

Advances in Intelligent Systems and Computing 1299

Chhabi Rani Panigrahi ·
Bibudhendu Pati ·
Binod Kumar Pattanayak ·
Seeven Amic · Kuan-Ching Li *Editors*

Progress in Advanced Computing and Intelligent Engineering

Proceedings of ICACIE 2020

 Springer

Advances in Intelligent Systems and Computing

Volume 1299

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

Advisory Editors

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing,
Universidad Central de Las Villas, Santa Clara, Cuba

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

Hani Hagras, School of Computer Science and Electronic Engineering,
University of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University,
Gyor, Hungary


Vladik Kreinovich, Department of Computer Science, University of Texas
at El Paso, El Paso, TX, USA

Chin-Teng Lin, Department of Electrical Engineering, National Chiao
Tung University, Hsinchu, Taiwan

Jie Lu, Faculty of Engineering and Information Technology,
University of Technology Sydney, Sydney, NSW, Australia

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute
of Technology, Tijuana, Mexico

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro,
Rio de Janeiro, Brazil

Ngoc Thanh Nguyen , Faculty of Computer Science and Management,
Wrocław University of Technology, Wrocław, Poland

Jun Wang, Department of Mechanical and Automation Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Indexed by DBLP, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST).

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/11156>

Chhabi Rani Panigrahi · Bibudhendu Pati ·
Binod Kumar Pattanayak · Seeven Amic ·
Kuan-Ching Li
Editors

Progress in Advanced Computing and Intelligent Engineering

Proceedings of ICACIE 2020

 Springer

Editors

Chhabi Rani Panigrahi
Department of Computer Science
Rama Devi Women's University
Bhubaneswar, Odisha, India

Bibudhendu Pati
Department of Computer Science
Rama Devi Women's University
Bhubaneswar, Odisha, India

Binod Kumar Pattanayak
Department of Computer Science
and Engineering
S 'O' A Deemed to be University
Bhubaneswar, Odisha, India

Seeven Amic
Faculty of Information
and Communication Technology
Université des Mascareignes
Pamplemousses, Mauritius

Kuan-Ching Li
Department of Computer Science
and Information Engineering
Providence University
Taichung, Taiwan

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-33-4298-9

ISBN 978-981-33-4299-6 (eBook)

<https://doi.org/10.1007/978-981-33-4299-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This volume contains the papers presented at the 5th International Conference on Advanced Computing and Intelligent Engineering (ICACIE) 2020: The 5th ICACIE (www.icacie.com) is held during June 25–27, 2020, at Université des Mascareignes (UdM), Mauritius, in collaboration with Rama Devi Women’s University, Bhubaneswar, India, and S ‘O’ A Deemed to be University, Bhubaneswar, India. There were a total of 294 submissions, and each qualified submission was reviewed by a minimum of two Technical Program Committee (TPC) members using the criteria of relevance, technical quality, originality, and presentation. The TPC committee accepted 74 full papers for oral presentation at the conference, and the overall acceptance rate is 25%.

ICACIE 2020 was an initiative taken by the organizers which focuses on research and applications on topics of Advanced Computing and Intelligent Engineering. The focus was also to present the state-of-the-art scientific results to disseminate modern technologies and to promote collaborative research in the field of advanced computing and intelligent engineering.

Researchers presented their work in the conference through virtual as well as online mode due to COVID-19 pandemic and had an excellent opportunity to interact with eminent professors, scientists, and scholars in their area of research. All participants were benefitted from discussions that facilitated the emergence of innovative ideas and approaches. Many distinguished professors, well-known scholars, young researchers, and industry leaders participated in making ICACIE 2020 an immense success. We had many invited talks by professors, research scholars, and industry personnel in emerging topics of advanced computing, sustainable computing, and machine learning.

