty percent of the entire dataset is used for testing, and the rest eighty percent of the dataset is used for training. Twenty percent of training dataset is used for validation which is used to tune hyper parameters.

4.2 Model Architecture

We are using RoBERT or recurrence over BERT as a base model, thus TRoBERT (Tamil - Recurrence over BERT). The main difference is that instead of using BERT, we are using mBERT which allows us to use it for Tamil. Let $D = d_1, d_2, d_3, ..., d_n$ be a Tamil document, where d is a token in D and n is the number of tokens. k split with l overlap is done on the document which is described in Eq. (1).

$$D_{\text{split}} = \begin{bmatrix} d_1 & d_2 & d_3 & \dots & d_k \\ d_{k-l} & d_{k-l+1} & d_{k-l+2} & \dots & d_{2k-l} \\ d_{2k-2l} & d_{2k-2l+1} & d_{2k-2l+2} & \dots & d_{3k-2l} \\ \dots & \dots & \dots & \dots & \dots \\ d_{n-l} & d_{n-l+1} & d_{n-l+2} & \dots & d_n \end{bmatrix} \tag{1}$$

where k is the split amount and l is overlap amount. However, the document will only split without leftover only if the following equation described in Equation (2) is satisfied.

$$\frac{n + (a - 1) * l}{ak} = 1 \tag{2}$$

where $a \in N$ is a constant. If this equation is not satisfied, some padding tokens could be used until the matrix is filled up. An example is given below:

ஐபிஎல் 5வது சீசன் போட்டியில் சூதாட்டத்தில் ஈடுபட்ட 5 வீரர்களை தனியார் சேனல் ஒன்று சிக்க வைத்தது .

if we assume k to be 5 l to be 2, then the following sentence will be split which is given below.

ஐபிஎல் 5வது சீசன் போட்டியில் சூதாட்டத்தில்
போட்டியில் சூதாட்டத்தில் ஈடுபட்ட 5 வீரர்களை
5 வீரர்களை தனியார் சேனல் ஒன்று
சேனல் ஒன்று சிக்க வைத்தது [PAD].

Although in practice, we would tokenize then split because BERT uses subword tokenization.

Each row of the matrix D_{split} is then used independently as the input to the BERT$_{\text{base}}$ model. The pooled output for one row in BERT$_{\text{base}}$ is then stacked with other rows and then is used as an input to a small LSTM layer. The output of LSTM layer is then used in a small linear layer with softmax activation function, and output of this is used for classifying a Tamil document D. The high-level architecture is given in Fig. 1.

5 Model Training

We then tuned for the number of epochs and learning rate. Due to resource limitations, we were only able to set batch size of about 8. We found that the optimal learning rate is around $2 * 10^{-5}$. Also training was done for 4 Epochs. After training, we then use testing dataset to compare our model with two other commonly used machine learning algorithms—logistic regression, support vector machines with both polynomial kernel and RBF kernel, and random forest.

Term frequency-inverse domain frequency (TF-IDF) is used as a pre-processing technique for machine learning algorithms. TF-IDF is a vectorization method which

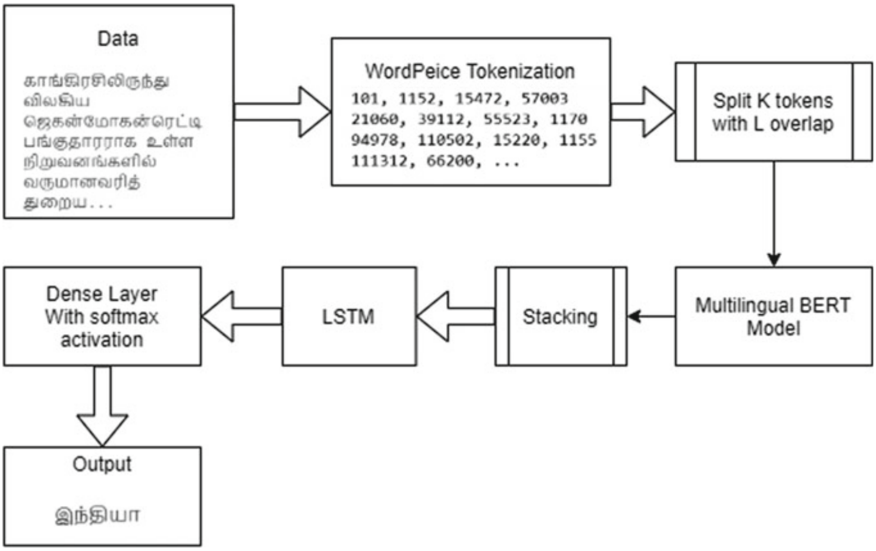


Fig. 1 High-level architecture diagram

Table 3 $F1$ scores of TRoBERT, compared with other machine learning algorithms

Class	Logistic Reg.	SVM (RBF)	SVM (poly)	Random forest	TRoBERT
இந்தியா	0.82	0.84	0.84	0.74	**0.90**
உலகம்	0.62	0.67	0.66	0.03	**0.88**
குற்றம்	0.77	0.79	0.78	0.72	**0.80**
தமிழகம்	0.79	0.81	0.81	0.72	**0.86**
தலையங்கம்	0.73	0.78	0.72	0.36	**0.92**
விளையாட்டு	**0.97**	**0.97**	0.96	0.95	**0.97**
Average (weighted)	0.80	0.82	0.81	0.68	**0.88**

Bold indicates highest $F1$ score in a class

vectorizes a document based on the frequency of the words. TF-IDF is calculated using a vector multiplication of term frequency, a measure of the frequency of words in a document, and inverse document frequency which is used to scale down the most frequent words as they provide very little to the document's meaning. The $F1$ score for each algorithm for each class is given in Table 3.

6 Experimental Result Analysis

$F1$ scores of each category are given in Table 3. Our model was able to outperform substantially when compared to other machine learning algorithms like logistic regression, SVM, and random forest. For SVM, we have used both radial basis function (RBF) and polynomial (Poly) kernels. For polynomial kernel, we have used degree of 2. $F1$ score of விளையாட்டு (sports) is about equal for all algorithms ($F1$ score of about 0.97). This could be because of words like “cricket” and “stadium” which are usually used in sports journalism which are not widely used in other categories, thus allowing many of the algorithms to perform well. TRoBERT was able to outperform significantly in உலகம் (world) category with an $F1$ score of 0.88 when compared to logistic regression ($F1$ score of 0.62), SVM with RBF kernel ($F1$ score of 0.67), SVM with polynomial kernel ($F1$ score of 0.66), and random forest ($F1$ score of 0.03). Overall the worst performing algorithm was random forest with $F1$ score of 0.68.

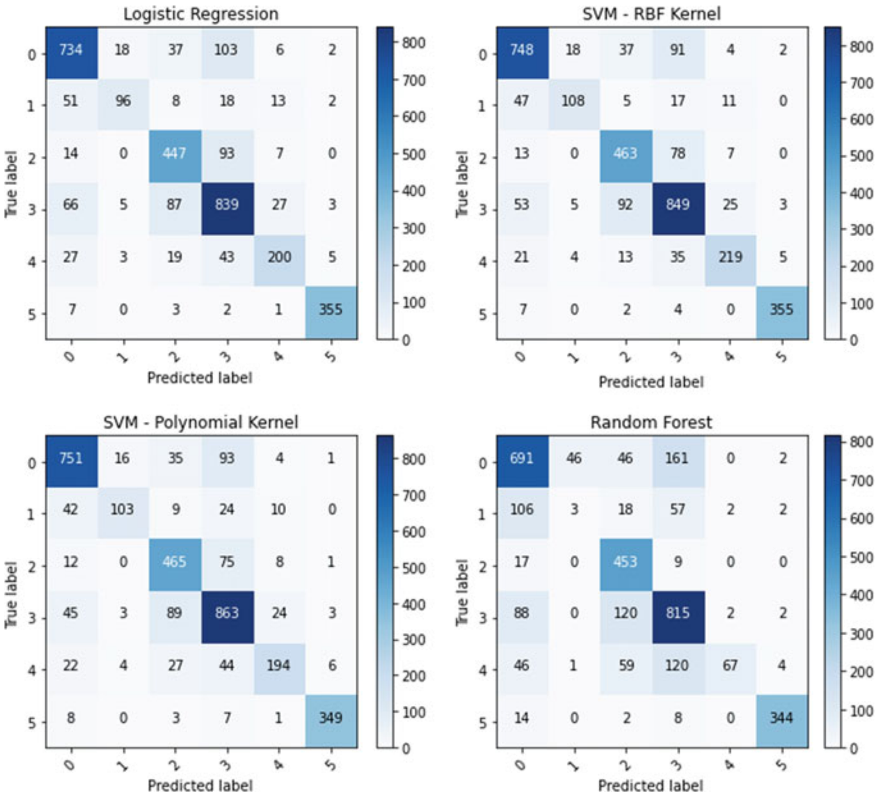


Fig. 2 Confusion matrices for all machine learning algorithms. Top left: logistic regression. Top right: SVM with RBF kernel. Bottom left: SVM with polynomial kernel. Bottom right: random forest

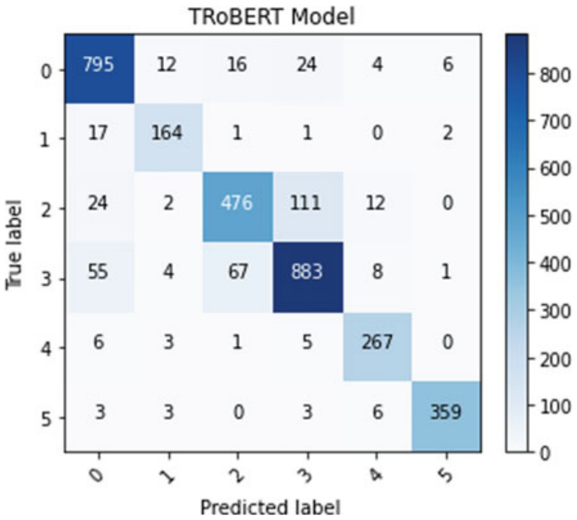


Fig. 3 Confusion matrix for TRoBERT model

Confusion matrix of each of the models is given in both Figs. 2 and 3. From the confusion matrix, we can see that all models were having hard time classifying between class குற்றம் (crime) and தமிழகம் (Tamil Nadu) on all models. Also random forest was completely unable to classify உலகம் (world) class, it was only able to correctly classify 3 out of 188 in the test dataset, although this could change if more data is given. Some of the documents from தமிழகம் (Tamil Nadu) were wrongly classified as இந்தியா (India) in our model, but it was less common in both SVM-based models but it was not able to properly differentiate between குற்றம் (crime) and தமிழகம் (Tamil Nadu).

7 Conclusion and Future Work

We have shown that hierarchical architecture on multilingual BERT has substantially increased $F1$ scores for all classes when compared to standard machine learning algorithms for Tamil language with an weighted average score of 0.88. TRoBERT uses hierarchical transformer architecture which uses stacking and LSTM and combines with mBERT. Stacking and LSTM allow us to implement transfer learning for large documents, without pruning or pooling, so the entire document can be used for document classification, thus allowing the model to gain more context in a particular document.

One of the main disadvantages of this approach is the increase in inference time. But this could be alleviated by using smaller models like DistillBERT [13], or a smaller BERT model [14]. We could improve the metrics by using pre-trained models specifically trained on Tamil language or using models which is trained on fewer languages like IndicBERT [15], which was only trained specifically for Indian Languages. We could also experiment with gated recurrent unit (GRU) instead of LSTM to reduce the number of parameters.

References

1. Comrie, B.: Languages of the World, chap. 2, pp. 21–38. John Wiley & Sons, Ltd. (2017)
2. Daniels, P.T.: Writing Systems, chap. 5, pp. 75–94. John Wiley & Sons, Ltd. (2017)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4171–4186 (2019)
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
5. Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., Dehak, N.: Hierarchical transformers for long document classification. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 838–844. IEEE (2019)
6. Adhikari, A., Ram, A., Tang, R., Lin, J.: Docbert: Bert for document classification. arXiv:1904.08398 (2019)
7. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)
8. Rao, G., Huang, W., Feng, Z., Cong, Q.: Lstm with sentence representations for document-level sentiment classification. Neurocomputing **308**, 49–57 (2018)
9. Rajan, K., Ramalingam, V., Ganesan, M., Palanivel, S., Palaniappan, B.: Automatic classification of tamil documents using vector space model and artificial neural network. Expert Syst. Appl. **36**(8), 10914–10918 (2009)
10. Reshma, U., Barathi Ganesh, H., Anand Kumar, M., Soman, K.: Supervised methods for domain classification of tamil documents. ARPN J. Eng. Appl. Sci. **10**(8), 3702–3707 (2015)
11. Sanjanasri, J., et al.: A computational framework for tamil document classification using random kitchen sink. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1571–1577. IEEE (2015)
12. vijayabhaskar, J.: Tamil news classification dataset (tamilmurasu) (Jan 2020). https://www.kaggle.com/vijayabhaskar96/tamil-news-classification-dataset-tamilmurasu
13. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108, http://arxiv.org/abs/1910.01108 (2019)
14. Turc, I., Chang, M.W., Lee, K., Toutanova, K.: Well-read students learn better: on the importance of pre-training compact models. arXiv:1908.08962 (2019)
15. Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M.M., Kumar, P.: IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4948–4961. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.445, https://www.aclweb.org/anthology/2020.findings-emnlp.445

Analysis of Hybrid MAC Protocols in Vehicular Ad Hoc Networks (VANET) for QoS Sensitive IoT Applications

Nadine Hasan, Arun Kumar Ray, and Ayaskanta Mishra

Abstract Vehicular Ad Hoc Networks (VANET) represent the base for future Intelligent Transportation system (ITS). The technology proposed to face different traffic problems related either to human safety, general safety, or Internet-related services. MAC protocols and routing protocols were developed to play a vital role in VANET communication. MAC protocols represent the time organizer to utilize the channel efficiently. This paper introduces the concept of TDMA efficiency working under hybrid protocols. It discusses recent hybrid MAC schemes and their performance in different proposed scenarios as well as in the term of TDMA frame efficiency. Efficiency of the frame serves in the enhancement and incrementing the bits in one TDMA frame. Where the efficiency is the amount of useful data to be transmitted within an allocated time can be expressed as the efficiency of access method utilized in MAC layer. Efficiency of TDMA frame represents the ratio of useful information bits to the total bits in a frame. This paper discusses hybrid MAC protocols with respect to QoS parameters. It also provides a comparison between the efficiency of hybrid MAC protocols.

Keywords VANET · QoS · MAC · TDMA · CSMA · Hybrid protocols · Efficiency

1 Introduction

Ad hoc networks as an emerged communication technology are considered as the future of the information networks. Mobile Ad Hoc Networks (MANET), VANET, Flying Ad Hoc Networks (FANET), and Airborn Ad Hoc Networks (AANET) are the main classes of ad hoc networks. Second and third layers of the TCP/IP stack play the main role in network performance. In the context of FANET [1] has proposed a hybrid secure routing protocol inspired from bee colony optimization. The proposed

N. Hasan · A. K. Ray · A. Mishra (✉)
School of Electronics Engineering, Kalinga Institute of Industrial Technology, Deemed to be University, Bhubaneswar 751024, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_56

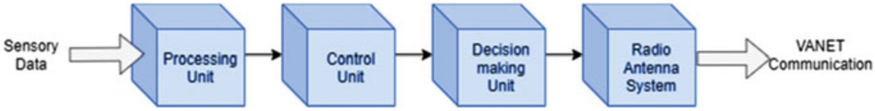


Fig. 1 Functionality of on-board unit (OBU)

protocol provided enhancement in QoS aspects of FANET. From the aspect of non-flying ad hoc networks, VANET is a basic part of ITS [2] as an infrastructure of smart cities campus. It is a subclass of MANET with expanding sensory networks in the nodes (vehicles). VANET trend is to integrate with different technologies such as 5G and millimeter waves along with the developing communication domain. Enabling VANET communication in different directions is the research concern to provide high, reliable, and safety-related services. MAC protocols development represents an important aspect in designing VANET.

Communication in VANET was parameterized under vehicular to pedestrian (V2P) and vehicular to vehicular communication (V2V). These types of communication are classified as internal VANET communication class. In external communications, VANET implements fixed access points along the track of vehicles, Road Side Units (RSU), which must be distributed in a way that provides communication along the trajectory. Each vehicle is equipped with inner communication system under the name of On-Board Unit (OBU). OBU represents the adapter of the sensory data from a variety of sensors in the vehicle to the radio of the vehicle. Figure 1 describes the functionality of OBU.

Along with the IEEE802.11p standard for VANET, VANET is expanding to utilize different wireless technologies like LTE and Millimeter-wave technologies [3]. This comes as two main issues raised in IEEE802.11p and WAVE standard. In this design, the channel access time is utilized equally between one control channel (CCH) for safety-related messages and six service channels (SCH) for non-safety messages. This results in wasting the CCH interval and the bandwidth when no control data to be disseminated, and utilization of constraint bandwidth when increasing network size. An utilization of hybrid protocols with different schemes was proposed in an attempt to overcome such problems.

QoS is the key parameter to evaluate network performance. Ad hoc networks require scalability, fast-tracking of topology changes along with efficient wireless access to achieve better QoS. Differentiating services in VANET comes as an aspect to be considered when talking about QoS especially when it comes to safety application. VANET provides two main types of services. Safety-related services which are the main concern of VANET and non-safety-related services. In this paper, we present the QoS concept in hybrid MAC protocols, how hybrid protocols consider QoS aspects when designing, and introduce the concept of frame efficiency.

Rest of the paper is organized as follows. Section 2 provides the related works. Section 3 gives a brief knowledge of DSRC standard. Section 4 discusses MAC protocols for VANET. Section 5 presents a comparative study between different hybrid protocols related to QoS parameters and the efficiency of the TDMA frame in the hybrid protocol with discussions. Paper is concluded with future scope is Sect. 6.

2 Related Work

Self-control vehicles (Autonomous vehicles) are the future of transportation system. VANET was introduced to be the cornerstone of ITS. Hybrid infrastructure (centralized and decentralized) is the main character of VANET. Vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) are main blocks of VANET. Vehicle to everything (V2X) communication has extended the idea of vehicular communication to be included in the world of IoT. V2X redistributes VANET connectivity to other network components. RSU plays the central component connecting internal and external VANET communication. The implementation of safety and non-safety application increases the functionality of VANET. In the mean of different situations, QoS must be supported as a key parameter to evaluate the network [4]. QoS in a general concept defines the degree of fairness of services provided to the users under their satisfaction requirement. In VANET domain, it represents fast and high dissemination of a dedicated service in critical situations (large network sizes, and constraint resources). Protocols from different layers, non-fixed network size, high mobility, and infrastructure less network all are factors that affect the degree of services. At the MAC layer, CSMA/CA is proposed to work in DCF or PCF models in the IEEE802.11p standard. Meanwhile, TDMA is based on the organized allocation of the users' data in more organized mode compared with CSMA/CA. Therefore, it is mostly used for cluster protocols utilized in VANET.

VANET demands fast and reliable dissemination of data information regarding safety-related services. Implementation of IEEE 802.11p with PCF (Point Coordination Function) in the context of the V2X network was introduced by Mishra et al. [5]. The work has resulted in ensuring better performance of IEEE802.11p in the count of IEEE 802.11a regarding different routing protocols. MAC, IEEE802.11p has given better performance in PDR, throughput, and less delay compared with IEEE 802.11a in the proposed design for both V2X and data unit specified for it [6]. Proposed a novel adaptive hybrid MAC protocol based on dynamic splitting of the data frame. One portion of the frame will be accessed by CSMA/CA and the remaining part will be used by STDMA. The portion percentage depended on channel access delay and packet drop probability. The authors carried out their scheme by varying the usage percentage of CSMA/CA from zero to 100%(complete access interval). This hybrid model provided a satisfying performance in the domain of delay and packet successful ratio. For improving system throughput, Bilstrup et al. [7] outlined the flexibility of the MAC protocols used in VANET (STDMA and CSMA). Through the analytical study and the simulation study, slot allocation was based on vehicle position which has resulted in better performance compared with contention-based access scheme to prevent concurrent transmission between nodes.

The scheduling of STDMA supports higher network load and minimized interference effects [8]. It proposed an effective model to adjust the data frame in hybrid mac protocols (Effective Broadcast Frame Algorithm EBFA). The model was built on selecting a relying node as a header node. The header node will redistribute the time slots between neighbor nodes based on a set of rules. This algorithm has increased PDR compared with other suggested algorithms Dedicated Multi-Channel MAC (DM-MAC) and Hybrid Efficient and Reliable MAC (HER_MAC). We notice that most works focus on calibrating the performance of MAC schemes in terms of some QoS aspects. Efficiency of the TDMA frame did not take that attention. This aspect can be used to evaluate the hybrid protocol since efficiency can be defined as the number of useful bits in a frame. In this paper, we introduce the concept of TDMA frame efficiency and compare hybrid MAC schemes in terms of its efficiency.

3 DSRC—A Brief Overview

OBU and RSU are the basics of VANET. Each autonomous vehicle contains fixed communication subsystem called On-Board Unit (OBU). OBU is a central processing unit to control the sensory data. LiDAR sensors, cameras, GPS subsystems and testamentary subsystems feed the data to the OBU. OBU establishes either V2V or V2R communication. RSU performs the bridge between V2V or V2I communication. RSU links VANET components with the Internet. Dedicated Short Range Communication (DSRC) is proposed to adapt the heterogeneous communication in VANET. (DSRC) indicates the communication range for VANET communication. This can be considered as limitation compared with cellular systems. Specification of DSRC standards is provided in Table 1.

Federal Communication Commission (FCC) has dedicated the spectrum of 5.9 GHz to be used for VANET through a 75 MHz band in the band of 5.850–5.925 GHz. This band is divided into 7 channels of 10 MHz with 5 MHz guard

Table 1 DSRC specifications

Parameter	DSRC
Spectrum	5.9 GHz
No. of channels	7
Bandwidth	10 Mbps
Bit rate	3–27 Mbps
Modulation	OFDM
Antenna	Omni-directional
Transmission range	<300 m
Standard	802.11p
Cost	Cheap

band. Supporting variety of safety-related messages using one channel (CCH). Infotainment and Internet-related applications are supported by six service channels. Each service type utilizes the wireless channel equally (50% of the channel access time). The transmission range is dedicated to support both V2V and V2I communication. This range is small compared with conventional mobile communication. MAC IEEE802.11p is the base for channel access in VANET communication. IEEE 802.11p supports the hybrid structure of VANET. Also the complete architecture of DSRC communication is cheap compared with millimeter wave and LTE communication. The stack of DSRC along with WAVE standard and BSM structure are shown in Fig. 2.

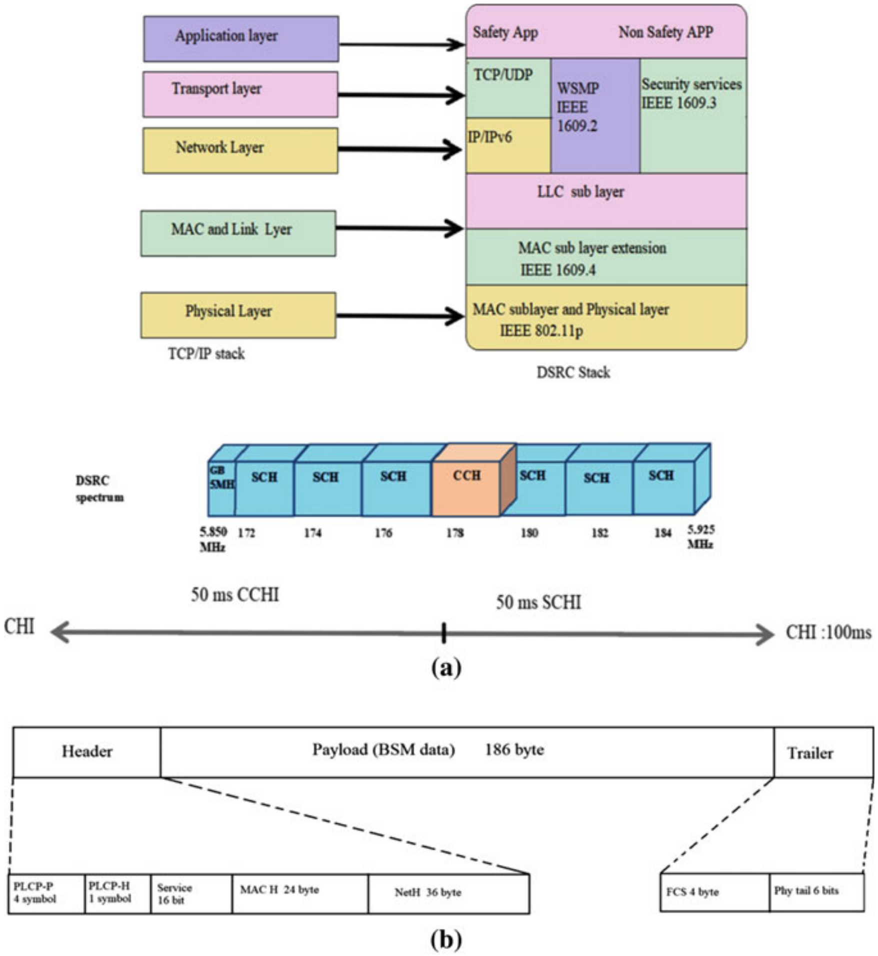


Fig. 2 **a** DSRC stack with channel assignment and intervals, **b** BSM structure

Along with the channel assignment distribution, DSRC proposed a modified TCP/IP stack which include the IEEE1609.x standard to provide a new variety of services including Wireless Access utilization in Vehicular environments (WAVE) [9]. DSRC stack provides flexibility to cope with the dynamic parameters of VANET. However, MAC layer mission is still dedicated to track the topology changes and achieve connectivity within a minimum delay. For these reasons, many researchers have proposed different working mechanisms in the context of MAC protocols. Discussion of MAC protocols is provided in the next section.

4 MAC Protocols for VANET

In the domain of QoS requirements, MAC protocols play a vital role in manipulating several network parameters. Delay, throughput, packet delivery, manage the limited resources such as bandwidth and time, controlling the overhead is affected by the MAC layer protocol utilized. These requirements must serve the priority of messages from the upper layers. Messages in VANET are categorized under three main types, periodic messages, event-driven messages, and data messages which are (non-safety messages). The priority of these messages varies according to their importance in VANET where safety messages employ priority of the first level. Dynamic network size creates issues related to broadcast and hidden terminals in addition to increasing contention and congestion. MAC protocols are supposed to minimize or avoid these issues. Therefore different MAC models were proposed [10]. Three main categories of MAC protocols work in VANET. MAC protocols based on fair access, MAC protocols based on random access, and hybrid protocols. We will discuss these protocols with some examples. We have also considered few of the predominant variants of above families of MAC based on their performance for QoS sensitive application framework.

4.1 MAC Protocols Based on Fair Access

These protocols utilize the access time of the channel by portioning channel access into multiple time frames. Each frame is divided into time slots. Each time slot will be assigned to one user (vehicle). Time division multiple access (TDMA) is the fair access MAC protocol. This scheme is based on transmitting different signals in the same frequency channel. This model reduces the interference and prevents contention by providing fairness in time utilization between nodes though time slots may not be utilized efficiently in the existence of low network loads. Utilizing fair access protocols supposes that VANET working under this type of protocol is requested to maintain synchronization by GPS subsystem.

4.1.1 Vehicular Self-organized MAC (VeSOMAC)

A TDMA-based MAC protocol. This protocol utilizes the geographical location to assign time slots to each vehicle. It enables slot reuse within a non-interference range of nodes. In these protocols, each node will be assigned into a time frame to select a free time slot to transmit its data information. Depending on the position and the track, vehicles pretend to achieve less delay with sequential occupation of the time slots. When the node doesn't find any free time slot, it utilizes an occupied time slot used by a node far away from this transmitting node. Concurrent transmissions caused by same back-off time will be reduced in this process [11]. Non-centralized time slot allocation is one advantage of this protocol, while external vehicle communication (V2I) is not addressed in this protocol.

4.1.2 Cluster-Based MAC Protocols

MAC protocols utilizing cluster method work on portioning the network into multiple clusters. This method is utilized to preserve both energy and lifetime of the network. Any cluster is consisted of cluster head and cluster members. Cluster head functions as a helper for assigning the time slots to the cluster members within a frame interval. Cluster formation determines cluster members and selecting the cluster head. In cluster-based protocols, the helper node is the node selected according to some parameters such as speed, location, ID, node connectivity, and traffic flow. Vehicles assign their data between the time slots in a fair base. Cluster head functionality is to manage different communication in the cluster [12]. Provides a comparative study between different clusters-based MAC protocols with respect to different parameters such as energy consumption and bandwidth utilization. TDMA cluster-based MAC for VANET (TC-MAC) is proposed in [13]. It provides enhancement in network performance compared with DSRC utilization.

4.2 MAC Protocols Based on Random Access

Sensing the channel randomly for data transmission is the basis of these protocols. Vehicles utilize CSMA/CA to avoid contention. VANET differs from MANET in that the MAC model has to adapt with the priority of the messages. This fact supposes that safety-related messages don't require any ACK mechanism. The random access ensures the packet is at the wireless medium only. Consequently, no retransmission takes place in the case of safety messages. An example of random access protocols is CSMA with priority and polling.

CSMA with priority and polling (PP_CSMA). AS the name suggests this protocol works based on two factors prioritizing the messages (P) and the distance to the closest receiving vehicle (D). Based on mapping between these two factors (P&D), the packet will be transmitted. Short distance packets will have higher priority. Each

Table 2 CSMA versus TDMA

Methodology	Random access	Fair access
1	CSMA	TDMA
2	Contention-based	Contention-free
3	Low transmission reliability	High transmission reliability
4	Low packet arrival rate	increase overhead due to synchronization
5	Small slot length	Longer time slot length
6	No synchronization	Depends on GPS for synchronization

node adjusts the back-off time according to the importance of the messages. It is important to notice that emergency messages have a higher priority in all the schemes. Table 2 compares MAC protocols based on random and fair access.

4.3 Hybrid Protocols

Hybrid protocols are based on a scheduling scheme between two access methodologies. CSMA provides contention-free access, but collision may occur when two nodes pick the same back-off time when having same contention window size. Contention reduces network throughput and causes more packets loss. Meanwhile, TDMA utilization results in wasting time slots when nodes are in sleep or idle mode and network size is small enough to utilize all the slots in the frame. To improve conventional MAC protocols performance, hybrid protocols were proposed. Hybrid protocols combine the advantages of two schemes by enhancing their advantages and minimizing disadvantages to improve the system performance. Utilizing hybrid protocols in different scenarios has proven its efficiency compared with conventional MAC protocols.

4.3.1 Multi-Priority Time Division—CSMA (MPTD-CSMA)

To utilize the advantages of both CSMA and TDMA, an algorithm combines these two schemes. The algorithm is based on the following aspects. First, considering the vehicles are in the same direction. Second, the road is combined of clusters. Third, each cluster is of a radius equal to the diameter of the transmission range (R). Fourth, each area contains multiple segments assigned with a particular frequency. These segments are repeated in the adjacent areas. This algorithm [14] proposed MPTD-CSMA scheme with a superframe slot of $T_s = 13$ µs (as determined in IEEE802.11p standard). One-time slot is dedicated for the preamble (head) and one-time slot for the tail.

TD-CSMA is based on the event schedule in both source and destination. At the source, if the number of a packet transmitted is less than the threshold, then the access is done based on CSMA. Otherwise, the TDMA mode is utilized. With CSMA, the throughput will decrease when the arrival rate increases above a threshold (determined to be 2). This threshold is used further to determine the type of access method at the receiver side. Due to this result, at the receiver side, a scheduling mechanism is also used in the context of TD-CSMA. The number of information packets is first determined to be equal or less than the threshold (is equal to 2), then the 1-CSMA mode is used. Otherwise, TDMA is used. To design MPTD-CSMA, the authors utilized priority levels to access the channels. This implementation reduces the probability of channel collision.

The priority levels are distributed among the nodes. Nodes with higher order have higher priority access level to access different channels. This number of accessible channels is equal to the node order. So we can conclude the working principle of the protocol as the following:

- Determine the arrival rate of the information packets using Poisson distribution.
- Determine the number of packets arrived based on Poisson distribution.
- Select the access methodology (TDMA or CSMA) which is based on the ratio of the arrived packets.
- Allocating the data into the time slots in TDMA is done according to the threshold of arrival rate, and the priority level.
- At the receiver, the arrival rate of the received packets determines the access scheme for further processing.
- Same access scheme will be used according to the arrival rate of the packets.

4.3.2 QoS Aware Centralized Hybrid Protocol (QCH-MAC)

The frame is divided into two periods, Reservation Period and Transmission Period (RP and TP). No time slots are in RP. While time slots are dividing the TP. The reservation period is only used by the new node (Vehicle). Two groups of access categories are used in this protocol. The first access group contains the two higher access priorities (AC3 && AC2), and the second group is for the other two access categories (AC1 && AC2). Safety messages have high priorities in this group, while non-safety messages are assigned in the second access group. MAC parameters are specified for each group by CW, back-off time, and schedulers. The RP is utilized by HELLO messages from RSU and RESREQ from the vehicle. This operation will make VANET based on the clusters mechanism, where each RSU will take the responsibility to assign the time slots to the vehicles in the area. Then transmission process will take a place within time slots allocated to each vehicle. At each new BSS/RSU entry, a vehicle will send two reservation messages safety and non-safety RESREQ. This process is repeated at each new RSU (BSS) entry. Once the RSU receives the RESREQ, it will allocate a time slot based on the priority of the RESREQ. Whenever the high priority reservation request arrives at RSU, it is mandatory to assign a time slot for this REREQ. In this process, the busy time slots were divided into two statuses.

- The first access group occupies the busy frame completely. In this case, RSU will free a time slot of the oldest vehicle and assign it to the new RESREQ.
- Both access groups' data are utilizing the frame. In this case, RSU releases a time slot of non-safety data and assigns it to the new request. In both statuses, the RSU send two types of messages, RESREPLY to the requesting node, and cancel reservation (CANCRES) to the node owning the released slot [15].

4.3.3 BH-MAC: Bitmap-Based Hybrid MAC Protocol

This protocol utilizes TDMA and CSMA with modification on the BSM packet to be sent on TDMA. Each time frame is divided into N synchronization intervals (SI). SI is subdivided into N time slots (N), and each time slot is allocated to a vehicle for BSM transmission. Slot size depends on the time required for the packet to be transmitted. Depending on Markov Chain, the back-off interval will change its values when sensing the business of channel in the SIFS.

A modification is done on the content of the TDMA packet. Error Bitmap (EB), Reserved Slot (RS), and TDMA Slot Allocation Bitmap (TSAB) are added to the BSM packet. TDMA slot allocation bit map (TSAB) determines the reserved slots from 1-hop neighbors TSAB information. The length of TSAB is N. The content of TSAB is updated periodically with the free slots of the old transmitting vehicles. EB is also N bits each bit represents a corresponding time slot. The value of 1 in the bit refers to collision occurrence in the corresponding time slot. This case means the two transmitters must go for new SIFS to select a new time slot. CSMA is utilized only to reserve a time slot after determining the free one from the 2-hops TSAB. After selecting the time slot, the transmission process will take place [16].

5 Efficiency and Comparative Study

The type of service and the network parameters represent factors that affect QoS. The dynamic topology of VANET represents the first obstacle for safety message dissemination. Safety-related messages aim to reduce traffic issues and accidents. Intelligent transportation systems are based on extending this new automotive technology for the future of smart cities. Minimizing dissemination delay, increasing the PDR of the safety-related messages are goals to be considered for QoS. From an aspect of increasing the amount of useful information in a time unit can be considered when come to network performance. This concept is called frame efficiency. In this section, we provide frame efficiency in hybrid protocols to determine the percentage of useful information in a TDMA frame when works with another access method.

Frame efficiency is the ratio of the data bits transmitted to the total no of bits in the frame. We can represent it as following [17].

Table 3 Efficiency of hybrid protocols

Protocol	BH-MAC	EFBA	(DSS-MAC)	MPTD-CDMA
Frame efficiency (Ȩ)	69%	71%	97%	90%

$$\text{Efficiency}(\%) = \left(\frac{N_b}{F_b}\right) \times 100 = \left(\frac{F_b - \text{OH}_b}{F_b}\right) \times 100 \qquad (1)$$

where

N_b is number of information bits in the TDMA,

F_b is total number of bits in the frame,

OH_b represents the number of bits used for synchronization, tail, and headers data.

Comparison of different hybrid protocols with respect to their efficiency is given in Table 3.

We take into consideration the following aspects while calculating the efficiency of each frame in MAC hybrid protocol.

- The packet size is shown in Fig. 2a unless modified according to the protocol requirements.
- The standard frame duration is $T_{\text{f}} = 1$ s unless the protocol requirements and simulation design use another value.
- Time slot is $T_{\text{s}} = 13$ µs unless the protocol design has defined another time interval.
- The number of time slots in the frame is

$$\text{No. of time slots}(N_s) = \frac{T_f}{T_s} \qquad (2)$$

- Calculating the efficiency is done for safety-related messages (BSM) and based on the parameters defined in each MAC protocol scheme.
- When both CSMA and TDMA were used, we calculated the efficiency concerning TDMA using the previous formula. Figure 3 shows hybrid protocols efficiency.

Figure 3 shows that DSS_MAC has the highest efficiency. The utilization of the time slots in DSS_MAC provides high ratio of useful information to be carried within a frame. While BH-MAC provides less efficiency compared with its counterpart. This is due to the use of more control data in the frame. In BH-MAC three types of control data were added to the BSM packet that are Error Bitmap (EB), Reserved slot (RS), and TDMA Slot Allocation Bitmap (TSAB).

From other QoS parameters, one or two parameters took the concern when designing any hybrid scheme. Each scheme provides its privilege regarding delay, PDR or throughput. Throughput has taken more attention in hybrid MAC protocols compared with PDR and Delay. DSS_MAC and QCH-MAC have provided enhancement in delay reduction and PDR comparing with the two conventional

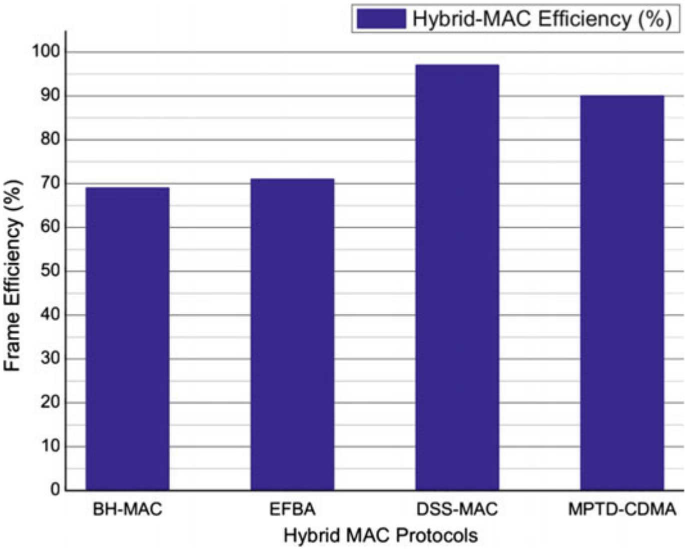


Fig. 3 Efficiency of hybrid MAC protocols

MAC schemes TDMA and CSMA. MPTD-CSMA has enhanced the throughput of the system compared with individual MAC scheme CSMA. EFBA outperforms DM-MAC and HER-MAC in the context of Packet Delivery Ratio (R/S). Even though still there is a need to address QoS parameters in Hybrid protocols for futuristic VANET services.

6 Conclusion

Hybrid protocols can optimize both time and bandwidth. Fair access mode along with random access modes combines a strong protocol that supports future VANET. By increasing the useful information in one synchronized transmission interval supports QoS of the proposed network. Taking into consideration any combination of access schemes requires more additional configuration as well as increasing the complexity of the scheme compared with conventional MAC protocols, and also requires high accuracy when designing and utilizing. In MPTD-CSMA, the threshold between CSMA and TDMA must be chosen accurately. This protocol provides good utilization of TDMA along with CSMA in the context of throughput comparing with CSMA. DSS-MAC provides dynamic utilization of both CSMA and STDMA without any threshold between the access models. This can be considered as a reason for providing higher efficiency compared with other models. Still the utilization of this protocol is constrained to packet drop probability and channel access time. Both EFBA and BH-MAC increase the controlling bits in the account of the useful information that is why the efficiency of these protocols is less. Still PDR is enhanced by EFBA as well as throughput in QCH-MAC. But From QoS aspects, QoS parameters such as PDR and delay need to be addressed more in the context of MAC hybrid protocols. Extending the utilization of hybrid protocols for other technologies needs to be addressed as well as for non-safety application. Non-safety applications need to be included for different hybrid protocols. Efficiency of hybrid protocols can be used as a factor to improve network performance in term of both safety and non-safety application under the aim of QoS enhancement. Future work will focus on the cross-layer optimization aspects of Mobile Ad Hoc Networks for QoS sensitive IoT application using fuzzy Logic system.

References

1. Raj, J.S.: A novel hybrid secure routing for flying ad-hoc networks. J. Trends Comput. Sci. Smart Technol. (TCSST) **2**(03), 155–164 (2020)
2. Bestak, R.: Intelligent traffic control device model using ad hoc network. J. Inf. Technol. **1**(02), 68–76 (2019)
3. Kong, L., Khan, M.K., Wu, F., Chin, G., Zeng, P.: Millimeter wave wireless communication for IoT cloud supported autonomous vehicles: overview, design, and challenges. IEEE Commun. Mag. 62–68 (2017). https://doi.org/10.1109/MCOM.2017.1600422CM
4. Rashid, S.A., Audah, l., Hamdi, M.M., Alani, S.: An overview on quality of service and data dissemination in VANET. IEEE Explore 1–5 (2020)
5. Mishra, A., Chaki, S., Sadhuskhan, S., Sinha, S.: A Novel VPDU Based Frame Work for Internet of Vehicles (IoV) and Performance Evaluation of VANET in Urban Scenario, pp. 1–8 (2020)
6. Cao, L., Xu, W., Lin, X., Lin, J.: A CSMA/TDMA dynamic splitting scheme for MAC protocol ın VANET. IEEE Explore 1–6 (2012)
7. Bilstrup, K.S., Uhlemann, E., Strom, E.G.: Scalability Issues of the MAC Methods STDMA and CSMA of IEEE 802.11p When Used in VANETs, IEEE Xplore, Sweden, pp. 1–5 (2010)
8. Nguyen, V.D., Oo, T.Z., Tran, N.H., Hong, C.S.: An efficient and fast broadcast frame adjustment algorithm in VANET. IEEE Commun. Lett. **21**(7), 1589–1592 (2017)
9. Arena, F., Pau, G., Severino, A.: A review of an IEEE 802.11p for intelligent transportation system. J. Sens. Actuators Netw. 1–11 (2020). https://doi.org/10.3390/jsan9020022
10. Dharsandiya, A.N., Patel, R.M.: A review of MAC protocols for vehicular ad hoc networks. In: IEEE WISPNET Conference, pp. 1040–1045 (2016)
11. Cutting Edge Directions, 2nd edn., pp. 1–25. Wiley, New York (2013)
12. Tarunpreet, K., Dilip, K.: TDMA-based MAC protocols for wireless sensor networks: a survey and comparative analysis. IEEE Paper, pp. 1–6 (2016)
13. VANETs (TC-MAC), 978-1-4673-1239-4/12, IEEE Paper (2012). pp. 1–6
14. Ding, H., Lu, X., Wang, L., Bao, L.W., Yang, Z., Liu, Q.: Hardware Implementation and performance analysis of MPTD-CSMA protocol based on field programmable gate array in VANET. IET Commun. J. **14**(16), 2769–2779 (2020)
15. Boulila, N., Hadded, M., Loauiti, A., Saidane, L.: QCH-MAC: a QoS aware centralized hybrid MAC protocols for VANET. In: IEEE 32nd International Conference on Advanced Information Networking and Application, pp. 55–62 (2018). https://doi.org/10.1109/AINA.00021

16. Kumar, S., Kim, H.W.: BH-MAC: an efficient hybrid MAC protocol for vehicular communication. In: 12th International Conference on Communication System and Networks (COMSENTS), pp. 362–370 (2020)
17. Timothy, P., Charles, B., Jeremy, A.: Satellite Communication, 2nd edn., pp. 237–238. Wiley, India (2008)

Programming with Natural Languages: A Survey

Julien Joseph Thomas, Vishnu Suresh, Muhammed Anas, Sayu Sajeev, and K. S. Sunil

Abstract Programming with natural language is a research area that has a wide range of applications including basic programming, robotics, etc. Factors like preserving the meanings, handling the ambiguity, etc. have to be considered while converting a natural language text to programming language statements. Many developments have been taken place in this area over the past few years. Different types of CFG parsers were used initially for converting the natural language texts to programming language statements. The developments in technologies based on AI have a huge impact in this area and efficient models like GPT-3 are created. These models are capable of converting natural language to target programming language more precisely. In this paper, we do a detailed and systematic study of the developments that happened in this area and list some of the relevant research works among them.

Keywords Keywords · NLP · GPT-3 · fuSE · CCG parser · CFG parser · KIF

1 Introduction

Programming languages are used in computer programming to implement algorithms. Thousands of different programming languages have been created, and more are being created every year. Many programming languages are written in an imperative form (i.e., as a sequence of operations to perform) while other languages use the declarative form (i.e., the desired result is specified, not how to achieve it). Our current trend in programming mainly focuses on structured programming techniques where only a professional programmer or anyone with a good programming background can only do the coding. These techniques use a lot of syntax and semantics constraints to express our ideas in a computer-understandable form.

Focusing on current programming strategies, we can infer that the difficulties arise because the system is not understanding the user's language and the user tries

J. J. Thomas · V. Suresh · M. Anas · S. Sajeev · K. S. Sunil (✉)
Government Engineering College Idukki, Idukki, Kerala, India
e-mail: sunilks@gecidukki.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_57

to convey his idea to the system in a language that is suitable and understandable by the computer in which the user is not much comfortable with the machine understandable language. Thus, anyone who wants to write programs has to study the whole twists and turns in the programming language that takes even many weeks to months of training. If we focus on any organization we can see that they train their employees for about 4–6 months before assigning them any programming tasks. We can say that the difficulties we face to train a new programming language to anyone are mainly because of the unavailability of any programming strategy which is of general understandable behavior. From the beginning onwards we are trying to learn computer understandable language and then trying to implement our ideas through these languages. Whenever new programming languages are introduced in the market most of the firms accept those languages if it is good enough. In order to keep up with the updated technologies, here most of the programmers are forced to learn new languages and it takes almost 4–6 months to learn a new language. Do you think it is simple? Do we have any alternatives for this?

When we go through the current paradigm of programming we can see that if we can introduce a programming strategy where we can express our ideas in our spoken language in a structured format and the computer understands it, then it reduces the training/learning time. If we come up with such a programming technology that eventually results in a revolution in the field of software engineering. A lot of research has taken place in the area of programming through natural languages, and thereby Natural Language Processing(NLP) also evolved. Nowadays, there are even Artificial Intelligence(AI) tools that can generate code from English text. It gives us hope that natural language can be used as input to computers in the future. In this paper, we discuss the most prominent research ideas in the area of natural language programming over these years.

This paper mainly focuses on the developments of programming with natural languages. We did a systematic study of natural language programming. We start with the feasibility study of natural language programming in Sect. 2. Handling ambiguity is the main concern here. We discuss the methods and techniques to deal with the ambiguity in Sect. 3. Converting the natural language statements in to programming language statements is the major task in natural language programming. There are different techniques available for this conversion. We give a detailed review and comparison of some of these techniques in Sect. 4. Natural language programming has a wide range of applications including robotics. Section 5 elaborates the usage of natural language programming in robotics and related areas. We conclude the paper with our inferences in Sect. 6.

2 Feasibility of Natural Language Programming

How natural, a natural language programming need to be? Miller et al. [12] focus on the problems and issues in the increasing usage of natural language interfaces. They also focus on the different styles of programming used by different programmers. Here they are expecting numerous obstacles while implementing a system that can interpret natural language. Even though they are expecting difficulties in style, semantics, and world knowledge, they are not measuring the magnitude of such errors. Considering the first, there are only a few options to choose a different style from one another. Hence the difference in programming language style is very less. Concerning semantics, it is difficult to find the extent to which the meaning is dependent. It can be connected with immediate and prior context. The extent to which the knowledge understanding among the people is also limited. Inclusive of the negatives, this study permits anyone who knows English to be a programmer. Here the programmer only needs to be capable of describing the methods to develop computer programs. Although here the results focus mainly on the negative side, the same results will give some hope and expectations about two important things: (1) several constraints are there in implementing a natural language interface, and (2) more features that help in programming can be included while modifying programming languages. In conclusion, they illustrate the difficulties opposing a natural language interface implementation.

Programming in natural language might seem infeasible because it would appear to require profound knowledge and understanding of natural language. Liebernman, et al. [10] in 2005 studied the feasibility of programming in natural language. They say that several developments might now make programming in natural language feasible. The main among them are *improved language technology*. Here they attain the system to understand the natural language by using improved broad coverage language parsers and semantic extraction to attain partial understanding. The second approach is *mixed-initiative dialogs* for meaning disambiguation. They claim that if the user of the system has a clear conversation, it will develop a system that is capable of understanding the natural language. Here if the system finds any difficulty in translation, it guesses the command by supplying some arbitrary English words. The third method is called *programming by example*. Here We can see that they adopt an interactive two-way communication methodology, which is a combination of examples in relation to demonstrations and descriptions in natural language. Often it is easier to demonstrate than to explain using words. The user can command the system, their needs, so that the system can verify what they meant by the command and in the case of an extreme breakdown of the more sophisticated techniques, the user is allowed to type the code. Thus, the user can communicate with the system very easily.

A study of utilizer errors in a natural language is mainly focused by Bruckman et al. [3] in 1999. They had done online interaction with 16 children who issued 35047 commands on MOOSE crossing and analyzed the occurrences of the errors. They found 2970 natural language errors in total. From these, a total of 314 natural language errors were visually examined. In most of those errors, the child was able to redress the situation after putting in some effort. In order from most to least frequent, they include syntax errors, guessing a command name by supplying an arbitrary English word. This work opens up the risks of programming with natural language.

Heidron et al. [7] proposes an idea to develop an automatic programming system that can accept natural language dialogs. NPGS (Naval Postgraduate School) introduced a method for this, an NLP that uses a data structure which is a form of semantic network. It consists of a collection of objects called records, each of which is just a list of attribute-value pairs. These records represent things such as concepts, words, physical entities, and probability distributions. The attributes of a particular record depend upon what it represents. The NLP has some rules called encoding and decoding. Before a dialog, NLP is given a set of about 300 "named" records containing information about some words and concepts relevant to simple queuing problems. So After giving the dialog to the system, it uses the information it obtained from the user to build a subnetwork called the Internal Problem Description (IPD). Here IPD describes the flow of entities and also this system allows the user to ask questions and use these answers are used to make the sentence more understandable. IPD is represented as a graph where nodes represent records. The labels of outgoing edges of a node represent the attribute names. The grammar for the system is embedded in the encoding/decoding rules. Based on this grammar each sentence is converted into statements in a programming language.

The job of a software developer is to analyze or determine customer requirements and implement a program that satisfies them. Many of us have used google translator at least once for translating a language into another language. Recently, Recurrent Neural Networks(RNN) are used for this translation. RNN worked well on this translation by using knowing correct data to train and that raised the question, why can't we create a program from known data?. Ernst et al. [6] proposed an approach to convert English specifications of file system operations into corresponding bash commands. They trained RNN on 5000 [bash, text] that were manually collected from web pages such as bash tutorial and stack overflow. The file system they created consists of more than 200 flags and 17 system utilities, etc. Tellina, the system they introduced produced an accuracy of 69 % and the structure of command was 88 %. Tellina produces the correct result most of the time some time that produces incorrect results. But using Tellina, the programmers spend less time while completing more file system tasks. Sometimes Tellina's output is wrong due to the command line flag that the programmer didn't know and they found that no neural network language technique achieves perfect accuracy.

3 Handling Ambiguities

Ambiguity is a situation in which a word can be understood in many ways. This is one of the major issues to be considered carefully while converting natural language text to programming language statements. Computational semantics deals with the study of how to represent meaning in computer understandable form. The study made by Patrick Blackburn et al. [1] discusses how to represent the meaning of natural language in logic form. This book gives some basic ideas on three inference tasks: *querying, consistency checking,* and *informativity checking*. The logical concepts of satisfiability in a given model for a given formula are represented by these tasks.

We have already seen the scenario of using a predefined format for storing or taking input in natural language. The same base idea is used in Prasad et al. [14]. Here a predefined natural language format is used for taking input and is called *codified knowledge*. Based on these particular formats, the input data is classified into different snippet files. The snippet files are stored based on the nomenclature. Each of these snippet files will be related to a particular interrogatory. These snippet files will be provided as the answer to the interrogatory. The interrogatory includes conditions like having extensions like *.what, .how, .which, .when, .who*, etc. The input may be represented in many different perspectives and views. So the results will be affected by how anything is documented. The output may have many differences even though the inputs are similar. The data flow also has much influence on this approach and each input is verified at each level. Rather than verifying, the system identifies unassigned entities as well, if any.

Based on the authoring tool and conditions for both inputs, output the results may differ. Also whenever data is found missing it will be notified. It means that if the system needs one more data or variable to process a given input, it will be notified to the user. This is identified using the query "What data is need at which point in the process flow?". Accordingly, the output data will be modified and chooses what has to be displayed on the screen. So the system has almost all over control over the input. Since the inputs are almost well managed the outputs will be also under some sort of control. The main advantage is that the middle layers do not have to suffer much. The main security of this system depends on these keywords. This way is expected to be having the most structured and synced way of conversion. This method was expected to be working just like in a search engine. Fortunately, this approach was not mutually exclusive with the current programming standards or approaches. With the existing software, it is possible to co-exist along with this new approach. One of the main limitations of this kind of growth was snippets. It is difficult to include and manage all kinds of data inputs to snippets. Even if that is made possible the mapping will be much more complicated. As we have already seen many times, this approach also ended up with some kind of hope for the future.

4 Natural Language to Programming Language

Nowadays many people are interacting with computers in their daily life. So, as computers have become one part of everyone's life they must be capable of performing more tasks as well. We all know that repetitive or specialized tasks often require the creation of certain programs. Program synthesis using natural language was another work on the same by Desai et al. [5]. The problem here is that the end users are still struggling with the use of Domain Specific Languages(DSL). The main focus of this work is to avoid such problems. The presented idea includes generating a general framework for constructing a program that takes natural language input and produces the expressions in target DSL. The basic idea sounds much similar to that of Dalal et al. [4]. Here each of the natural languages is paired with some targeted

DSL. These kinds of pairs are limited in numbers and are being used for training purposes as well. From the given training pairs a synthesizer is being constructed on the basis of optimal weight and classifiers. The classifiers will rank the outputs of keyword programming based on translation. The frameworks are being applied to three domains: *editing of respective text, intelligent tutoring system* and *queries regarding flight information*. As we have seen earlier the outputs generated were not at all 100%. But the main advantage of this feature was that the output generated over here was given to the user based on its ranking. The ranking system helps here to provide much more accurate results when compared to the previous results.

Even though this approach is being said to be one of the best so far, based on the basic idea used, it reveals some of its inabilities as well. Here it is said that even the data set used is limited, the unknown variables are also managed using temporary variables, qualifiers, and natural language abstraction. This approach will not work well for automata-related problems. This limitation is due to the absence of direct correspondence. The lack of correspondence between natural language and DSL uses nontrivial logical reasoning during the process of conversion and the effectiveness of the system. It also faces some kind of efficiency issues due to the less impact of dictionary size on the runtime corresponding to the input.

Mihalcea et al. [11] proposes a method to convert a natural language to a programming language. They implemented a small version of the system. In procedural programming, a computer program is composed by a developer that contains sequences of action statements that expressing the operations to be performed on various data structures. Similarly, procedural natural language programming is targeting to generate a program from a natural language. Here the system focused on generating skeletons for a computer program that is used as a starting point for creating procedural computer programs. The system first analyzes the sentence and using three components to create skeletons for natural language procedural programming. The first component is *the step finder* which is used to identify natural language text and find action statements that are to be converted into programming statements. For example, if a sentence starts with the natural language text "you should count how many times each number is generated and write these counts on the output screen". So here two main steps should be identified: (1) count how many times each number is generated, and (2) write these counts out to the screen. Here the step finder also identifying all verbs that have chance turned into program functions, such as read, write, count, etc. The next component is *the loop finder*. The role of the loop finder component is to identify repetitive statements from natural language structures. For example, if a sentence starts with each, every, all like that. The third component is *the comment identification* which marks the steps as a comment if that starts with a word like "for instance", or "for example", etc. After finding all these the text is converted into computer program skeletons.

Instead of programming in a high-level language, it is always better to program in a natural language. Vadas et al. [15] clearly describes that an interface using natural language like English makes programming more interesting and creative. It is the best way of writing code too. Here we can see the studies about how people write instructions in English and the methods by which how we can process the natural language. A prototype system that can translate some English instructions into executable python is also introduced in this paper. Using modern parsing techniques such as Combinatory Categorial Grammar(CCG) parser, the system is accepting natural language as input and after the successful transformation of the input through different phases like syntactic phase, semantic phase, and functional identification phase, it finally outputs the code in Python.

Comparing with the older systems, using unrestricted syntax will allow us to the usage of a wide range of syntactic and semantic methods to get results from user's instructions. The natural language which is accepted by their system will be parsed by the CCG parser in the syntactic analysis phase. Then the system translates the output of CCG parser which is a predicate-argument representational structure to a first-order logical representation of Dynamic Reasoning Systems(DRS) predicates. The DRS representation gives us a more general form of the sentence rather than the input words written by the user. In the functional identification stage, the extracted functional verb and its arguments from the semantic analysis stage are mapped onto a function. The simple technique used in the current prototype system has a list of primitives, which consists of a mapping between a specific verb and the corresponding number of arguments. If the primitive matches perfectly with the semantic information extracted, then the equivalent Python code is generated.

5 Natural Language Programming in Robotics

From the early stage of robotics, most of the researchers are trying to communicate with the robots through natural language. The main object was to do tasks by communicating with the robot. At the same time, computer scientists were also trying to use natural language as input to the computer. As the research is going parallel, there was an idea that was proposed by Winograd [17], an American professor who designed a single eye, single-arm robot which can be controlled using instructions given in English in 1971 at the Department of mathematics at MIT. Here the text is passed using basic Context Free Grammar(CFG) parsers using the language LISP. But most of the English sentence possesses some ambiguity problems. They spent most of the time solving the ambiguity problem and the robot was only able to do tasks in a very restricted environment. The solution to this problem robot needs to learn its current environment by itself which was not practical at that time.

Stenmark et al. [13] proposed a robot arm that works on the instructions given in the English text in 2013 at the International Symposium on Robotics (ISR). This paper is the practical implementation of the idea presented by Winograd [17]. They used a Knowledge Integration Framework (KIF). The English text is given as an input to KIF. As the first step, KIF sends it to a natural language parser module within itself. Natural Language parser (NL parser) sends back the semantics of text to KIF. Then the best semantics which is more likely to the input text will be chosen using statistical techniques. The KIF then sends the output to Engineering System (ES) and the ES convert to a program that can perform in the robotic arm. This arm movement is just the starting of the idea, how to perform tasks using a robot with English. Generally speaking, KIF has a role in this arm movement. NL parser is a set of parsers available in it. All this functionality is pre-packed with a software called RobotStudio. Input to the system is the raw English text. The RobotStudio will execute the sentence in a pipeline manner. The NL parser takes about 100 milliseconds for parsing a sentence. KIF is also capturing and storing the input and output in its memory. This speeds up the process when the same text is used again. Since the natural language is used to control the arm it is much user friendly to give commands to the robot. Based on this project, there are more industrial robots designed. Recent industrial robots are capable of capture human movements and acting the same as that. The main drawback of the system is that it cannot use the voice as input.

Artificial Intelligence (AI) is changing our lives daily. In the case of Robotics, AI is the brain and all the robotics companies are researching the developments of AI and machine learning-related areas for a long time. Weigelt et al. [16] proposed an idea in June 2020 at the Association for Computational Linguistics about using the ML model to solve the problem of natural language programming. They proposed a model called Function Synthesis Executor(fuSE), which is capable of extract procedures from the input English sentence. fuSE used to take input from the speech itself. It mainly consists of three sections- *classification of teaching efforts*, *classification of the semantic structure*, and *method synthesis*. In the first stage, they discover utterances from the input English speech. To improve the accuracy the authors used pre-trained networks based on Deep Neural Network(DNN) architecture for the classifier. The pre-trained model used was the Bidirectional Encoder Representations from Transformers (BERT) by Google which gives 97.7% accuracy on the testing data. The second stage which is also a classifier discovers the semantic parts from the output of the first stage. The custom RNN gives more accuracy than BERT at this stage. In the final stage, the output from the second stage is converted to a procedure style that can be run by a high-level language. This procedure contains some API for appropriate functionalities to perform the actions given by means of voice. These are the overall stages and functionalities of fuSE. The F1-score of the fuSE API calls is 79.2%. fuSE is an example of natural language programming, but it is not as accurate to give a better result from the input. The input of the fuSE is the direct voice of the user, and then it extracts the logic into some procedure calls. The procedures are executed in the background and show the output.

fuSE is not only the DNN that can produce the code from natural language but there are some other models that can also do the same task. OpenAI, an AI research laboratory in the USA published an article by Brown et.al. [2] in July 2020 about a DNN, called Generative Pre-trained Transformer 3 (GPT-3) which is an autoregressive language model that can solve most of the problems associated with natural language translation. GPT-3 is the third generation of the GPT series. Microsoft's Turing NLG introduced in February 2020 was the best-known NLP DNN before the release of GPT-3. GPT-3 produces sentences more like a human, where people cannot identify that these are AI generated. GPT-3 has 175 billion ML parameters and this model was modified from its own older versions. GPT-3 is capable of answering general questions on a given sentence. A video that shows the problem-solving capability of GPT-3 was posted on the internet in July 2020 (https://youtu.be/SboKeK6FFHQ). In that video GPT-3, produces Python code from the given English text. OpenAI's GPT-3 is capable of generating codes in CSS, JSX, Python, and some other languages.

6 Conclusion

Natural language programming is used in many areas including robotics. Different models like CFG parsers, CCG parsers, fuSE, GPT-3 are available to convert a natural language text to statements in programming languages. Each model has its advantages and disadvantages. Table 1 summarizes an overview of the break-through results in this area. We did a detailed and systematic study of the developments that happened in this area over the past years. A brief overview of some of the relevant results are given in this paper. Among the ideas discussed, we found that AI-based models are more efficient than others. For example, GPT-3 like models are capable of generating Python code from the English text. The recent developments in this area give us hope that we will be able to communicate with computers using natural language in near future. In that case, we do not have to learn any programming language to write programs and that will be a revolution in software engineering.

Table 1 Milestones in natural language programming

S. No.	Author & Title	Year	Contributions
1	Winograd [17] *Procedures as a representation for data in a computer program for understanding natural language*	1971	Designed a single eye, single-arm robot which can be controlled using instructions given in English. Here the text is passed using basic context free grammar (CFG) parsers using the language LISP. They spent most of the time solving the ambiguity problem and the robot was only able to do tasks in a very restricted environment. The solution to this problem robot needs to learn its current environment by itself which was not practical at that time
2	Prasad et al. [14] *Computer knowledge representation format, system, methods, and applications.*	1971	A predefined natural language format is used for taking input and is called *codified knowledge*. Based on these particular formats, the input data is classified into different snippet files. The snippet files are stored based on the nomenclature. Each of these snippet files will be related to a particular interrogatory

(continued)

Table 1 (continued)

S. No.	Author & Title	Year	Contributions
3	Heidron et al. [7] *Automatic programming through natural language dialog: a survey*	1976	Proposes an idea to develop an automatic programming system that can accept natural language dialogs. They use some encoding and decoding rules for NLP
4	Miller et al. [12] *Natural language programming: styles, strategies, and contrasts.*	1981	Discussed the problems and issues in the increasing usage of natural language programming interfaces
5	Bruckman et al. [3] *Should we leverage natural-language knowledge? An analysis of user errors in a natural-language-style programming language*	1999	They did a study of utilizer errors in a natural language through online interaction with 16 children who issued 35,047 commands on MOOSE crossing and analyzed the occurrences of the errors. They found 2970 natural language errors in total. From these, a total of 314 natural language errors were visually examined. This work open ups the risks of programming with natural language
6	Blackburn et al. [1] *Representation and inference for natural language—a first course in computational semantics.*	2005	Discusses how to represent the meaning natural language in logic form. This book gives some basic ideas on three inference tasks: *querying, consistency checking,* and *informativity checking*
7	Vadas et al. [15] *Programming with unrestricted natural language.*	2005	A prototype system that can translate some English instructions into executable python is also introduced in this paper
8	Liebernman et al. [10] *Feasibility studies for programming in natural language*	2006	They say that several developments like *improved language technology*, *mixed-initiative dialogs* and *programming by example* might now make programming in natural language feasible
9	Mihalcea et al. [11] *NLP (natural language processing) for NLP (natural language programming)*	2006	They implemented a small version of a system to convert a natural language to a programming language. Their procedural natural language programming system is targeting to generate programs from a natural language

(continued)

Table 1 (continued)

S. No.	Author & Title	Year	Contributions
10	Dalal et al. [4] *Computer program product and computer system for language enhanced programming tools.*	2011	Here each of the natural language is paired with some targeted Domain Specific Languages(DSL). From the given training pairs a synthesizer is being constructed on the basis of optimal weight and classifiers. The classifiers will rank the outputs of keyword programming based on translation
11	Stenmark et al. [13] *Natural language programming of industrial robots*	2013	Proposed a robot arm that works on the instructions given in the English text which is a practical implementation of Winograd [17]. They used a Knowledge Integration Framework (KIF)
12	Li et al. [9] *The NLP engine: A universal turning machine for NLP.*	2015	Attempted to develop a general framework and methodology for computing the informational and/or processing complexity of NLP applications and tasks. They developed a universal framework *akin* to a Turning Machine that attempts to fit most of the NLP tasks into one paradigm.
13	Desai et al. [5] *Program synthesis using natural language*	2016	The main focus of this work is to avoid problems of using domain specific languages (DSL) in programming. The presented idea includes generating a general framework for constructing a program that takes natural language input and produces the expressions in target DSL
14	Ernst et al. [6] *Natural language is a programming language: applying natural language processing to software development*	2017	Proposed an approach to convert English specifications of file system operations into corresponding bash commands. Recurrent neural networks (RNN) are used for this translation
15	Iacob et al. [8] *NLCP: towards a compiler for natural language*	2017	Present an NLCP (natural language compiler) that gives a new programming environment, which features a strong type system and support for a natural language interface. Additionally, it provides an interactive interpreter that can assist in the process of designing an algorithm

(continued)

Table 1 (continued)

S. No.	Author & Title	Year	Contributions
16	Weigelt et al. [16] *Programming in natural language with fuse: synthesizing methods from spoken utterances using deep natural language understanding.*	2020	They proposed a model called function synthesis executor (fuSE), which is capable of extract procedures from the input English sentence. fuSE used to take input from the speech itself. It mainly consists of three sections—*classification of teaching efforts*, *classification of the semantic structure*, and *method synthesis*
17	Brown et al. [2] *Language models are few-shot learners*	2020	Discussed about a DNN, called generative pre-trained transformer 3(GPT-3) which is an autoregressive language model that can solve most of the problems associated with natural language translation. GPT-3 is the third generation of the GPT series. GPT-3 produces sentences more like a human, where people cannot identify that these are AI generated

References

1. Blackburn, P., Bos, J.: Representation and Inference for Natural Language—A First Course in Computational Semantics. CSLI Studies in Computational Linguistics. CSLI Publications, 2005
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. CoRR. abs/2005.14165, 2020
3. Bruckman, A.S., Edwards, E.: Should we leverage natural-language knowledge? An analysis of user errors in a natural-language-style programming language. In: Williams, M.G., Altom, M.W. (eds.) Proceeding of the CHI ’99 Conference on Human Factors in Computing Systems: The CHI is the Limit, Pittsburgh, PA, USA, May 15-20, 1999, pp. 207–214. ACM, 1999
4. Dalal, B., Kalra, M.: Computer program product and computer system for language enhanced programming tools. United States Patent, 12/475468, 2011
5. Desai, A., Gulwani, S., Hingorani, V., Jain, N., Karkare, A., Marron, M., Sailesh, R., Roy, S.: Program synthesis using natural language. In: Dillon, L.K., Visser, W., Williams, L.A. (eds.) Proceedings of the 38th International Conference on Software Engineering, ICSE 2016, Austin, TX, USA, May 14–22, 2016, pp. 345–356. ACM, 2016
6. Ernst, M.D.: Natural language is a programming language: applying natural language processing to software development. In: Lerner, B.S., Bodík, R., Krishnamurthi, S. (eds.) 2nd Summit on Advances in Programming Languages, vol. 71, pp. 4:1–4:14, SNAPL 2017, May 7–10, 2017, Asilomar, CA, USA, LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017
7. Heidorn, G.E.: Automatic programming through natural language dialogue: a survey. IBM J. Res. Dev. **20**(4), 302–313 (1976)

8. Iacob, R., Rebedea, T., Trausan-Matu, S.: NLCP: towards a compiler for natural language. In: 21st International Conference on Control Systems and Computer Science, CSCS 2017, pp. 252–259, Bucharest, Romania, May 29–31, 2017. IEEE, 2017
9. Li, J., Hovy., E.H.: The NLP engine: a universal turing machine for NLP. CoRR. abs/1503.00168, 2015
10. Lieberman, H., Liu, H.: Feasibility studies for programming in natural language. In: Lieberman, H., Paternò, F., Wulf, V. (eds.) End User Development, Human-Computer Interaction Series, pp. 459–473. Springer, 2006
11. Mihalcea, R., Liu, H., Lieberman, H.: NLP (natural language processing) for NLP (natural language programming). In: Gelbukh, A.F. (ed.) Computational Linguistics and Intelligent Text Processing, 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19–25, 2006, Proceedings, vol. 3878 Lecture Notes in Computer Science, pp. 319–330. Springer, 2006
12. Miller, L.A.: Natural language programming: styles, strategies, and contrasts. IBM Syst. J. **20**(2), 184–215 (1981)
13. Stenmark , M., Nugues, P.: Natural language programming of industrial robots. In: Proceedings of the 44th International Symposium on Robotics, IEEE ISR 2013, pp. 1–5, Seoul, Korea (South), October 24–26, 2013. IEEE, 2013
14. Tangirala, V.P.: Computer knowledge representation format, system, methods, and applications. A Ph.D. Thesis, Submitted to Roskilde University, Denmark, 1971
15. Vadas, D., Curran, J.R.: Programming with unrestricted natural language. In: Baldwin, T., Curran, J.R., van Zaanen, M. (eds.) Proceedings of the Australasian Language Technology Workshop, ALTA 2005, pp. 191–199, Sydney, Australia, December 10–11, 2005. Australasian Language Technology Association, 2005
16. Weigelt, S., Steurer, V., Hey, T., Tichy, W.F.: Programming in natural language with fuse: synthesizing methods from spoken utterances using deep natural language understanding. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, pp. 4280–4295, Online, July 5–10, 2020. Association for Computational Linguistics, 2020
17. Winograd, T.: Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. MIT Libraries, Computer Science and Artificial Intelligence Lab (CSAIL) (1971)

An Exhaustive Exploration of Electromagnetism for Underwater Wireless Sensor Networks

Nebu Pulickal and C. D. Suriyakala

Abstract The influence of underwater sensor network (UWSN) has major role in the development of many areas like oil and gas, pollution, surveillance and marine life. The major success of UWSN is its ability in real-time monitoring, measurement, analysis, storage and transmission of the collected data to the outside world. The real-time monitoring of the measured parameters and the ability of transmitting huge amount of data from underwater to the terrestrial wireless sensor network is possible with the backup of technology development like Internet of submerged things (IoST). One of the factors that determines the quality of network is being reliability, and the deployment of sensors in dynamic condition is the major challenge in UWSN. Through this paper authors perform a thorough examination of the possibility of RF communication through the dynamic RF UWSN.

Keywords Sensors · IoT · IoST · UWSN · Underwater RF · Electromagnetism

1 Introduction

The underwater wireless communications possess unique and several distinct challenges when compared with terrestrial communication. Unlike in wired or wireless communications through the atmosphere, underwater communication is influenced by several phenomena like temperature, pressure, turbidity, ocean currents, salt concentration, amount of light, winds and their effects on waves. The traditional method in underwater communication relays on sensors that can record the data during the monitoring phase. These sensors are later recovered for the collected data [1]. The collected data from underwater sensor network plays a vital role in exploring the unexplored oceanic environment which can be used to obtain early information about natural disasters like earthquakes, tectonic plate movement which can cause tsunamis and floods, water pollution, mineral exploration and so on. Till date underwater wireless sensor communication mainly deals with the transmission of data in

N. Pulickal · C. D. Suriyakala (✉)
School of Ocean Engineering and Underwater Technology, Kerala University of Fisheries and Ocean Studies (KUFOS), Kochi, India
e-mail: dr_soeut@kufos.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_58

the form of light, sound and electromagnetic (EM) waves. Each of these techniques has its own benefits and limitations.

Acoustic communication is known to be a proven technology for underwater communication due to the low attenuation of acoustic waves in the medium [2, 3]. This is actually true in thermocline region only. For shallow water settings, the propagation of sound waves can be adversely affected by temperature, salinity and water depth. Another important concern in acoustic communication is in the mitigation of the Doppler effect and ISI that can occur due to multipath propagation [4]. This imposes the need for bulky, costly and high-power consuming transceivers. However, the long-range propagation (up to 20 km) makes the acoustic communication a favourable technology for underwater.

The optical transmission, which can provide an ultra-high bandwidth is another attraction for researchers [5, 6]. It can provide high data transmission rates up to 1 Gbps. But the requirement for line-of-sight transmission, along with scattering, dispersion, temperature variations, turbidity and the presence of suspended organic matter within the transmission medium [7] limits the use of optical waves for underwater communication. Despite of these limitation, the optical technology can be useful in some specific applications like oil rig maintenance, linking submarines to land, etc. [8].

If higher frequency range in electromagnetic spectrum (MHz range—RF) is considered, underwater wireless networks suffer from short range as well as electromagnetic interference [9] whereas in terrestrial applications, RF has quick response and can provide efficient communication. Among the technologies used so far, the RF technology exhibits certain explicit features like the ability to pass from one medium to another (water to air) without causing any attenuation or disturbance, tolerance against instability that can be caused due to tides or any other means, will not affect the marine life, immunity to acoustic noise [7].

RF can also be considered as a perfect candidate for hybrid communication which requires the signal propagation between terrestrial base station and underwater communication node. In such applications, signal transmissions can be made possible through antennas on the floating devices and for long distance and harsh environment, communication might be possible with the help of satellites, i.e. signals from terrestrial base station can be send to satellites and then to a floating buoy. There are several other applications in which the RF technology outperforms the first two. Best example is the sensor network deployed to control the coastal erosion through seabed sediments monitoring [10]. This shows that RF communication is more reliable for underwater sensor communication over short distances which is reflected in Table 1.

1.1 Objective and Motivation

Wireless communication technologies can be deployed for every application according to the requirements and demands. Underwater communication gives a

Table 1 Comparison of UWC technologies

Features	RF waves	Acoustic waves	Optical waves
Communication range	Up to 10 m	Up to 20 km	Up to 100 m
Operating frequency range	Up to 300 Hz	Up to 100 Hz	Tera Hz
Factors affecting speed of propagation	Frequency Salinity Conductivity Temperature	Depth Salinity Temperature	Turbidity Frequency Salt ions
Benefits	Possible short distance Communication in air–water channel Propagation in highly turbid water up to 100 Mbps data rate low propagation delay Unaffected by turbidity, pressure High bandwidth	Up to 20 km propagation range Low attenuation	Data rate in Gbps Very high bandwidth Moderate propagation range
Limitation	Short propagation range High energy consumption	Data transmission rate in kbps Strong reflection at air–water boundary Marine life affected	Not possible to cross boundary between air and water LOS required Short distance propagation

better idea about the underwater environment and is combined with different wireless techniques to trigger up the future of UWC. With the implementation of 5G technology, users will be able to get high end experience in terms of data rate and bandwidth. Filter bank multicarrier (FBMC) and generalized frequency division multiplexing (GFDM) integrated with 5G networks will become the future trend in IoST applications.

Supplying power for sensor nodes in underwater environment is one of the major challenges. Several researches are going on with the wireless power transfer, underwater piezoelectric method, simultaneous wireless information and power transfer (SWIPT), etc. Hybrid technologies are also an area of interest of the research. Through this paper, authors provide a survey on different wireless UWC technologies and development of UWSN framework and IoST along with future perspective.

This paper is organized as Introduction section followed by Sect. 2 in which a detailed analysis of the EM propagation in water is made. Section 3 addresses the underwater sensor network and its different architectures. The design challenges for an underwater RF antenna design are summarized in Sect. 4, followed by concluding Sect. 5.

2 Analysis of EM Propagation Underwater

As the energy of propagating EM waves lies in the electromagnetic fields, when the permittivity/conductivity of the media changes, it will strongly influence the propagation of EM waves [11] in terms of attenuation. The propagation of EM waves is also influenced by other parameters like temperature, salinity, turbidity and various kinds of noises. Not only that, transition frequency also plays a vital role in the propagation of EM waves underwater. Under this section, analysis of EM waves for freshwater and sea water in terms of different frequencies, propagation range and propagation loss, application with propagation ranges, and losses with depths is dealt.

Table 2 shows the first analysis change in various parameters that affects the propagation of EM waves with frequency in sea water as well as freshwater. From the typical values of the Table 2 reflects that attenuation for EM waves are more in sea water than in freshwater. The propagation velocity and propagation distance of EM waves at 100 Hz and 10^5 Hz are 3.18×10^5 m/s and 5.8×10^3 m and 1.01×10^7 m/s and 1.84×10^2 m in freshwater and 1.81×10^4 m/s and 3.25×10^2 m and 4.84×10^5 m/s and 8.81 m, respectively, in sea water. This can also be substantiated with a channel attenuation model [12] with the help of mathematical Eq. (1) as given.

$$a(f) = \sqrt{\pi \sigma \mu_0 f} = mf \tag{1}$$

In which '$a(f)$' is the channel attenuation/meter, 'f' is the carrier signal frequency in Hz, 'μ_0' represents the permeability ($\approx 4\pi \times 10^{-7}$ H/m) and 'σ' is the water conductivity.

Even though both the freshwater and sea water have almost similar permeability, the conductivity (σ) is a function of temperature and salinity. The sea water conductivity is found to be approximately 4 mhos/m, which is double that of freshwater conductivity. So, it can be said that the attenuation of the radio frequency signals is more in sea water than in freshwater.

The border behaviour of EM waves in a medium is defined by the transition frequency, which is the ratio of electrical conductivity to dielectric permittivity (σ/ε). For sea water, it is found to be 888 MHz. This means that the signals with frequencies

Table 2 Performance of EM waves for different frequencies (Hz) [7]

Frequency →		100	1000	10,000	100,000	1 million
Propagation velocity (m/s)	Freshwater	3.18×10^5	0.98×10^6	3.19×10^6	1.01×10^7	3.23×10^7
	Sea water	1.81×10^4	4.9×10^4	1.48×10^5	4.84×10^5	1.54×10^6
Wavelength (m)	Freshwater	3.18×10^3	1.03×10^3	3.18×10^3	1.04×10^3	3.18×10^3
	Sea water	1.7×10^2	4.9×10^1	1.52×10^1	4.9	1.5
Propagation distance (m)	Freshwater	5.8×10^3	1.8×10^3	5.8×10^2	1.84×10^2	5.8×10^1
	Sea water	3.25×10^2	8.94×10^1	2.3×10^1	8.81	2.8

below the transition frequency no longer behaves like a wave, rather it acts like a diffusion field. For frequencies above transition frequency (Eq. 1), the absorption loss for sea water is directly proportional to the frequency. This makes the EM propagation literally impractical in sea water. In case of freshwater, the transition frequency is about 14 MHz. The absorption loss (α) for freshwater is defined using mathematical expression (using Eq. 2) through which it is evident that α depends only on conductivity (σ), permeability (μ), and permittivity (ε) but not on frequency.

$$\alpha = \frac{\sigma}{2}\sqrt{\frac{\mu}{\varepsilon}} \tag{2}$$

Through this observation it is very clear that EM propagation at 14 MHz is possible underwater. The above analysis makes us come to a conclusion that a heteromedium communication is possible which will help to establish a communication link between underwater and terrestrial transceiver. The maximum data rate of EM propagation for the specific applications with the propagation ranges are shown in Table 3 (second analysis).

In third analysis, as in any communication system, the total power loss that can happen during the multipath propagation of waves (promising characteristics of underwater RF) in underwater is of much importance. Author describes a plane wave model [13] for underwater EM propagation assuming the depth of water as infinity. Based on this, the total power loss can be calculated as the sum of transmission and propagation loss. Mathematical expression for a plane wave incident on an air water interface at normal incidence with e_i as the incident electric field, the incident power (Eq. 3) can be calculated as:

$$P_i = \mathrm{Re}\left\{\frac{|e_i|^2}{2\eta_0}\right\} \tag{3}$$

where 'η_0' is the intrinsic impedance of air which is given by $\sqrt{\frac{\mu_0}{\epsilon_0}}$.

Table 3 Bandwidth of RF signal propagation in different water medium [4]

Propagation range	RF—Freshwater	RF—Sea water	Underwater RF Applications
Up to 10 km	1 bps	1 bps	Telemetry-deep water
Up to 2 km	10 bps	10 bps	Telemetry
Up to 200 m	5 Kbps	1 Kbps	Sensor networks
Up to 50 m	100 Kbps	5 Kbps	Diver communications, sensor array
Up to 10 m	1 Mbps	100 Kbps	AUV data download from sensor network
Less than 1 m	Up to 100 Mbps	Up to 100 Mbps	Wireless connector and AUV docking

The mathematical expression for the power transmitted by considering all the above media parameters [14] can be expressed through Eq. (4);

$$P_T = \mathrm{Re}\left\{\frac{|e_t|^2}{2\eta_1^*}\right\}|\tau|^2 \mathrm{e}^{-2\alpha t} \tag{4}$$

'τ' is the transmission coefficient which is the ratio of transmitted electric field to incident electric field, 'η_1' is the intrinsic impedance of water given by $\sqrt{\frac{\mu_1}{\varepsilon_0 \overline{\varepsilon_r}}}$ in which '$\overline{\varepsilon_r}$' is the complex permittivity of water and t is the penetration depth.

'α_T' is the transmission loss is expressed though Eq. 5.

$$\alpha_T = 10\log_{10}\left(|\tau|^2 \mathrm{Re}\left(\frac{\eta_0}{\eta_1^*}\right)\right) \tag{5}$$

For calculating the propagation loss inside water, the attenuation α [15] and is given by Eq. 6.

$$\alpha = w\sqrt{\mu\varepsilon}\left\{\frac{1}{2}\left[\sqrt{1+\left(\frac{\sigma}{w\varepsilon}\right)^2} - 1\right]\right\}^{1/2} \tag{6}$$

The propagation loss is given by Eq. (7)

$$\alpha_p = 10\log_{10}\left(\mathrm{e}^{-2\alpha d}\right) \tag{7}$$

Table 4 shows the variation of propagation loss for various frequencies at different depths [13], and it is very clear that as the depth increases the propagation loss increases whereas the transmission loss remains the same.

The total power loss P expressed in dB (Eq. 8) is the sum of transmission and the propagation loss, and it depends on the propagation depth and on the complex permittivity of water.

$$P = \alpha_T + \alpha_P = 10\log\left(\frac{P_T}{P_i}\right) \tag{8}$$

Table 4 Propagation loss at various propagation depth [13]

	0.5 m	2 m	10 m	50 m
100 kHz	0.3	1	5.8	28
1 MHz	0.7	2.9	15	70
10 MHz	0.9	3.8	19	99

3 Internet of Submerged Things (IoST)

IoST is a class of IoT which comprises of a network of smart interconnected underwater sensor nodes. This will enable numerical practical applications such as environment monitoring, underwater exploration and disaster prevention. IoST is different from terrestrial wireless networks in terms of bandwidth, data rate, reliability, etc. These differences bring up lot more challenges in implementing IoST compared to terrestrial wireless sensor network. The main components of IoST are underwater sensors. Each sensor node consists of number of sensors equipped with modem which are distributed sparsely in either shallow or deep water. These sensor nodes have the capability to sensing data like pressure, temperature, water quality which can affect aquatic life.

3.1 The Structure of Underwater Sensor Node

The emerging technologies (Table 5) in UWC points out the need for understanding the architecture of an underwater sensor node. From the architecture of a sensor node shown in Fig. 1, it is more evident that apart from sensing, network nodes also have to perform storing, processing and transmission of the collected data. The network node includes onboard controllers, memory units, sensors, modems and

Table 5 Emerging technologies in UWC

Technology	Application	Characteristics
IoST	Monitoring marine life Study of natural hazards Disaster prevention Exploration	Early warning Sustainable exploiting of marine resources Improving food production
Non-orthogonal multiple access UWC	Underwater sensor data extraction	High bandwidth High system throughput Low latency
mm wave UWC	5G terrestrial to underwater data transmission Highly secure and surveillance data application	High transmission bandwidth Improved communication performance High data rate
MIMO UWC	Monitoring application Audio, video transmission	High bandwidth Real-time high-quality data transmission Energy efficient communication
Energy harvesting from UWC	Powering up of sensor nodes and accessories Charging super SWIPT application	Improved power consumption efficiency

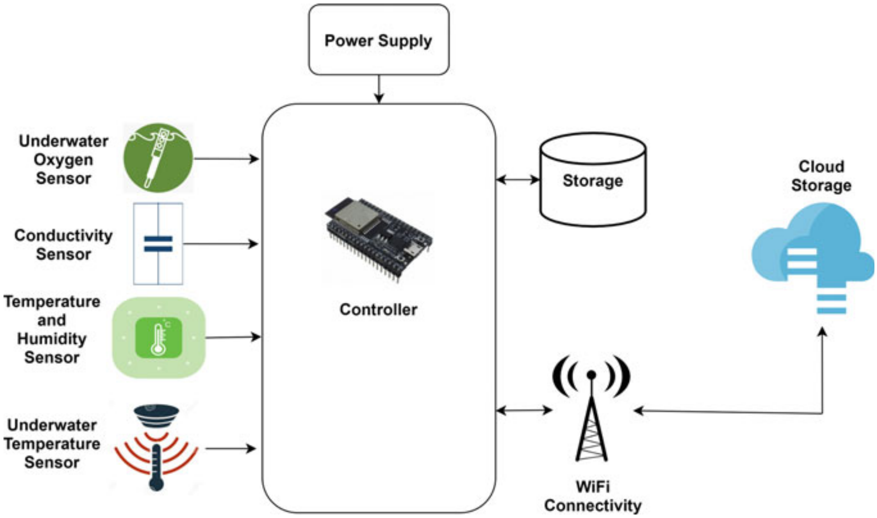


Fig. 1 Architecture of an underwater sensor node

the supporting circuitries [16]. The onboard controllers collect the sensed data from sensors and process it. The entire circuitry is usually protected within a polyvinyl chloride (PVC) housing and is mounted on a frame [17, 18].

The collected data is then stored in memory or buffer registers before the transmission. Using centralized (one node to the main unit) or multihop communication (one node to another node), the collected data is transmitted to water surface by means of [9] EM waves. Rather than using centralized communication it is more practical to rely on multihop communication which will reduce the energy level to reach the central node which is at a very far distance.

UW sensors can be of various types which can measure different properties of water. The DS18B20 and SEN0189 are two among them. The former is a 1-wire programmable temperature sensor, widely used to measure temperature in harsh environments like in mines, soil, solutions, etc., they are capable of measuring a wide range of temperature (−55 to +125 °C) with a high accuracy. The construction of these sensors is rugged and the water proof packaging helps in easy mounting of these sensors even under water. The unique address of each sensor helps to transfer the data using a single pin of microcontroller unit (MCU) [19]. The latter one is a turbidity sensor that measures the level of turbidity to detect the water quality. It is based on the principle that scattering rate and light transmittance changes with the presence of total suspended solids (TSS) to detect the suspended particles in water. The turbidity level of liquid increases with the increase in TSS [20]. These sensors are much needed in the measurement of wastewater and effluent, quality of water in streams and rivers, sediment transport and laboratory. The other parameters which can be measured are pH, salinity, pressure, flow, conductivity, acidity, etc. Other than these normal sensors disposable sensors, silicate sensors, DNA microarrays, voltametric sensors, hydrothermal sulphide sensors, amperometry microsensors, gold-amalgam electrode sensors have also been developed [16].