We express our sincere gratitude to the Patrons Prof. Radhakrishna Somanah, Director General, Université des Mascareignes (UdM), Mauritius; Prof. Padmaja Mishra, Vice-Chancellor, Rama Devi Women’s University, Bhubaneswar, India; and Prof. Ashok Kumar Mahapatra, Vice-Chancellor, S ‘O’ A Deemed to be University, Bhubaneswar, India, for allowing us to organize ICACIE 2020 and their unending timely support toward organization of this conference. We would like to extend our sincere thanks to Prof. Binod Kumar Pattanayak, Mr. Seeven Amic,

and Dr. Bibudhendu Pati, General Chairs of ICACIE 2020 for their valuable guidance during review of papers as well as other aspects of the conference. We thank all the Technical Program Committee members and all reviewers/sub-reviewers for their timely and thorough participation during the review process. We appreciate the time and efforts put in by the members of the local organizing team at Université des Mascareignes (UdM), Mauritius, and administrative staff, who dedicated their efforts to make ICACIE 2020 successful. We would like to extend our thanks to Er. Subhashis Das Mohapatra and Mr. Sanjeev Cowlessur for designing and maintaining ICACIE 2020 website and extending their support for managing the sessions in virtual as well as online mode.

Bhubaneswar, India
Bhubaneswar, India
Bhubaneswar, India
Pamplemousses, Mauritius
Taichung, Taiwan

Chhabi Rani Panigrahi
Bibudhendu Pati
Binod Kumar Pattanayak
Seeven Amic
Kuan-Ching Li

Contents

AI and Machine Learning Applications

Forecasting Price of Indian Stock Market Using Supervised Machine Learning Technique	3
Mohit Iyer and Ritika Mehra	
Speaker Recognition Using Noise Robust Features and LSTM-RNN	19
Mohit Dua, Pawandeep Singh Sethi, Vinam Agrawal, and Raghav Chawla	
Brain Tumor Segmentation Using Random Walks from MRI Images	29
Shiv Naresh Shivhare and Nitin Kumar	
Factors Accountable for Diabetes Using Artificial Intelligence in Medico-Care	43
Karuna Babber and Shruti Wadhwa	
An Evaluation of Deep Learning Networks to Extract Emotions from Yelp Reviews	55
Yasser Chuttur and Leevesh Pokhun	
Classification of Brain Tumor MRIs Using Deep Learning and Data Augmentation	69
Gulshansingh Bhagbut and Zahra Mungloo-Dilmohamud	
Deep Neural Network Optimization for Handwritten Text Recognition	85
Chahat Goel, Aishwarya Chaudhary, S. Indu, and Sudipta Majumdar	
Consolidating Online Real Estate Data Using Image Analysis and Text Processing	95
Yasser Chuttur and Haydar Mahadooa	

Time Series Visualization of Customer Emotions Using Artificial Neural Network	109
Yasser Chuttur and Nandishta Rawoteea	
Sentiment Analysis Using Deep Learning for Recommendation in E-Learning Domain	123
Rawaa Alatrash, Hadi Ezaldeen, Rachita Misra, and Rojalina Priyadarshini	
Selection of Best K of K-Nearest Neighbors Classifier for Enhancement of Performance for the Prediction of Diabetes	135
Subhash Chandra Gupta and Noopur Goel	
Phishing Website Prediction: A Machine Learning Approach	143
Anjaneya Awasthi and Noopur Goel	
COVID-19 Sentimental Analysis Using Machine Learning Techniques	153
Chhinder Kaur and Anand Sharma	
Multimodal Music Mood Classification Framework for Kokborok Music	163
Sanchali Das, Sambit Satpathy, and Swapan Debbarma	
Forecasting of Daily Demand's Order Using Gradient Boosting Regressor	177
Tansif Anzar	
Improving Impulse Noise Classification Using Ensemble Learning Methods	187
Kunaraj Kumarasamy, S. Maria Wensch, S. Balaji, L. J. Jenifer Suriya, A. Jerlin, and S. Robert Rajkumar	
Image Data Preservation with Fractional Sine Transform and Dual Chaotic Sequence	201
Sharad Salunke, M. Venkatadri, Md Farukh Hashmi, and Bharti Ahuja	
Enhancing Deep Learning Capabilities with Genetic Algorithm for Detecting Software Defects	211
Kajal Tameswar, Geerish Suddul, and Kumar Dookhitram	
Application of Classifier for Breast Cancer Cell Detection	221
Ashutosh Mishra, Unnati Mantry, Dibyasha Garhnayak, Sourav Panda, Rasmita Rautray, Rasmita Dash, and Rajashree Dash	
Apriori-Backed Fuzzy Unification and Statistical Inference in Feature Reduction: An Application in Prognosis of Autism in Toddlers	233
Shithi Maitra, Nasrin Akter, Afrina Zahan Mithila, Tonmoy Hossain, and Mohammad Shafiu Alam	