The compact integrated MCU, designed to perform a specific operation in the sensor node, consists of a processor, memory unit and input/output (I/O) peripherals on a single chip. This core unit process the collected data from underwater sensors and controls all the functionalities within a sensor node such as data sensing interval, mode of transmission and data representation. For better functioning of sensor nodes, microcontrollers are selected based on parameters like security, power efficiency, processing power, temperature tolerance, memory, hardware architectur and, cost. For optimum operation of sensor nodes, sometimes a trade-off between the specific parameters has to be addressed. For example, a multicore processor will be faster with the cost of more energy consumption. Generally, controllers like ATtiny85, which are having high performance and low power consumption capability are well suited for these applications [21].

The power consumption in a sensor node is a critical issue that needs to be addressed. UWSN nodes can be powered using rechargeable or stand-alone batteries. The energy requirement for transmission and reception of data in UWSN is high, hence the rechargeable batteries are said to be a perfect candidate for this. The case of stand-alone batteries is that once it drains out, the network node seems to be useless. This is because the positions of these UWSN nodes may change based on the flow of water as well as need for larger energy requirement demand the need for frequent charging. At the same time, energy can also be harvested from the EM waves which is used for communicating with other nodes.

Acoustic modem technology or the acoustic transmissions is best suited for longer distance application and it also outperforms EM in vertical ranges. But it has several limitations [8, 11] like frequency spreading, time spreading, Doppler spread, bandwidth limitation, etc. For variety of reasons like turbidity dependence and propagation distance, the optical modem links seem to be impractical [9]. The first underwater RF modem—(S1510)—was released in 2006 by wireless fibre system with a data rate of 100 bps [22]. In 2007, underwater broadband RF modem was also introduced with a speed of 1 Mbps [23] and a communication range of 1 m.

A UWSN differs from the terrestrial sensor networks in many aspects like implementation, power requirements, cost and memory. Because of this lot of researches are progressing to find an optimal solution using innovative technologies to make UWSN communication [16] for effective.

3.2 UWSN Topology

This section briefly describes underwater RF sensor network topologies both in two- and three-dimensional. The popularity of a network topology is decided by various factors like its ability to perform communication without depending on a fixed infrastructure, ease of deployment, signal transmission, data collection, etc. [24]. As it is known, the main resources of any underwater communication system are energy and capacity, the design of the network should be done in such a way that, the network provides maximum capacity and it should also be energy efficient. The backbone of any UWSN architecture is the sensors. These can be mainly classified into two: Embedded sensors and the floating sensors. Embedded sensors are those which are deep seated within the sea bed, whereas the latter one floats on the surface of the waterbodies, and these are mainly used in areas where the environment is so rigid else and they can be easily detected by the enemies and can maltreat them.

2D UWSN Architecture. Figure 2 shows a referenced 2D UWSN [11]. This consist of group of sensor nodes interconnected to one or more underwater gateways (UW-gateways) in a wireless fashion. These nodes are also anchored deep underwater with the help of deep ocean anchors. The UW-gateways are capable of relaying data from UW network to surface stations [25]. In order to make this transfer possible, the UW-gateways are assembled with horizontal and vertical transceivers. The communication with the sensor nodes, i.e. in order to send the commands, configure data and to collect the monitored data, is made possible with the help of horizontal transceivers whereas the vertical transceivers are responsible for the communication with surface station, hence it must be of long range. The surface station is also equipped with a transceiver which is capable of full duplex communication [24].

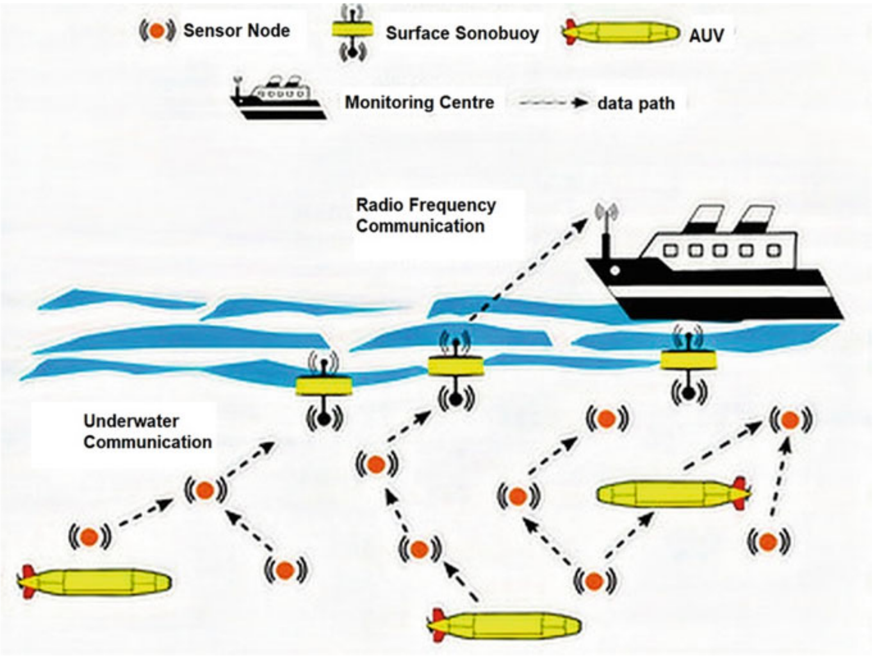


Fig. 2 2D underwater sensor network architecture

As mentioned earlier, the sensors send the monitored data to UW-gateways through direct links or through multihop paths. With a cost of complex routing, more capacity and energy efficiency is attained with the help of multihop paths.

3D UWSN Architecture. Three-dimensional UWSN is mainly used to obtain information about certain phenomena, whose adequate information cannot be attained with the help of 2D sensor networks. These networks consist of floating sensors at different water levels. One of the simplest solutions for such a deployment is to attach UW sensors to surface buoy. The depth of these sensors can be controlled by the length of the wires through which they are connected to surface buoy. But this technique has a demerit of easily detectable by the enemies or multiple buoys can affect the ship navigation [11]. Another approach which overcomes this demerit and maintains sensors at different depths is shown in Fig. 3. In this topology, each sensor is fitted with a floating buoy and is anchored to the bottom of the water body. The floating buoy can be inflated with the help of a pump and this pushes the sensors to the surface. The depth required for each sensor can then be electronically controlled by adjusting wire length (used to connect the anchor to sensor) with the help of an electronically controlled engine [25].

Sensing and the communication coverage needed for the sensors in such an architecture, in order to obtain a 3D visualization on the phenomena are the major challenges that arise with 3D architecture.

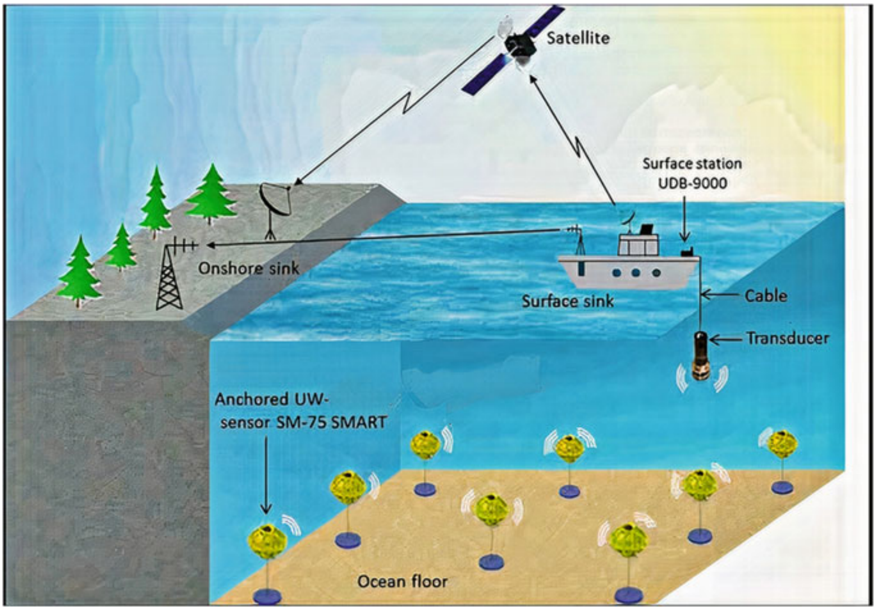


Fig. 3 3D underwater sensor network architecture

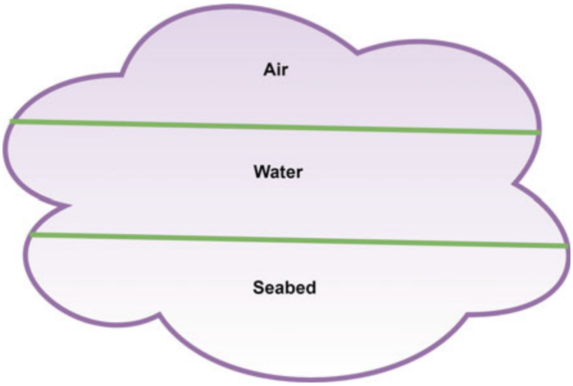
4 Underwater RF Sensor Node Antenna

Transducers help in the conversion of analogue data into electrical signals or vice versa in wireless communication through an antenna in air. Electromagnetic waves cannot travel through conductors. Unlike in terrestrial applications, EM wave propagation through water depends on two factors, conductivity of water and frequency of radio wave [26–28]. The conductivity of water in real-time application cannot be controlled hence the parameter that affects the UWC is frequency, which affects the antenna dimension. The conductivity of water mainly depends on the geographical location, temperature, etc. Hence, for efficient use of an antenna, different matching circuits must be designed based on the water conductivity [29, 30]. More commonly EM waves can be used for shallow water communication compared to optical and acoustic communication in coastal environment.

There are several applications that work at low frequency (in KHz) range and many researches are going on at high frequencies. For a sensor network, data can be stored in sensor node and can be collected at periodic intervals by sweeping over the nodes. In such scenarios, the high-frequency antennas will perform well and large amount of data from sensor node can be collected by using peer-to-peer communication technique.

The need for deploying large number of sensor nodes underwater, which is associated with the short-range communication, is another main challenge in RF antenna design. For EM wave propagation through underwater, large-size antenna is required for proper functioning. Hence, mounting the antenna on floating devices will be a challenge [4]. Survey shows that there are work [31] which model the shallow water coastal environment as a three-layer dielectric medium as shown in Fig. 4 with water body as a middle layer sandwiched between air and seabed layers. Authors recommending this model for the design of RF antennas for shallow water environments.

Fig. 4 Three-layer model of shallow water environment



5 Conclusion

Through this paper, authors are analysing an outline of underwater sensor networks and its technologies. The need for fast and efficient underwater data handling reveals the potential of high-frequency electromagnetic communication. The intensive study here reiterates that unlike in terrestrial application, why an RF antenna design for underwater communication needs more attention in analysis and design. Authors have discussed the major challenges in UWSN while deploying sensors. From the comparative analysis, propagation velocity, wavelength and propagation distance for the frequency ranges varying from 100 Hz to MHz, freshwater is outperforming sea water. Similarly, from the analysis of different bandwidth with respect to propagation range and application, better bandwidth of 1 Mbps is found to be more suitable for underwater sensor networks at the propagation range of 10 m. This observation is found to be the best option. At this juncture, authors would like to highlight that the propagation range may be increased by increasing the domain area with very valid support of EM propagation underwater. The next finding is selection of propagation depth in terms of frequency, shows that even though the transmission loss is constant, too much depth will affect severely the propagation loss. With the conventional basic architecture of sensors and network topology suitable for underwater is also taken into consideration. At the same time, the water conductivity can affect various parameters of the sensor node antenna. After thorough study and analysis by considering all the parameters authors strongly highlight, RF-based communication networks will be more suitable for shallow freshwater communication.

References

1. Proakis, J.G., Rice, J.A., Sozer, E.M., Stojanovic, M.: Shallow water acoustic networks. In: Proakis, J.G. (ed.) Encyclopedia of Telecommunications. Wiley, New York (2003)
2. Riksfjord, H., Haug, O.T., Hovem, J.M.: Underwater acoustic networks—survey on communication challenges with transmission simulations. In: Third International Conference on Sensor Technologies and Applications, Athens, Glyfada (2009)
3. Au, W.W.L., Nachtigal, P.E., Pawloski, J.L.: Acoustic effects of the ATOC signal (15 Hz, 195 dB) on dolphins and whales. J. Acoust. Soc. Am. **101**(5), 2973–2977 (1997)
4. Ali, M., Jayakody, D.N.K., Chursin, Y.A., et al.: Recent advances and future directions on underwater wireless communications. Arch. Comput. Methods Eng. (2019)
5. Hanson, F., Radic, S.: High bandwidth underwater optical communication. Appl. Opt. **47**, 277–283 (2008)
6. Arnon, S.: An underwater optical wireless communication network. In: Proceedings of SPIE 7464, Free-Space Laser Communications IX (2010)
7. Che, X., Wells, I., Dickers, G., Kear, P., Gong, X.: Re-evaluation of RF electromagnetic communication in underwater sensor networks. IEEE Commun. Mag. **48**(12), 143–151 (2010)
8. Lanbo, L., Jun-Hong, Z.S.C.: Prospects and problems of wireless communication for underwater sensor networks, Wiley WCMC special issue on underwater sensor networks. Wirel. Commun. Mob. Comput. **8**(8), 977–994 (2008)

9. Gussen, C.M., Diniz, P.S., Campos, M., Martins, W.A., Costa, F.M., Gois, J.N.: A survey of underwater wireless communication technologies. J. Commun. Inf. Syst. **31**(1), 242–255 (2016)
10. Hunt, K.P., Niemeier, J.J., Kruger, A.: RF communications in underwater wireless sensor networks. In: IEEE International Conference on Electro/Information Technology, pp. 1–6 (2010)
11. Akyildiz, I.F., Pompili, D., Melodia, T.: Underwater acoustic sensor networks: research challenges. J. Ad Hoc Netw. **3**(3), 257–279 (2005)
12. Zoksimovski, A., Sexton, D., et al.: Underwater electromagnetic communications using conduction–channel characterization. In: 7th ACM International Conference on Underwater Networks Systems, Los Angeles, CA, USA (2012)
13. Jiang, S., Georgakopoulos, S.: Electromagnetic wave propagation into fresh water. J. Electromagn. Anal. Appl. **3**(7), 261–266 (2011)
14. Hasted, J.B.: Aqueous Dielectrics. Chapman and Hall, New York (1973)
15. Balanis, C.A.: Advanced Engineering Electromagnetics. Wiley, New York (1989)
16. Cayirci, E., Tezcan, H., Dogan, Y., Coskun, V.: Wireless sensor networks for underwater surveillance systems. Ad Hoc Netw. **4**(4), 431–446 (2006)
17. Codiga, D.L., Rice, J.A., Baxley, P.A.: Networked acoustic modems for real-time data delivery from distributed subsurface instruments in the coastal ocean: initial system development and performance. J. Atmos. Oceanic Tech. **21**(2), 331–346 (2004)
18. Sendra, S., Lloret, J., Rodrigues, J.J.P.C., Aguiar, J.M.: Underwater wireless communications in freshwater at 2.4 GHz. IEEE Commun. Lett. **17**(9), 1794–1797 (2013)
19. Maxim Integrated, USA, DS18B20-Programmable Resolution 1-Wire Digital Thermometer. Available: https://datasheets.maximintegrated.com/en/ds/DS18B20.pdf (2019)
20. Turbidity sensor SKUSEN0189 (2019). Available: https://wiki.dfrobot.com/TurbiditysensorSKUSEN0189
21. Atmel Microcontroller, USA, ATtiny85 (2018) Available: https://www.microchip.com/wwwproducts/en/ATtiny85
22. RF Exynos RF Series (2018) Available: https://www.samsung.com/semiconductor/minisite/exynos/products/modemrf/exynosrf-ic-series
23. Farr, N., Chave, A., Freitag, L., Preisig, J., White, S., Yoerger, D., Titterton, P.: Optical modem technology for seafloor observatories. In: Proceedings of OCEANS, pp. 1–6 (2006)
24. Bhambri, H., Swaroop, A.: Underwater sensor network: architectures, challenges and applications. In: International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 915–920 (2014)
25. Pompili, D., Melodia, T., Akyildiz, I.F.: Three-dimensional and two-dimensional deployment analysis for underwater acoustic sensor networks. Ad Hoc Netw. **7**(4), 778–790 (2009)
26. Abdou, A., Shaw, A., Mason, A., Al-Shamma, A., Cullen, J., Wylie, S.: Electromagnetic (EM) wave propagation for the development of an underwater wireless sensor network (WSN). IEEE Sens. Proc. (2011)
27. Partan, J., Kurose, J., Levine, B.N.: A survey of practical issues in underwater networks. ACM SIGMOBILE Mob. Comput. Comm. Rev. **11**(4), 23 (2007)
28. Pozzebon, A.: Bringing near field communication under water: short range data exchange in fresh and salt water, pp. 152–156 (2015)
29. Incio, S.I., et al.: Dipole antenna for underwater radio communications. In: IEEE Third Underwater Communications and Networking Conference (UComms), Lerici, pp. 1–5 (2016)
30. Incio, S.I., et al.: Antenna design for underwater radio communications, OCEANS, Shanghai (2016), pp. 1–6
31. Cella, U.M., Johnstone, R., Shuley, N.: Electromagnetic wave wireless communication in shallow water coastal environment: theoretical analysis and experimental results. In: Proceedings of the 4th ACM International Workshop on UnderWater Networks, WUWNet (2009)

Two-Stage Feature Selection Pipeline for Text Classification

Vinod Kumar, Abhishek Sharma, Anil Bansal, and Jagnur Singh Sandhu

Abstract Text classification is one of the significant fields in NLP. It has numerous applications in the business world such as e-mail spam filtering, fraud detection, recruitment and various other tasks. The past three decades have witnessed an exponential rise in the amount of information. This has made text classification all the more challenging. There is a greater need for optimal feature subset selection which in turn enhances classification performance. In this study, we discuss a two-stage feature selection pipeline which combines conventional filter methods (Chi square and information gain) with evolutionary algorithms (particle swarm optimization and genetic algorithm). Subsequently, we aim to compare the results from these pipelines with the results from evolutionary algorithms over existing classifiers. The experiments are performed on two popular data sets—20 newsgroups and IMDB reviews. Results obtained show that the pipeline-based feature selection methods perform better than their respective wrapper methods.

Keywords Text classification · Feature selection · Evolutionary algorithms · Particle swarm optimization · Information gain · Chi square · Filter methods · Naïve Bayes

1 Introduction

With the advent of the Internet, the increasing amount of unstructured and semi-structured information available in documents has contributed to the need for text mining. Of all the data generated in the world, 90% is not structured. Text mining is the process of extraction of useful and vital information from unstructured data. Text mining includes tasks such as text classification, entity extraction, text clustering and entity summarization [1].

V. Kumar · A. Sharma · A. Bansal (✉) · J. S. Sandhu
Delhi Technological University, New Delhi, Delhi 110042, India

V. Kumar
e-mail: vinod_k@dtu.ac.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_59

Text classification (TC) is the most important task of text mining. Text classification is the process of assigning a class/category to a document, from a set of given classes. Studying all the documents in order to extract the useful one is a highly infeasible task for humans, hence the need for automatic text classification.

Text classification can further be of two types: single-label text classification and multilabel text classification. In single-label text classification, only one class is assigned to a document, or it can be said that the document belongs to a single class. In multilabel classification, more than one class can be assigned to a document, in other words a single document can belong to more than one class. In this paper, only single-label classification is considered.

The compelling need for efficient text classification techniques and their applications across various fields makes this an extensively researched area of NLP. Over the course of time, the problem of high dimensionality has inspired many feature selection methods. The use of TF-IDF [2] for text classification was among the initial works along with the comparative study of feature selection metrics such as Chi2 and IG [3]. More recently, nature-inspired optimization algorithms are being employed for feature subset selection; for instance, particle swarm optimization (PSO) [4, 5] and genetic algorithm (GA) [6–9].

A few notable hybrid approaches have also been used; for instance, a filter method followed by GA with latent semantic indexing [7]; two-stage feature selection based on IG with principal component analysis and GA [6]; a hybrid approach based on enhanced-GA with filter methods for classifying Arabic texts [10].

These hybrid/multistage studies, though effective in dimensionality reduction, are limited in scope as most of them employ only GA with the feature selection metrics on a select few classifiers (maximum two). Our study aims at increasing this scope by evaluating the performance of pipeline-based feature selection, i.e. feature ranking methods followed by feature subset selection using GA as well as particle swarm optimization. We observe that these pipeline-based methods consistently outperform GA and PSO, respectively, over two data sets and two classifiers.

The vocabulary of the complete data set will contain tens of thousands of unique words. Not all the words will help in class prediction. Some of the words will be irrelevant and others redundant. Also, using all the features can make the training computationally infeasible. It is important to remove these features before training a model in order to achieve best performance. To achieve this, we have used a set of standard methods in data pre-processing which includes removing stopwords and other irrelevant words. To select the best feature subset out of vocabulary, we believe that two-stage feature selection or pipeline works better than historic single stage feature selection methods to reduce high dimensionality of feature space. To test our hypothesis, we have created four pipelines of Chi2 and GA, Chi2 and PSO, IG and GA, IG and PSO. These pipelines are then compared with their respective single stage algorithms (PSO and GA). The resulting feature subset is then used to train machine learning models, particularly linear SVC and Naive Bayes to do a performance analysis on two widely used data sets (IMDB, 20 newsgroups).

2 Related Work

GA [11] is an example of an evolutionary algorithm and is based on Darwin's theory of evolution through natural selection. By presenting a candidate solution in the form of bit strings and applying three operators namely selection, crossover and mutation on a population of these strings, GA conducts a search for an optimal feature subset over the search space. Tan et al. [8] used GA to limit the feature subset size by searching for an optimal subset in a feature pool. This feature pool was created by applying entropy-based feature ranking as well as T-statistics method, thus maximizing classification accuracy (NB and associative classification) and minimizing feature subset size. A GA based on biological evolution (BGA) was proposed by Tsai et al. [12]. The idea was to let organisms allocate resources among themselves efficiently after long-term evolution. The introduction of novel mechanisms such as elite reserve and migration helped reduce time taken in subset selection and achieved slightly better performance on k-NN and SVM classifiers.

Particle swarm optimization algorithm is a meta-heuristic algorithm introduced by Eberhart and Kennedy [12] in 1995. PSO is based on the swarming behaviour of flocks of birds. PSO starts by initializing a swarm of potential solutions, called particles. Along with velocity, each particle also has a position (which defines the solution represented by the particle). Both, the position and the velocity are initialized randomly in the beginning. A particle moves in the search space updating its velocity based on its own experience and the information from its neighbouring particles. This updated velocity along with the current position is then used to find the new position for the particle. Thus, the particles converge to the best position, i.e. the position with the best fitness value. The following mathematical equations govern the change in velocity and position:

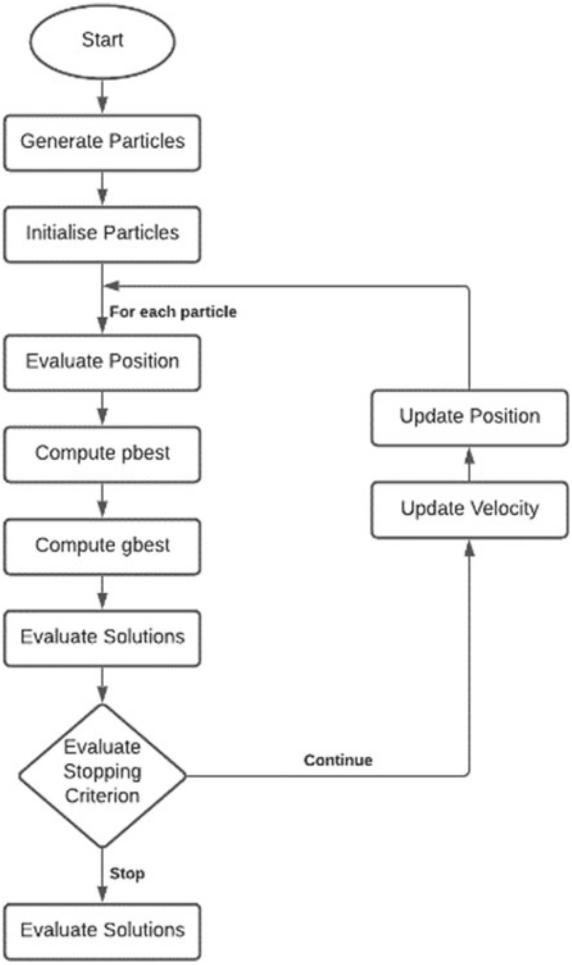
$$\begin{aligned} V_i(n+1) = w.V_i(n) + c_1.r_1(n).[P_i(i) - X_i(n)] \\ + c_2 r_2(n).\left[P_g(n) - X_i(n)\right] \end{aligned} \tag{1}$$

$$X_i(n+1) = X_i(n) + V_i(n+1) \tag{2}$$

Here, i is the particle index, n is the iteration number, w is the weight inertia constant. $V_i(n)$ is the velocity of ith particle in nth iteration. c_1 represents self-confidence and c_2 represents swarm confidence. r_1 and r_2 are random values between 0 and 1. $P_i(n)$ is the personal best for ith particle at nth iteration. $P_g(n)$ is the global best. $X_i(n)$ represents ith particle's position at nth iteration (Fig. 1).

A variant of the original PSO, the binary particle swarm optimization (BPSO) algorithm with KNN was employed by Chantar et al. [4] for feature subset selection for categorization of Arabic documents. BPSO deals with discrete search spaces in contrast to the original PSO which dealt with continuous search spaces. Here, a particle's position is represented by a binary vector of components, where each component is an Arabic word. kNN was used to test the quality of the feature subset

Fig. 1 Flowchart for PSO algorithm



on the training data. After classification, the delivered results compared well with the existing studies in the field. In another research by Aghdam and Heidari [13], it was proposed to use the classifier performance and length of the selected subset as the heuristic. Results showed that stochastic methods like PSO have the ability to converge quickly and determine the minimal feature subset efficiently. The proposed method outperformed IG and chi square as it achieved better performance on fewer numbers of features.

Hybrid feature selection approach is a combination of filter methods such as IG, Chi2, DF as the first stage and wrapper methods such as GA and other swarm evolutionary algorithms as the second stage. Uguz [6] in 2011 proposed a hybrid feature selection method. His method used a combination of IG as filter and GA and PCA as wrapper. To test the model, KNN and DT were used. Roulette wheel selection was used for selection of chromosomes. The results showed that IG and GA, IG and PCA gave better results with a smaller feature subset as compared to when using IG alone. To further research, this area Gunal [7] in 2012 used four filter methods namely IG, DF, MI and Chi along with GA as a second-stage wrapper method. SVM and DT classifiers were used for evaluation. Lei [14] 2012 also worked in the field of hybrid feature selection using GA as second-stage feature selection. Both these researches showed that hybrid feature selection reduces dimensionality and improves performance of text classification.

Filter methods choose the features depending on their relevance which is based on univariate statistics instead of cross-validation performance which is the basis of wrapper methods. Filter methods give faster results than wrapper methods and are the most common feature selection methods. Some of the filter methods commonly used are IG and Chi2.

In 2012, Shang Lei [15] worked on an improved feature selection method based on information gain. Information gain represents the amount of information provided by a feature item for a category. Higher the information gain, more is the contribution of that feature in determining the category of document. IG helps us realize the importance of a given attribute in the feature vector. The formula is given below:

$$\begin{aligned} G(D, x) = &-\sum_{i=1}^{m} P(C_i)\log(P(C_i) \\ &+ P(x)\sum_{i=1}^{m} P(C_i|x)\log(P(C_i|x)) \\ &+ P(\bar{x})\sum_{i=1}^{m} P(C_i|\bar{x})\log(P(C_i|\bar{x})) \end{aligned} \tag{3}$$

Here, C represents the collection of documents in which the feature 'x' is absent.

In 2018, Yujia and Song [16] researched a feature selection technique for text classification based on chi square. Chi square method gives a relevance between a feature x and category c. Higher the chi square value more dependent the category is on the feature. The experiments on comment corpus having around 8 k data points and SVM classifier showed chi square outperforming information gain.

$$\chi^2(t, c_i) = \frac{N(AD - CB)^2}{(A\ + C) * (B + D) * (A + B) * (C + D)} \tag{4}$$

where N represents a number of documents in total and A, B, C and D denote a term t and a category c_i coexist, t exists without c_i, c_i exists without t, and neither c_i nor t exist, respectively.

For n classes, every term value will have n correlation value, the average value computation for a class as follows:

$$CHI(t) = \sum_{i=1}^{n} P(c_i)\log(CHI(t, c_i)) \tag{5}$$

3 Methodology

3.1 Pre-processing

Before proceeding with model training and feature selection, data cleaning is performed which includes removing of stopwords [17], stemming [18] and parsing html and normalization. Words such as ‘a’, ‘an’, ‘the’, occur frequently in all documents and thus are of no significant use in text classification. For this research, the existing list of stopwords has been used for stopword removal. To reduce the dimensionality, words are reduced to their root form using stemming. Porter stemmer in sklearn is used for this purpose. The cleaned documents are then converted to a feature vector also called term weighing using countvectorizer for different pipelines.

$$\text{data} = \{w_1, w_2, w_3, \ldots, w_{|t|}\}$$

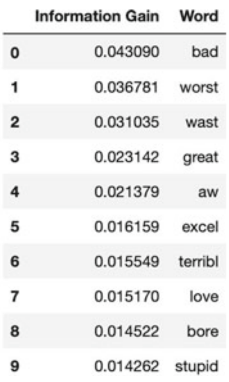
Here, w_i is the weight of the feature i in the document.

3.2 Filter Methods

This is the first stage of our pipeline-based feature selection method. For this study, we have employed two feature ranking methods namely IG and Chi Square. As the name suggests, these methods are key to filtering out those features which are less likely to contribute towards greater accuracy in classification. Once we allocated scores to respective features, they were sorted in descending order of these scores. From this large set of features, a subset consisting of top N features was selected, later to be fed to the wrapper methods. Figures 2, 3, 4 and 5 show top 10 features ranked on the basis of their IG score and Chi2 score.

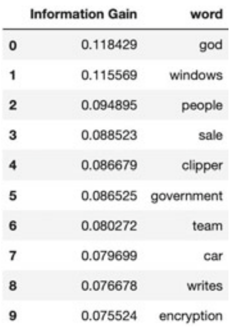
3.3 Genetic Algorithm

The subsets, obtained by ranking the words using the above-stated feature selection metrics, contain informative features but the dimensionality is still high. Thus, this subset is fed into the GA which is the next stage in the pipeline.



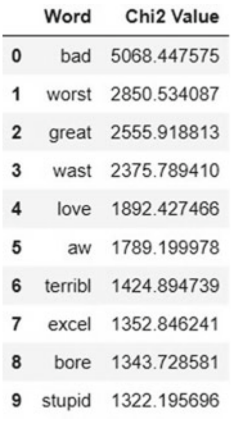
	Information Gain	Word
0	0.043090	bad
1	0.036781	worst
2	0.031035	wast
3	0.023142	great
4	0.021379	aw
5	0.016159	excel
6	0.015549	terribl
7	0.015170	love
8	0.014522	bore
9	0.014262	stupid

Fig. 2 IG scores for top features



	Information Gain	word
0	0.118429	god
1	0.115569	windows
2	0.094895	people
3	0.088523	sale
4	0.086679	clipper
5	0.086525	government
6	0.080272	team
7	0.079699	car
8	0.076678	writes
9	0.075524	encryption

Fig. 3 IG scores for top features

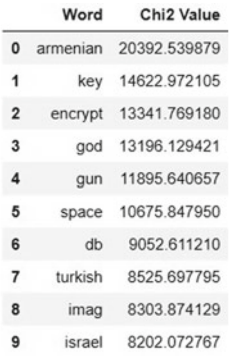


	Word	Chi2 Value
0	bad	5068.447575
1	worst	2850.534087
2	great	2555.918813
3	wast	2375.789410
4	love	1892.427466
5	aw	1789.199978
6	terribl	1424.894739
7	excel	1352.846241
8	bore	1343.728581
9	stupid	1322.195696

Fig. 4 Chi2 scores for top features

Population. An individual (chromosome) represents a candidate solution and is evaluated by a fitness function. A set of these chromosomes constitutes the population. We have allocated a population size of 25 for this study.

Fig. 5 Chi2 scores for top features



	Word	Chi2 Value
0	armenian	20392.539879
1	key	14622.972105
2	encrypt	13341.769180
3	god	13196.129421
4	gun	11895.640657
5	space	10675.847950
6	db	9052.611210
7	turkish	8525.697795
8	imag	8303.874129
9	israel	8202.072767

Fitness function. This function acts as a measure of the fitness of a candidate solution. In our method, it is the classification accuracy of the current individual. A score is assigned to the individual using the following formula:

$$\text{Fitness}(x) = c(x) - \alpha \frac{\text{features}(x)}{N} \tag{6}$$

where $c(x)$ is the classification accuracy, features(x) represents the number of features selected, N is the total number of features and alpha is a hyperparameter which represents trade-off between the two criteria (alpha = 0.01).

This score forms the basis for the selection phase.

- Selection: In this phase, the fittest individuals are selected and their genes (parameter values) are passed off to the next generation. The strategy that we have used for selection is tournament selection with a tournament size of 3.
- Crossover: Crossover is a significant operation of GA as it contributes to the diversity in population. A new generation of individuals is produced in this process due to swapping of features of two parent subsets. We have used a single point crossover method, which involves exchange of features after an arbitrary index in the array of individuals (parents). We have assigned a crossover probability of 0.5, which implies a 50% chance that exchange of features will happen between the parents.
- Mutation: This is another important operation that is used on a single individual. It involves flipping the values of a random number of features. This is another way of ensuring diversity and avoiding premature convergence. For our study, we have decided on a relatively low mutation rate or probability of 0.2.

3.4 Particle Swarm Optimization

The filter methods in the previous step rank the features using scoring metrics and select the top features. PSO here works as a wrapper method, with the aim of selecting the best feature subset which can accurately determine the correct category. At this stage, we obtain feature vector corresponding to each document, selected by the filter methods. An instance of discrete binary PSO is initialized. We have initialized a swarm of 30 particles. In the swarm, each particle has a velocity and a position. Here, position is a binary vector, where 1 indicates the inclusion of the corresponding feature in the subset and vice versa.

For fitness function, we have used an equation from the works of Vieira et al. [19].

$$f(X) = \alpha(1 - P) + (1 - \alpha)\left(1 - \frac{N_f}{N_t}\right) \tag{7}$$

Here α is a hyperparameter which serves as a trade-off between the feature subset size N_f and the classifier performance P. In this study, we have used $\alpha = 0.9$. N_t here is the total number of features. The classifier we have used to guide PSO here is Naive Bayes. The classifier performance is measured in terms of accuracy. We have used 0.9 as the inertia weight constant. The particles converge on the best available position. The final position is a binary vector, from which all the features corresponding to '0' are dropped. Thus, we obtain the optimal feature subset which will later be used by classifiers for classification.

3.5 Classification

In this study, to compare the performance of hybrid feature selection and single-stage feature selection, two classifiers support vector machine (SVM) and Naive Bayes (NB) are used. The models are trained on 80% of the data set. The remaining 20% is used for testing.

NB text classifier is based on Bayes theorem [20], which computes conditional probability of two events from the probabilities of occurrence of individual events. Because of its simplicity and high classification accuracy in text-related fields, it has found its application in text classification.

SVM [21] is among the most powerful classification algorithms in machine learning. It has two versions: linear and nonlinear. In this study, we have used linear SVC for classification. This classifier uses hyperplanes to separate classes. 'L2' penalty for regularization and squared hinge loss as loss function are used as SVM parameters. One vs. the rest scheme is used to handle multi class classification (Table 1).

Table 1 Classification accuracy on all models

Model	20 Newsgroups		IMDB reviews	
	Naive Bayes (%)	Linear SVC (%)	Naive Bayes (%)	Linear SVC (%)
PSO + IG	65.4	58.9	86.5	84.1
PSO + Chi2	68.6	57.8	83.4	84.1
PSO	53.3	52.2	77.1	80.8
GA + IG	61.2	52.3	84.1	84.8
GA + Chi2	60.8	52.2	84.5	85.2
GA	60.4	51.5	83.7	84.4

4 Results

4.1 20 Newsgroups

PSO. A comparison of pipelines with PSO on two classifiers, NB and SVC, has been depicted in Figs. 8 and 9. Classification accuracy has been used as a metric for measuring the performances of the respective text classification methods. On analysing the performance of PSO, we observe that it underperforms in terms of classification accuracy on both the classifiers. It reaches its peak of approx. 52% accuracy for 14,000 features which is below par as far as this study is concerned. However, if we apply a filter method as the first stage in our pipeline-based feature selection, a considerable jump in the graph is observed (refer to PSO + Chi2, PSO + IG in Figs. 8 and 9). For instance, in Fig. 8, PSO + Chi2 with NB scores a maximum of about 68.6% for 14,000 features. Between the two pipelines, accuracy trends indicate that though PSO + IG pipeline achieves greater accuracy for much less features as compared to PSO + Chi2. However, the slope for PSO + IG graph is more gradual than PSO + Chi2.

GA. Following on the lines of PSO pipeline on 20 newsgroups Figs. 6 and 7 throw light on the comparative performance of GA pipeline with standard GA. We observe that standard GA performs better on initial population but as the population size increases, pipeline surpasses standard GA. GA achieves its peak accuracy of about 60% on Naive Bayes and 51.5% on linear SVC, whereas pipeline achieves 61.2% and 52.2%, respectively.

4.2 IMDB Reviews

PSO. Figures 12 and 13 show similar trends as observed in Figs. 8 and 9 (20 newsgroups). PSO + IG continues to give better accuracy for lesser features whereas PSO + Chi2 witnesses a steeper increase in accuracy as the numbers of features are

Fig. 6 Comparison of pipeline with GA on 20 newsgroups using Naïve Bayes

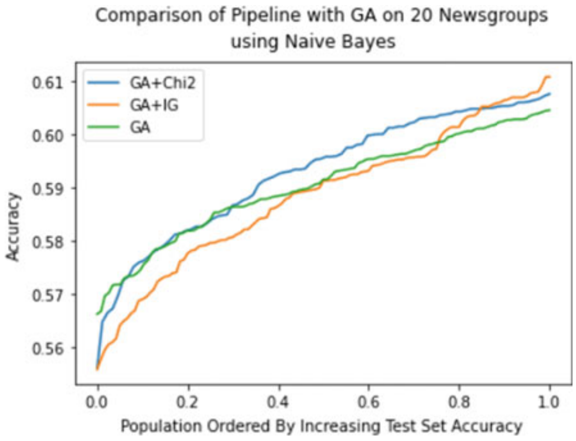
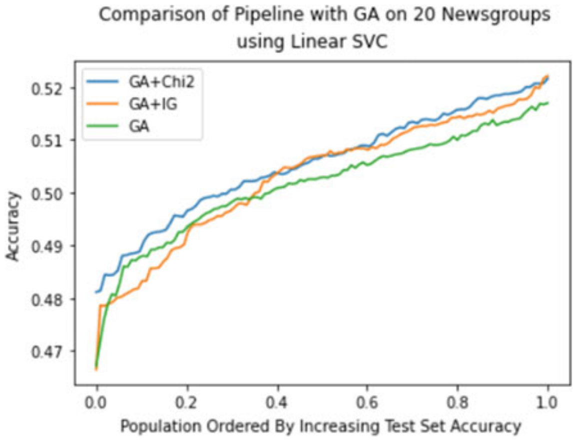


Fig. 7 Comparison of pipeline with GA on 20 newsgroups using linear SVC



increased. PSO + IG achieves a maximum accuracy of 86.5% for 14,000 features (Figs. 10 and 11).

GA. Figures 10 and 11 compare GA with the two pipelines on the basis of accuracy achieved by Naive Bayes and linear SVC, respectively. As we can see in Fig. 10, GA + Chi2 and GA + IG move neck to neck with GA at initial populations. As the population increases, it can be observed that both the pipelines are performing better than GA. Best accuracy, 84.5%, is achieved by GA + Chi2. Figure 8 also follows a similar trend as Fig. 10 in the beginning population, the difference being GA + Chi2 emerges as the better performer as the population increases, highest accuracy being 85.2%. Here also we observe that both the pipelines perform better than GA.

IMDB reviews being a binary class data set is giving significantly higher performance as compared to 20 newsgroups which is a multiclass data set.

Fig. 8 Comparison of pipeline with PSO on 20 newsgroups using Naïve Bayes

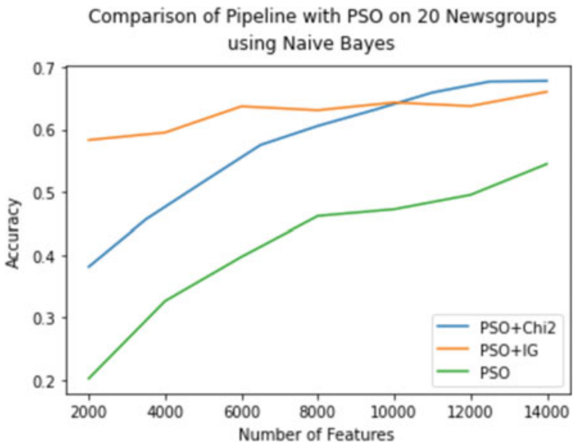


Fig. 9 Comparison of pipeline with PSO on 20 newsgroups using linear SVC

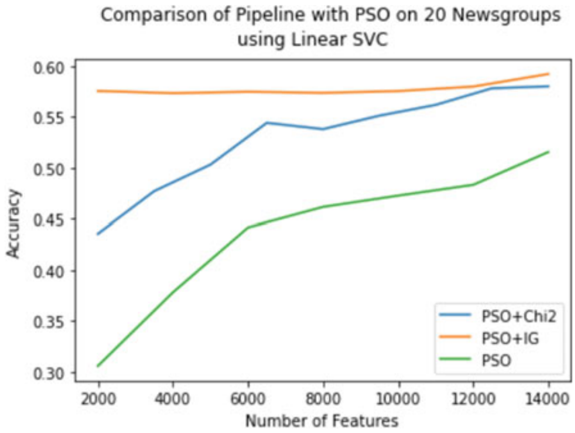


Fig. 10 Comparison of pipeline with GA on IMDB reviews using Naïve Bayes

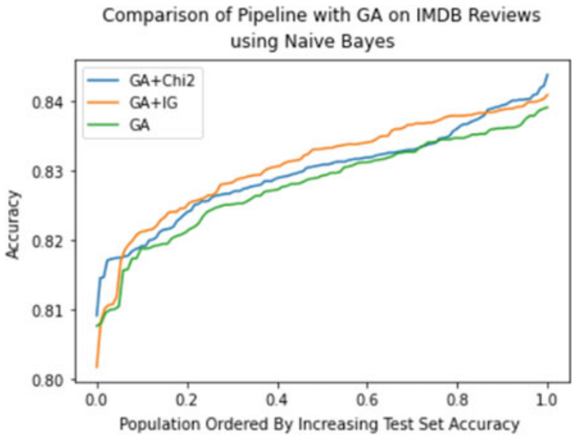


Fig. 11 Comparison of pipeline with GA on IMDB reviews using linear SVC

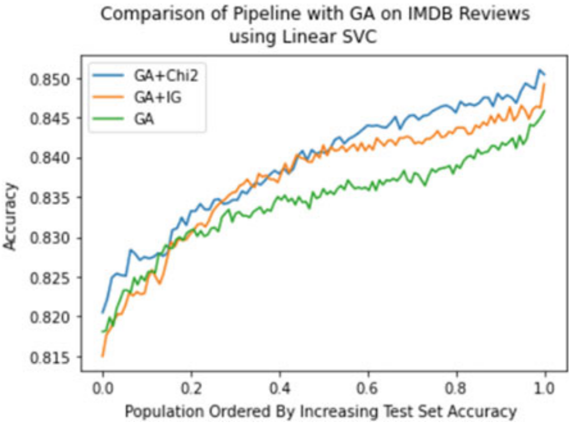


Fig. 12 Comparison of pipeline with PSO on IMDB reviews using Naïve Bayes

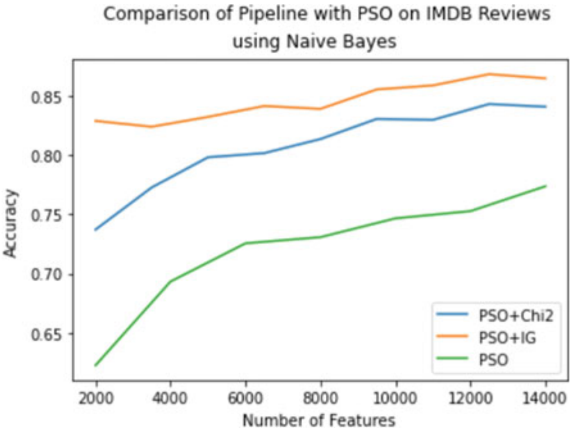
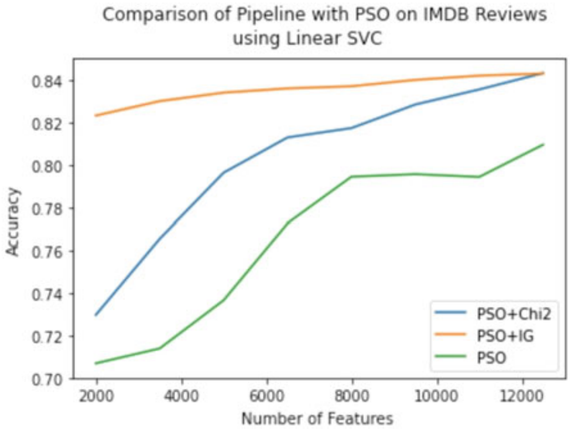


Fig. 13 Comparison of pipeline with PSO on IMDB reviews using linear SVC



5 Conclusion

This study explores the prospect of two-stage feature selection methods and whether they perform better in terms of classification accuracy than their respective wrapper algorithms. To test our hypothesis, we created four pipeline-based models namely GA with Chi2 and IG, PSO with Chi2 and IG and compared their performance with GA and PSO, respectively. The comparisons were based on the performance of models on two classifiers Naive Bayes and linear SVC and two data sets 20 newsgroups and IMDB reviews. After experimentations, we observe that the pipeline-based feature selection methods give better accuracy than individual wrapper methods. While the pipelines consistently perform better, the difference is much more significant in case of PSO-based pipelines. We believe that deep learning can give fresh insights into the pipeline-based feature selection methods, which is the future scope of this research.

References

1. Text mining from Wikipedia, the Free Encyclopedia
2. Joachims, T.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning. Nashville, Tennessee, USA, 1997
3. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML ‘97: Proceedings of the Fourteenth International Conference on Machine Learning (1997)
4. Chantar, H.K., Corne, D.W.: Feature subset selection for Arabic document categorization using BPSO-KNN. Third World Congress on Nature and Biologically Inspired Computing, 2011
5. Zahran, B.M., Kanaan, G.: Text feature selection using particle swarm optimization algorithm. World Appl. Sci. J. **7**(Special Issue of Computer & IT) (2009)
6. Uguz, H.: A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowl.-Based Syst. (2011)
7. Gunal, S.: Hybrid feature selection for text classification. Turk. J. Electr. Eng. Comput. Sci. (2012)
8. Tan, F., Fu, X., Zhang, Y., Bourgeois, A.G.: A Genetic Algorithm-Based Method for Feature Subset Selection. Soft Computing- Springer, 2008
9. Tsai, C.F., Chen, Z.Y., Ke, S.W.: Evolutionary instance selection for text classification. J. Syst. Softw. (2014)
10. Ghareb, A.S., Bakar, A.A., Hamdan, A.R.: Hybrid feature selection based on enhanced genetic algorithms for text categorization. Expert Syst. Appl. (2016)
11. Goldberg, D.E.: Genetic Algorithm in Search, Optimization and Machine Learning. Addison-Wesley, Reading (1989)
12. Kennedy, J., Eberhart, R. C.: Particle swarm optimization. In: Proceedings of IEEE ICNN (Perth, Australia). IEEE Press (1995)
13. Aghdam, M.H., Heidari, S.: Feature selection using particle swarm optimization in text categorization. J. Artif. Intell. Soft Comput. Res. **5**(4) (2015)
14. Xie, L.J., Lei, J., Xie, W., Gao, X., Shi, Y., Liu, X.: Novel hybrid feature selection algorithms for diagnosing erythemato-squamous diseases. In: International Conference on Health Information Science (2012)
15. Lei, S.: A feature selection method based on information gain and genetic algorithm. In: International Conference on Computer Science and Electronics Engineering (2012)

16. Zhai, Y., Song, W., Liu, X., Liu, L., Zhao, X.: A chi-square statistics-based feature selection method in text classification. In: IEEE 9th International Conference on Software Engineering and Service Science (ICSESS) (2018)
17. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM (1975)
18. Porter, M.: An algorithm for suffix stripping. Program: electronic library and information systems, 1980
19. Vieira, S.M., Mendonca, L.F., Farinha, G.F., Sousa, J.M.C.: Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Appl. Soft Comput. (2014)
20. Zhang, H., Li, D.: Naive bayes text classifier. In: IEEE International Conference on Granular Computing, 2007
21. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intell. Syst. Appl. **13**(4) (1998)

A Systematic Approach of Analysing Network Traffic Using Packet Sniffing with Scapy Framework

S. H. Brahmanand, N. Dayanand Lal, D. S. Sahana, G. S. Nijguna, and Parikshith Nayak

Abstract Many professionals and network engineers face many issues where the available resources have been very unsuccessful in solving the problem. Tools which are available will provide better outcomes; however, it generates large amount of information chunks that requires time and effort to remove those unwanted data. Therefore, learning some scripting often helps to solve problems, analysis and carry on with automation helps to save time, expense and effort. This research work discusses the development in packet sniffer through Python along with Scapy framework. Packet sniffer is a network traffic and data interception, tracking, and analysis software tool. Including a sniffer research work shed light on packet elements, packet layers, designing elements, dissection and sniffing along with further exploration.

Keywords Packet sniffer · Scapy framework · Protocols: TCP · UDP · ICMP

1 Introduction

Computer networks have continued to expand in size and functionality which depend on the number of users over the past five decades. With the technology development, the network traffic and data are increased, it is essential to maintain the network into smooth and economically efficient by controlling, maintaining and monitoring

S. H. Brahmanand · N. D. Lal · D. S. Sahana (✉)
Department of Computer Science and Engineering, GITAM School of Technology, Bengaluru, Karnataka, India
e-mail: ssanthos@gitaml.edu

S. H. Brahmanand
e-mail: bsavadat@gitam.edu

G. S. Nijguna
Department of Information Science and Engineering, S E A College of Engineering and Technology, Bengaluru, Karnataka, India

P. Nayak
Department of Computer Science and Engineering, GITAM, Deemed To Be University, Bengaluru, Karnataka, India
e-mail: nparikshh@gitam.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_60

the network. A packet sniffer is used for this purpose [1]. The Internet and its innovative applications connect every individual and organization. Every day, computer networks' strength, flexibility increases; however, it is difficult for the administrator to protect network operations.

In a modern, complex network, monitoring and measurement of the network have become increasingly important. Also, the administrators could control limited devices due to bandwidth limitations in wired and wireless networks. It is essential to monitor the network data flow and traffic so that efficiency of the network could be maintained. Also, it is essential to monitor the network breaches. So administrators must be specialized in monitoring and analysis then only reliability of the network can be maintained by avoiding network failures. Packet sniffing is the process of identifying the movement of each packet in the network. If the information belongs to other network users is sniffed by a user, then it is termed as packet sniffing.

Depends on the user and application preferences, packet sniffing can be used either as administrative tool or for malicious operations. They will be used for network traffic management and to validate by network administrators. Packet sniffers are, fundamentally, softwares [2]. The world's Internet penetration rate is accelerating at a significant rate. It is both extremely difficult and important to analyse and evaluate network traffic [3]. Cyber-attacks are growing and have emerged as a major challenge for coordinating diverse backgrounds in parallel with the growth in Internet penetration rates. Therefore, to avoid these accidents, information technology engineers and specialists have a sensitive responsibility.

Design is the most difficult-after capability for data protection experts and for the networks. Production activities, software build can be attained through scripting. Packet sniffers for network and security professionals are extremely useful tool. Sniffers allow network packets to be analysed, filtered and monitored by security professionals. In this research work, a packet sniffer model is presented that sniffs the packets based on TCP, UDP and ICMP protocols. By relating the statistics, the growth of Internet and its penetration level can be examined [4].

Scapy is a shell command-based interpreter, and it is not a programme. The Python loop interpreter is used to evaluate the instances, classes and functions. In Scapy, class instances are used as packet-specific algorithms. Whenever an object is instantiated, i.e. a package creation and manipulation refer to changing the parameters or invoking the instance object process. By introducing packets as artefacts, single line code is sufficient for a packet. Whereas it could take several lines in C code. Layers are the basic elements in a packet and multiple packets are added to generate a complete packet. Layering facilitates realistic representation and execution of code [5]. Group of packets can be identified in Scapy, and packets responses such as how the packets are sent, sniffed, the method of showing tables and pairs are always the same and are supported by Scapy. Scapy's brief work as an interactive packet manipulation tool with steps is illustrated in Fig. 1.

Steps:

1. Build packets by taking the data type into account, MAC address of source and MAC address of destination.

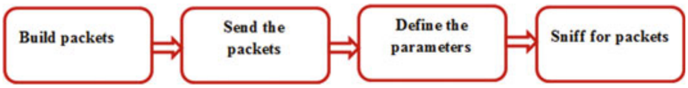


Fig. 1 Scapy's work flow

2. Send the packets that were generated to the destination by specification of parameters.
3. Defining the parameters.
4. Sniff for packets.

The focus of this research work is to discuss the features of packer sniffer methods developed in Python and Scapy, and by using layer composition, all packets are identified. All the incoming and outgoing packets from interfaces are sniffed by the host machines. Based on TCP, UDP and ICMP protocols, packets will be categorised. They are split on every protocol classification as outgoing packets and incoming packets. Source and destination IP, Port, time stamp, geolocation, etc. details are extracted on the device. There is no GUI interface for the packet sniffer, and it is executed from the command console. The primary objective of this research work is to develop a packet sniffer model so that it could assist some professionals in information security. These devices are extremely helpful for security professionals. Some of the main topics that this paper will discuss are:

- Packet sniffers over view and their use cases
- Packet sniffer creation with sequential steps on a characteristics in comparison.

2 Literature Survey

According to Qadeer et al. [6], they have been working on the behaviour of current sniffer applications like Wireshark, tcpdump and snort. All of these programmes offer various characteristics and shortcomings. The writers offer an overview of the packet sniffers that only provide data logs must be analysed by a network administrator in order to detect an error or a network adapter attack. Present systems are only capable of viewing packet logs. They are not up to date with the new network technology generation and are thus inadequate. Excessive processing time is the major limitation of protocol-based analysis.

Authors Dabir et al. [7] illustrates about the bottlenecks associated with packet captured from local area networks (LANs) and without information loss using commodity hardware. Using the Wireshark packet sniffer, tests were carried out to write captured packets directly to disc on a fast Ethernet network with different test set-ups. These experiments involved the processing of large packets at a rate almost linear. They also experimented with different sizes of the kernel level buffer associated with the packet capture socket. Also, for the capturing application, user-level buffers are used as a multithread architecture so that experiments are progressed to obtain the solution. Experimental results demonstrate that increased buffer dramatically boost capture efficiency in the application level or at the kernel.

Splanger [8] describes packet sniffing as an innovative, technical tool for local networks and Internet connection which need to be analysed, debugged, managed and monitored. It collects the information that passes through dialup link or Ethernet network card, analyses this information and project into a simple manner so that anybody can understand. Any security experts, network managers, network application developers and those who required details of traffic flow in the network can use Soft Ideal Network Protocol Analyser. The results of Soft Perfect Network Protocol Analyser and its network analysis are simple and comfortable. It also helps to defray network packets and reassemble them into streams.

Patel et al. [9] describes a packet tapping method where the model focuses on each packets passing through the network, and it is treated as a packet sniffing. The other network user data can be sniffed using the proposed technique, and it is successful for non-switched and switched networks. It is based on the intent of the user. This research work explores the sniffing process in switched network and hub and also discussed the various sniffing methods in detail used by Anti-Sniff to detect these sniffing programmes.

3 Methodology

3.1 Packet Sniffer Development

The flow of data in the network is monitored and recorded using packet sniffing process. Each individual packet is analysed through predetermined metrics so that thorough tracking of networks and bandwidth measurements can be made in the packet sniffing process. However, in order to be able to understand the importance of the data being tracked, it needs greater awareness of networks and their internal functions [10]. In two very distinct criteria, a packet sniffer may typically be set-up:

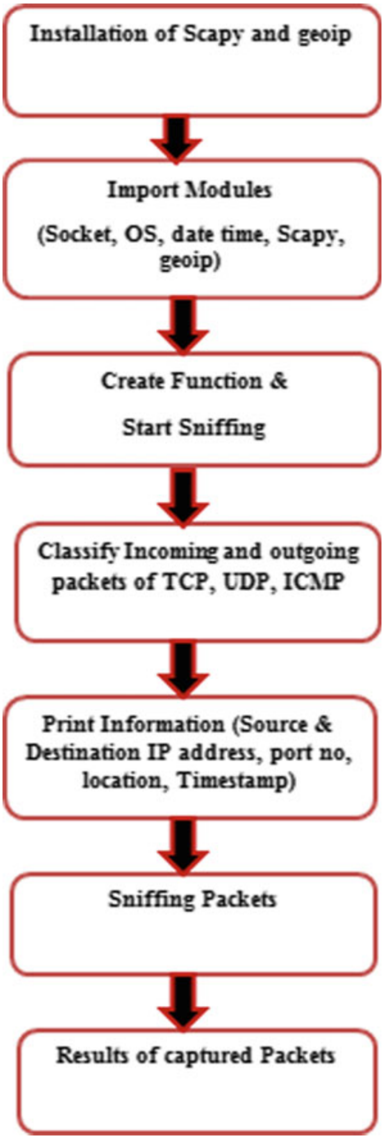
1. **Criteria of Unfiltered Packets**—This configuration catches all packets created throughout the network
2. **Criteria of filtered packets**—This configuration catches only certain packets; they tend to contain specific versions of elements of the data [11].

3.2 Packet Sniffer Scripting on Python and Scapy

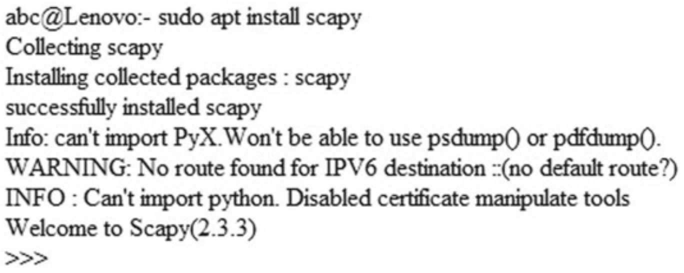
Figure 2 illustrates the procedure of packet sniffer using Scapy framework with steps. Scapy installations on operating system of Linux have been used to know the sniffing of packets over the network.

Step 1: Installing Scapy

Fig. 2 Procedure of packet sniffing using python and Scapy framework


Installation of Scapy and geoip
Import Modules
(Socket, OS, date time, Scapy, geoip)
Create Function &
Start Sniffing
Classify Incoming and outgoing packets of TCP, UDP, ICMP
Print Information (Source & Destination IP address, port no, location, Timestamp)
Sniffing Packets
Results of captured Packets

To do sniffing, several tools can be used, but most of them only have fixed features. Scapy is different: it can be used to create other sniffing instruments not only as a tool, but also as a building block, i.e. can do with incorporating the Scapy features into our own software. Usage of Scapy has been done for a set of assignments. Scapy is installed for Python3 in the current version of the SEED VM. To install Scapy for Python3, we can use the following instruction which is shown in Fig. 3.



```
abc@Lenovo:- sudo apt install scapy
Collecting scapy
Installing collected packages : scapy
successfully installed scapy
Info: can't import PyX.Won't be able to use psdump() or pdfdump().
WARNING: No route found for IPV6 destination ::(no default route?)
INFO : Can't import python. Disabled certificate manipulate tools
Welcome to Scapy(2.3.3)
>>>
```

Fig. 3 Installation of Scapy

Step 2: Install Geoip using command as shown in Fig. 4, since it is not pre-installed in python.

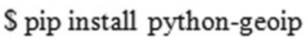
Step 3: Import all the modules such as socket, os, scapy, geoip and time data under python file as shown in Fig. 5.

Step 4: Once the necessary modules are imported, a function is created using python as shown in Fig. 6 to initialize the packet sniffing. In order to sniff the packets continuously, prn parameter is used and its count value is set into 1 so that single packet gets sniffed at a time.

Step 5: In order to verify whether a packet has protocol layers, Scapy has a built-in feature to identify packets, i.e. layer (TCP) or packet has layer (UDP) or other Scapy-supported protocols. Packet grouping into incoming and outgoing packets. There are ipv4 network addresses used on ISP. Depending on the location, the incoming and outgoing packets can be classified for TCP, UDP and ICMP and it is depicted in Figs. 7, 8 and 9, respectively, and sniffing is illustrated in Fig. 10.

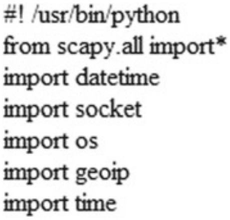
1. **Classifying packets into TCP and to know incoming and outgoing packets**
2. **Classifying packets into UDP and to know incoming and outgoing packets**
3. **Classifying packets into ICMP**
4. **Sniffing of packet**

Fig. 4 Geoip Installation



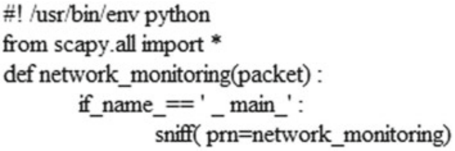
```
$ pip install python-geoip
```

Fig. 5 Importing modules

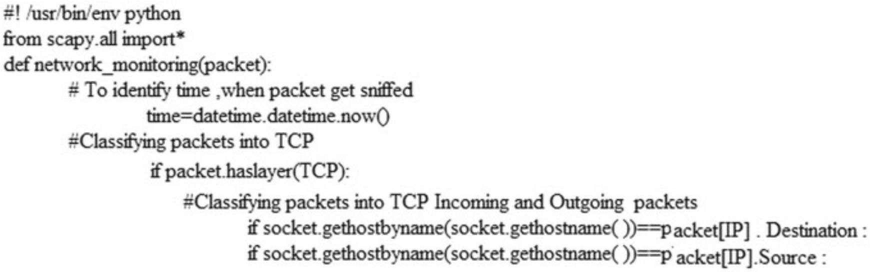


```
#! /usr/bin/python
from scapy.all import*
import datetime
import socket
import os
import geoip
import time
```

Fig. 6 Initializing packet sniffing

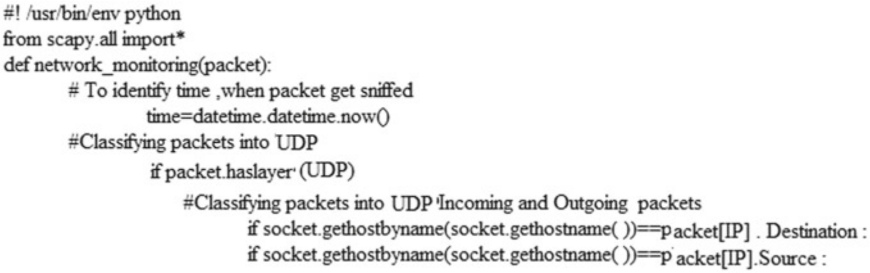


```
#! /usr/bin/env python
from scapy.all import *
def network_monitoring(packet) :
        if_name_== ' _ main_' :
                sniff( prn=network_monitoring)
```



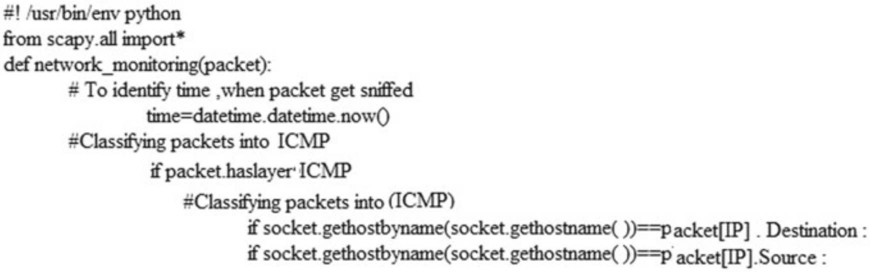
```
#! /usr/bin/env python
from scapy.all import*
def network_monitoring(packet):
        # To identify time ,when packet get sniffed
                time=datetime.datetime.now()
        #Classifying packets into TCP
                if packet.haslayer(TCP):
                    #Classifying packets into TCP Incoming and Outgoing  packets
                            if socket.gethostbyname(socket.gethostname( ))==packet[IP] . Destination :
                            if socket.gethostbyname(socket.gethostname( ))==packet[IP].Source :
```

Fig. 7 Incoming and outgoing packets classification function for TCP



```
#! /usr/bin/env python
from scapy.all import*
def network_monitoring(packet):
        # To identify time ,when packet get sniffed
                time=datetime.datetime.now()
        #Classifying packets into UDP
                if packet.haslayer' (UDP)
                    #Classifying packets into UDP 'Incoming and Outgoing  packets
                            if socket.gethostbyname(socket.gethostname( ))==packet[IP] . Destination :
                            if socket.gethostbyname(socket.gethostname( ))==packet[IP].Source :
```

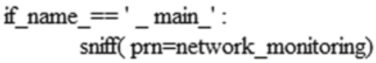
Fig. 8 Incoming and outgoing packets classification function for UDP



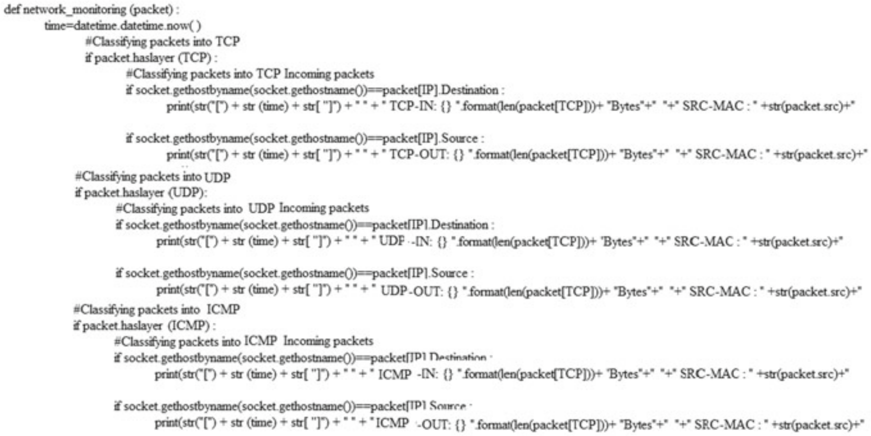
```
#! /usr/bin/env python
from scapy.all import*
def network_monitoring(packet):
        # To identify time ,when packet get sniffed
                time=datetime.datetime.now()
        #Classifying packets into  ICMP
                if packet.haslayer' ICMP
                    #Classifying packets into (ICMP)
                            if socket.gethostbyname(socket.gethostname( ))==packet[IP] . Destination :
                            if socket.gethostbyname(socket.gethostname( ))==packet[IP].Source :
```

Fig. 9 Incoming and outgoing packets classification function for ICMP

Fig. 10 Sniffing of packets



```
if_name_== ' _ main_' :
        sniff( prn=network_monitoring)
```



```
def network_monitoring (packet) :
        time=datetime.datetime.now( )
                #Classifying packets into TCP
                if packet.haslayer (TCP) :
                        #Classifying packets into TCP Incoming packets
                        if socket.gethostbyname(socket.gethostname())==packet[IP].Destination :
                                print(str("[") + str (time) + str[ "]") + " " + " TCP-IN: {} ".format(len(packet[TCP]))+ "Bytes"+"  "+" SRC-MAC : " +str(packet.src)+"

                        if socket.gethostbyname(socket.gethostname())==packet[IP].Source :
                                print(str("[") + str (time) + str[ "]") + " " + " TCP-OUT: {} ".format(len(packet[TCP]))+ "Bytes"+"  "+" SRC-MAC : " +str(packet.src)+"
              #Classifying packets into UDP
              if packet.haslayer (UDP):
                      #Classifying packets into  UDP Incoming packets
                      if socket.gethostbyname(socket.gethostname())==packet[IP].Destination :
                              print(str("[") + str (time) + str[ "]") + " " + " UDP -IN: {} ".format(len(packet[TCP]))+ 'Bytes"+"  "+" SRC-MAC : " +str(packet.src)+"

                      if socket.gethostbyname(socket.gethostname())==packet[IP].Source :
                              print(str("[") + str (time) + str[ "]") + " " + " UDP-OUT: {} ".format(len(packet[TCP]))+ "Bytes"+"  "+" SRC-MAC : " +str(packet.src)+"
              #Classifying packets into  ICMP
              if packet.haslayer  (ICMP) :
                      #Classifying packets into ICMP  Incoming packets
                      if socket.gethostbyname(socket.gethostname())==packet[IP].Destination :
                              print(str("[") + str (time) + str[ "]") + " " + " ICMP  -IN: {} ".format(len(packet[TCP]))+ 'Bytes"+"  "+" SRC-MAC : " +str(packet.src)+"

                      if socket.gethostbyname(socket.gethostname())==packet[IP].Source :
                              print(str("[") + str (time) + str[ "]") + " " + "ICMP  -OUT: {} ".format(len(packet[TCP]))+ "Bytes"+"  "+" SRC-MAC : " +str(packet.src)+"
```

Fig. 11 Sample showing the printing of required values from packet

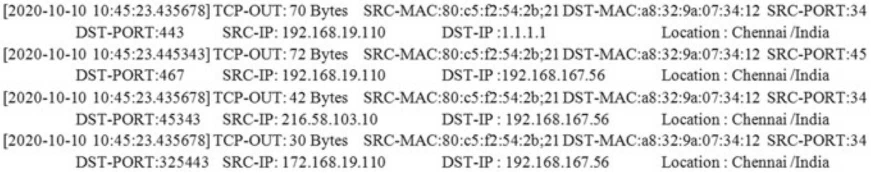
Step 6: Last step is carried out to take up the challenge of packet sniffer development which provides all the necessary information such as source and destination IP, Port, time stamp and geolocation as shown in Fig. 11.

Step 7: Packet Sniffing: The process is completed in this stage and packets are sniffed using the sudo code in python with the respective filename with extension of ‘py’.

4 Results

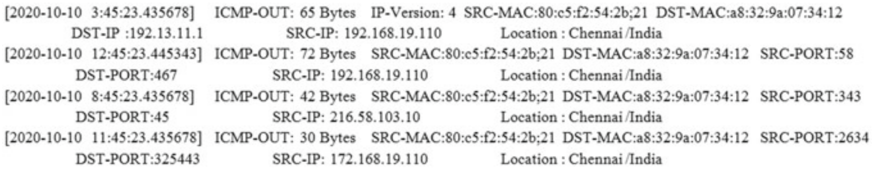
Case 1: Incoming and outgoing TCP and UDP packets are captured and depicted in Fig. 12.

Case 2: Capturing ICMP incoming and outgoing packets by pinging is shown in Fig. 13.



```
[2020-10-10 10:45:23.435678] TCP-OUT: 70 Bytes   SRC-MAC:80:c5:f2:54:2b;21 DST-MAC:a8:32:9a:07:34:12  SRC-PORT:34
          DST-PORT:443       SRC-IP: 192.168.19.110          DST-IP :1.1.1.1                  Location : Chennai /India
[2020-10-10 10:45:23.445343] TCP-OUT: 72 Bytes   SRC-MAC:80:c5:f2:54:2b;21 DST-MAC:a8:32:9a:07:34:12  SRC-PORT:45
          DST-PORT:467       SRC-IP: 192.168.19.110          DST-IP :192.168.167.56           Location : Chennai /India
[2020-10-10 10:45:23.435678] TCP-OUT: 42 Bytes   SRC-MAC:80:c5:f2:54:2b;21 DST-MAC:a8:32:9a:07:34:12  SRC-PORT:34
          DST-PORT:45343     SRC-IP: 216.58.103.10           DST-IP : 192.168.167.56          Location : Chennai /India
[2020-10-10 10:45:23.435678] TCP-OUT: 30 Bytes   SRC-MAC:80:c5:f2:54:2b;21 DST-MAC:a8:32:9a:07:34:12  SRC-PORT:34
          DST-PORT:325443    SRC-IP: 172.168.19.110          DST-IP : 192.168.167.56          Location : Chennai /India
```

Fig. 12 Incoming and outgoing TCP and UDP packets



```
[2020-10-10  3:45:23.435678]   ICMP-OUT: 65 Bytes   IP-Version: 4  SRC-MAC:80:c5:f2:54:2b;21  DST-MAC:a8:32:9a:07:34:12
         DST-IP :192.13.11.1              SRC-IP: 192.168.19.110          Location : Chennai /India
[2020-10-10  12:45:23.445343]  ICMP-OUT: 72 Bytes   SRC-MAC:80:c5:f2:54:2b;21  DST-MAC:a8:32:9a:07:34:12  SRC-PORT:58
          DST-PORT:467               SRC-IP: 192.168.19.110           Location : Chennai /India
[2020-10-10  8:45:23.435678]   ICMP-OUT: 42 Bytes   SRC-MAC:80:c5:f2:54:2b;21  DST-MAC:a8:32:9a:07:34:12  SRC-PORT:343
          DST-PORT:45                SRC-IP: 216.58.103.10            Location : Chennai /India
[2020-10-10  11:45:23.435678]  ICMP-OUT: 30 Bytes   SRC-MAC:80:c5:f2:54:2b;21  DST-MAC:a8:32:9a:07:34:12  SRC-PORT:2634
          DST-PORT:325443            SRC-IP: 172.168.19.110           Location : Chennai /India
```

Fig. 13 Console showing ICMP incoming and outgoing packets

5 Conclusion

A technique to detect packets by way of packet sniffing is proposed in this paper. The packet sniffer enables all site visitors passing through its group link to be monitored and analysed by the computer network. It merely makes a copy of each packet flowing via the network interface and finds the packets' Ethernet addresses for supply and destination. It is used by many computer administrators or community administrators to track and troubleshoot community visitors. The packet sniffer developed here can able to track the packets such as UPD, TCP and ICMP. In each part, the source and destination IP, Port, time stamp and geolocation are printed. The developed script was helpful in collecting packets from the terminal. For information security and network professionals, this tool is extremely useful for capturing packets by running a small script.

References

1. Asrodia, P., Patel, H.: Network Traffic Analysis Using Packet Sniffer
2. Ailawadhi, A., Bhandari, A.: Literature Review on an Approach to Detect Packets Using Packet Sniffing
3. Bhandari, A., Gautam, S., Koirala, T., Islam, M.: Packet Sniffing and Network Traffic Analysis Using TCP—A New Approach (2018). https://doi.org/10.1007/978-981-10-4765-7_28
4. Siswanto, A., Syukur, A., Kadir, E.A.: Network Traffic Monitoring and Analysis Using Packet Sniffer (2019). https://doi.org/10.1109/COMMNET.2019.8742369
5. Rohith, R., Moharir, M., Shobha, G.:SCAPY—a powerful interactive packet manipulation program. In: 2018 International Conference on Networking, Embedded and Wireless Systems (ICNEWS), Bangalore, India, 2018, pp. 1–5. https://doi.org/10.1109/ICNEWS.2018.8903954
6. Qadeer, M.A., Zahid, M., Iqbal, A., Siddiqui, M.R.: Network traffic analysis and intrusion detection using packet sniffer. In: ICCSN 10 Second International Conference, pp. 313–317 (2010)
7. Dabir, A., Matrawy, A.: Bottleneck Analysis of Traffic Monitoring using Wireshark, pp. 158–162 (2007). https://doi.org/10.1109/IIT.2007.4430446
8. Splanger, R.: Packet sniffing detection with Anti Sniff. University of Wisconsin-Whitewater, May 2003
9. Patel, N., Patel, R., Patel, D.: Packet sniffing: network wiretapping. In: IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6–7 Mar 2009, pp. 2691–2696 (2009)

10. https://en.wikipedia.org/wiki/PRTG_Network_Monitor
11. https://computer.howstuffworks.com/firewall1.html

Detecting Ransomware Attacks Distribution Through Phishing URLs Using Machine Learning

B. N. Chaithanya and S. H. Brahmananda

Abstract During the last few years, phishing and ransomware seem to be the most widespread type of threat and a rapidly growing hazard to organizations. Nowadays, ransomware is injected via phishing emails, poorly informed users appear without much thought to click on links and attachments, believing that the emails are valid hence any defence action must concentrate on e-mail security. Ransomware is a malicious code or software that encrypts files in the victim's system in a simplified way, and the perpetrators are demanding a lot of ransom for it. Ransomware can inflict enormous harm to companies, resulting in a lack of production and sometimes economic difficulties. Most importantly, there is a lack of documents that may reflect hundreds of hours of work or customer information that is important to the effective operation of the organization. Traditional defence methods that depend on malware signatures and fundamental protection rules have been found to be ineffective towards ransomware threats. In reality, attackers are developing their malware to compromise conventional Web and email security set-ups that are vulnerable. A thorough review of the organization's defensive measures should address the ransomware problem and learn how they are truly capable of reacting to the current threats. The paper goal is to see how using this approach helps to prevent malicious malware and uses it as a self-defence tool through machine learning techniques. Research was done on several uniform resource locator (URL) samples and the findings show that we can say the difference between malicious and benign URL.

Keywords Malware attacks · Ransomware attacks · Hacking · Emails · Phishing Emails

B. N. Chaithanya (✉) · S. H. Brahmananda
Department of CSE, GITAM School of Technology, Doddaballapura, India
e-mail: cnagaraj@gitam.edu

S. H. Brahmananda
e-mail: bsavadat@gitam.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_61

1 Introduction

In modern days, emails are used, which is a way to connect. Although, there are still several cases wherein the email can be misused or attacking the victim becomes a benefit for the attack. [1] Hacking, phishing, malware, spam, ransomware are some of the prevalent threats from emails. In 2019, ransomware assaults made a massive comeback, but attackers moved from corporate targets to SMBs, unlike in past years. In 2019, one out of five SMBs was struck by a ransomware threat, according to Datto [2]. Although several authorities remain strict regarding the origin of threats, many have acknowledged that 67% of ransomware attacks were triggered by employees clicking on malicious links [3]. Understanding the strategies used by assailants to distribute ransomware is crucial to better preventing this threat. Figure 1 explains the most common approaches for ransomware attacks, which are: silent exploit kit infections, email phishing attachments, phishing emails links and vulnerabilities in software.

1.1 Medium of Attacks to the End-Users

Identifying how ransomware evolves is essential to stopping victims from attacking. Ransomware may use a variety of attack vectors to target machines or servers. These are the four main ways that ransomware can invade its victims.

1.1.1 Phishing Emails

Phishing emails are easy to create. Creating a phishing email does not necessitate a high level of skill. Hackers imitate the company's feel and look by using advertising

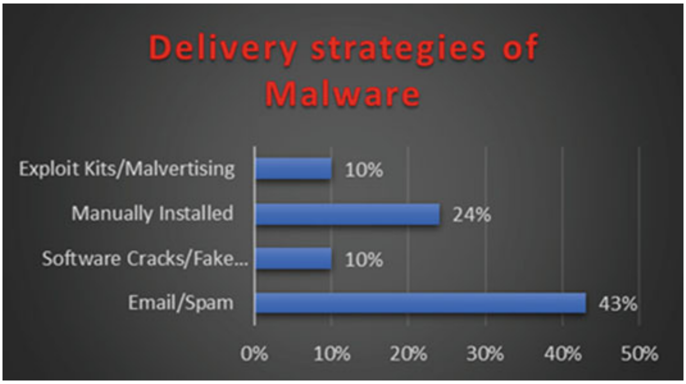


Fig. 1 Delivery strategies for malware

pictures and logos from the real brand’s website or Google Images to create the appearance of authenticity. Hackers may use display name spoofing to create a false email address by applying their chosen name to any email address. Invoice phishing is the most popular way of spreading ransomware-laden attachments. The client takes the assumption that an invoice has been submitted by a colleague or organization. At the time of clicking, the link in the attachment installs the malware. In many other instances, the ransomware update starts automatically while the attachment is accessed, often by Word documents and PDF macro or fraudulent scripts in.zip files. Phishing kits are bought online to make it easier for the attacker. A typical phishing kit includes all of the phishing attacks needed to make the web page appear legitimate and avoid the identification, including a false web page and tools. Some kits also define goals, create phishing emails and collect information. Attackers have a variety of tools to bypass an email firewall, all of which are generally online. A URL shortener, Bitly, may be used to create a phished URL alias to check for blocked URLs by exploiting content.

Attackers use thoroughly designed phishing emails to direct the user to open a file or click on a malicious link. The file can be delivered in many formats, including a PDF, ZIP, Word, JavaScript, or other forms. When an attacker opens a Word document, the user is mostly tricked into “Activating Macros”. It causes a code that installs and runs a malicious executable file (EXE) via an external Web server to be executed by the intruder. The EXE will provide the requisite functions for encrypted data on the computer of the suspect, as shown in Fig. 2. When the files are encrypted

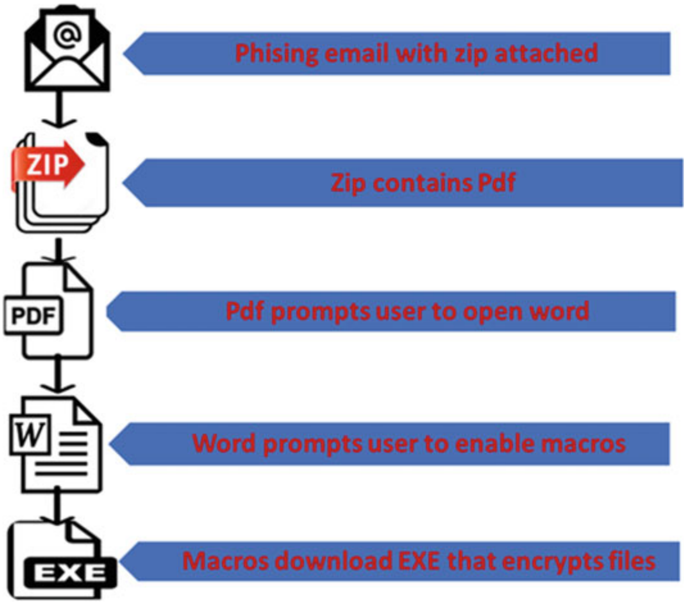

Phising email with zip attached
ZIP
Zip contains Pdf
PDF
Pdf prompts user to open word
W
Word prompts user to enable macros
EXE
Macros download EXE that encrypts files

Fig. 2 How phishing email works

and ransomware has a significant presence on one computer, the more sophisticated ransomware versions are expanded to other network computers. It needs to open a vulnerable attachment in a phishing message for an individual and compromise a whole organization. Common phishing ransomware leveraging victims involves Locky, Cerber and Nemucod.

1.1.2 Remote Desktop Protocol

Remote desktop protocol (RDP) is an increasingly common method for hackers to attack victims. Remote desktop protocol was developed to allow IT administrators to access or use the computer remotely safely. Currently, RDP runs across port 3389. While opening computer doors for legitimate use has many advantages, it also gives a bad actor the chance to use it for illicit purposes. In 2017, over 10 million computers were reported as having port 3389 available on the public internet, i.e. they run RDP more than 3389 computers. Hackers may identify computers that are prone to contamination on search engines such as Shodan.io. Hackers typically have access by brute-forcing the password to log in as an admin after target machinery has been identified—open-source software for password cracking helps to attain this aim. Cybercriminals can easily and instantly obtain many passwords using standard software, including Caín and Able, John the Ripper and Medusa. Hackers will launch a ransomware encryption operation while they are in as an administrator and monitor the system. Some hackers disable protection endpoint software on the computer or uninstall Windows file backups until ransomware is operating to build further losses. This gives the user even more reason to pay the demand, as there will no longer be Windows recovery options. Common RDP ransomware leveraging victims involves SamSam, LowLevel04 and CrySis.

1.1.3 Drive-By Downloads

Another way attackers use ransomware is by so-called drive-by downloads from a compromised website. Once you visit a compromised website, you can unknowingly download malicious programs. Attackers also start drive-by downloads by using established vulnerabilities in legal website applications. These vulnerabilities are exploited to inject malicious code into a website or to redirect the user to another website that they operate. Exploit kits allow hackers to search the visiting computer for its unique vulnerabilities silently and, where identified, execute background code without the user clicking. Suddenly, a ransom note will confront the virus's unsuspected user warning and request.

1.1.4 USB and Removable Media

Another way to enter a network is through a USB computer ransomware. The USB drives disguised as a Netflix advertising program, then implemented ransomware on the naive user's device until accessed. The strong Spora ransomware has also introduced the ability to duplicate itself (in secret file formats) on USB and removable media drives, risking subsequent computers where the USB unit plugged in payment for returned files. Common drive-by downloads ransomware leveraging victims involves CryptoWall, PrincessLocker, CryptXXX.

Ransomware has been the go-to assault of choice to earn capital for cybercriminals. It is easy to buy from ransomware-as-a-service (RaaS) on the dark network, and attacks are pretty easy to initiate through one of the methods above. Companies must understand how their networks can be targeted and proactively protect themselves and ensure their business operation quality through a layered security strategy.

1.2 Preventive Methodologies

In email filtration, the two widely used approaches are information engineering and machine learning. In the information engineering approach, a set of rules must be established from which emails are labelled as malicious or not. The operator of the filter, or some other authority, should decide on a group of such procedures. No positive benefits mean that the laws must be regularly updated and followed, which is a waste of resources and effort and difficult for most people to do. The machine learning approach is more powerful than the information engineering approach, and no guidelines are needed [4]. These samples are a collection of pre-classified email addresses rather than a sequence of training samples. The classification rules for these emails can then be deduced using a specific process. The approach of machine learning has been extensively investigated, and a variety of approaches could be used in email filtering.

2 Related Work

Shad and Sharma [5] identified phishing sites using multiple machine learning techniques and then evaluated the performance of the various methods. Their experimental findings showed that the random forest algorithm has the best precision, recall and accuracy. Their research further highlights the critical attributes used for the identification of phishing.

Sönmez et al. [6] proposed a method to classify phishing attacks. Focussed on the UCI Irvine ML database, attribute extraction was performed from various sites. The analysis was performed using MATLAB, and it was observed that the intense learning algorithm had the maximum accuracy of 95.39%.

In 2019, Williams et al. [7] analysed how three of such elements that have not been extensively studied in earlier research affect email belief and clear choices, specifically utilizing misfortune and reward-based effect mechanisms, actual structure indications and connection to remarkable current advancements. Developed a technique in progress to describe a method for computer-based recognizing evidence of human effect standards of social engineering within phishing attacks [8].

The effects of pre-processing phases on the study of supervised spam classification techniques have been addressed, according to Ruskanda [9]. The following are two commonly used supervised spam classifier algorithms that were put to the test: NB and SVM were introduced, as well as THEMIS, a phishing email recognition model that is used to display emails at the email header, email body, character level and word level together at the same time. Based on the recent progress in investigations about machine learning for big data analytics and diverse approaches for multiple social implementations concerning actual processing circumstances. [10].

Authors in [11] identified and evaluated REDFISH, an algorithm of detection for ransomware's varieties that encrypt distributed masses of network data. The algorithm operates for a traffic clone, with no regular impact performance for consumers. It defines ransomware based on its specific actions like files read, compose and delete. They saw the algorithm parameters could be modified to identify 100 per cent ransomware with 19 different ransomware detections. The fake positive is exceptionally rare: one false alert opens, reads and writes 10 million documents in a real business situation of more than 4800 consumers who work for a full day and have access to more than 1500 connected volumes network. The conclusions are consistent with the study of the false positive rate model given. In 10 GB/s data traffic, the algorithm is implementable, using a minimal number of CPU cores, without affecting the host CPUs since they do not have any apps installed. The whole identification system is out of the development network, and no one will strike it. No one will hit the whole identifying system because it is no longer part of the development network. A form of malware that has the potential to deactivate it.

Attacks on windows and the android environment; Monika Zavarsky and Lindskog [12] have developed experiments that analyse the ransomware samples, which extensively attacks the android operating system and windows. It has been demonstrated that ransomware may be identified using Windows MD5 validation significance evaluations on any sample monitored and registry processes, as well as suspicious file structures. In Android, more significant consideration has been given to Android apps' privileges. In comparison, comprehensive research by Choi et al. [13]. Police, which focuses on social credibility, Web pages of television networks that displayed this way of life online, and certain significant considerations led to ransomware's victimization. It proposed that the education program was established to guide social media utilization programs for the public against ransomware assault.

3 Methodologies

3.1 System Model Design

Attack may happen through various ways, one of the methods is when the users who have not been conscious of the malicious link included in the email open it to the fraudulent site given by the assailants. In this model, the attributes of the website are collected using a word bag technique. The URL is split into three parts, such as protocol, domain and route, to extract information. These classes emphasise the important features in the given dataset that provides the classifier model with an input to identify a URL as a legitimate or malicious URL [14]. Machine learning techniques proved to be an effective tool to categorize malicious behaviours as spam emails or phishing sites. Many of these techniques would require training data, fortunately, there are indeed a number of phishing websites samples for the training of a machine learning model. Framework for malicious URL detection is shown in Fig. 3.

Efficient systems for identifying such malicious URLs in a timely basis will significantly help to address a vast amount and diversity of cyber threats. Machine learning methods, using a collection of URLs as training data and relying on mathematical properties, develop a predictive feature to identify the URL as suspicious, phishing or legitimate. When the training data is obtained, the next objective is to extract the descriptive features so that they adequately explain the URL and can be represented mathematically by machine learning techniques. For e.g. it may not be possible to learn a successful prediction model by using the URL string directly. Therefore, it

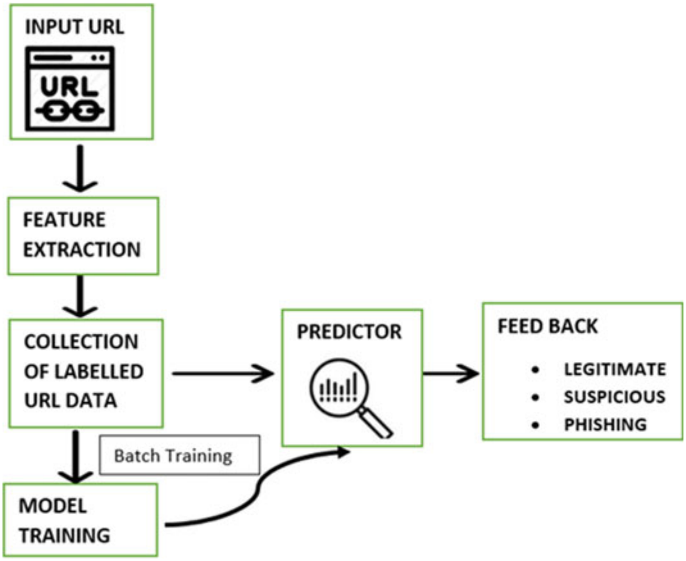


Fig. 3 Framework for malicious URL detection using machine learning

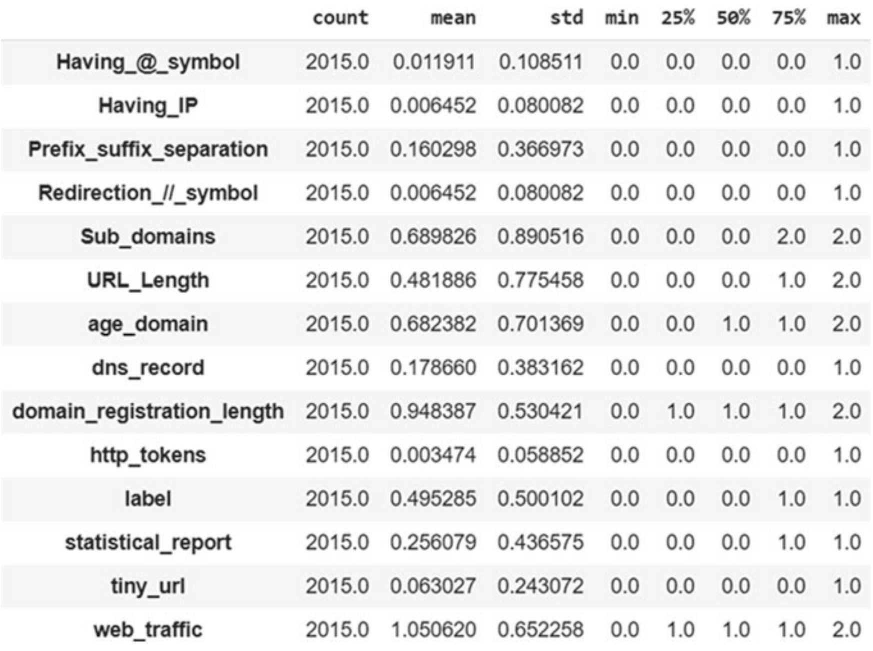
would be essential to extract proper functionality based on specific concepts or heuristics to achieve a better URL representation. These attributes, after being collected, must be translated into an appropriate format (e.g. a numerical vector) so that they are being incorporated for a machine learning approach for the training stage. The potential of such features to provide meaningful information is vital to successive machine learning. The fundamental principle of machine learning (classification) methods is that features, suspicious and legitimate URLs have distinct distributions. The consistency of the URL description function is crucial to the effectiveness of the process's malicious URL predictive model taught through machine learning. Using the training data with the required representation of the features, the next step in constructing the predictive models is the model's real training. Several different classifiers can be explicitly used over the training results. These methods aim to examine the URL's details and its related websites or websites, extract successful URL representations, and build a predictive model on the training of both suspicious and legitimate URLs. There are two kinds of features—static features and dynamic features which can be used. We evaluate a webpage in the static analysis based on knowledge accessible without executing the URL. The extracted features include lexical URL attributes, host descriptions, and often even HTML and JavaScript content. These strategies are better than the dynamic approaches since no executing is needed. The simple premise is that these functions are applied differently to suspicious and legitimate URLs. A prediction model can be developed with this distribution information, which can foresee new URLs. Implementing machine learning techniques has widely studied static analysis methods because of the comparatively safer atmosphere for collecting valuable information and the potential to generalize to all forms of threats.

3.2 DataSet

The dataset which we used for our studies is quite well analysed and evaluated by some research teams. The dataset from Kaggle comprises more than 10,000 sample websites, 15 percent of which were utilised in the testing stage. Description table of dataset and Correlation of features in datasets are shown in Figs. 4 and 5, respectively.

4 Results and Discussion

Identification of phishing is a classification problem hence we have to use a binary classification algorithm, considering "1" as a phishing and "0" as a legitimate. For the identification of phishing sites, we used machine learning techniques in our research, which are decision tree, support vector machine, random forest, neural networks, and XGBoost. We assess the precision, accuracy, recall, F1 score and confusion matrix of these modelling techniques and use various feature selection approaches to get the best results. Recall evaluates the number of phishing URLs the



	count	mean	std	min	25%	50%	75%	max
Having_@_symbol	2015.0	0.011911	0.108511	0.0	0.0	0.0	0.0	1.0
Having_IP	2015.0	0.006452	0.080082	0.0	0.0	0.0	0.0	1.0
Prefix_suffix_separation	2015.0	0.160298	0.366973	0.0	0.0	0.0	0.0	1.0
Redirection_//_symbol	2015.0	0.006452	0.080082	0.0	0.0	0.0	0.0	1.0
Sub_domains	2015.0	0.689826	0.890516	0.0	0.0	0.0	2.0	2.0
URL_Length	2015.0	0.481886	0.775458	0.0	0.0	0.0	1.0	2.0
age_domain	2015.0	0.682382	0.701369	0.0	0.0	1.0	1.0	2.0
dns_record	2015.0	0.178660	0.383162	0.0	0.0	0.0	0.0	1.0
domain_registration_length	2015.0	0.948387	0.530421	0.0	1.0	1.0	1.0	2.0
http_tokens	2015.0	0.003474	0.058852	0.0	0.0	0.0	0.0	1.0
label	2015.0	0.495285	0.500102	0.0	0.0	0.0	1.0	1.0
statistical_report	2015.0	0.256079	0.436575	0.0	0.0	0.0	1.0	1.0
tiny_url	2015.0	0.063027	0.243072	0.0	0.0	0.0	0.0	1.0
web_traffic	2015.0	1.050620	0.652258	0.0	1.0	1.0	1.0	2.0

Fig. 4 Dataset description

model attempts to identify. Precision evaluates the extent to which phishing URLs observed are actually phishing. The F1 score is the mean value of precision and recall. The classification analysis visualisation shows the accuracy, recall, F1, and support scores of the method. Support is the number of real instances of the category in the dataset listed. Table 1 shows the classification report of decision tree, support vector machine, random forest, neural networks and XGBoost models. Among all random forest gives the 83% of accuracy and which proves to be efficient in detecting phishing URLs.

5 Conclusion

In this study, we carried out and analysed five classifications on the phishing Website dataset, consisting of 1612 legitimate URLs and 403 phishing URLs. The classifications studied include decision tree, support vector machine, random forest, neural networks and XGBoost. However, according to our finding in Table 1, we have very good output for random forest and XGBoost in terms of accuracy. The findings inspire future work to incorporate more functionality to the data collection, which will increase the accuracy of these models so that it could integrate machine learning models with other ransomware and phishing detection approaches in order to achieve

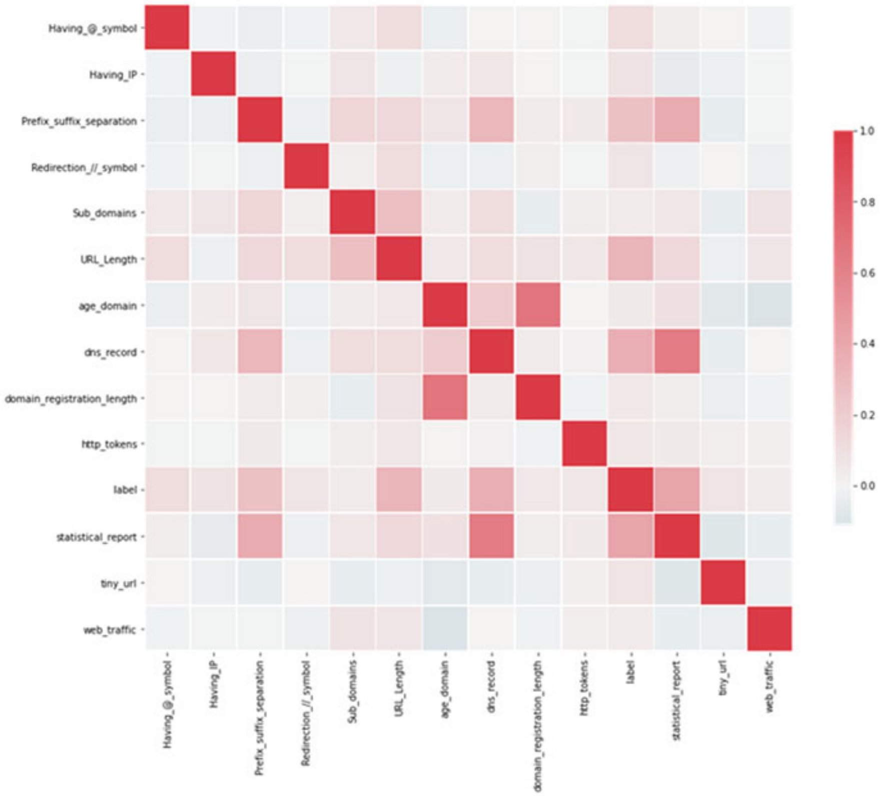
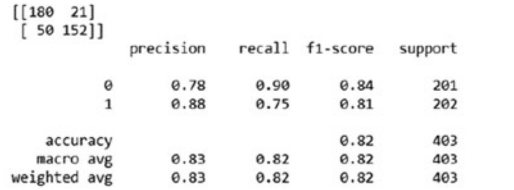


Fig. 5 Correlation of features in datasets

improved efficiency. In addition, we will discuss the possibility of proposing and developing a new mechanism to retrieve new functionality from the URLs and keep up with new tactics in ransomware and phishing attacks.

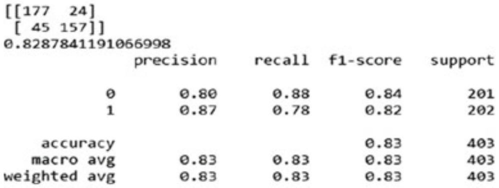
Table 1 Classification report of models

<table>
<tr>
<td>Classification report of Decision tree</td>
<td>



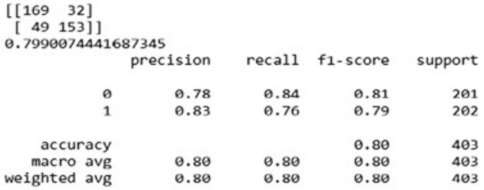
```
[[180  21]
 [ 50 152]]
              precision    recall  f1-score   support

           0       0.78      0.90      0.84       201
           1       0.88      0.75      0.81       202

    accuracy                           0.82       403
   macro avg       0.83      0.82      0.82       403
weighted avg       0.83      0.82      0.82       403
```

</td>
</tr>
<tr>
<td>Classification report of Random Forest</td>
<td>



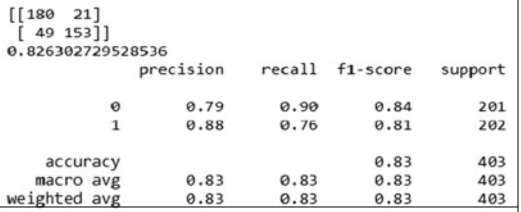
```
[[177  24]
 [ 45 157]]
0.8287841191066998
              precision    recall  f1-score   support

           0       0.80      0.88      0.84       201
           1       0.87      0.78      0.82       202

    accuracy                           0.83       403
   macro avg       0.83      0.83      0.83       403
weighted avg       0.83      0.83      0.83       403
```

</td>
</tr>
<tr>
<td>Classification report of SVM</td>
<td>



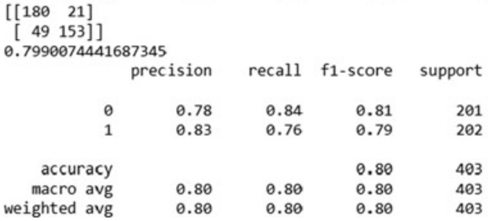
```
[[169  32]
 [ 49 153]]
0.7990074441687345
              precision    recall  f1-score   support

           0       0.78      0.84      0.81       201
           1       0.83      0.76      0.79       202

    accuracy                           0.80       403
   macro avg       0.80      0.80      0.80       403
weighted avg       0.80      0.80      0.80       403
```

</td>
</tr>
<tr>
<td>Classification report of XGBoost</td>
<td>



```
[[180  21]
 [ 49 153]]
0.826302729528536
              precision    recall  f1-score   support

           0       0.79      0.90      0.84       201
           1       0.88      0.76      0.81       202

    accuracy                           0.83       403
   macro avg       0.83      0.83      0.83       403
weighted avg       0.83      0.83      0.83       403
```

</td>
</tr>
<tr>
<td>Classification report of neural networks</td>
<td>



```
[[180  21]
 [ 49 153]]
0.7990074441687345
              precision    recall  f1-score   support

           0       0.78      0.84      0.81       201
           1       0.83      0.76      0.79       202

    accuracy                           0.80       403
   macro avg       0.80      0.80      0.80       403
weighted avg       0.80      0.80      0.80       403
```

</td>
</tr>
</table>

References

1. Lokuketagoda, B., Weerakoon, M.P., Kuruppu, U.M., Senarathne, A.N., Abeywardena, K.Y.: R-Killer: an email based ransomware protection tool. In: The 13th International Conference on Computer Science & Education (ICCSE 2018), pp. 735–741. IEEE (2018)
2. Gendre, A.: Ransomware Attacks: Why Email Is Still the #1 Delivery Method. www.vadesecure.com. Available at: https://www.vadesecure.com/en/blog/ransomware-attacks-why-email-is-still-the-1-delivery-method (n.d.). Accessed 9 Mar 2021

3. Jin, X., Osborn, S.L.: Architecture for Data Collection in Database Intrusion Detection Systems. Lecture Notes in Computer Science, pp. 96–107 (n.d.)
4. Wu, C.-H.: Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert Syst. Appl. **36**(3), 4321–4330 (2009)
5. Shad, J., Sharma, S.: A novel machine learning approach to detect phishing websites Jaypee Institute of Information Technology. In: 5th International Conference on Signal Processing and Integrated Networks (SPIN), 2019, pp. 425–430, 2018
6. Sönmez, Y., Tuncer, T., Gökal, H., Avci, E.: Phishing web sites features classification based on extreme learning machine. In: 6th International Symposium on Digital Forensics and Security ISDFS 2018—Proceeding, vol. 2018–Janua, pp. 1–5, 2018
7. Williams, E.J., Polage, D.: How persuasive is a phishing email? The role of authentic design, influence, and current events in email judgements. Behav. Inf. Technol. **38**(2), 184–197 (2019)
8. Ferreira, A., Teles, S.: Persuasion: how phishing emails can influence users and bypass security measures. Int. J. Hum Comput. Stud. **125**, 19–31 (2019)
9. Ruskanda, F.Z.: Study on the effect of preprocessing methods for spam email detection. Indones. J. Comput. (Indo-JC) **4**(1), 109–118 (2019)
10. Tripathy, H.K., Acharya, B.R., Kumar, R., Chatterjee, J.M.: Machine learning on big data: a developmental approach on societal applications. In: Big Data Processing Using Spark in Cloud, pp. 143–165. Springer, Singapore (2019)
11. Morato, D., Berrueta, E., Magaña, E., Izal, M.: Ransomware early detection by the analysis of file sharing traffic. J. Netw. Comput. Appl. **124**, 14–32 (2018). Available at https://www.sciencedirect.com/science/article/pii/S108480451830300X. Accessed 16 Dec 2019
12. Monika Zavarsky, P., Lindskog, D.: Experimental analysis of Ransomware on windows and android platforms: evolution and characterization. Proc. Comput. Sci. **94**, 465–472 (2016). https://doi.org/10.1016/j.procs.2016.08.072
13. Choi, K.-S., Scott, T.M., Leclair, D.P., Ks, C., Tm, S., Dp, L.: Ransomware against police: diagnosis of risk factors via application of cyber-routine activities theory virtual commons citation ransomware against police: diagnosis of risk factors via application of cyber-routine activities theory. Int. J. Forensic Sci. Pathol. **4**(7), 253–258 (2016). https://doi.org/10.19070/2332-287X-1600061
14. Blum, A., Wardman, B., Solorio, T., Warner, G.: Lexical feature based URL detection using online learning. In: Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security, pp. 54–60 (2010)

A Framework for APT Detection Based on Host Destination and Packet—Analysis

R. C. Veena and S. H. Brahmananda

Abstract In cybersecurity, advanced persistent threats have gained more attention. Even after that, a variety of techniques, like change control, sandboxing, and internet traffic analysis, are often used to identify APT attacks. Even so, 100% accomplishment wasn't achievable. APTs use a variety of sophisticated methods to overcome several types of detection, as per recent research. This paper examines the most standard strategies, techniques, and mechanisms used by adversaries to describe and evaluate APT problems. It also outlines the vulnerabilities and capabilities of current security technologies that are in usage from the time the threat was detected in 2006 till now. Besides, this study introduces a different mechanism to eliminate the issue through APT with internet traffic by host destination and packet inspection.

Keywords Packet inspection · Advanced persistent threats · Network traffic

1 Introduction

McAfee and Kaspersky recently published a survey that reports top cybersecurity threats and their impacts and advanced persistent threats (APT) are one among them [1, 2]. A few weeks, or maybe even months, normally get around until this kind of threat strain is detected [3]. APT is being used by the many advanced attackers on the Network, in which this kind of activity involves a significant amount of expertise and persistence to obtain private and public sector data. The objective of APT's is typically public or private organizations, like healthcare providers, financial firms, universities, and government departments.

The phrase APT describes the three terms advanced, persistent, and threat. Advanced implies the hackers. These are the people using APT as broad degree of expertise. Using the new technologies of intrusion capture, the industry's network

R. C. Veena (✉) · S. H. Brahmananda
Department of Computer Science and Engineering, GITAM University, Bengaluru, India
e-mail: vchalapa@gitam.edu

S. H. Brahmananda
e-mail: bsavadat@gitam.edu

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_62

and its data are the main objectives of these kinds of users. Persistent is consistent assault, typically carried out utilizing a "low and steady" strategy that relies on a persistent attack duration that can extend few months to years [4].

Several strategies are often used to identify APT attempts, like change control, sandboxing, and network traffic examination. Even so, each of these techniques [5] is divided into two types.

1. Signature-based strategies have significant structures that are previously installed in the virus protection software.
2. Behavior-related approaches that are related to malware behavior analysis during an assault.

APT parties utilizing zero-day malware that is not available in the security software repository. As a result, the unidentified advanced persistent threat will simply escape on detection [6–8]. Whereas, approaches based on behavior have three constraints: increased false-positive rate, failure to sense certain polymorphic threats [9], and overall complexity [10]. Conventional methods fail to detect APTs.

Despite these constraints, studies have endeavored to search for a new strategy that can identify APT threats. Studies also found two vulnerabilities that can be used in this identification process. First, most APT attacks express the steps taken on assault, indicating several steps of the attack before hitting a target [11]. Second, a data breach cannot be impossible to detect; it generally requires outgoing traffic. It's an effective way to explore the APT attacks [12]. Consequently, the research work presented here highlights the issues due to APTs and a suitable method to detect attacks.

2 Literature Review

Internet-connected threats on a particular target, typically a public sector or corporate organization, are identified as APTs. The primary objective of these cyber-attacks is typically to steal crucial data contained in all of these organizations' databases [13]. The threat of an APT attack is a major issue for data security and networking technologies [14]. APT attacks are being grouped with shareware or even other types of software available for download. That several kinds of APTs had no trouble clearing the system's firewall. To avoid traditional detection models, advanced evasion technologies are used in APTs [15]. Traditional attacks such as malware, trojan horses, ransomware, worms, and backdoors are less sophisticated than APTs (Table 1).

3 Phases of an APT Attack

Everything should be specifically formulated and performed in an APT attack. The attackers' sequence of steps has been persistent since the phrase was invented by the

Table 1 Traditional and APT attacks comparative analysis

Attacks	Traditional attacks	APT attacks
Target	Unidentified target, Commonly individual's systems	Organizations and institutions in the public and private sectors are being targeted
Attacker	Majority of the time attacker is a person	The group has a high level of organization that containing adequate resources and sophisticated technologies
Purpose	Benefits in terms of money and exhibiting skills	Strategic benefits and competitive advantages
Approach	Short time, single-run, "smash and grab" monetary advantages,	Frequent infiltration efforts, maintaining a slow and low appearance to avoid detection, long-term

United States Air Force (USAF) in 2006. The different steps of APT are from 4 to 8 steps that have been described in earlier studies [16–21]. Most APTs, nevertheless, hold the characteristic of progressing through the same stages to achieve their goal. The attack of an APT is illustrated in Fig. 1, and each step is summarized below:

1. Intelligence gathering.
2. Initial Exploitation.

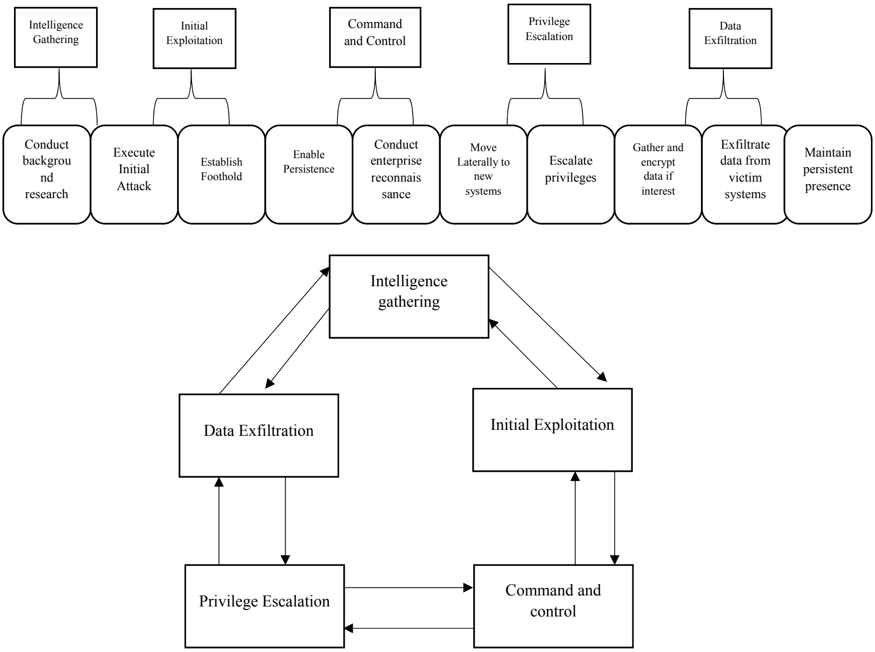


Fig. 1 APT attack process. *Source* [22]

3. Command and control.
4. Privilege Escalation.
5. Data Exfiltration.

4 Groups and Tools for APT

There are several different names for APTs [23], but almost all of them belong to a certain threat category. On these categories, each organization nominates its names. APT 29 does have more than 12 various names. Furthermore, during the hacking process, such groups frequently utilize the same resources. [23] showed that only 174 classes are used, even though they have a large number of names.

According to [23], the origins of these entities are restricted to some few nations, with China leading the way with 73, Russia with 16, North Korea with 8, Iran with 17, Israel with 2, NATO with 2, the Middle East with 11, Other Actors with 20, and Unmapped Actors with 25. The tables in [23] addressed over 174 APT classes and operations that have existed in various parts of the globe. Total 40 tools are used, and its limit could not be exceeded.

This study proposed a new approach for overcoming this attack, which is focused on APT activity in data traffic and packet inspection. According to previous studies, most researchers concentrated on one or two steps when it came to identifying APT threats [16].

5 APT Attack Detection

The task of detecting APTs is difficult and time-consuming. According to previous studies, the three most commonly used methods for detecting APT attacks are

(a) Change Controlling
(b) Sandboxing
(c) Network traffic analysis.

(a) Change Controlling
Change controlling is used to track any changes to significant and critical aspects of a network or computer. Unless the alteration remains unauthorized, it is notified and the respective action will be taken. It helps to detect variety of malwares, and it doesn't need to know the malware features. However, it has limitation as the malware operate through two or more results and alter the system state until it is verified. Extra tool requirements due to poor memory management are found as another limitation of this detection method. More for large number of features the performance is significantly slower.

(b) Sandboxing

Sandboxing is a method of creating a virtual environment under the control of malware so that it can run it, and then examine its actions to see whether the files are malicious or not. As a consequence, any files that aren't trusted are blocked. It offers long-term malware protection, but such a mechanism is difficult to automate, and manual work is costly.

(c) Network traffic analysis

In order to control victim's computer remotely a command and control channel will be created by the attackers. It owns transmitting data and transmitting commands. PC Share, Ghost, remote access tool (RAT), and Poison Ivy are only a few examples of malware that utilize traffic inbound and outbound. Traffic analysis approach is the one of the oldest and widely used models that define the traffic bounds. The outbound traffic analysts detect attacks by utilizing feature series related to APT attacks. The attack can be identified irrespective of how the computer was compromised (although the APT intrusion is recent). Even then, analyzing traffic in a vast network is too difficult. Moreover, detecting APT malware in large network is a difficult process [24].

To mask the attack, the intruder can use single system as entry to target all the remaining computers in the organization [25, 26]. To minimize the impact of an APT attack, the organization must identify APT-infected machines as soon as possible.

APT events can be detected by tracking and examining data traffic, as per the researchers [12]. They examined Enfal, Taidoor, Sykipot, and IXESHE, as well as other APT campaigns that were conducted as targeted attacks. Through HTTP protocols, the command and control server establish connection as malicious programs and it is usually designed to use three ports such as 443, 80, and 8080. Another way to predict APT threats is to keep track of the volume and timing of internet traffic.

Backdoor samples, that are widely used during APT events, have been used in the experimental process in another research [27, 28]. The results of these studies verified the results of the primary studies, suggesting that packet monitoring efficiently detects APT attacks. The experimentation demonstrates that the proposed approach is extremely fast as compared to other intrusion detection methods and antivirus.

6 Proposed Method

APT threats have two vulnerabilities that can be identified. Initially, a command and control channel will be created by the attackers. It is difficult to detect the data breach as it generally requires outbound traffic. Second, every APT threat should set up a command and control channel. APT events can also be detected by tracking and analyzing network traffic, according to the researchers. As a result, in our analysis, we found that traffic analysis was the most effective method for detecting APT risks. The issue, on the other hand, examines the large volume of traffic data. So, this research work proposed a destination host filter unit and blacklist of host destinations to resolve the above issues. The origins of these categories are restricted to a few nations.

The proposed concept for this article is based on various functional points.

1. Previous research on APT attacks has shown that the intruder must interact with the attacker multiple times, i.e., at the time of entry and at the time of transferring from one target to other. Since for several timestamps, the communication will be established with control commands so that tracking the unusual activity can be identified easily.
2. There is no incoming traffic in this case where the method of attack only uses outgoing flow, such as APT backdoors that only allow outbound traffic [29]. Everything we have to do now keeps an eye on repeated packets in the outbound to check whether all the packets are sent to the same destination. Based on previous research [28], we used the technique of a "Packets Detection System" in our proposed framework. Studies have shown that when compared to other antivirus software and IDs, the proposed approach is extremely fast. This method can also identify zero-day and encrypted malware.
3. However, tracking outgoing large number of packets is difficult. In order to reduce the difficulty, it is essential to concentrate on a small number of suspect hosts to reduce the massive data packets that are tracked [30]. Although the load is not verified, this approach works even for encrypted connections. Furthermore, the approach is extensible and most analyses can be run in parallel. The proposed design is depicted in Fig. 2. This methodology has been used as a "Destination Host Filter device" in our proposed system.

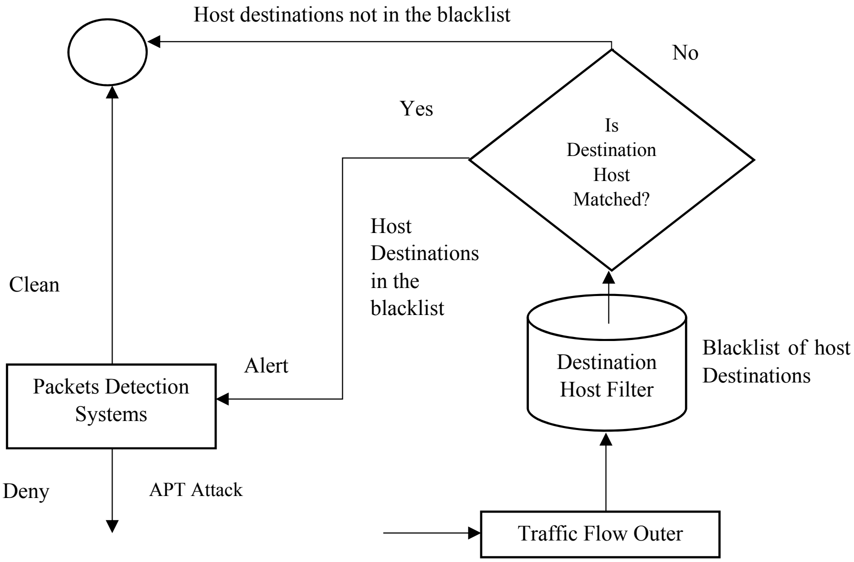

Host destinations not in the blacklist
No
Yes
Is Destination Host Matched?
Host Destinations in the blacklist
Clean
Blacklist of host Destinations
Destination Host Filter
Alert
Packets Detection Systems
Deny
APT Attack
Traffic Flow Outer

Fig. 2 Proposed framework for APT Identification

The proposed framework transfers the flow to the destination host filter unit so that it will be compared when the transfer is initiated. If the flow is matched, then the stream is converted to detection system, and final decision is made in this stage. Based on the matching the transfer will be initiated to the external network else the packets will remain in the same source. This process eliminates the process of continuous tracking of large number of packets.

7 Conclusion

The APT threat is a big challenge in front of cybersecurity, networking technologies. It is not challenging for several kinds of APTs to get through the network firewall. Compared to conventional threats, APTs are more complex as it employs specialized avoidance strategies to escape from detection systems. Research work demonstrates and analyzes using common methods and tools that are used by intruders. Also, research work discusses feature benefits, disadvantages of current defense strategies. The article outlines more than 174 APT classes and operations across the globe. This report has made an important contribution to analysts by including a concise overview of malware used to execute APT phases. Besides, the study introduced a new method that repels the challenge focused on APT behavior with internet traffic by packet processing.

References

1. Zimba, A., Chen, H., Wang, Z.: Bayesian network-based weighted APT attack paths modeling in cloud computing. Future Gener. Comput. Syst. (2019)
2. Han, W.: MalInsight: a systematic profiling-based malware detection framework. J. Netw. Comput. Appl. **125**, 236–250 (2019)
3. Ugarte-Pedrero, X., Graziano, M., Balzarotti, D.: A close look at a daily dataset of malware samples. ACM Trans. Priv. Secur. (TOPS) **22**(1), 6 (2019)
4. Jasek, R., Kolarik, M., Vymola, T.: APT detection system using honeypots. In: Proceedings of the 13th International Conference on Applied Informatics and Communications (AIC'13), WSEAS Press (2013)
5. Sonawane, S., Prasad, G., Pardeshi, S.: A survey on intrusion detection techniques. World J. Sci. Technol. **2**(3) (2012)
6. Balzarotti, D.: Efficient detection of split personalities in malware. In: Network and Distributed System Security Symposium (NDSS) (2010)
7. Radmand, A.: A ghost in software [cited 21 Sept 2013; Course] (2009). Available from: http://cs.columbusstate.edu/cae-ia/StudentPapers/radmand.azadeh.pdf.
8. Hamed, T., Ernst, J.B., Kremer, S.C.: A Survey and taxonomy on data and preprocessing techniques of intrusion detection systems In: Computer and Network Security Essentials, pp. 113–134. Springer (2018)
9. Maarof, M.A., Osman, A.H.: Malware detection based on hybrid signature behaviour application programming interface call graph. Am. J. Appl. Sci. **9** (2012)
10. Idika, N., Mathur, A.P.: A survey of malware detection techniques. Purdue University, p. 48 (2007)

11. Slot, T., Kargl, F.: Detection of apt malware through external and internal network traffic correlation. Master Thesis, University of Twente March (2015)
12. Villeneuve, N., Bennett, J.: Detecting apt activity with network traffic analysis. Trend Micro Incorporated Research Paper, pp. 1–13 (2012)
13. Sharma, P.K.: DFA-AD: adistributed framework architecture for the detection of advanced persistent threats. Clust. Comput. **20**(1), 597–609 (2017)
14. Chakkaravarthy, S.S., Vaidehi, V., Rajesh, P.: Hybrid analysis technique to detect advanced persistent threats. Int. J. Intell. Inf. Technol. (IJIIT) **14**(2), 59–76 (2018)
15. Haq, T., Zhai, J., Pidathala, V.K.: Advanced persistent threat (APT) detection center. Google Patents (2017)
16. Alshamrani, A.: A survey on advanced persistent threats: techniques, solutions, challenges, and research opportunities. IEEE Commun. Surv. Tutor. (2019)
17. Lim, Y.: Review on the cyber attack by advanced persistent threat. Korean Assoc. Terrorism Stud. **6**(2), 158–178 (2013)
18. Tankard, C.: Advanced persistent threats and how to monitor and deter them. Network Secur. **2011**(8), p. 16–19 (2011)
19. Chen, P., Desmet, L.,Huygens, C.: A study on advanced persistent threats. In: IFIP International Conference on Communications and Multimedia Security, Springer (2014)
20. McWhorter, D.: Mandiant Exposes APT1—One of China's Cyber Espionage Units and Releases 3,000 Indicators. Mandiant, 18 Feb 2013
21. De Vries, J.: Towards a roadmap for development of intelligent data analysis based cyber attack detection systems (2012)
22. Burazer, R.: Cybersecurity: Issues and Isaca's Response, 6 Jan 2019. Available from: https://csa-cee-summit.eu/archive/wp-content/uploads/2015/03/Renato-Burazer.pdf (2015)
23. Stirparo, P.: APT Groups and Operations, 2 July 2019. Available from: https://docs.google.com/spreadsheets/u/1/d/1H9_xaxQHpWaa4O_Son4Gx0YOIzlcBWMsdvePFX68EKU/pub html?cv=1
24. Zhao, G.: Detecting APT malware infections based on malicious DNS and traffic analysis. IEEE Access **3**, 1132–1142 (2015)
25. Alperovitch, D.: Revealed: Operation Shady RAT, vol. 3. McAfee (2011)
26. Liu, S.T., Chen, Y.M., Hung, H.C.: N-victims: An approach to determine n-victims for apt investigations. In: International Workshop on Information Security Applications. Springer (2012)
27. Alminshid, K., Omar, M.N.: Detecting backdoor using stepping stone detection approach. In: Second International Conference on Informatics and Applications (ICIA). IEEE (2013)
28. Al-Minshid, K.A.A.: Backdoor Attack Detection Based on Stepping Stone Detection Approach. Universiti Utara Malaysia (2014)
29. Welch, V.: Security at the cyber border: exploring cybersecurity for international research network connections (2012)
30. Marchetti, M.: Analysis of high volumes of network traffic for advanced persistent threat detection. Comput. Netw. **109**, 127–141 (2016)

A Trust-Based Handover Authentication in an SDN 5G Heterogeneous Network

D. Sangeetha, S. Selvi, and A. Keerthana

Abstract The fifth generation (5G) of wireless network paves the way for the development of new technologies to overcome the existing challenges in a heterogeneous network (HetNet). 5G supports huge data traffic with fastest and reliable network access. The main challenge in the existing 5G HetNet is the presence of different types of cells installed in the same geographical area. The frequent handover and authentication among different small cells gives rise to security challenges like access point insecurity, handover vulnerability and attacks. To overcome the existing challenges in such vulnerable network, software-defined networking (SDN) is introduced which is found to reduce the complexity of 5G networks and construction cost. Hence, this paper proposes a SDN-based handover authentication to enable efficient handover authentication and to enhance the security in a 5G mobile communication. The proposed algorithm helps to achieve mutual authentication using a three-way handshaking protocol. Moreover, the security of the proposed authentication scheme is evaluated using a trust value algorithm with clustering mechanism. From the experimental results, it is found that the performance of the proposed mechanism is better in terms of throughput, packet delivery ratio, and reduced delay when compared to the existing system.

Keywords SDN · Handover authentication · HetNets

1 Introduction

The next generation of wireless network that is the 5G will run applications that require data with more speed on high demand. One of the solutions to solve the requirement of data rate is to create a heavy network by deploying small cells in

D. Sangeetha (✉) · A. Keerthana
Department of Information Technology, MIT Campus, Anna University, Chennai, India
e-mail: dsangeetha@mitindia.edu

S. Selvi
Department of Computer Science and Engineering, Government College of Engineering, Krishnagiri, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_63

the network. Heterogeneous network (HetNet) is used for mobile communication networks which are actually composed of distinct mobile nodes and various small cells with several access techniques to provide nodal communication. It comprises low power cells including femtocells and picocells and high power macro-cells. Macro-cells are used to supply network provision for a larger domain and to provide high efficiency output. Mobile users utilize small cells to expand the coverage area of the service, thereby strengthen the network volume. HetNet supports low mobility and high rate traffic. It also helps to reduce transmission power of mobile nodes and base stations. 5G HetNet is a layered network imposed with more number of micro-cells and different access points that are employed within the macro-cell region. Since there is deployment of many distinct kinds of access points (APs) in the network and the complication in the network topology is more in 5G heterogeneous networks (HetNets), many new difficulties in managing the network with security and mobility aspects including the lack of security and the occurrence of periodic handovers among various kinds of nodes and cells in the HetNet architecture take place. The introduction of SDN technique will help to reduce the complication of network topology, minimize the construction and deployment cost, and to assist the management of the network using the applications of software [13]. By enabling the SDN technique, users can get the service for their network at any time in any place through different kinds of wireless access technologies until the agreements between the user and the operator exist [5]. The main focus is on the authenticity of handover in heterogeneous network using SDN to assure the network strength and reliability and also to reduce the management cost with the help of SDN controller. An authentication scheme for the handover in HetNet is proposed based on trust value of the nodes and SDN technique. Thus, the malicious nodes in the network are detected with the help of trust values using the controller to ensure security.

1.1 Heterogeneous Network

Heterogeneous network (HetNet) is utilized in mobile communication networks which consist of distinct kinds of nodes with distinct access mechanisms. It comprises many low power micro-cells and high power macro-cells. Macro-cells are used to serve the client nodes with network service for a larger zone and to give high efficiency output. Mobile users use micro-cells like femtocells and picocells to lengthen the coverage zone of the service and to improve the network capacitance. HetNet supports low mobility and high rate traffic. It also helps to reduce transmission power of mobile nodes and base stations. 5G HetNet is a multi-layered network which consist of more number of overlaid small cells including picocells and femtocells that are deployed within the domain of macro-cells which cover larger area with different access points. The 5G network is supposed to assist for huge number of wireless connections and mobile data traffic. 5G network is also sought to provide better energy efficiency, low communication delay, high reliability and security. It also provides increased bandwidth for all users thereby achieves faster speed.

1.2 Software-Defined Networking

Software-defined networking (SDN) is a type of network framework which uses software applications to control and manage the network with central controller. With the abstraction from hardware and all the limitations that a hardware-bound network once had, SDN creates networks that can provide new services, reduce the cost of the construction, and enable innovation with flexibility. The main components of SDN are controller, forwarding devices, and communication protocols between them. SDN describes a networking system that segregates the control plane and data plane. The controller in SDN is considered as the brain of the network. The controller knows all the available path in network for packet transmission so that it can direct packets based on traffic requirements. It also informed about congested links to the network operators. SDN applications assist to substitute and enlarge on functions that are carried out via hardware devices in a conventional network. SDN provides benefits such as central network provisioning, more granular security, low operating costs, cloud abstractions, and guaranteed content delivery. There is a lot of issues in 5G HetNet since different types of large number of micro-cells are located in the one network domain and the network is accessed through different access mechanisms of distinct operators by the users. Using SDN technique in HetNet will enable the network to highly decrease its complexity by reducing the construction and maintenance cost and manages the network with the help of central controller it provided.

2 Related Work

Different handover authentication methods and techniques have been proposed lately. In [6], Fondo-Ferreiro and others proposed a fast decision algorithm for efficient access point assignment in SDN-controlled wireless access networks. A new AP assignment approach is proposed which takes quick decisions on user actions based solely on the previous history. It is shown that a centralized allocation algorithm focused on predicted decisions reaches efficient network implementation levels using real data. Finally, the expired time is quantified due to the occurrence of user traffic event until an AP is allocated to its terminal when needed. In [1], Alezabi and others improvised the EAP-AKA protocol and introduced the intra-re-authentication and inter-re-authentication protocols in the LTE-WLAN intertwining framework for distinct mobility scenarios. With the proposed method, an entrusted authentication idea is utilized to change the EAP-AKA protocol where 3GPP server assigns its re-authentication tasks to the local server. For LTE wireless networks, Qiu and others proposed a handover authentication technique based on proxy signature [11]. The proposed scheme uses elliptic curve cryptography algorithm to provide mutual authentication for handover mechanisms. The correctness of the proposed method is proved using BAN logic and a software verification tool called SPIN. Using the ticket mechanism, Fu and others introduced a quick and reliable authentication technique for handover that occurs in Wi-Fi and WiMAX HetNets [7]. The key agreement and shared authentication between the user equipment and base stations (BS) or access point (AP) can be achieved using a certified ticket produced by the already visited target stations without the server inclusion. Although the proposed method cannot accomplish the protection for identity privacy, it notably minimizes the authentication cost of the handover. In [9], Jing and others proposed a privacy preserving authentication technique for handover in wireless networks using extensible authentication protocol. The proposed method uses proxy signature technique for handover authentication between mobile node and access point without the intrusion of any third party. From the experimental analysis, it is found that communication and computation overheads are minimized. Sun et al. also proposed an authentication technique for mobile users between Wi-Fi and WiMAX networks [12]. The working time for regenerating the key is highly minimized during the handover using key reuse technique. This technique is used to reutilize a key for authentication when the user roams to the visited network again. The results of the analysis show that the authentication cost is still more when the user roams to a target base stations or access points without the key. Wang and others proposed a SDN-based authentication technique for handover among mobile edge computing nodes. It introduced an efficient SDN-based handover authentication scheme for the mobile networks (SHAS) [13]. A handover authentication module in the SDN controller is employed for authentication provision and key allocation using three-way handshaking protocol to attain shared authentication between user and access points. The SHAS scheme shows that it provides robust anti-attack ability for the network using the security analysis methods such as formal validation mechanism called automated validation of Internet security protocols and applications (AVISPAs) and BAN logic. In [3], Cao and others introduced an invariable handover authentication technique using identity-based cryptography (IBC) without pairing function for different access networks. By the proposed technique, shared authentication is achieved by implementing key agreement between a user and the target access point through a three-way handshake process without connecting with other parties. Cao et al. proposed a capability-based privacy protection handover authentication mechanism for SDN-based heterogeneous networks in 5G [2]. It integrates user competency and software-defined network technique to attain the key agreement and mutual authentication between user nodes and target stations in the network. It also largely reduces the cost of handover authentication. The proposed approach can give safe and secure protection for security by utilizing security testing techniques such as the formal verification tool called Scyther and BAN logic. Duan et al. proposed a handover authentication and privacy protection using software-defined networking in 5G Hetnets [5]. Introducing SDN platform into 5G helps to provide efficient handover authentication and to secure the privacy. The objective is to remove the complexities from handover authentication mechanism in 5G HetNets. It is done through sharing of security context information that is dependent on user among associated access points. It is shown that the centralized control capability delivered by SDN platform provides efficient security solutions which is necessary for low delay communications in 5G. In [14], Yang et al. proposed a rapid and unified handover authentication that depends on link signatures in SDN-based HetNets. The proposed method extracts data for handover authentication using wireless link signatures determined by user locations to attain fast and unified handover authentication approach in the HetNet. The secure context information is collected by using the characteristics of distinct wireless route between the source access point and a user node. It is then sent to the destined AP to decide whether the user is the legal one or not. Then the performance of the authentication is analyzed by using multiple attributes. Yazdinejad et al. proposed a handover authentication with proficient privacy protection based on blockchain technique in SDN-based networks [15]. Blockchain and SDN techniques are used in the proposed authentication approach to eliminate the unnecessary authentication in frequent handover among different types of cells. This technique is also configured to give less latency by replacing the cells with low delay among different cells or clusters in the network using their private and public keys and to protect their privacy. The keys are granted by the devised blockchain component. The overhead in the signaling and energy consumption are also reduced. Jia and others introduced a method in which dynamic cluster heads are selected for wireless sensor network [8]. This method provides efficient mechanism compared to other wireless sensor network clustering algorithms in order to resolve the issue of the unreasonable cluster head (CH) selection that may cause coverage overlapping and unbalanced energy consumption in the cluster communication. From the experimental results, it is shown that the dynamic CH selection scheme balances the nodal energy of the network in two phases compared with the existed algorithms. The survival time of the network is also longer than that of deterministic clustering algorithm which helps to balance energy and adaptive energy optimized clustering algorithm. Leu et al. proposed a new clustering scheme called regional energy aware clustering with isolated nodes for wireless sensor networks (REAC-IN) [4]. This method selects the cluster heads based on weight of the nodes which is calculated with respect to the remaining energy of each nodes and the regional average energy of all nodes in each cluster of the network. The configured algorithms for the clustering method separate the nodes from cluster heads. By utilizing large amount of energy, the separated nodes interact with the sink. The distance between the sensors and sink and the regional average energy are utilized to check whether the data of the secluded nodes are sent to a cluster head or to the sink in order to extend the lifespan of the network topology. Pallavi and others proposed a clustering method based on trust value of a node in mobile ad hoc networks [10]. The clustering method uses trust value-based clustering algorithm to form a cluster with trustworthy nodes. The selection of the nodes is also based on weight and residual energy of the nodes. The proposed method executes better performance in terms of throughout with the help of trusted clusters.

Based on the already defined handover authentication schemes and techniques, a SDN-based handover authentication algorithm is proposed to provide efficient handover process in 5G HetNets and to overcome the challenges provided in the defined schemes. The algorithm uses three-way handshaking protocol to provide mutual authentication and secure key exchange between nodes. On considering the challenges of security and network performance in the defined schemes, a trust-based clustering algorithm is proposed for cluster head selection that selects the cluster head using the trust values calculated, helps to detect the malicious nodes present in the network, and increases the network stability.

3 Proposed Work

The proposed model consists of authentication handover module which act as a controller to provide secure interaction among the nodes in the network with the help of cluster head using handover authentication algorithm. When the user starts handover, controller checks for key agreement between the user and terminal and provides mutual authentication for security and to protect their privacy. The outline of handover authentication method is shown in Fig. 1. Three algorithms are proposed for the handover authentication model. One is trust value algorithm to calculate trust values of the nodes which also helps to identify the malignant nodes. The other one is cluster head selection algorithm which helps to select the cluster head in the network with the help of trust values associated with each node. The next is enhanced handover authentication algorithm (EHAA) which uses three-way handshaking protocol to provide security in the network by ensuring mutual authentication between nodes in key exchange and message exchange for communication.

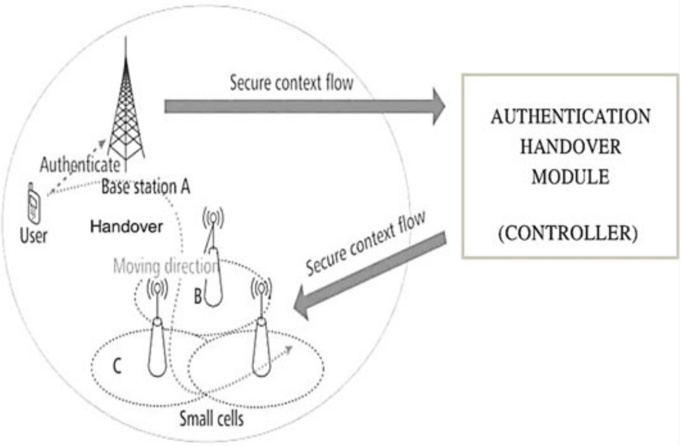


Fig. 1 SDN-based handover authentication

3.1 Cluster Head Selection

A trust-based clustering algorithm is proposed for cluster head selection that selects the cluster head using the trust values calculated in the network and also helps to differentiate between malicious and trustworthy nodes. The cluster head selection method depends on the calculated trust value of nodes. Initially, all nodes of wireless network have zero trust value. If a node does not involve in packet drop, then the trust value of the node will get incremented each time. Based on the proposed clustering algorithm, each node sends hello messages along with the calculated trust values at regular intervals. Each node computes the trust value of the neighboring nodes following the reception of hello messages. Each node compares its own trust value with the neighbor node's trust value to select the cluster head of the network. The cluster head then acts as a local controller by managing the nodes in its cluster and communicates with main controller to exchange nodes information for secure communication and to detect attacker node. Each node is in any one of the three states, state 0 for not decided, state 1 for cluster head, and state 2 for cluster member. The state of the node is decided by the trust value.

3.2 Handover Authentication

The handover authentication algorithm uses three-way handshaking protocol to provide secure key exchange between nodes in order to attain mutual authentication. The algorithm consists of two modules such as key agreement to authenticate the nodes with the help of controller and to provide secure key exchange between nodes, handover request and handover authentication to authenticate the nodes to be communicated with the help of key agreement and to provide secure message exchange mechanism. The overall message exchange scheme of the enhanced handover authentication algorithm (EHAA) is given in Fig. 2 and the notations of the EHAA scheme are explained in Table 1.

Key agreement helps to provide secure key exchange between nodes using controller and cluster heads of the network. Handover authentication method provides authentication between the nodes using the controller to start the communication.

4 Implementation and Simulation Results

The proposed handover authentication scheme is simulated on Ubuntu 14 platform using the NS-2 network simulator. The source and destination pair of the nodes is randomly spread across the network. The simulation parameters of the wireless network topology are shown in Table 2.

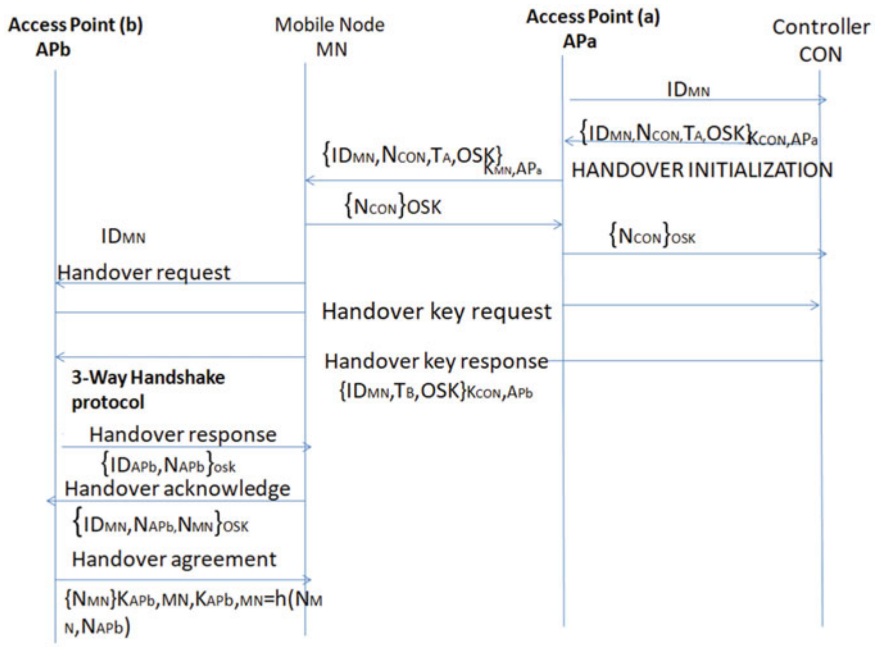


Fig. 2 Message exchange in EHAA

Table 1 Notations of EHAA

Attribute	Content
N_x	Nonce value generated randomly by x
ID_x	Identity value of x
K_{AB}	A and B share the symmetric key K
$\{M\}_k$	Message M encrypted with symmetric key k
H(m)	Hash function of message m
OSK	One time session key

The network is formed with three clusters consisting of mobile nodes and one node to perform as a controller at the center as shown in Fig. 3. The cluster head for each cluster is selected with the help of cluster head selection algorithm and the key exchange between nodes takes place using EHAA as given in Figs. 4 and 5. The malignant nodes in the network are detected using trust value algorithm as shown in Fig. 6.

Table 2 Simulation parameters

Parameters	Settings
Simulation area	1350 * 1100
No. of nodes	48
Channel type	Wireless/Channel
Antenna model	Antenna/omni antenna
Energy model	Battery
Interface queue type	CMU priqueue
Link layer type	LL
Initial energy	100
Routing protocol	AODV
Traffic source	Constant bit rate (CBR)
MAC type	IEEE 802.11
Packet size	500

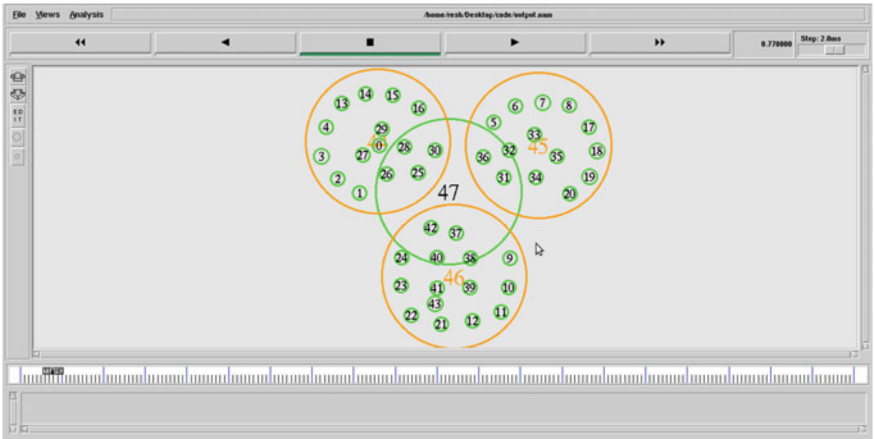
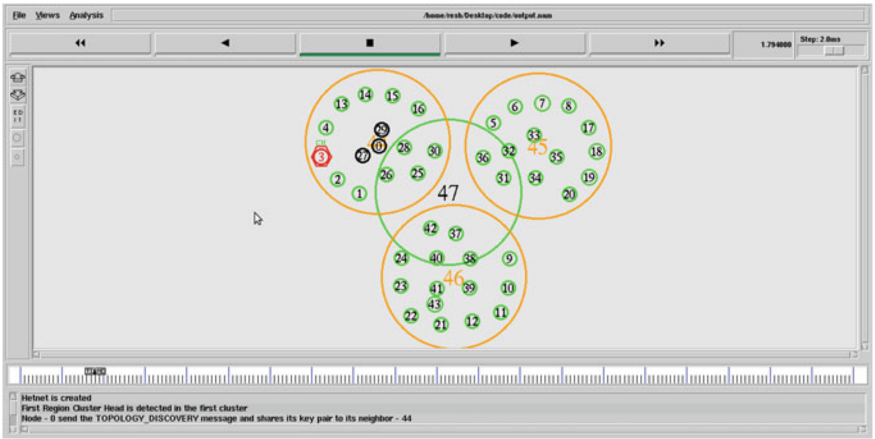


Fig. 3 Network formation

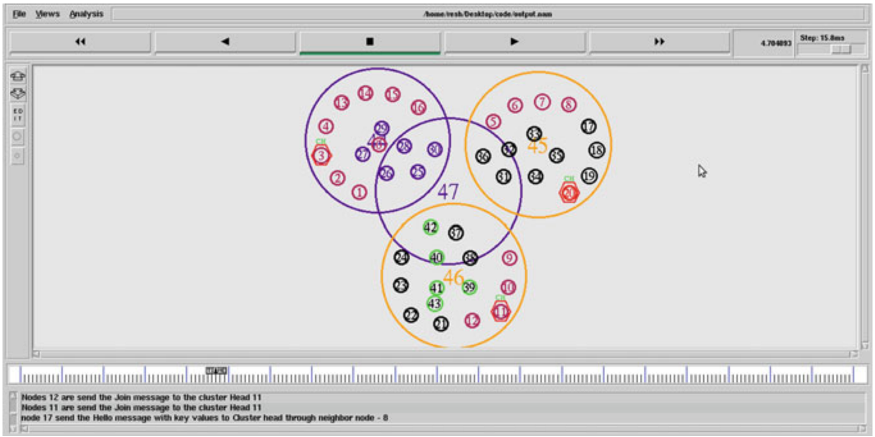
5 Conclusion and Future Work

The proposed trust-based handover authentication mechanism is deployed in SDN which decreases the maintenance and construction cost of the 5G heterogeneous network. The proposed scheme used a handover authentication algorithm which is based on three-way handshaking protocol. This ensures mutual authentication between the heterogeneous nodes in the 5G network. Overheads in the communication are considerably reduced by utilizing clustering method in the network topology. Using the proposed trust value algorithm and cluster head selection algorithm, the cluster heads are selected for the network which helps to perform routing process and mobility, and more efficient resource allocation. The cluster head selection algorithm



```
Hetnet is created
First Region Cluster Head is detected in the first cluster
Node - 0 send the TOPOLOGY_DISCOVERY message and shares its key pair to its neighbor - 44
```

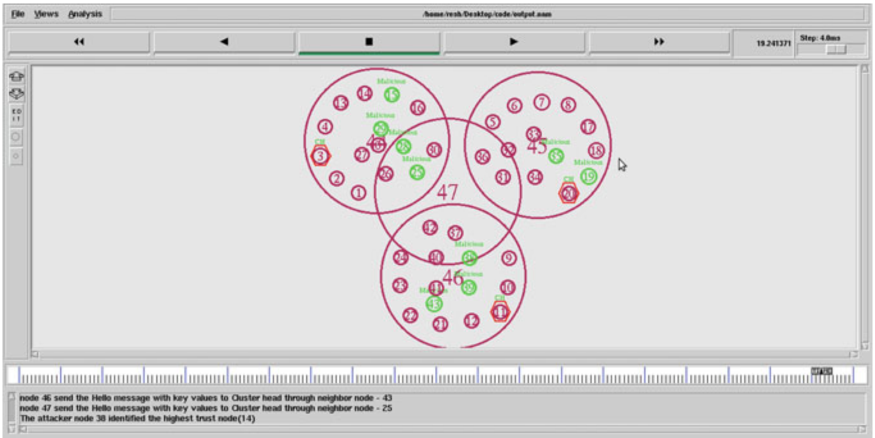
Fig. 4 Cluster head selection



```
Nodes 12 are send the Join message to the cluster Head 11
Nodes 11 are send the Join message to the cluster Head 11
node 17 send the Hello message with key values to Cluster head through neighbor node - 8
```

Fig. 5 Key exchange

is constructed by considering mobility of nodes. The nodes determine whether they become cluster head by themselves using trust value calculated using trust value algorithm. The packet drop attack in the wireless network topology is detected using trust value algorithm so that packets can be transferred through some alternative paths. Thereby the malicious nodes in the network are detected using trust values associated with each node based on the trust value algorithm. The experimental results prove that the proposed handover authentication scheme reduces the latency in authentication with minimum usage of the cryptographic techniques. It also shows that the proposed mechanism delivers high throughput and packet delivery ratio compared to the existing approach. The latency is also found to be considerably low in the network topology. To ensure the 5G network security, a wireless network with packet drop



```
node 46 send the Hello message with key values to Cluster head through neighbor node - 43
node 47 send the Hello message with key values to Cluster head through neighbor node - 25
The attacker node 38 identified the highest trust node(14)
```

Fig. 6 Malicious node identification

attack is created and simulated in NS2 simulator using AODV routing protocol and its effect is examined. As a future work, different routing protocols can be used in the simulations which may be helpful to compute varied and interesting conclusions in employing the best routing protocol. This would in turn increase the performance and reduce the packet drop in the SDN-based 5G HetNets.

References

1. Alezabi, K.A., Hashim, F., Hashim, S.J., and Ali, B.M.: On the authentication and re-authentication protocols in LTE-WLAN interworking architecture. Trans. Emerg. Telecommun. Technol. **28**(4) (2017)
2. Cao, J., Ma, M., Fu, Y., Li, H., Zhang, Y.: CPPHA: Capability-based privacy-protection handover authentication mechanism for SDN-based 5G HetNets. IEEE Trans. Dependable Secur. Comput. (2019)
3. Cao, J., Ma, M., Li, H.: An uniform handover authentication between E-UTRAN and Non-3GPP access networks. IEEE Trans. Wireless. Commun. **11**(10), 3644–3650 (2012)
4. Chiang, T., Leu, J.: Regional energy aware clustering with isolated nodes in wireless sensor networks. In: 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), Washington, DC (2014)
5. Duan, X., Wang, X.: Authentication handover and privacy protection in 5G hetnets using software-defined Networking. IEEE Commun. Mag. **53**(4), 28–35 (2015)
6. Fondo-Ferreiro, P., Mhiri, S., López-Bravo, C., González-Castaño, F.J., Gil-Castiñeira, F.: Fast decision algorithms for efficient access point assignment in SDN-controlled wireless access networks. IEEE Trans. Netw. Serv. Manage. **16**(3), 1059–1070 (2019)
7. Fu, A., Zhang, G., Zhu, Z., Zhang, Y.: Fast and secure handover authentication scheme based on ticket for WiMAX and WiFi heterogeneous networks. Wireless Person Commun. **79**(2014), 1277–1299 (2014)
8. Jia, D., Zhu, H., Zou, S., Hu, P.: Dynamic cluster head selection method for wireless sensor network. IEEE Sens. J. **16**(8), 2746–2754, Apr 15 2016

9. Jing, Q., Zhang, Y., Fu, A., Liu, X.: A privacy preserving handover authentication scheme for EAP-based wireless networks. In: 2011 IEEE Global Telecommunications Conference—GLOBECOM. Houston, TX, USA, pp. 1–6 (2011)
10. Khatri, P., Tapaswi, S., Verma, U. P.: Clustering based on trust of a node in mobile ad-hoc networks. Secur. Comput. Commun. **377** (2013)
11. Qiu, Y., Ma, M., Wang, X.: A proxy signature-based handover authentication scheme for LTE wireless networks. J. Netw. Comput. Appl. **83**, 63–711 (2017)
12. Sun, H.M., Chen, S.M., Chen, Y.H., Chung, H.J., Lin, I.H.: Secure and efficient handover schemes for heterogeneous networks'. In: Proceedings of IEEE Asia-Pacific Services Computing Conference. Yilan, Taiwan, pp. 205–210 (2008)
13. Wang, C., Zhang, Y., Chen, X., Liang, K., Wang, Z.: SDN-based handover authentication scheme for mobile edge computing in cyber–physical systems. IEEE Internet Things J. **6**(5), 8692–8701 (2019)
14. Yang, J., Ji, X., Huang, K., Chen, Y., Xu, X., Yi, M.: Unified and fast handover authentication based on link signatures in 5G SDN-based HetNet. IET Commun. **13**(2), 144–152 (2019)
15. Yazdinejad, A., Parizi, R. M., Dehghantanha, A., Choo, K. R.: Blockchain-enabled authentication handover with efficient privacy protection in SDN-based 5G networks. IEEE Trans. Netw. Sci. Eng. (2019)

Intrusion Detection for Vehicular Ad Hoc Network Based on Deep Belief Network

Rasika S. Vitalkar, Samrat S. Thorat, and Dinesh V. Rojatkar

Abstract There has been continued to be a lot of research into self-driving and semi-self-driving in the last few years, which has to the creation of the vehicular ad hoc networks, but has become more vulnerable to potential attacks due to the misuse of networks. The proposed model of deep learning algorithm, namely deep belief network is used for detecting intrusion in the vehicular ad hoc network (VANET). Deep belief network algorithm gives more accuracy for intrusion detection in the network than existing methodologies such as machine learning algorithms or another deep learning algorithm. Nowadays, automation is more important in all fields, similarly automatic vehicles, i.e., driverless cars. These types of vehicles will come to market and all these vehicles are connected through a wireless network. All the vehicles are communicating with each other by sending some informative packets but there is an attacker who accesses that data and changes the data which may affect the security of the vehicle and also damage the system responsible for the accident. So, intrusion detection system for the vehicular ad hoc network is important with maximum accuracy. For this purpose, we used the updated CICIDS2017 dataset for training, testing and evaluation process. Experimental results using a deep belief network for intrusion detection mechanisms proved that the proposed model could have good results on multiclass and binary classification accuracy 90% and 98% respectively.

Keywords Deep belief network · Vehicular ad hoc network · Intrusion detection · Deep learning · Distributed DOS attacks · Multiclass classification · Binary classification

R. S. Vitalkar (✉) · S. S. Thorat · D. V. Rojatkar
Department of Electronics Engineering, Government College of Engineering, Amravati, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_64

1 Introduction

The vehicular ad hoc network is one of the types of mobile ad hoc network (MANET), because the communication node is a vehicle and an important part of intelligent transportation systems [1]. There are two types of communication systems for exchanging information between nodes in VANET. One is vehicle-to-vehicle and the other is vehicle-to-infrastructure [2, 3]. Deployed by interconnected vehicles and infrastructure, VANET extends the security vulnerabilities derived from wireless communication system, especially in Distributed DOS attacks [4]. Variety of services has been designed for VANET, which are classified into two categories: commercial and security services. Most of them depend on a variety of collected data or transmitted to vehicular nodes. Making the VANET network more secure has become a major challenge as it has become easier for attackers to manage vehicles.

Extensive research is underway to secure network systems and to control the intrusions. Hoppe et al. [5] An Intrusion Detection System (IDS) was proposed in the vehicle. Significant attack patterns such as increasing message numbers and missing message IDs can be identified. Larsen et al. [6] proposed feature-based techniques for detecting IDS attacks. This proposed technique compared the behavior of the current specification system with pre-defined patterns. Zaidi et al. [7] proposed the intrusion detection system based on detecting false information using statistical analysis on VANET. Using this approach reduces the network message congestion. Sedjelmaci et al. [8] suggested the mechanism for intrusion detection called as ELIDV for VANET. In this approach, designed various set of rules for malicious vehicle detection. Schmidt et al. [9] suggested the mathematical model for intrusion detection based on spline function. Ghaleb et al. [10] proposed the misbehavior-aware on-demand collaborative intrusion detection system using distributed ensemble learning technique on NSL-KDD dataset. The random forest algorithm is used as classifier, to aggregate the data by voting scheme. This mechanism is very effective for reducing the communication overhead. Ali Alheeti et al. [11] suggested the mechanism for intrusion detection on VANET. Their mechanism extracts the minimum feature from trace file and analyzes the normal or abnormal behavior of vehicle. The artificial neural network and fuzzy logic were used to detect the attack.

Network bandwidth and other resources are being used to damage the resources of the target system by adopting new methods of intruders to carry out DDOS attacks, and the number is increasing every year. Thus, DDOS attacks pose a threat to the main system. To provide high availability of VANET, a scalable, reliable and robust network intrusion detection system should be developed to effectively reduce DDoS attacks.

Therefore, our aim to propose a strong and competent security mechanism to protect such networks against intruders, such as the use of network traffic monitoring and management services. This article proposed a deep learning approach to identify intrusions by studying recent research. Deep learning has been studied extensively in machine learning research, including signal processing, image processing, speech recognition, and more and widely used for practical applications. Once the system features are trained, the proposed system monitors the exchange packets in the network of vehicles to decide whether the system is being attacked. Since DNN takes less time to make a decision, the system responds quickly to an attack.

2 Proposed Methodology

2.1 Dataset

This research used the CICIDS2017 dataset available from https://www.unb.ca/cic/datasets/ids-2017.html which is related to the real world. According to Sharafaldin et al. [12], the CICIDS 2017 dataset contains eight different five-day files and traffic data of the Canadian Institute of Cyber Security. Only 83 statistical features are extracted from the total dataset for network traffic. All the packets in the network flow from source to destination or destination to source.

2.2 Pre-processing

All machine learning algorithms are correlated with the data in the dataset, and to get accurate results, this data must be preprocessed or cleaned. It normalizes all the values from the dataset and removes the features which have zero values in the dataset and are not required to train or test the system. First, we identify rows in a dataset that has lost values, infinite values, and meaningless values. This step is important to maintain the reliability of the dataset and avoid noise, so choosing the method has to be done carefully. Finally, checked and removed all duplicate rows. As a result of cleaning and feature removal methods, we end up with 2,414,417 examples and a dataset of 79 features (Fig. 1).

2.3 Deep Belief Network Model

For these proposed works deep belief network algorithm of deep learning is used to train the system with some tuning parameter. It creates some hidden layers and visible layers to train the system. The model has three layers; Input, hidden, and output. Each layer has assigned various neurons with weight. Select the hidden layer with its parameter using the selective method for processing. After the processing data is transformed from the next layer for further processing. The mathematical defined as

$$A = N^p \times N^q \tag{1}$$

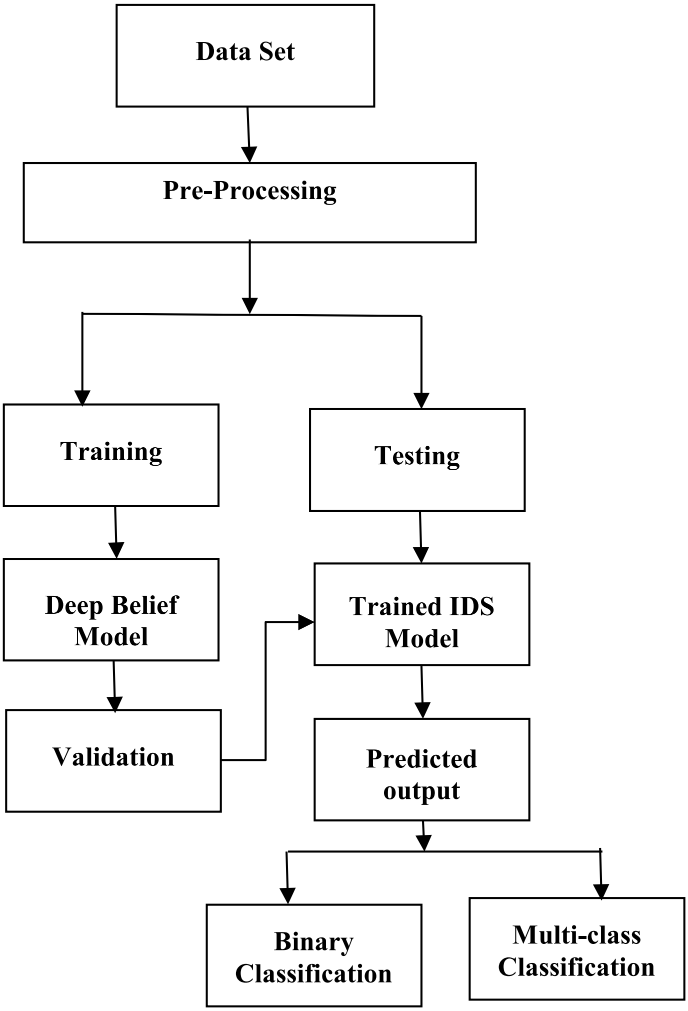


Fig. 1 Block diagram for proposed system

where, p is the input $m = m_1, m_2, m_3, \ldots m_p$.

q is the output of A(m), The Numerical representation of each layer is defined as

$$h_i(m) = f\left(w_i^T m + c_i\right) \tag{2}$$

where, $h_i : N^{d_i-1} \rightarrow N^{d_i}$.

$$f : N \rightarrow N,\ w_i \in N^{d \times d_i - 1}, b \in R^{di} \tag{3}$$

d_i represent the size of input f is the nonlinear function which has sigmoid value (0,1).

In a classification of multiclass, our DBN model used the softmax function as a nonlinear function. The Softmax function expects the probability of each class and selects the largest of the probability values to give a more accurate value.

Mathematical representation of Sigmoid, Softmax, and Tangent function is as follows

$$\text{Sigmoid} = \frac{1}{1 + e^{-x}} \tag{4}$$

$$\text{Tangent} = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{5}$$

$$\text{Softmax} = \frac{e^{mi}}{\sum_{j=1}^{m} e^{mi}} \tag{6}$$

For many hidden layer, DBN is defined as

$$H(m) = H_i(H_{i-1}(H_{i-2}(\ldots(H_i(m))))) \tag{7}$$

This way of stacking hidden layers is typically called deep Belief networks. The DBN model has a more advanced feature with each hidden layer with a strong activation function like ReLU. The ReLU has good capability as compared to other nonlinear functions for trained the model [13]. The hidden layer has several layers with maximum neurons represent the width of DBN.

2.4 Loss Function

In the proposed model, includes loss function by finding optimal parameters for better performance. The loss function is used to measure the difference between predicted and target values [14]. The mathematical representation can be defined as follows

$$d(t1,\ p1) = |t1 - p1| \tag{8}$$

where $t1$ represent the targeted value.

$p1$ represent the predicted value.

For multiclass classification used negative probability with $t1$ as targeted value class and $p1(pad)$ probability as follows

$$d(t1,\ p1(pd)) = -\log(p(pd))t \tag{9}$$

Model received the various input and output for training So decrease the loss mean is defined as follows

$$\text{Loss(Input, output)} = \frac{1}{m}\sum_{i=1}^{m} d(d(t1,\ p1),\ h_i(m)) \tag{10}$$

2.5 Validation

After trained the model validation is required to check whether the training for the model using a deep belief network is accurate or not. For both binary and multiclass classification validation result is given by confusion matrix. Before training, we have to select the classification type.

2.6 Trained Intrusion Detection System Model

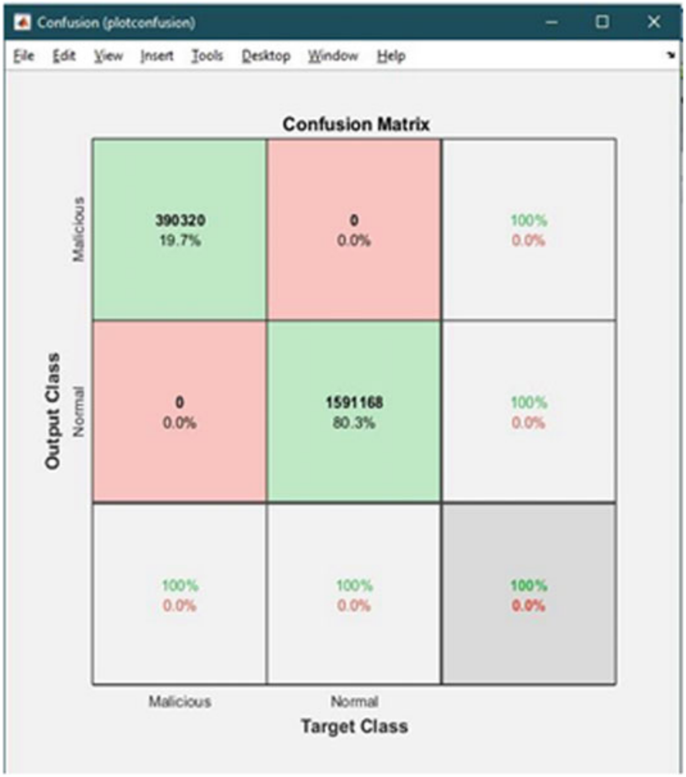
If the validation result is proper then save that model for testing and then test data is tested using that save intrusion detection model which gives the output.

2.7 Predicted Output

The output is in two forms binary and multiclass if we select binary then it shows output the data is malicious or normal and for multiclass classification it shows the output as the name of attack such as Denial of Service Attack (DoS), Distributed Denial of Service Attack (DDoS), PortScan Attack, Patator Attack, Web Attack, Botnet, and Normal.

3 Results and Discussion

The simulation and performance of proposed model can design in MATLAB software with necessary system configuration MS Win-10 OS, Intel Core i3 CPU, 8-GB RAM, 2-GB graphic cards, etc. (Figs. 2 and 3).


Confusion (plotconfusion)
File Edit View Insert Tools Desktop Window Help

Confusion Matrix

Output Class \ Target Class	Malicious	Normal	
Malicious	390320 19.7%	0 0.0%	100% 0.0%
Normal	0 0.0%	1591168 80.3%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%

Fig. 2 Validation confusion matrix for binary classification

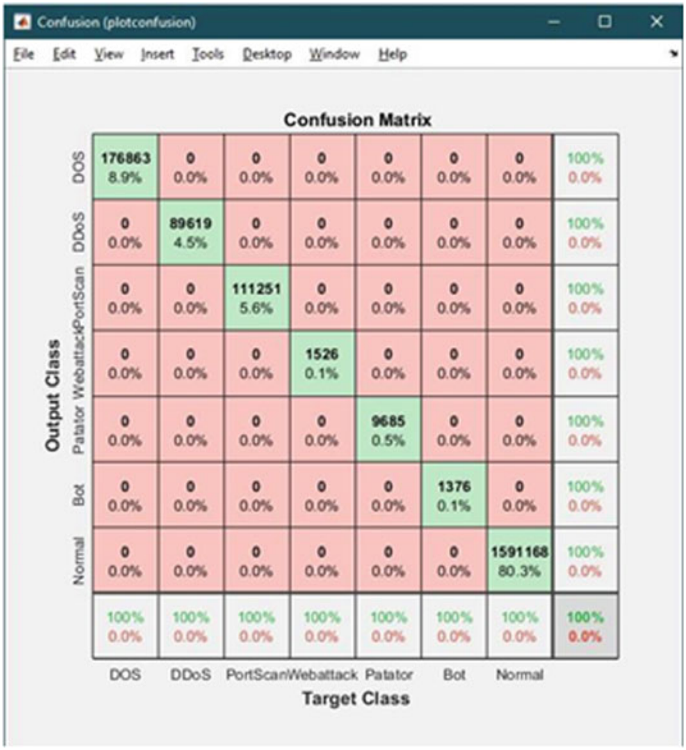
3.1 *Confusion Matrix*

Confusion Matrix for both Binary Classification and for Multiclass Classification gives 100% accuracy it states that training for each classification is accurate and proper.

Number of live nodes at 600 s, 700 s and 800 s.

3.2 *Confusion Matrix Result*

To show the performance of proposed methodology confusion matrix is used. A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. Each row in a confusion matrix represents an actual class, while each column represents a predicted class. The confusion matrix gives you a lot of information, such as accuracy, precision, sensitivity, and specificity (Figs. 4 and 5).


Confusion (plotconfusion)
File Edit View Insert Tools Desktop Window Help

Confusion Matrix

Output Class	DOS	DDoS	PortScan	Webattack	Patator	Bot	Normal	
DOS	176863 8.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
DDoS	0 0.0%	89619 4.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
PortScan	0 0.0%	0 0.0%	111251 5.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Webattack	0 0.0%	0 0.0%	0 0.0%	1526 0.1%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Patator	0 0.0%	0 0.0%	0 0.0%	0 0.0%	9685 0.5%	0 0.0%	0 0.0%	100% 0.0%
Bot	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1376 0.1%	0 0.0%	100% 0.0%
Normal	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1591168 80.3%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Target Class

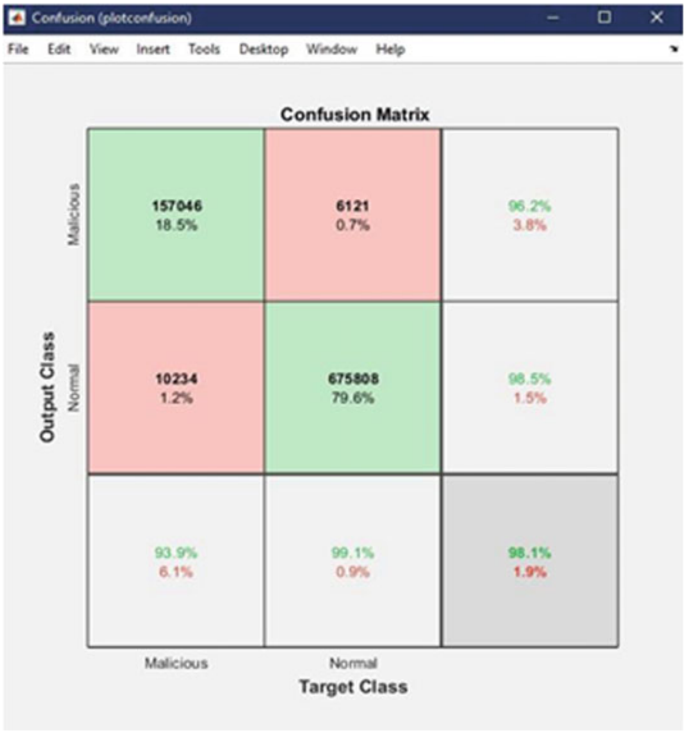
Fig. 3 Validation confusion matrix for multiclass classification

3.3 Binary Classification Result

From the confusion matrix, we get true positive (TP), false positive (FP), true negative (TN), false negative (FP) for binary classification. From these all values we calculate the parameters value such as accuracy, specificity, and sensitivity, which all given in Table 1 (Figs. 6 and 7).

3.4 Multiclass Classification Result

From confusion matrix for multiclass classification, we get all parameter values for each attack type so the accuracy, specificity, sensitivity, *F*-score is different for all attacks, i.e., for DoS attack, DDoS attack, Web Attack, etc. (Table 2).


Confusion (plotconfusion)
File Edit View Insert Tools Desktop Window Help

Confusion Matrix

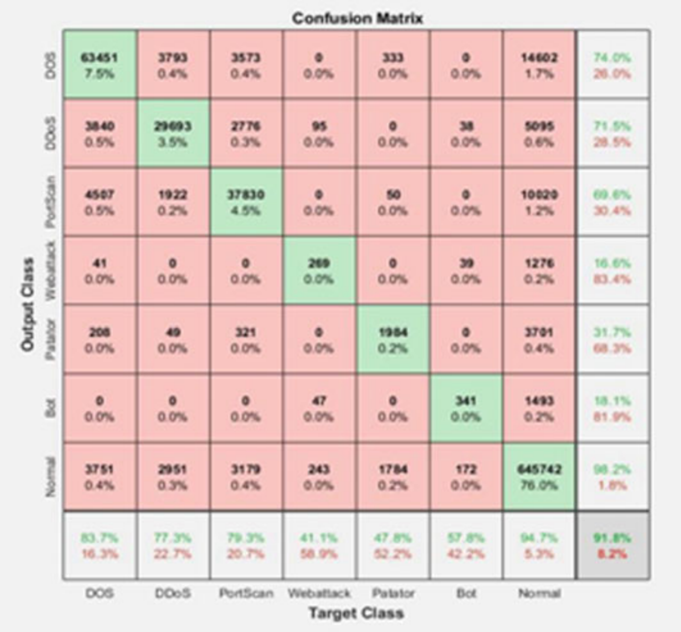
Output Class	Malicious	Normal	
Malicious	157046 18.5%	6121 0.7%	96.2% 3.8%
Normal	10234 1.2%	675808 79.6%	98.5% 1.5%
	93.9% 6.1%	99.1% 0.9%	98.1% 1.9%

Target Class

Fig. 4 Confusion matrix for binary classification

4 Conclusion and Future Scope

From this complete work conclude that deep belief network gives better accuracy or parameter values than the existing methodologies. For binary classification, it gives good accuracy but for multiclass classification accuracy is low because the availability of data for particular attack is low it may improve if same number of data will available or real-world data will available. The future scope of this work is that use the vehicular ad hoc network dataset which having normal and malicious data for automatic vehicles which gives proper result.



Confusion Matrix

Output Class

	DOS	DDoS	PortScan	Webattack	Patator	Bot	Normal	
DOS	63451 7.5%	3793 0.4%	3573 0.4%	0 0.0%	333 0.0%	0 0.0%	14602 1.7%	74.0% 26.0%
DDoS	3840 0.5%	29693 3.5%	2776 0.3%	95 0.0%	0 0.0%	38 0.0%	5095 0.6%	71.5% 28.5%
PortScan	4507 0.5%	1922 0.2%	37830 4.5%	0 0.0%	50 0.0%	0 0.0%	10020 1.2%	69.6% 30.4%
Webattack	41 0.0%	0 0.0%	0 0.0%	269 0.0%	0 0.0%	39 0.0%	1276 0.2%	16.6% 83.4%
Patator	208 0.0%	49 0.0%	321 0.0%	0 0.0%	1984 0.2%	0 0.0%	3701 0.4%	31.7% 68.3%
Bot	0 0.0%	0 0.0%	0 0.0%	47 0.0%	0 0.0%	341 0.0%	1493 0.2%	18.1% 81.9%
Normal	3751 0.4%	2951 0.3%	3179 0.4%	243 0.0%	1784 0.2%	172 0.0%	645742 76.0%	98.2% 1.8%
	83.7% 16.3%	77.3% 22.7%	79.3% 20.7%	41.1% 58.9%	47.8% 52.2%	57.8% 42.2%	94.7% 5.3%	91.8% 8.2%

Target Class

Fig. 5 Confusion matrix for multiclass classification

Table 1 Parameters value for binary classification

Parameters	Values
True positive	10,234
True negative	6121
Accuracy	1.9259
Sensitivity	99.1024
F-score	96.2486
Negative predictive rate	6.1179
False positive rate	80.786
Rate of positive prediction	93.8665
True positive	10,234
True negative	6121
Accuracy	1.9259
Sensitivity	99.1024

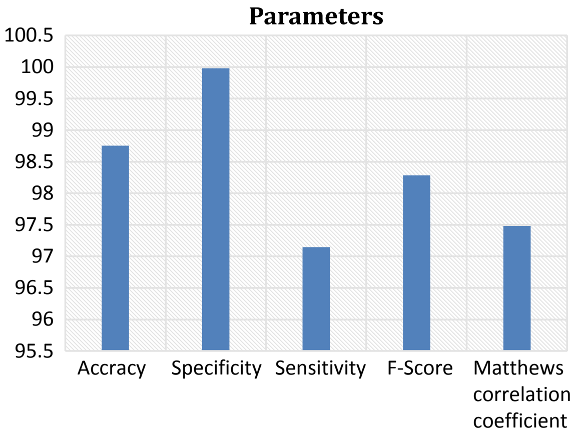


Fig. 6 Predicted result for binary classification

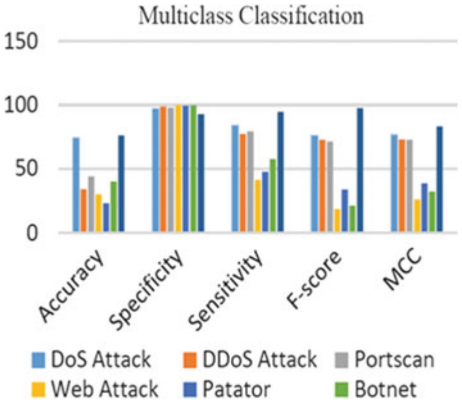


Fig. 7 Predicted result for multiclass classification

Table 2 Parameters value for multiclass classification

Parameter	DoS	DDoS	Portscan	Web attack	Patator	Bot	Normal
True positive	63,451	29,693	37,830	269	1984	341	645,742
False negative	12,347	8715	9849	385	2167	249	36,187
True negative	75,110	798,957	785,031	847,199	840,779	847,079	155,200
False positive	22,301	11,844	16,499	1356	4279	1540	12,080
Accuracy	7.47178	3.49655	4.45473	0.031676	0.233629	0.040155	76.0404
Error rate	2.6261	1.3947	1.9429	0.15968	0.050388	0.18135	1.4225
Sensitivity	83.7107	77.3094	79.3431	41.31315	47.7957	57.7966	94.6934
Specificity	97.1166	98.5392	97.9416	99.8402	99.4936	99.8185	92.7786
F-score	75.7523	72.5791	71.3787	18.8007	33.9691	21.0131	97.4494
Positive predictive rate	73.9936	71.4857	69.6313	16.5538	31.6781	18.1287	98.1636
Negative predictive rate	98.3828	98.921	98.7209	99.9546	99	99.9706	81.0922
False negative rate	16.2893	22.6906	20.6569	58.8685	52.2043	42.2034	5.30656
False positive rate	2.8835	1.4608	2.0584	0.1598	0.50636	0.18147	7.2214
Rate of negative predictions	89.9021	95.1087	93.6024	99.8086	99.2625	99.7785	22.5371
Rate of positive prediction	10.0979	4.89126	6.3976	0.191355	0.73751	0.2215	77.4629
Matthews correlation Coefficient	76.4851	73.077	72.7026	26.0073	38.5471	32.2923	83.2627

References

1. Gao, Y., Wu, H., Song, B., Jin, Y., Luo, X., Zeng, X.: A distributed network intrusion detection system for distributed denial of service attacks in vehicular ad hoc network. IEEE Access **7**, 154560–154571 (2019). https://doi.org/10.1109/ACCESS.2019.2948382
2. Wang, X., Ning Z., Zhou, M., Hu, X., Wang, L., Zhang, Y., Yu, F.R., Hu, B.: Privacy-preserving content dissemination for vehicular social networks: Challenges and solutions. IEEE Commun. Surv. Tutor. **21**(2), 1314–1345, Second quarter (2019)
3. Hu, X., Zhao, J., Seet, B., Leung, V.C.M., Chu, T.H.S., Chan, H.: Saframe: agent-based multi-layer framework with context-aware semantic service for vehicular social networks. IEEE

Trans. Emerg. Top. Comput. **3**(1), 44–63 (2015)

4. Parkinson, S., Ward, P., Wilson, K., Miller, J.: Cyber threats facing autonomous and connected vehicles: future challenges. IEEE Trans. Intell. Transp. Syst. **18**(11), 2898–2915 (2017)
5. Hoppe, T., Kiltz, S., Dittmann, J.: Security threats to automotive can networks practical examples and selected short-term countermeasures. Reliab. Eng. Syst. Saf. **96**(1), 11–25 (2011)
6. Larson, U.E. Nilsson, D.K Jonsson, E.: An approach to specification-based attack detection for in-vehicle networks. In: IEEE Intelligent Vehicles Symposium, Proceedings (2008)
7. Zaidi, K., Milojevic, M.B., Rakocevic, V., Nallanathan, A., Rajarajan, M.: Host-based intrusion detection for VANETs: a statistical approach to rogue node detection. IEEE Trans. Veh. Technol. **65**(8), 6703–6714 (2016). https://doi.org/10.1109/TVT.2015.2480244
8. Sedjelmaci, H., Senouci, S.M.: A new intrusion detection framework for vehicular networks. In: 2014 IEEE International Conference on Communications (ICC), Sydney, NSW, Australia, pp. 538–543 (2014). https://doi.org/10.1109/ICC.2014.6883374.
9. Schmidt, D.A., Khan, M.S., Bennett, B. T.: Spline based intrusion detection in vehicular Ad Hoc networks (VANET). In: 2019 Southeast Conference, Huntsville, AL, USA, pp. 1–5 (2019). https://doi.org/10.1109/SoutheastCon42311.2019.9020367
10. Ghaleb, F.A., Saeed, F., Al-Sarem, M., Ali Saleh Al-rimy, B., Boulila, W., Eljialy, A.E.M., Aloufi, K., Alazab, M.: Misbehavior-aware on-demand collaborative intrusion detection system using distributed ensemble learning for VANET. Electronics **9**, 1411 (2020). https://doi.org/10.3390/electronics9091411
11. Ali Alheeti, K.M., Gruebler, A., McDonald-Maier, K.D.: An intrusion detection system against malicious attacks on the communication network of driverless cars. In: 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, USA, pp. 916–921 (2015). https://doi.org/10.1109/CCNC.2015.7158098
12. Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: 4th International Conference on Information Systems Security and Privacy (ICISSP), Purtogal (2018)
13. Toupas, P., Chamou, D., Giannoutakis, K.M., Drosou, A., Tzovaras, D.: An intrusion detection system for multi-class classification based on deep neural networks. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, pp. 1253–1258 (2019). https://doi.org/10.1109/ICMLA.2019.00206
14. Vinayakumar, R., Alazab, M., Soman, K.P., Poornachandran, P., Al-Nemrat, A., Venkatraman, S.: Deep learning approach for intelligent intrusion detection system\. In: IEEE Access **7**, 41525–41550 (2019). https://doi.org/10.1109/ACCESS.2019.2895334; Soreanu, P., Volkovich, Z.: Energy-efficient circular sector sensing coverage model for wireless sensor networks. In: 2009 Third International Conference on Sensor Technologies and Applications, Athens, Glyfada, pp. 229–233 (2009). https://doi.org/10.1109/SENSORCOMM.2009.45
15. Panwar, S.S., Negi, P.S., Panwar, L.S., Raiwani, Y.P.: Implementation of machine learning algorithms on CICIDS-2017 dataset for intrusion detection using WEKA. Int. J. Recent Technol. Eng. (IJRTE). ISSN: 2277–3878, **8**(3) (2019)

Highly Secured Steganography Method for Image Communication using Random Byte Hiding and Confused & Diffused Encryption

S. Aswath, R. S. Valarmathi, C. H. Mohan Sai Kumar, and M. Pandiyarajan

Abstract This paper propounds a new notion for a secured method for image communication using random byte hiding (RBH) technique with confused and diffused data embedding technique. This entire method is based on dual keys for embedding as well as for retrieving. Security through obscurity followed by Kerckhoff's principle is the main ideology of this method. Because of two keys, a large keyspace is needed, which can be effigiated only by brute-force attack. The cover image can be retrieved by using one key, and the secret image can be retrieved by using another key. Among the other dual-key encryption techniques, this method upholds its advantage in the form of security since the dual-key concept is used for two encryptions, which are very difficult to predict and break. The main motivation of the proposed technique is to reduce the time consumption along with increasing the security. The RBH algorithm reduces the encryption time and decryption time by 87.9% and 67.04%, respectively, compared to the conventional LSB steganography. The data hiding rate can also be improved to an extent of 98.33% compared to the conventional LSB technique.

Keywords Chaotic image · Confused and diffused algorithm · LSB technique · Random byte hiding (RBH)

1 Introduction

In digital communication, data security is one of the most significant factors. Since the rise of the Internet, the data security has been a major addressing centre. In order to protect the data, cryptography was established. Different techniques in cryptography

S. Aswath (✉) · R. S. Valarmathi · C. H. Mohan Sai Kumar · M. Pandiyarajan
Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India

R. S. Valarmathi
e-mail: drrsvalarmathi@veltech.edu.in

C. H. Mohan Sai Kumar
e-mail: chmohansaikumar@veltech.edu.in

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_65

were developed to maintain a high secured communication where the secret data is protected properly.

Often keeping the intended message confidential alone is not enough, but the presence of the message can still need to be kept secret. Below figure shows the general type of steganography. Steganography is a technique of concealing a letter inside an envelope where only the envelope is visible and the letter inside the envelope is hidden. Steganos "means" "cover", and "Graphie" means writing. It is derived from Greek terms [1]. Stenography is an ancient art, and it is possible to trace its origins back to 440 BC. Steganography aims to hide a secret message inside another message which itself a meaningful message, without leaving a trace of the original message which is hidden already. Although steganography resembles cryptography, but many differences exist between them. Cryptography is tied in with camouflaging the entire segments of the message, though the scrambled information package is itself the proof of the presence of significant data. By using steganography, we can make the original intended message to be not visible to the illegitimate user. Watermarking and fingerprinting are two different advancements that are firmly identified with steganography, the two of them related with the protection of licensed innovation [2], yet steganography is related to concealing content as data, for example, picture, text, sound and video over another information (Figs. 1 and 2).

Image steganography is a method of hiding an input image inside a cover image so that the final steg image will look like the cover image.

The image steganography is the method that is considered here. Normally the secret data to be secured is encrypted, and it will be embedded in the image information. The resulting image is called as stego image, but here the input image which embeds the secret data can be easily decrypted as there is no other security. Conventional encryption methods cannot be used for multimedia objects whereas it is exclusively used for text and numerical data. Therefore, Lian et al. [3] have proposed a class of partial encryption techniques for encrypting multimedia objects such as image and video. The significant segment in the information is encrypted or

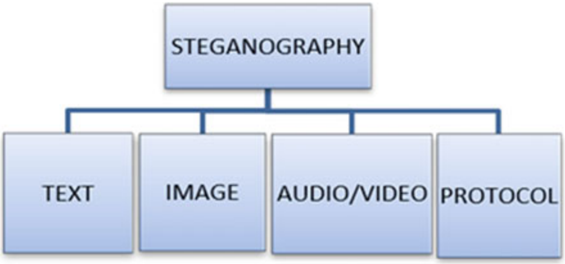


Fig. 1 Different types of steganography

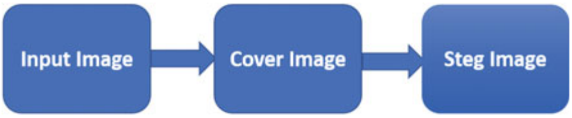


Fig. 2 Image steganography

encoded with standard encryption techniques, whereas non-crucial components are either kept unencrypted or encrypted with some low standard methods of encryption. Yekkala and Venimadhavan [4] have suggested lightweight encryption using lossless compression. Several methods are available for data hiding in an image. Even we can use audio information to hide in a reversible data manner. The interloper can easily decode the data if he knows the data hiding pattern. Phadte et al. [2] stated that dual security can be implemented to enhance the security for encryption standards. Those dual encryptions discussed in [2] include both steganography and cryptography. This paper aims at building an algorithm that implements double encryption only for steganography. To enhance the security and to minimize the security issues here two concepts were discussed, one is random byte hiding and the other is confused and diffused encryption. The confused and diffused encryption is otherwise called as chaotic encryption. Chaotic image encryption's safety can also be increased by using a hybrid optimized schema [5]. In this paper, the dual encryption is more of steganography alone. A secret image is hidden inside another image (cover image). Then that cover image which already has a secret image inside it will be encrypted by using the strongest encryption method called as chaotic encryption method.

Pan et al. [6] discussed the procedures of chaotic encryption and double chaotic encryption. It is indeed possible to use another robust technique called a random byte hiding technique (RBH) as a steganography technique. The detailed procedure of its implementation is discussed in [7] and is implemented for video encryption but this paper employs it for image encryption as the video is a sequence of images. LSB method [8] gives the exceptionally essential thought of steganography in a very simple manner. Here all the secret message bits were supplanted into the LSB of every pixel in the image. The binary information alone is supplanted into the LSB of each pixel of image with a minute change of the pixel value either with increment or decrement of 1 was done [9]. This technique proved powerless against some attacks because the technique follows a predefined sequence of hiding the data in LSBs, and by just selecting the LSBs, the interloper can get to the information. The information in the LSB can be destroyed due to quantization [10].

Thus, an interloper can effortlessly decode the information and is likewise not resistant to the compression strategies and external noise. Instead of using the regular LSB technique, the random byte hiding technique can be used because LSB encryption has a lot of disadvantages [7, 11]. Yun-Qing Shi et al. [12] have suggested few methods for hiding the data in reversible manner. Here also the security issue has to be addressed as the interloper once decodes the pattern then he can easily take out the hidden data as all the data is hidden in reversible manner only. Thus, the algorithm presented in this paper holds a high accountability in security.

1.1 Image

An image is a two-dimensional signal which has both intensity information and colour information. An image is a collection of pixels. Colour image will have three planes namely red, blue and green planes. The combination of values in all three planes gives the colour details. In each plane, the pixel will have values ranging 0–255; therefore, a total of 16.7 million possible colours can be obtained. Practically we will not witness that many colours.

1.2 Plane Separation Process

To process a colour image, it needs a lot of time and energy as there are 16.7 million colour combinations available to process mathematically. In a grayscale digital image, each value of pixel determines the brightness of that pixel, that is, it carries only intensity information.

Grayscale image differs from black and white, as black-and-white images have only the pixel values as 0 and 1, but the grayscale image will have different shades between black and white. The black colour will have a pixel value of 0 and white will have a pixel value of 255 and all the remaining colours range from 1 to 254. The plane separation process provides a very important role in conveying the information in a reduced space. To reduce the complexity, here grayscale image is preferred rather than a colour image. The below image shows the RGB planes of the Lenna image (Figs. 3 and 4).

2 Proposed Method

The proposed method involves two-step encryption process.

1. The cover image will be embedded with the secret image's pixel values using RBH technique.
2. Resultant image from the previous step is put through the confused and diffused encryption algorithm.

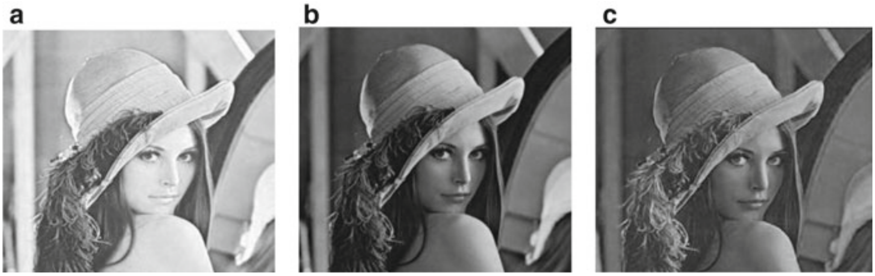


Fig. 3 RGB plane separation. **a** Red plane. **b** Green plane. **c** Blue plane

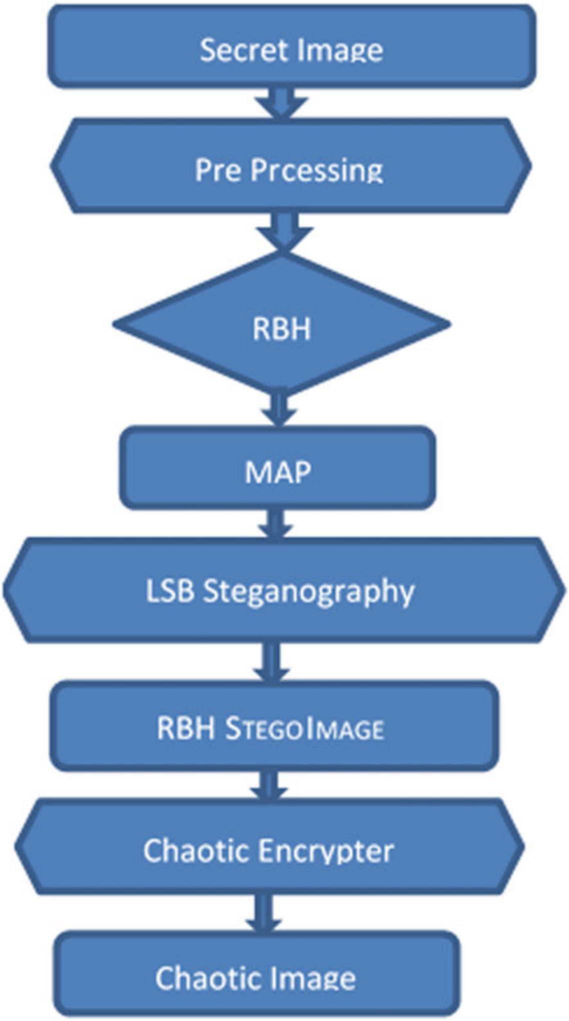


Fig. 4 Flow chart of the proposed method

After the second step, the resultant output image which has the input secret image embedded onto a cover image whose pixels are further displaced is obtained.

2.1 Random Byte Hiding Technique

Here, the input/secret image's binary information is stored in different locations of the cover image. Let "*x*" be the value of a pixel in the secret image, then that pixel information is stored in *y*-"*x*" location of the cover image. Only the legitimate user knows the value of "*y*". "*y*" value should be higher than 256. One bit higher than 255 (the highest pixel value of a grayscale image) so that *y*-"*x*" does not go negative. The "*y*" value is embedded in the cover image's first pixel. So, the legitimate user can find the "*y*" value easily during decryption. Therefore, sharing of key is not needed in this method. The same technique is implemented for video steganography [13].

The same technique can be used on columns too. The lossless hiding in steganography requires putting away the concealed data to a particular location and also it requires a handy amount of time to execute the entire process because it should locate the exact position of the hidden message bits.

Therefore, it is very tough to implement lossless steganography in real time since the execution time and the system specifications are strongly correlated. Higher the system specifications then the lesser execution time of the algorithm can be achieved. Lossy steganography deals with storing the information at LSB location of pixels or at some other particular pixel location. This method suits well for real-time implementation. Therefore, here LSB method is preferred. LSB steganography is the method which is been followed in most of the cases because of its low complexity design. In the event of decryption, the process of finding the payload location fails because it assumes that fixed size of payload will be carried by each image. The payload size can be varied so that this assumption can be broken. In those cases, the information to order the payload logically for recovering it from the encrypted message lies in the mean residuals. Since the data is stored in LSB it can be decrypted easily, but here even though it is decrypted the exact location of the pixel is unknown. Even to make it tough the encrypted image is subjected to further encryption called as confused and diffused encryption. The proposed method's novelty lies in increasing the level of security by adding the security level. First level is implemented by executing RBH, and then the second level of security is done by executing confused and diffused encryption technique, but by adding the level of security the complexity of the design should not be increased so keeping both the things in mind this proposed methodology is designed.

2.2 Confused and Diffused Encryption Technique

It is otherwise called a chaotic encryption technique. Here the image's pixels will be displaced without any order so that it looks like a noise image rather than a meaningful image. In 1989, a chaotic encryption technique was introduced [14]. A variety of chaos algorithms for image encryptions are discussed in [15–25]. Majority old chaotic systems are designed with the base of baker chaotic map [22], logistic chaotic map [25], tent chaotic map [26], 1D-discrete chaotic maps [20, 27] and so on. The chaotic encryption algorithm discussed in [28] provides a good result in both security as well as in the time consumption. For generating pseudorandom numbers, this paper employs Lorenz chaotic system [28]. The Lorenz system is generally given as

$$\begin{cases} \dot{x} = & \mathrm{SP}_1 * (y - x), \\ \dot{y} = & \mathrm{SP}_3 x - y - xz, \\ \dot{z} = & xy - \mathrm{SP}_2 z \end{cases} \quad (1)$$

where SP_1, SP_2 and SP_3 are system parameters.

If $\mathrm{SP}_1 = 10$, $\mathrm{SP}_2 = 8/3$ and $\mathrm{SP}_3 = 28$, then the system has strange chaotic attracter, by inspiring from [29], the two pseudorandom number generators are generated and tested in [27]. The same generators had been considered and applied to our cover image which already had a secret image data in it.

$$S1_i = \mathrm{mod}\big(\mathrm{round}\big((x_i + y_i) * \ 10^{12}, \ 2\big)\big); \ i = 1, \ 2, \ldots \quad (2)$$

$$S2_i = \mathrm{mod}\big(\mathrm{round}\big(z_i * \ 10^{12}, \ 256\big)\big); \ i = 1, \ 2, \ldots \quad (3)$$

where {xi}, {yi} and {zi} are the Lorenz chaotic system's sample sequences over the interval of sampling $T = 0.1$.

Moreover, the sampling interval was chosen in such a way that based on the samples itself the chaotic signal can be totally depicted [29]. To assess the efficacy of both pseudorandom number generators, a standard NIST SP800-22 test was conducted and the test results were summarized [27]. Because of the very good randomness both the pseudorandom number generators can be used without a second thought.

Let us consider an image "*I*" of $M \times N$ size. The below equation represents the vector form of the image I

$$I = \{I_1, I_2, I_3, \ldots I_{MN}\} \quad (4)$$

where $I_1, I_2, I_3, \ldots, I_{MN}$ are the values of pixel in the image I.

Simply it can be written as I_i where $i = 1,2,3, \ldots MN$ and i denotes the position of the pixel.

Switch control mechanism: The two chaotic sequences generated using Lorenz chaotic system [27] are given below

$$\begin{aligned} R &= \{R_1, \ R_2, R_3, \ \ldots \mathrm{RM}\} \\ L &= \{L_1, \ L_2, \ L_3, \ldots \mathrm{LN}\} \end{aligned} \quad (5)$$

Then these chaotic sequences R and L are sorted into two sets:

$$\begin{aligned} \mathrm{SR} &= \{\mathrm{SR1}, \mathrm{SR2}, \mathrm{SR3}, \ldots \mathrm{SRM}\} \\ \mathrm{SL} &= \{\mathrm{SL1}, \mathrm{SL2}, \mathrm{SL3}, \ldots \mathrm{SLN}\} \end{aligned} \quad (6)$$

At last, to mark the positions of every points in the first groupings R and L and also in the sequences, two arbitrary permutations can be used as follows:

$$\begin{gathered}\mathrm{TR} = \{\mathrm{TR1}, \mathrm{TR2}, \mathrm{TR3}, \ldots \mathrm{TRM}\} \\ \mathrm{TL} = \{\mathrm{TL1}, \mathrm{TL2}, \mathrm{TL3}, \ldots \mathrm{TLN}\}\end{gathered} \tag{7}$$

The switch control mechanisms were designed as follows:

Step 1: Choose the first M points from $S1$ and let $\theta = S1$, $i = 1, 2, 3, 4, 5, 6, 7\ldots, P$, where P represents the maximum value of row or column.

Step 2: As per the switch control rule move the pixels of the picture I along with "$S1$" the pseudorandom sequence. (i.e.) transformation of columns will be done if $\theta \neq 0$, transformation of rows will be done if $\theta = 0$ on the plain image. Equation (8) governs the switch control mechanism.

Step 3: After $M \times N$ times of transformations a new matrix "$\overline{\mathrm{I}}_{\mathrm{i}}$" would be obtained. On the off chance that any pixel in the picture has not been permutated totally, then dispose of the main MN points in S1 and rehash the initial two stages until the outcomes perform well.

The switch control mechanism is governed by the following equation

$$\overline{I} = \begin{cases} f1(I), & If S1_i = 0, \\ f2(I), & If \ S1_i = 1, \end{cases} \tag{8}$$

where I and $\overline{I}$ represent the original input plain image and the scrambled image, respectively.

Therefore, anyone who retrieves the cover image from the output chaotic image will think that it is the original input image. But this technique embeds double encryption.

The process of decryption: The backward cycle of image dissemination expects to get the diffused image into its unique value which is the converse of the encryption part. In the Lorenz chaotic framework, the keys utilized in the decryption process are almost similar to the keys utilized in the encryption to get three yielding namely $\{x_{\mathrm{i}}\}$, $\{y_{\mathrm{i}}\}$, $\{z_{\mathrm{i}}\}$ where I represents 1, 2, …, N. Then, a similar technique as utilized in above is used to compute $S1$ and $S2$.

The decryption equation is

$$\overline{\mathrm{Im}}_i = \ \mathrm{mod}\ (\ C_i \oplus S2\ - C_{i-1}\ -\ I_{i-1},\ 256) \tag{9}$$

where C_i is present encrypted value, C_{i-1} is previous encrypted value, $\overline{\mathrm{Im}}_i$ is the value of present shuffled image, $\overline{\mathrm{Im}}_{i-1}$ is the value of previous shuffled image. By using third equation, the value of $S2$ can be found out. The initial value of I_0 is set as zero.

To get θ, extract the P points from $S1$. Transformation of columns will be executed if θ does not equals zero otherwise the transformation of rows will be executed. Also the image transform mode will be dictated by the arbitrary permutation TR and TL.

It is very important to note that the round of permutation part utilized here should be equivalent to the one planned in the encryption cycle. Thus, the cover image can be retrieved.

The secret image (or) original input image can be extracted from the cover image by utilizing a shared key called a chaotic key. Remember the payload area just uncovers the original message bits, not the exact message. In order to get the full message, the payload which was found should be correctly organized in its correct order. That order can be found out by using a key map. This map will guide the decrypters to identify the location of the pixels in proper order. The overall map can be converted as a data stream and that data stream is a mutual key to the receiver. By using the chaos key,s the cover image can be recovered which in turn has to go on another decryption where the original secret image will be recovered.

2.3 Quality Measurement

In terms of measuring the signal quality, the most popular terms used to evaluate the performance measure, especially on the reconstructed image quality of encryption standard, are peak signal-to-noise ratio (PSNR) and mean square error (MSE). The MSE is also called as reconstruction error variance σ_{q}^2. It can be calculated as:

$$\mathrm{MSE} = \frac{\sum_{J,K}[OI(j,k) - RI(j,k)]^2}{J * K} \tag{10}$$

where the number of rows and columns of the original or secret image is represented by J and K.

Equation 10 is used to calculate the error between the original and recovered image (i.e.) OI and *RI*. If the MSE value is lower, then the error rate will also be lower. The peak signal-to-noise ratio (PSNR) value will be mostly calculated in decibels. The PSNR can be calculated between two images. As the PSNR value gets higher, the quality of the recovered image also improves.

$$\mathrm{PSNR} = 10\ \log_{10}\left(\frac{M_i^2}{MSE}\right) \tag{11}$$

where M_i is the image's highest possible pixel value. Here, it is 255 because of 8-bit image. When PSNR equals or greater than 20 dB, the original (or) secret image and the recuperated images are almost identical.

The pixel correlation in the image is another measure which gives the idea about the visual perception too. The sequence relation can be checked in three different orders, viz. horizontal, vertical and diagonal.

3 Results and Analysis

Here, for both the secret image as well as the cover image, the grayscale image is used. All the images are scaled to a size of 256 * 256. The set of images used for comparing the PSNR and MSE values of AES and confused and diffused algorithm are shown (Figs. 5 and 6).

The above images are converted to grayscale and resized to 255 * 255.

Then the random byte hiding was done, and the final image is subjected to confusion and diffusion process to produce a chaotic image. Images shown above are hidden inside the cover image. Here for cover image, cameraman image is used. All the input images are hidden in the cameraman image (Figs. 7, 8 and 9).

The comparison of PSNR and MSE value of normal image between AES and confused and diffused algorithm is given below in Tables 1 and 2.

In encryption technique, mostly AES is described as the standard encryption format because of its enhanced security. Therefore, here the chaotic encryption is compared with AES, and from the above table, it is very evident that chaotic encryption provides a strong security affinity for all the images used in this work. The MSE and PSNR of the chaotic images are low and quite high, respectively, compared to the steg images obtained through AES technique. If the embedding rates are varied then the PSNR value is also varied. The following Table 3 shows the different PSNR values for varying embedding rates.

From Table 3, it is very evident that as the embedding rate is increased the PSNR value gets reduces; therefore, the quality of the image reduces. The choice of embedding rate completely lies with the user. Without affecting the quality of the image,

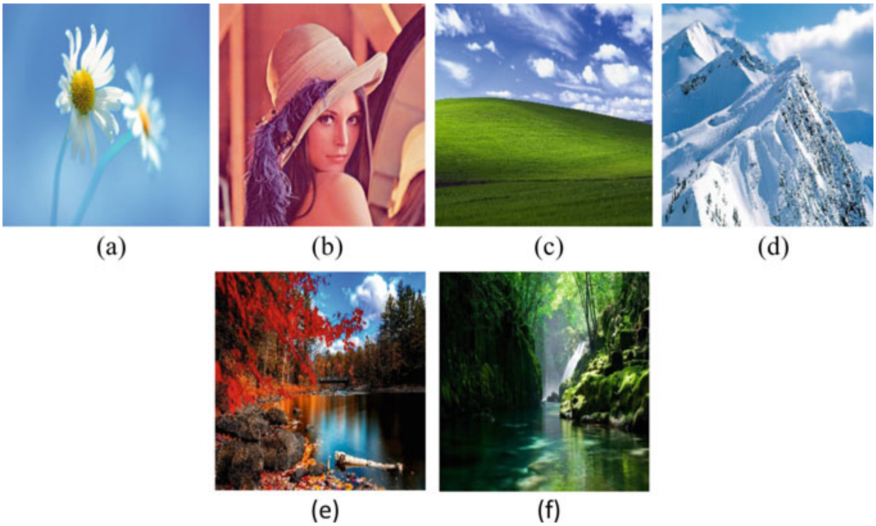


Fig. 5 Input images. **a** Flower. **b** Lena. **c** Hills. **d** Mountain. **e** Nature. **f** Water forest

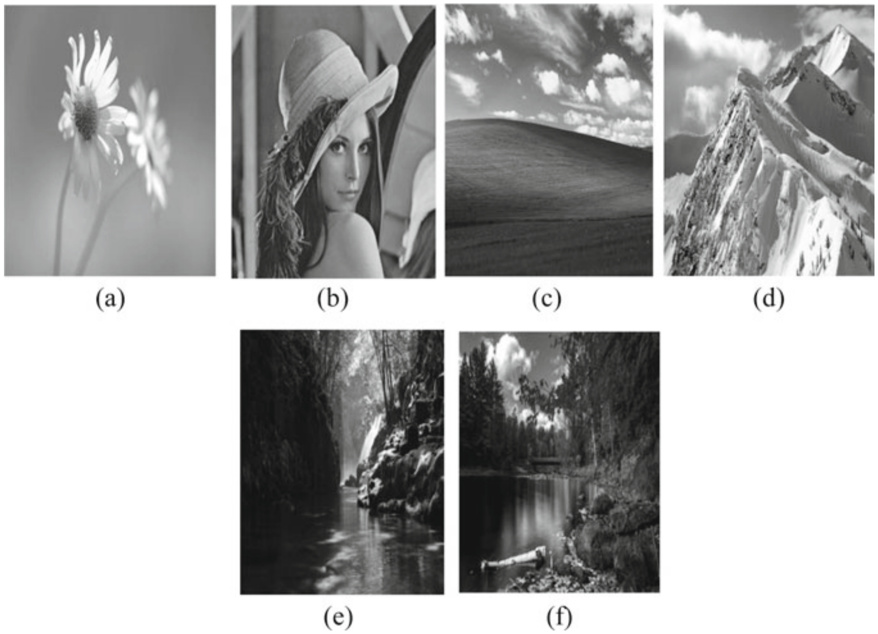


Fig. 6 **a**–**f** Input images converted and resized images

the process of encryption cannot be done easily. Time consumption comparison for the encryption and decryption process of the RBH and LSB methods for a grayscale image of dimension 255 * 255 is given below.

The time consumption comparison of RBH algorithm and LSB algorithm has been given in Figs. 10 and 11.

From the above figure, it is very evident that the time consumption taken by RBH is quite low compared to the LSB technique. Here, the time comparison is made between RBH and LSB since the complexity of both the techniques' design is quite low.

The data hiding ratio between RBH and LSB-based steganography for a frame size of 480 * 680 is discussed in [7] and it is represented in the below figure.

The hiding ratio is considerably reduced by using RBH method so there is a possibility of hiding more data than LSB. The pixel correlation of the input steg image and the chaotic encrypted image is shown in Table 4.

The input steg image has a high-level correlation with adjacent pixels since the correlation coefficient is pretty close to 1. To enhance the security, chaotic encryption is employed, and the correlation coefficient is reduced. In horizontal and vertical directions, the correlation between the adjacent pixels gone negative. The security has been improved a lot. The correlation coefficient for the input image and the decrypted image is given below (Table 5).

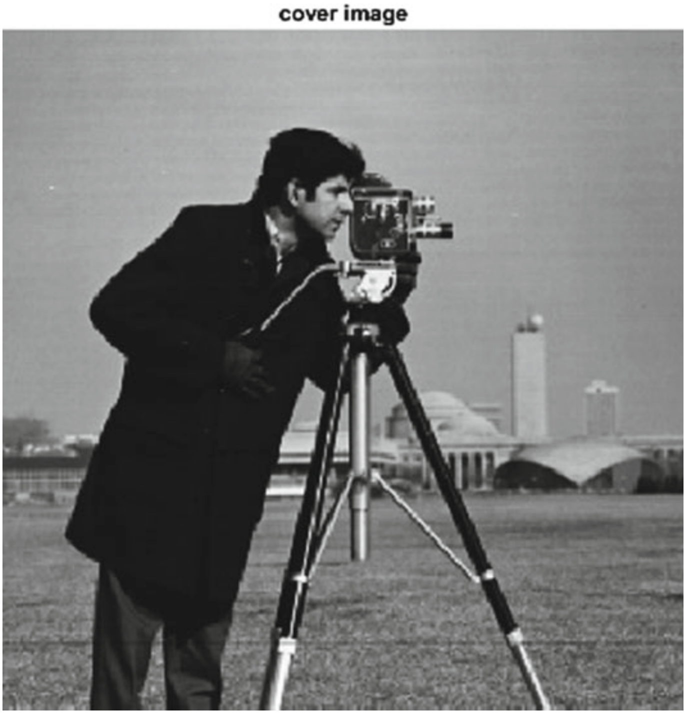


Fig. 7 Cover image

The correlation between the input image and the decrypt image for all images is almost 1. İt proves that the decrypted image is almost equal to the input image. So the recovery is done completely with a minimum of loss in data.

4 Advantages

1. The dual encryption is always advantageous as the security impact is pretty high.
2. Even if an intruder gets access to the image from the confused and diffused image, it is going to be the cover image, only the legitimate user knows the fact that the cover image holds another image which is the original message image, and that image is hidden randomly.
3. The implementation of this algorithm is less complex as well as time-efficient compared to the most followed existing algorithms.

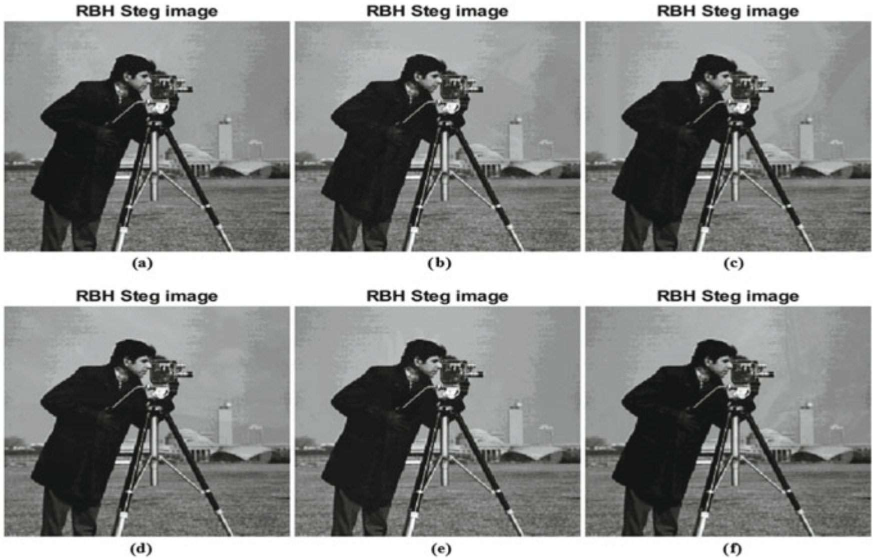


Fig. 8 **a** Flowers image in cameraman image. **b** Mountain image in cameraman image. **c** Lena image in cameraman image. **d** Hills image in cameraman image. **e** Water forest image in cameraman image. **f** Nature image in cameraman image

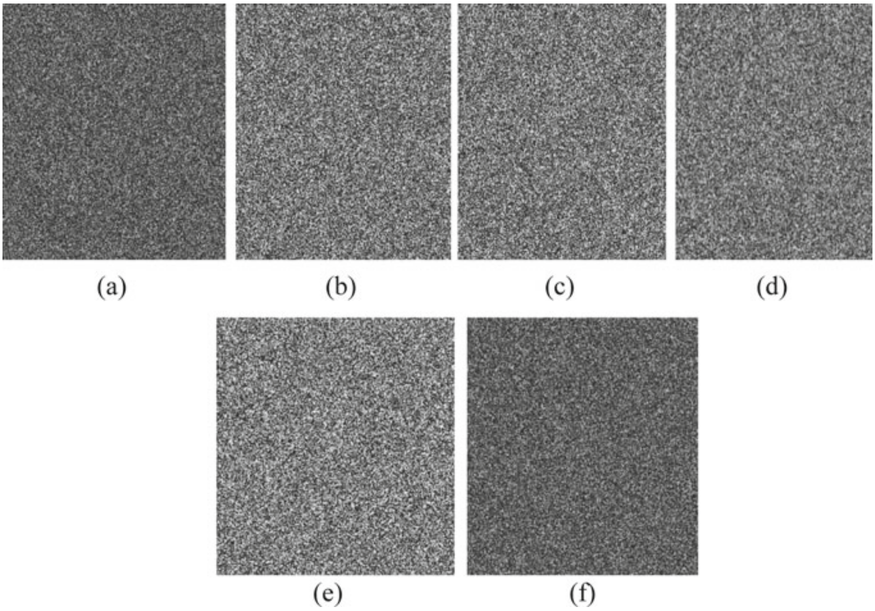


Fig. 9 **a** Chaotic image of flowers image in cameraman image. **b** Chaotic image of mountain image in cameraman image. **c** Chaotic image of lena image in cameraman image. **d** Chaotic image of hills image in cameraman image. **e** Chaotic image of water forest image in cameraman image. **f** Chaotic image of nature image in cameraman image

Table 1 Comparison of PSNR values of AES and chaos encryption

Input	AES	Chaos
Lena	58.8048	75.57
Flowers	72.578	73.69
Hills	70.9206	72.52
Mountain	71.1253	73.86
Nature	69.2735	74.62
Water forest	70.1983	72.67

Table 2 Comparison of MSE values of AES and chaos encryption

Input	AES	Chaos
Lena	0.08564	0.00180054
Flowers	0.0035915	0.00277710
Hills	0.0052605	0.00363159
Mountain	0.0050182	0.00267029
Nature	0.0076866	0.00224304
Water forest	0.0062122	0.00350952

Table 3 PSNR versus embedding rates

	0.05	0.1	0.2	0.3	0.4	0.5
Lena	74.72	72.56	68.79	66.52	62.68	49.04
Flower	77.38	74.10	71.98	69.20	68.67	50.78
Hills	77.70	74.78	71.56	69.26	68.56	50.78
Mountain	74.87	71.45	66.26	64.56	61.23	52.80
Nature	76.87	73.63	70.36	63.69	60.76	51.69
Water	74.46	72.13	67.58	64.36	61.25	50.46

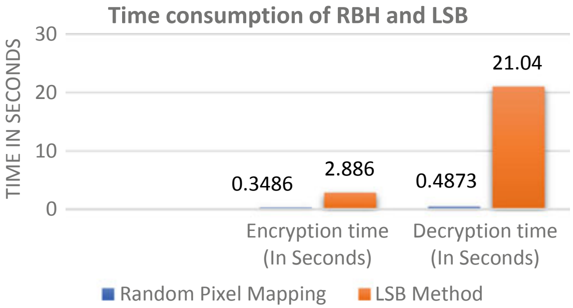


Fig. 10 Time comparison between encryption and decryption

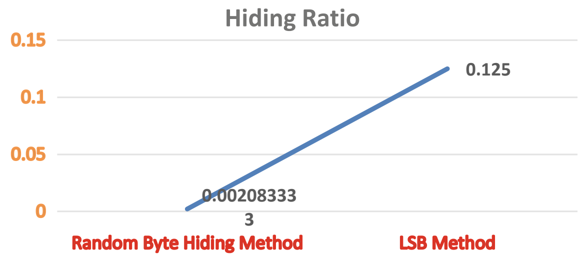


Fig. 11 Hiding ratio between RBH and LSB

Table 4 Pixel correlation of input steg image and chaotic image

Direction	Input steg ımage	Chaotic encrypted ımage
Flowers image in cameraman image		
Horizontal	0.9312 (H_{STI1})	-0.0075 (H_{CHI1})
Vertical	0.9565 (V_{STI1})	-0.0060 (V_{CHI1})
Diagonal	0.9066 (D_{STI1})	0.0029 (D_{CHI1})
Mountain image in cameraman image		
Horizontal	0.9309 (H_{STI2})	-0.0069 (H_{CHI2})
Vertical	0.9561 (V_{STI2})	-0.0061 (V_{CHI2})
Diagonal	0.9061 (D_{STI2})	0.0033 (D_{CHI2})
Lena image in cameraman image		
Horizontal	0.9300 (H_{STI3})	-0.0069 (H_{CHI3})
Vertical	0.9556 (V_{STI3})	-0.0063 (V_{CHI3})
Diagonal	0.9048 (D_{STI3})	0.0028 (D_{CHI3})
Hills image in cameraman image		
Horizontal	0.9332 (H_{STI4})	-0.0066 (H_{CHI4})
Vertical	0.9576 (V_{STI4})	-0.0069 (V_{CHI4})
Diagonal	0.9091 (D_{STI4})	0.0029 (D_{CHI4})
Water forest image in cameraman image		
Horizontal	0.9316 (H_{STI5})	-0.0057 (H_{CHI5})
Vertical	0.9563 (V_{STI5})	-0.0071 (V_{CHI5})
Diagonal	0.9069 (D_{STI5})	0.0029 (D_{CHI5})
Nature image in cameraman image		
Horizontal	0.9316 (H_{STI6})	-0.0056 (H_{CHI6})
Vertical	0.9561 (V_{STI6})	-0.0071 (V_{CHI6})
Diagonal	0.9069 (D_{STI6})	0.0030 (D_{CHI6})

5 Limitations

The only limitation of this method lies in RBH technique as the input images has to be resized to 255*255. If the input image is of some size greater than the specified size,

Table 5 Pixel correlation of input image and decrypted image

Image	Correlation coefficient
Flowers image in cameraman image	0.9575
Mountain image in cameraman image	0.9875
Lena image in cameraman image	0.9978
Hills image in cameraman image	0.9987
Water forest image in cameraman image	0.9658
Nature image in cameraman image	0.9985

then due to the resizing operation we may loss some information. Therefore, for some sensitive images this method is not suitable. This method gives fruitful advantages for the communication of the general images where only the visual quality alone is needed to be maintained.

6 Conclusion and Future Work

The algorithm implemented in this paper can provide a double layer of security for image communication. This combined or hybrid algorithm provides a good standin both MSE as well as in PSNR. The experimental results also validate the same. The time efficiency is also pretty good. For confused and diffused algorithm, the PSNR calculated for a basic of five images. It shows an average improvement of 3.4%. Mean square error also been reduced by an average of 20%. Therefore, the chaotic image provides a good improvement in terms of image quality compared to AES (the most used technique). By using the RBH algorithm the time taken for encryption and decryption also reduced by 87.9 and 67.04% compared to the conventional LSB technique. Instead of processing only grayscale images, the same technique can be used to process colour images as well since the world progress towards high resolution multicolour images. While processing in colour image the trade-off between complexity and time consumption has to be taken care properly to achieve a desired result. The same technique can be used for medical images as most of the scan images appear in grayscale range. So this technique will be very useful for the medical image steganography if the resizing operation finds any alternative as we tends to lose some information because of this resizing operation done in RBH algorithm. The correlation coefficient for the confused image is also very low and it is even in negative for horizontal and vertical.

Acknowledgements The authors like to thank all the listed reference papers' author for providing knowledge support to carry out this work successfully.

Availability of Data and Materials All images used in this work are available in MATLAB as well as windows desktop background images.

Competing Interests The authors declare that there are no competing interests.

Funding Not applicable.

Authors' Contributions Author SA formulated and implemented the methodology. SA, CHM, MP drafted the manuscript. RSV adjusted the methodology and supervised the entire work. SA, CHM carried out the survey related to the methodology. SA, MP carried out the result comparison with previous methodologies. RSV reviewed the manuscript and made the appropriate changes.

References

1. Moerland, T.: Steganography and steganalysis. Leiden Institute of Advanced Computing Science. www.liacs.nl/home/tmoerl/privtech
2. Phadte, R.S., et al.: Enhanced blend of image steganography and cryptography. In: International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (2016)
3. lian, S., et al.: On the design of partial encryption scheme for multimedia content. Math. Comput. Model. **57**(11–12), 2613–2624 (2013)
4. Yekkala, A., Venimadhavan, C.E.: Bit plane encoding and encryption. In: Proceedings of PReMI 2007; Lecture Notes in Computer Science, vol. 4815, pp. 103–110 (2007)
5. Hong, L.X., et al.: An ımage encryption schema based on hybrid optimized chaotic syst. IEEE Explore, 18 May 2020
6. Pan, H., et al.: Research on digital image encryption algorithm based on the double logistic chaotic map. EURASIP J Image Video Process (2018)
7. Bhole, A.T., Patel, R.: Steganography over video file using random byte hiding and LSB technique. In: 2012 IEEE International Conference on Computational Intelligence and Computing Research
8. Johnson, N.F., Jajodia, S.: Exploring steganography: seeing the unseen. IEEE Comput. **31**(2), 26–34 (1998)
9. Anderson, R.J.: Stretching the limit of steganography in information hiding. In: Springer Lecture Notes in Computer Science, vol. 1174, pp. 39–48 (1996)
10. Singh, P., et al.: Evaluating the performance of message hidden in first and second-bit plane. WSEAS Trans Inf. Sci. Technol. **2**(8), 1220–1222 (2005)
11. Verma, V., et al.: An enhanced least significant bit steganography method using midpoint circle approach. In: International Conference on Communication and Signal Processing (2014)
12. Shi, Y.Q., et al.: Reversible data hiding: advances in the past two decades. IEEE Access **4**
13. Aswath, S., et al.: Implementation of random byte hiding algorithm in video steganography. Int. J. Eng. Res. Technol. (2017)
14. Matthews, R.: On the derivation of a "chaotic" encryption algorithm. Cryptologia **13**(1), 29–42 (1989)
15. Chen, G., et al.: A symmetric image encryptionscheme based on 3d chaotic cat maps. Chaos Solitons Fractals **21**(3), 749–761 (2004)
16. Zhang, L., et al.: An image encryption approachbased on chaotic maps. Chaos Solitons Fractals **24**(3), 759–765 (2005), Accessed: 24 Nov 2020
17. Pareek, N.K., et al.: Image encryptionusing chaotic logistic map. Image Vis. Comput. **24**(9), 926–934 (2006)
18. Zhou, Y., et al.: A new 1d chaotic system for image encryption. Signal Process. **97**, 172–182 (2014)
19. El Assad, S., et al.: A new chaos-based image encryption system. Signal Process. Image Commun. **41**, 144–157 (2016)
20. Pak, C., et al.: A new color image encryption using combination of the 1d chaotic map. Signal Process. **138**, 129–137 (2017)

21. Li, C., et al.: An image encryption scheme based on chaotictent map. Nonlinear Dyn. **87**(1), 127–133 (2017)
22. Mao, Y., et al.: A novel fast image encryption scheme based on 3d chaotic baker maps. Int. J. Bifurcat. Chaos **14**(10), 3613–3624 (2004)
23. Ghadirli, H.M., et al.: An overview of encryption algorithms in color images. Signal Process. **164**, 163–185 (2019)
24. Mishra, D.C., Sharma, R.K., Suman, S., Prasad, A.: Multilayer security of color image based on chaotic system combined with RP2DFRFT and Arnold transform. J. Inf. Secur. Appl. **37**, 65–90 (2017)
25. Zhang, Y.-Q., et al.: A symmetric image encryption algorithm based on mixed linear-nonlinear coupled map lattice. Inf. Sci. **273**(8), 329–351 (2014)
26. Wang, X., Zhang, H.L.: A novel image encryption algorithmbased on genetic recombination and hyper-chaoticsystems. Nonlinear Dyn. **83**(1–2), 333–346 (2016)
27. Al-Maadeed, S., et al.: A new chaos-based ımage-encryption and compression algorithm. J. Electr. Comput. Eng. **2012**, Article ID 179693 (2012)
28. Xiao, S., Yu, Z., Deng, Y.: Design and analysis of a novel chaos-based ımage encryption algorithm via switch control mechanism. Hindawi Secur. Commun. Netw. **2020**, Article ID 7913061, 12 (2020)
29. Hu, H., et al.: Pseudorandom sequence generator based on the chen chaotic system. Comput. Phys. Commun. **184**(3), 765–768 (2013)

An Enhanced Energy Constraint Secure Routing Protocol for Clustered Randomly Distributed MANETs Using ADAM's Algorithm

Bandani Anil Kumar, Makam Venkata Subamanyam, and Kodati Satya Prasad

Abstract Mobile ad hoc network (MANET) contains association attributed to autonomous enforced dynamically changing mobile nodes that forms a dynamic network beyond having fixed network infrastructure. In the mobile ad hoc network, each and every mobile node will be operated with battery sources and each mobile node consists of limited energy. Various routing protocols are developed for routing, but every scheme suffers from the power constraint. The proposed scheme gives a solution for manipulating diverse paths among destination and source to reduce power constraints. The major intention is for finding best route among available routes between source and intermediate target mobile nodes. The energy model is considered to enhance the energy in routing protocol. Here, the ADAM's algorithm is devised for implementing power aware routing in the MANET environment. At the beginning, the dynamic mobile nodes are compiled and simulated in a randomly dispersed domain for persuading energy constrained secure routing, where the power in each and every mobile node is being computed. Once the energy levels of all the mobile nodes are evaluated, then the secure routing scheme is introduced and settled by using dynamic secure nodes. Thus, the secure routing scheme is developed by using the recommended ADAM's model. Thus, by considering the ADAM's algorithm, the most favorable way of information routing is continued in MANET ambiance.

Keywords MANET · Energy efficient model · Routing · ADAM's algorithm · AODV · Secure routing protocol

B. A. Kumar (✉)
B V Raju Institute of Technology, Narsapur, Medak District, Telangana, India

B. A. Kumar · K. S. Prasad
JNTU College of Engineering, Kakinada, Andhra Pradesh, India

M. V. Subamanyam
Shanthiram Engineering College, Nandyal, Kurnool District, Andhra Pradesh, India

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_66

1 Introduction

Mobile ad hoc network (MANET) contains association of self-governing enforced dynamic nodes that forms a dynamic network beyond having fixed network infrastructure. In ad hoc network, routing is treated as a big problem because of their changing topologies induced by the mobile nodes mutability in the dynamic mobile ad hoc network and also because of problems in node energy and link bandwidth. Due to the network framework defect, the nodes behave as router to allow association encompassed by the nodes which produces various paths from the source to the intermediate marked mobile nodes for a definite time sample. The traditional network routing schemes are not applicable to perform routing action because of undefined attributes of ad hoc network. Thus, the network routing has become a big controversy in ad hoc network. The data networking protocols engaged in physical wired networks are arranged based on distance vector method and link state routing method. To transmit the data over the mobile ad hoc network, both the algorithms require recurrent network routing publicity. In addition, the varying topologies produced by the movability of mobile nodes influence the power utilized for transmitting data. Thus, routing schemes, which contain methodologies that can handle the challenges sustained by the node's mobility, varying topologies, and power constraints, are required in MANETs. Moreover, the routing protocols must be effective in power consumption and quality of service (QoS) for guarantying the transmission of data through the wireless medium. ADAM optimization is an extension to stochastic gradient decent and can be used in place of classical stochastic gradient descent to update network weights more efficiently. The most beneficial nature of ADAM optimization is its adaptive learning rate.

As per the authors, it can compute adaptive learning rates for different parameters. This is in contrast to the SGD algorithm. SGD maintains a single learning rate throughout the network learning process. We can always change the learning rate using a scheduler whenever learning plateaus. But we need to do that through manual coding.

2 Review of Literature

Sarkar [1] initiated the energy-aware routing algorithm that overcomes the design issues of energy aware routing protocol by developing optimal and efficient energy aware routes. Based on energy consumption of each mobile node and its residual energies, energy-aware routing protocols are developed. While selecting a route, initial energy levels of mobile nodes and their residual energy levels are considered. To avoid link failure due to lack of energy levels in all the mobile nodes in the selected route, the considered parameters ensure better performance in energy awareness and reliability in MANETs.

Smith et al. [2] expressed the security using antecedent routing “SUPERMAN” method in MANET to address mobility, protection for route and data in the presence of confined network framework. This developed method achieves better mobile node attestation, security for route and communication and data access control.

Kumar et al. [3] introduced energy efficient clustering method for MANETs using AODV K-means routing protocols. By considering this method, the consumption of energy of each mobile node in MANET is reduced. A new type of energy consumption model is introduced with a fanatical routing protocol. The developed routing protocol builds upon the dynamic behavior of mobile nodes and energy levels of mobile nodes. Energy cost estimation is executed depending upon estimation of energy utilization levels of mobile nodes using K-means and AODV algorithms.

Sunitha et al. [4] introduced a predetermined mechanism “PEBER” which focuses to find the required energy by considering route length and sufficient energy levels required between mobile nodes. The method minimizes routing overhead, energy consumption and controls the data packet loss which occurs due to mobile node mutability.

Selvi and Balakrishna [5] expressed an efficient routing protocol approach to regain the route failure and efficient information distribution in ad hoc networks. This routing protocol method is very much useful for dependable data dispatch for varying mobile ad hoc environment. By considering this method, link failure is eliminated and data is transferred to receiver with better security.

Zapata and Asokan [6] expressed SAODV which is used to incorporate security mechanism into low power routing protocols for MANETs. This method is used to introduce security process to protect its routing information. To implement this method, two mechanisms were used to protect the AODV information; they are digital signatures and hash chains to protect mobile node count data. SAODV digital signatures are used to secure the integrity of the non-changeable data. SAODV hash chains are used to authenticate the mobile node count in such a way that allows not every mobile node that receives the message to verify that the mobile node count is reduced by an external hacker (Fig. 1).

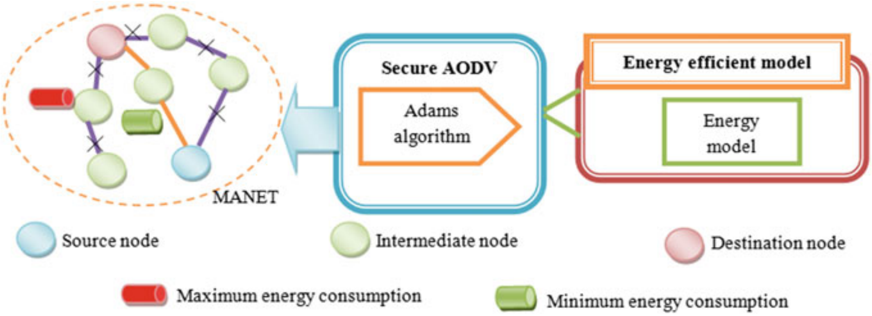


Fig. 1 Model diagram of energy-aware model

3 Proposed ADAM's Algorithm for Energy Constrained Routing in Mobile Ad Hoc Networks (MANET)

This proposed approach defines the ADAM's algorithm which provides energy attentive network routing method in mobile ad hoc network (MANET). In the beginning, dynamic moving nodes are executed in a randomly distributed network scenario to assure energy constrained secure network routing. Energy values of all the dynamic nodes are evaluated. Once the initial energy values of all the dynamic nodes are computed, the protected network routing method is introduced using the network protected nodes. The network protected routing scheme is implemented using developed ADAM's method. Therefore, the best approach of data routing is continued in MANET environment. Figure 2 presents the system diagram of energy aware routing scenario in MANET.

3.1 MANET System Model

The classification of ad hoc network (MANET), characterized in Fig. 2, explains the information transmission from transmitter node to destination mobile node by considering a single route as the optimal path for routing the data. Here, the gateway is a communicating device responsible for providing an interface among the intermediate networks for improving connectivity and coverage. The mobile nodes access the Internet connectivity for initiating the communication with wireless devices and to access the information. Assume a graph $S(X, Y)$ in a MANET, where $X = \left\{x_1, x_2, \ldots x_a, \ldots x_q\right\}$ represents a node routing scenario in MANET set, where q represents the total nodes present in the network where $1 \leq a \leq q$ and Y denotes a link set, where $y = \left\{y_1, y_2, \ldots, y_j\right\}$, which connects two nodes x_p and x_s where $1 \leq p \leq s \leq q$. The multi-objective functions considered for designing the MANET are reputation factor, power, distance, and delay. The links between the MANET nodes use the multi-objective functions for transmitting data between the two nodes x_p and x_s. Assume that, A represents the source node, which sends information toward the destination, expressed as B. During sending information from one mobile node to another mobile, energy is treated as a major criterion, which is used for determining the paths. The briefer illustration of energy model is described below.

3.2 Energy Model

Mobile node to another mobile node. The mobile network consists of various scattered dynamic mobile nodes operated with energy sources and energy level of each mobile node drains stirring the overall lifetime of entire mobile network because of

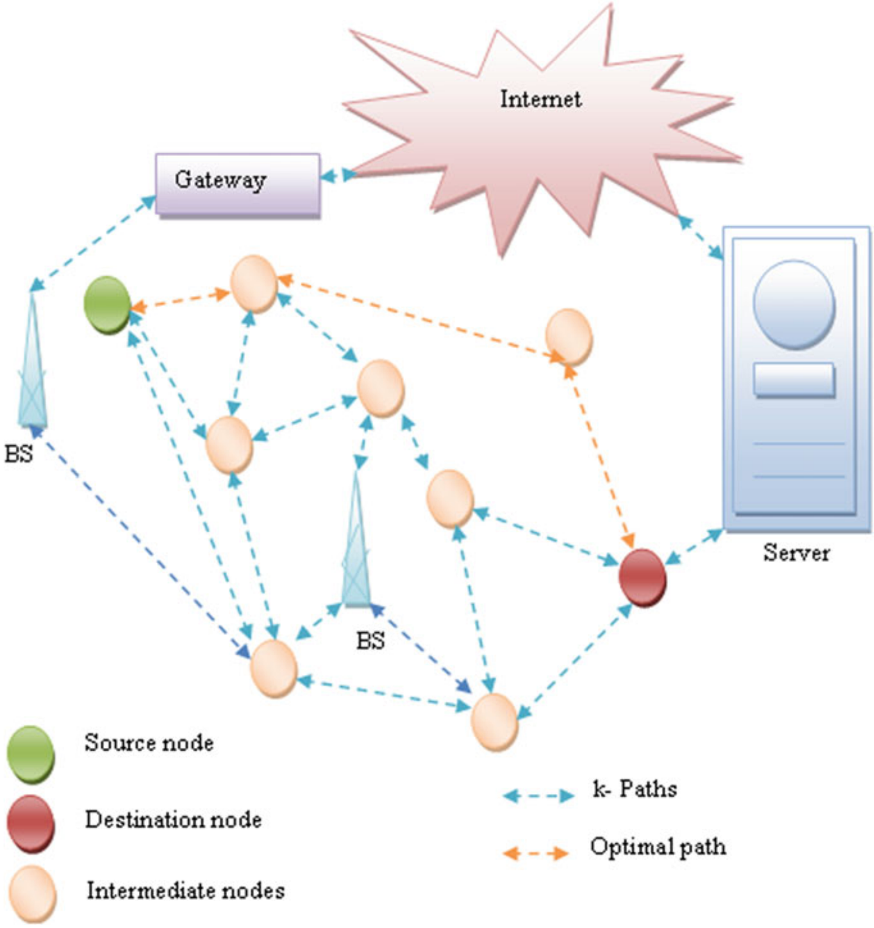


Fig. 2 System model of MANET

continuous functioning of the dynamic mobile network. Considering the early energy level of the mobile nodes is,ε_0 which specifies that the energy sources (batteries) are non-recoverable. When the information is transmitted from source to destination, some amount of power is reduced in the dynamic mobile nodes because of the distance between source and destination. The communication in the dynamic mobile network depends on the routing protocol and the power dissipation occurs because of existence of power amplifiers and radio electronics which are present at the transmitter. It is clear that based on distance from source to destination and dynamic nature of mobile nodes, power gratification is introduced into the mobile network. The power scheme is defined as follows: If P defined as total power of a mobile node, the residual power of the mobile node a at time slot t is defined as $\varepsilon_a^{rem}(t)$ and it is expressed by,

$$\varepsilon_a^{\text{rem}}(t) = \varepsilon_a^{\text{rem}}(t-1) - \varepsilon_a^P \times n(t-1,t) - \varepsilon_a^r \times n(t-1,t) \tag{1}$$

and, ε_a^P indicates the volume of power needed to transfer data packets, n specifies total count of data bits transmitted from t to $t-1$, and ε_a^r exemplify volume of power required to accept data packets. At a time instant value, $t = 0$, the residual power in the mobile node is completely adequate and it is expressed as $\varepsilon_a^{rem}(t) = P$. It is treated that residual power in a mobile node lies among two value ranges and it is defined by energy ratio; it is given by the equation stated below,

$$\text{energy ratio} = \frac{\varepsilon_a^{rem}(t)}{P} \tag{2}$$

where the remaining energy lays one of two in small range or in big range and $0 \leq \varepsilon_a^{rem} \leq 1$.

Case 1: If the parameter,energy ratio $\leq \tau$, where τ is uppermost threshold value of small range wherein the residual power is not sufficient to send information in the mobile ad hoc network.

Case 2: If energy ratio $> \tau$, then the residual power is sufficient to transmit data in mobile ad hoc network. Hence, the end user can identify the parameter value of threshold based on the need of user. The range of residual power of a particular mobile node supports to evaluate the power levels and status of each mobile node. Hence, the power $\varepsilon_{a,h}$ of mobile node a in hth route is evaluated using the residual power $\varepsilon_a^{\text{rem}}$. The behind category addresses the ADAM's method for introducing routing into mobile ad hoc network.

4 Differences Between ADAM's and EADAM's Algorithm

5 Routing in MANET Using ADAM's Algorithm

ADAM [7] is a first-order stochastic gradient-based optimization, which is extensively applied to the objective function that alters in contrast to the attributes. The major implication of the technique is computational efficiency and less memory requirements. In addition, the issues linked with the non-stationary objectives and the existence of noisy gradients are also managed in an effective manner. Moreover, the ADAM poses following benefits. Here, the magnitudes of updated parameters are invariant in contrast to the rescaling of gradient and also the step size is managed using hyperparameter, which works with the sparse gradients. Furthermore, the ADAM's algorithm is effectual in performing step size annealing.

Step 1: Loading

In this algorithm, the foremost step is loading of bias corrections, where $\hat{q}_l$ signifies corrected bias of first moment estimate and $\hat{m}_l$ symbolizes corrected bias of second moment estimate.

Step 2: Computation of the fitness function

In this algorithm, the fitness of bias is evaluated to choose best path in this dynamic routing. The fitness function is defined, as error activity which yields to global optimum solution. The function is termed as a minimization function and is expressed as,

$$\mathrm{Err} = \frac{1}{r}\sum_{k=1}^{r}\left(H_k - H_k^*\right)^2 \tag{3}$$

where r signifies total nodes, H_k symbolizes the output nodes generated from algorithm, H_k^* denotes the ground truth value.

Step 3: Determination of updated bias

ADAM's algorithm is used to improve optimization and behavior of convergence.

This technique generates smoother variation with effectual computational efficiency and less memory requirements. As per ADAM [7], the bias is expressed as shown below,

$$\theta_l = \theta_{l-1} - \frac{\alpha\hat{q}_l}{\sqrt{\hat{m}_l + \varepsilon}} \tag{4}$$

where α signifies step size, $\hat{q}_l$ represent corrected bias, $\hat{m}_l$ symbolize bias corrected second moment estimate, ε denote constant, and θ_{l-1} represents the parameter at previous time instant $(l - 1)$. The corrected bias of first-order moment is expressed as,

$$\hat{q}_l = \frac{q_l}{(1 - \eta_1^l)} \tag{5}$$

$$\hat{q}_l = \eta_1 q_{l-1} + (1 - \eta_1)G_l^1 \tag{6}$$

The corrected bias of second-order moment is represented as,

$$\hat{m}_l = \frac{m_l}{(1 - \eta_2^l)} \tag{7}$$

Step 4: Determination of optimal solution

The best solution is determined using error measure, and solution with the best solution is employed for finding the leaf disease classification (Fig. 3).

Step 5: Stopping criterion

The optimal weights are obtained in a repeated fashion, until the maximal iterations are attained.

5.1 Algorithm of the Proposed Energy Constrained Method

1. Start the Process.
2. Selection of Mobile Nodes.
3. Analyze the Energy Levels of Each Mobile Node.
4. Select the Routing Methodology based on ADAM's Algorithm.
5. Begin Node List Iteration based on Energy Level of all the Mobile nodes.
6. Find the shortest path between Source and Destination.
7. Start Packet Transmission.
8. End the Process.

Additionally, the simulation framework is accomplished by changing the count of mobile nodes as 10, 15, 20, 25, and 30. The media access control type is 802-11 with omnidirectional antenna-based system. This part will give a comprehensive glimpse of simulation outputs which are achieved by making use of "ADAM's" algorithm for affording efficient energy constrained and secure routing.

6 Experimental Results

The ADAM's algorithm is implemented and simulated in NS2 simulation environment. Simulation process framework is expressed in Table 1. The output extraction process, simulation process start time, and elapsed time are expressed as 0.001 to 50.000. By correlating the simulation results with the developed present methods, the conduct is evaluated by calculating the parameters of end-to-end delay, energy consumption, packet delivery ratio (PDR), data routing overhead, and throughput framework. The attainment parameters are given below.

- **Packet Delivery Ratio (PDR)**

It is defined as total no. of dynamic data packets accepted at the receiver to the total no. of dynamic data packets delivered from transmitter. It is expressed in Eq. (3).

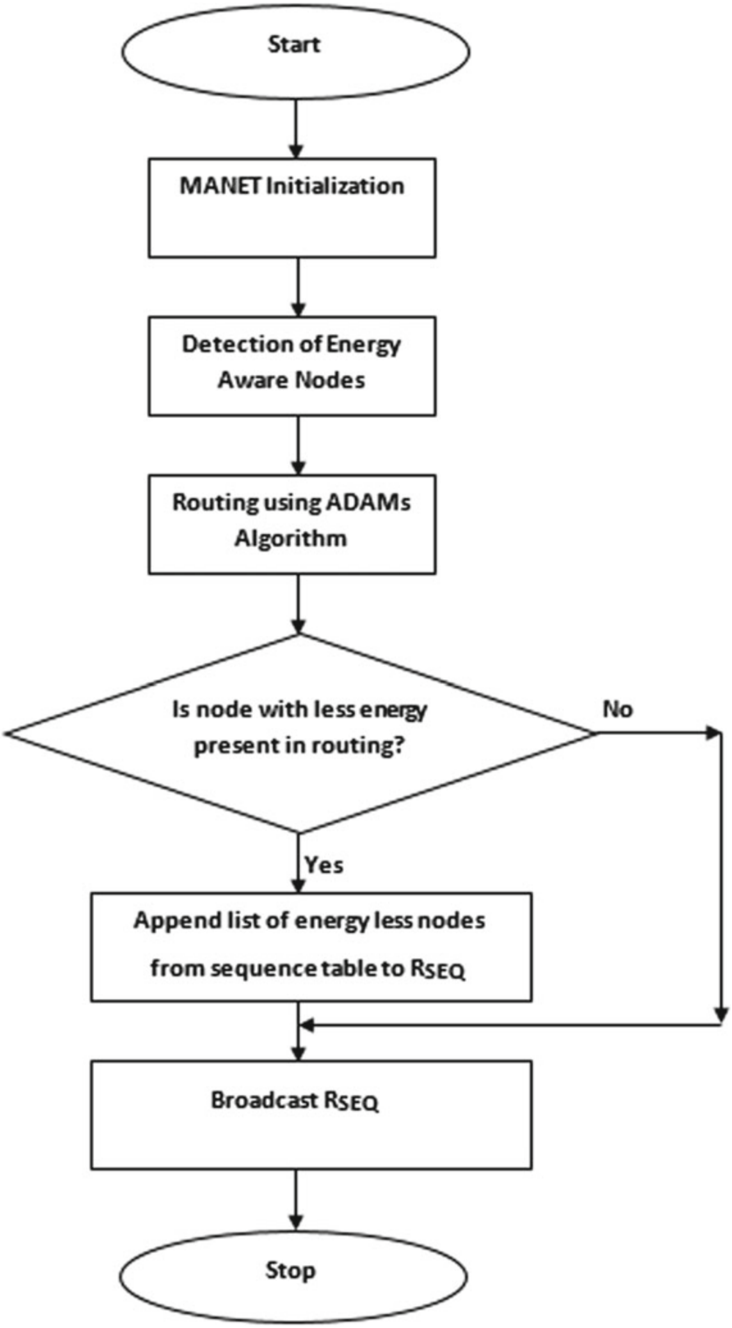

Start
MANET Initialization
Detection of Energy Aware Nodes
Routing using ADAMs Algorithm
Is node with less energy present in routing?
No
Yes
Append list of energy less nodes from sequence table to RSEQ
Broadcast RSEQ
Stop

Fig. 3 Flowchart of the proposed energy constrained method

Table 1 Diferences between ADAM's and EADAM's algorithm

ADAM's algorithm	EADAM's algorithm
Data traffic loss is high	Data traffic loss is medium
Collision warning is not present	Collision warning is present
High packet delay is present	Low packet delay is present
Low throughput is present	High throughput is present
High energy consumption	Low energy consumption

$$\text{PDR} = \frac{\text{Total no.of data packetssent}-\text{data packetslost}}{\text{Total no. of data packets delivered}} \times 100 \tag{8}$$

- **Energy Consumption**

It is expressed as total amount of power used by each and every mobile node for every packet transmission and reception to the total simulation time. It is defined in Eq. (4)

$$\text{Energy Consumption} = \frac{\text{Total Power for each data Packet}}{\text{Total Time taken for Simulation}} \times 100 \tag{9}$$

- **Total End-to-End Delay**

It is described as the disparity among sending time of data packets and receiving time of data packets.

$$\begin{aligned}\text{Delay} =& \text{Total amount of time spent on node1}\\ &+ \text{Total amount of time spent on node2}\\ &+ \cdots \text{Total amount of time spent on noden}\end{aligned} \tag{10}$$

- **Network Routing Overhead**

It is designated as total no. of data routing packets sent for preservation and also for path detection.

- **Toughput**

It is designated as ratio of data packets received by the target receiver and total no.of packets sent.

$$\text{Throughput} = \frac{\text{Data Packets accepted by the target}}{\text{Total no. of data packets transmitted}} \times 100 \tag{11}$$

Table 2 represents the end-to-end delay by varying dissimilar mobile nodes such as 10, 15, 20, 25, and 30 mobile nodes. Therefore, "EADAM's" method gives efficient simulation results when compared with SAODV and ADAM's.

Table 3 conveys the energy consumption framework by changing dissimilar mobile nodes such as 10, 15, 20, 25, and 30 mobile nodes. Therefore, "EADAM's" method gives efficient simulation results when compared with SAODV and ADAM's.

Table 4 indicates the packet delivery ratio by differing dissimilar mobile nodes such as 10, 15, 20, 25, and 30 mobile nodes. Therefore, "EADAM's" method gives efficient simulation results when compared with SAODV and ADAM's.

Table 5 demonstrates the routing overhead by varying dissimilar mobile nodes such as 10, 15, 20, 25, and 30 mobile nodes. Therefore, "EADAM's" method gives efficient simulation results when compared with SAODV and ADAM's.

Table 6 exhibits the throughput by changing dissimilar mobile nodes such as 10, 15, 20, 25, and 30 mobile nodes. Therefore, "EADAM's" method gives efficient simulation results when compared with SAODV and ADAM's.

The output extraction simulation parameters are displayed in Table 6. The process of simulation start and completion time is denoted as 0.001–50.000 correspondingly by altering the no. of immovable mobile nodes as 10, 15, 20, 25, and 30. The Media Access Control Type is 802-11 with Omni Antenna Directional System (Table 7).

Table 2 End-to-end delay

QoS	End-to-end delay (ms)				
Nodes	10	15	20	25	30
ADAM's	408	415	460	480	505
SAODV	408	410	440	460	480
EADAM's	318	380	420	430	440

Table 3 Energy consumption

QoS	Energy consumption (Joules)				
Nodes	10	15	20	25	30
ADAM's	3	3.1	3.15	4	5
SAODV	2.8	2.85	2.89	3.22	4.4
EADAM's	2	2.68	2.89	3.0	3.5

Table 4 Packet delivery ratio

QoS	Packet delivery ratio				
Nodes	10	15	20	25	30
ADAM's	400	410	430	475	480
SAODV	400	405	405	440	465
EADAM's	500	520	530	560	590

Table 5 Routing overhead

QoS	Routing overhead				
Nodes	10	15	20	25	30
ADAM’s	300	340	350	342	343
SAODV	340	340	330	340	343
EADAM’s	230	340	280	300	310

Table 6 Throughput

QoS	Throughput (kbps)				
Nodes	10	15	20	25	30
ADAM’s	510	530	612	710	730
SAODV	510	515	600	700	710
EADAM’s	610	612	628	790	795

Table 7 Simulation parameters of ADAM’s algorithm

Clustering technique	Fuzzy
Packet routing method	AODV
Optimization methods	Fuzzy optimization
Tool used or simulation	NS2
Start time of simulation	0.001 s
End time of simulation	50 s
Total no. of nodes	30
Antenna system used	Omni Directional Antenna
Speed	30 ms
Mobile network interface types	Wireless
MAC type	MAC-802-11
Introductory transmit power	0.75
Introductory initial receive power	0.75

6.1 Merits, Issues, and Challenges of Proposed Algorithm

There are some other benefits that make ADAM’s algorithm an even better optimizer. Most of the following points are directly referred to from paper.

- It is straightforward to implement without much tuning.
- ADAM is computationally efficient.
- It is memory efficient and has little memory requirements.
- ADAM works well in cases in large datasets and large parameter settings.

6.2 Issues and Challenges

- We can use only maximum threshold value; if it exceeds, packet collision will occur.
- Scale quantity is limited.
- Enlarged ultra-level source communication is not possible.
- Power efficiency is minimum.

7 Conclusion

The energy constrained routing technique, namely ADAM's algorithm, is proposed for the purpose of initiating secure data routing in mobile ad hoc networks (MANETs). The method evaluates the energy aware nodes with the help of energy model. After estimating the energy, the mobile nodes capitulating maximum power are used for routing. Energy model is the parameter adapted in ADAM's algorithm that offers energy constrained routing during transferring packets from one mobile node to another mobile node. At the beginning, dynamic mobile nodes are simulated in a randomly distributed environment, and power in mobile nodes is being calculated. Once power in all the mobile nodes is evaluated, making use of protected data carrying mobile nodes the immune routing scheme is entrenched. The immune routing scheme is continued by applying the newly adopted ADAM's model. Thus, in the MANET environment the optimal way of routing is advanced. We used an automatic optimization tool to address the optimal parameter tuning of the ADAM's routing technique for use in VANETs in this paper. We developed an optimization strategy for this work that is based on coupling optimization algorithms and the ns-2 network simulator (Figs. 4, 5, 6, 7 and 8).

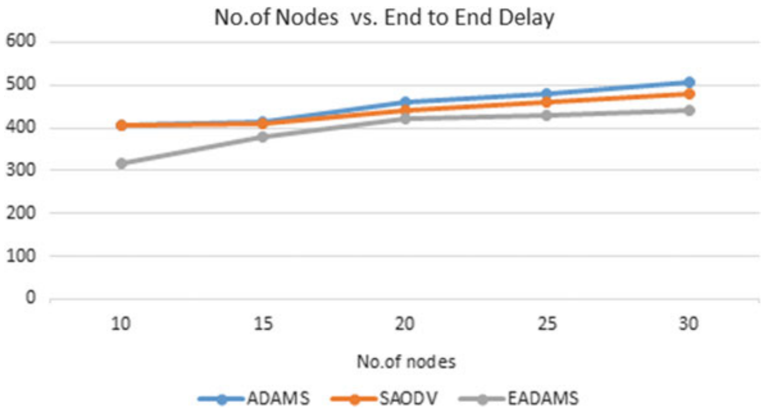


Fig. 4 No. of nodes versus end-to-end delay

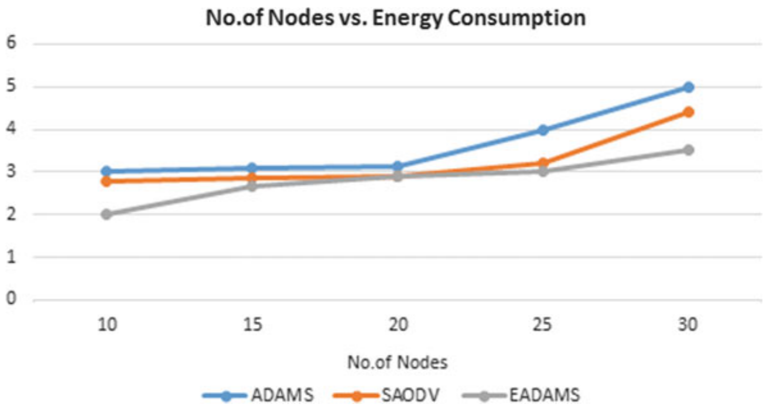


Fig. 5 No. of nodes versus energy consumption

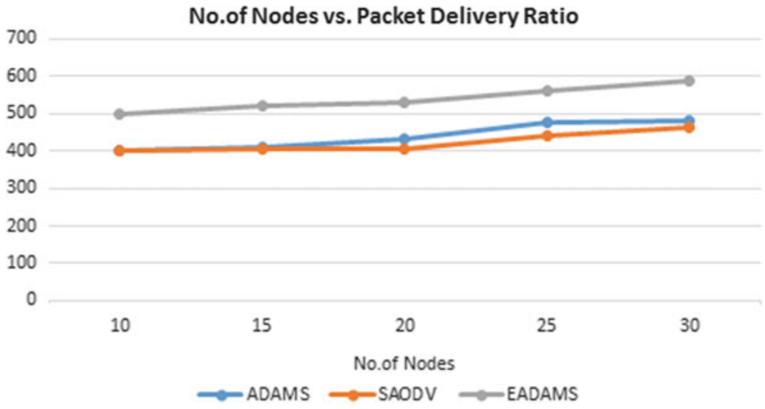


Fig. 6 No. of nodes versus packet delivery ratio

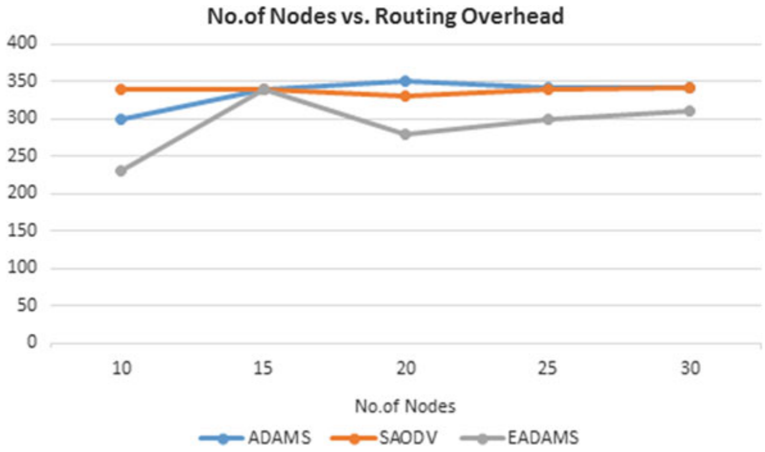


Fig. 7 No. of nodes versus routing overhead

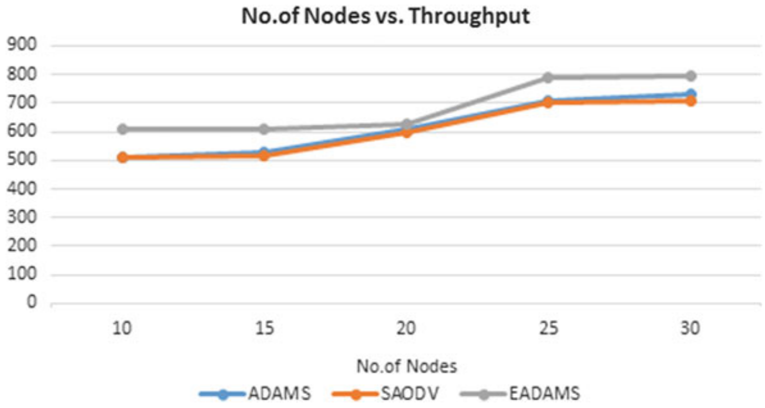


Fig. 8 No. of nodes versus throughput

References

1. Sarkar, S.: Reliable and energy aware routing in mobile ad hoc networks. Int. J. Wireles. Mob. Comput. **16**(2), 117–127 (2019)
2. Smith, D.H., Jodie, W., Andrew, A.: Security using pre-existing routing for mobile ad hoc networks: SUPERMAN. IEEE Trans. Mob. Comput. (2017)
3. Kumar, B.A., Subramanyam, M.V., Prasad, K.S.: An energy efficient clustering using K-means and AODV routing protocol in ad hoc networks. Int. J. Intell. Eng. Syst. (IJIES) **12**(2), 117–127 (2019)
4. Sunitha, M., Srinivas, P.V.S., Venugopal, T.: A predetermined energy based efficient routing mechanism for MANET. Adv. Wireless Mob. Commun. **10**(4), 623–638 (2017)
5. Selvi, M., Balakrishna, R.: A new method to recover the link failure and reliable data delivery in MANET. Int. J. Eng. Tech. **8**(8), 897–909 (2019)

6. Zapata, M.G., Asokan, N.: Securing Ad Hoc routing routing protocols. Wise 2002. Georgia, USA (2002)
7. Kingma, D.P., Jimmy, L.B.: ADAM: A method for stochastic optimization. II International Conference on Learning Representations. arxiv preprint arXiv:1412.6980 (2015)

Retraction Note to: Assessing Deep Neural Network and Shallow for Network Intrusion Detection Systems in Cyber Security

Deena Babu Mandru, M. Aruna Safali, N. Raghavendra Sai, and G. Sai Chaitanya Kumar

Retraction Note to: Chapter 52 in: S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies,* Lecture Notes on Data Engineering and Communications Technologies 75, https://doi.org/10.1007/978-981-16-3728-5_52

The Series Editor and the Publisher have retracted this chapter. An investigation by the Publisher found a number of chapters, including this one, with various concerns, including but not limited to compromised editorial handling, incoherent text or tortured phrases, inappropriate or non-relevant references, or a mismatch with the scope of the series and/or book volume. Based on the findings of the investigation, the Series Editor therefore no longer has confidence in the results and conclusions of this chapter.

The authors agree with this retraction.

The retracted version of this chapter can be found at
https://doi.org/10.1007/978-981-16-3728-5_52

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5_67

Author Index

A
Abbas, Heba Kh., 489
Abdul Razaq, Asmaa A., 489
Agarwal, Svarnim, 73
Aghbari, Zaher Al, 715
Aiswarya, E. S., 593
Akhil, Muvva, 451
Anand, R., 247
Anas, Muhammed, 767
Anisha, G. S., 325
Anjusha, P. P., 325
Aravind, S., 533
Arnaut, Uros, 1
Arul Sindiya, J., 33
Aruna Safali, M., 703
Asha, M., 627
Ashok, Amritha, 413
Aswath, S., 867
Avinash, T., 533

B
Bacanin, Nebojsa, 1
Bagade, Jayshree, 547
Balaji, S Ashwin, 99
Banerjee, Shreya, 147
Bansal, Anil, 795
Benny, Adheena Maria, 593
Bezdan, Timea, 1
Bhaskar, Anand, 439
Bhooshan, Sooraj, 313
Bindle, Abhay, 385
Brahmananda, S. H., 811, 821, 833
Briskilal, J., 741
Brown, Dane, 259
Burdak, Subhash, 343

C
Chaithanya, B. N., 821
Chamatagoudar, S. N., 333
Chandrakala, M., 399

D
Dadykin, Alex, 687
Dahiya, Abhishek, 131
Das, Ranjan Jyoti, 473
Das, Rik, 113
Deepika, A., 613
Desai, Digvijay, 547
Devi, M. Shyamala, 473
Dhanalakshmi, G., 399, 423
Dheepiga, S. R., 503
Divya Pai, R., 413
Dwivedi, Arpan, 47
Dwivedi, Prashant, 47

G
Ganesan, M., 533, 601
Gangwar, Siddhi, 343
Ghadge, Aniruddha, 547
Ghosh, Sanchita, 523
Govinda Rajulu, G., 247
Gulati, Tarun, 385
Gupta, Deepa, 651
Gupta, Priyansh, 99
Gupta, Uttam, 473

H
Hari Narayanan, A. G., 413
Hari Prasad, P., 533
Harish, K., 673

S. Smys et al. (eds.), *Computer Networks and Inventive Communication Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 75,
https://doi.org/10.1007/978-981-16-3728-5

Hasan, Nadine, 753
Hima Bindu, A., 305
Hridoy, Rashidul Hasan, 189

J
Jai Aakash, N. S., 533
Jain, Anubha, 581
Jain, Sambhav, 99
Jayalakshmi, S., 277
Jayashree, R., 147
Jeevana Jyothi, K., 423
Jindal, Rajni, 175

K
Kadam, Sharyu, 723
Kalaivani, M. S., 277
Kalyana Abenanth, G., 673
Kandikatla, Vyshnavi, 203
Karennavar, Rachana B., 147
Kaushik, Tavishi, 73
Kavita, 343
Kavitha, C. R., 73
Keerthana, A., 841
Khairnar, Vaishali D., 215
Khan, Abuzar Ahmed, 85
Khandelwal, Shekhar, 113
Khera, Amrit, 85
Kodali, Jyothirlatha, 203
Kumar, Anuj, 357
Kumar, Bandani Anil, 885
Kumar, Jogendra, 451
Kumar, Manoj, 59
Kumar, Neeraj, 287
Kumar, Preet, 287
Kumar, Shakti, 357
Kumar, Vinod, 131, 795

L
Lal, N. Dayanand, 811
Lavanya, R., 533

M
Mahesh, A. S., 369
Majumder, Anandaprova, 523
Malhotra, Ruchika, 85
Malhotra, Vinamra, 59
Mandru, Deena Babu, 703
Manjula, G., 247
Mareddy, Sahithya, 651
Marouf, Ahmed Al, 163
Mathur, Sumit, 439
Maurya, Satish, 17
Menon, Aswathi A., 369
Mishra, Ayaskanta, 753
Mohamad, Haidar J., 489
Mohankumar, N., 673
Mohan Sai Kumar, C. H., 867
Motwani, Dilip, 723

N
Nagati, Princy, 203
Nair, Shruti S., 413
Nair, Sreelekshmi M., 513
Namboothiri, Leena Vishnu, 593
Nandakumar, R., 313
Narkhede, Nandkishor, 439
Naveena, B., 423
Nayak, Parikshith, 811
Nerendla, Veena, 203
Nijguna, G. S., 811
Noble Mary Juliet, A., 503

P
Palaniappan, Y., 601
Pandiyarajan, M., 867
Patil, Neha, 215
Patnaikuni, Dinkar R. Patnaik, 333
Pattanaik, Manisha, 47
Paul, Trishna, 523
Prasad, Kodati Satya, 885
Prasanna, M., 305
Praveen Pai, R., 313
Pulickal, Nebu, 781
Pushpalakshmi, R., 33

R
Radha Krishna, Pisipati, 581
Radha, N., 613
Raghavendra Sai, N., 703
Raghuraman, Bhuvan, 741
Rahaman, Naziour, 163
Rajalakshmi, V. R., 513
Rajan, Rajesh George, 463
Rajendran, P. Selvi, 463
Rajesh, K., 399
Rajput, Amit Singh, 47
Rajput, Deependra Singh, 47
Rakshit, Aniruddha, 189
Ramakrishnan, S., 563
Ramesh, Santhosh Veeraraghavan, 473
Rasham, Noor H., 489

Rashid, Tarik A., 1
Ray, Arun Kumar, 753
Raziya Sulthana, P., 305
Reshma, R. S., 325
Riyaz Ahmed, M., 741
Rohith, V., 601
Rojatkar, Dinesh V., 853
Rubel, Salauddin, 163
Rukma Rekha, N., 233
Rushyendra, A., 673

S
Sachin, V., 673
Saeed, Mozamel M., 715
Sahana, D. S., 811
Sah, Mithila Bihari, 385
Sahu, Khomchand, 473
Sai Chaitanya Kumar, G., 703
Saini, Geetanjali, 131
Sajeev, Sayu, 767
Sam Rishi, R., 503
Sandhu, Jagnur Singh, 795
Sangeetha, D., 841
Santhosh, N., 601
Saranya, T., 369
Saxena, Jaya, 581
Selvi, S., 841
Seniaray, Sumedha, 175
Senthil Rajan, A., 563
Sharma, Abhishek, 795
Sharma, Yashna, 343
Shashankh, S., 73
Sheokand, Sahil, 131
Sinhmar, Abhinav, 59
Sirigeri, Prerana, 147
Sravani, M., 305, 451
Sreedevi, M., 203
Subamanyam, Makam Venkata, 885
Suba Rani, N., 503
Subba Rao, Y. V., 233
Subrahmanyam, Rolla, 233
Sundaresan, Arya, 513
Sunil, K. S., 767
Surekha, T. P., 627
Sureshbabu, Sheetal, 369
Sureshkumar, Pallapothu, 451
Suresh, Vishnu, 767
Suriyakala, C. D., 781
Swaminathan, J. N., 305

T
Tahini, Imad, 687
Tanwar, Lavi, 287
Thirukrishna, J. T., 247
Thomas, Julien Joseph, 767
Thorat, Samrat S., 853
Toppo, Kenneth, 287

U
Umamaheshwari, S., 305

V
Valarmathi, R. S., 867
Vasani, Meet J., 17
Veena, R. C., 833
Vignesh, O., 305
Vinod, Meghna, 513
Vishanth, V. A., 601
Vitalkar, Rasika S., 853

W
Wazare, Roshan, 547

Y
Yadav, R. K., 59, 99
Yasaswi, Valiveti, 451

Z
Zivkovic, Miodrag, 1