Developing a Framework for Generating Hotel Recommendation 255
 Md. Shafiqul Alam Forhad and Mohammad Shamsul Arefin

**Rumor Source Identification on Social Networks: A Combined
 Network Centrality Approach** 269
 Abhijit Das and Anupam Biswas

**Sentiment Polarity Detection on Bengali Book Reviews Using
 Multinomial Naïve Bayes** 281
 Eftekar Hossain, Omar Sharif, and Mohammed Moshikul Hoque

Real-Time Facial Emotions Analysis in Videos 293
 Babajee Phavish, Suddul Geerish, Armoogum Sandhya, and Foogooa Ravi

**An Extended Genetic Algorithm-Based Prevention System Against
 DoS/DDoS Flood Attacks in VoIP Systems** 301
 Sheeba Armoogum and Nawaz Mohamudally

Entropy Based Cluster Selection 313
 Arko Banerjee, Arun K. Pujari, Chhabi Rani Panigrahi,
 and Bibudhendu Pati

Human Activity Recognition Using Machine Learning: A Review 323
 Ankita Biswal, Sarmistha Nanda, Chhabi Rani Panigrahi,
 Sanjeev K. Cowlessur, and Bibudhendu Pati

Advanced Computer Networks and Algorithms

**Ensuring Secure Communication from an IoT Edge Device to a Server
 Through IoT Communication Protocols** 337
 Roshni Vidya S. Boolakee, Sandhya Armoogum, Ravi Foogooa,
 and Geerish Suddul

**A QoI Assessment Framework for Participatory Crowdsourcing
 Systems** 351
 Ashley Rajoo, Kavi Kumar Khedo, and Utam Avinash Einstein Mungur

**An Approach to Personalize VMware vSphere Hypervisor (ESXi)
 Using HPE Image Streamer** 363
 Richa and Jyoti Singh

**A Proposed IoT Architecture for Corals Research Using AI
 and Robotics** 371
 Wafiq Aumeer, Nabeelah Pooloo, and Rajeev Khoodeeram

**Voice Password-Based Secured Communication Using RSA
 and ElGamal Algorithm** 387
 Prashnatita Pal, Bikash Chandra Sahana, S. Ghosh, Jayanta Poray,
 and Amiya Kumar Mallick

UD-RMM: A Remote Monitoring and Management Tool Using PowerShell Universal Dashboard for Universities	401
Pranav Nagarajan and Jayavignesh Thyagarajan	
Performance and Resource-Aware Virtual Machine Selection using Fuzzy in Cloud Environment	413
Vikas Mongia and Anand Sharma	
Redefining Data Dimensionality Through Dynamic Linkages in Data-Space Continuum	427
Benard Alaka and Bernard Shibwabo Kasamani	
Implementation of Encryption Techniques in Secure Communication Model	437
Md. Sharif Hossen and Md. Shakhaowat Hossen	
A Group Decision Making Problem Involving Fuzzy TOPSIS Method	449
Prashanta Kumar Parida	
IoT-Based Smart Intravenous Drip Monitoring System	463
Muskan Jindal, Nidhi Gajjar, and Nehal Patel	
Cryptanalysis of Lightweight Ciphers Using Metaheuristics	469
Seeven Amic, K. M. Sunjiv Soyjaudah, and Gianeshwar Ramsawock	
Product Classification in E-Commerce Sites	485
Anannya Patra, V. Vivek, B. R. Shambhavi, K. Sindhu, and S. Balaji	
Real-Time Detection of Inter-Frame Video Forgeries in Surveillance Videos	497
Kshitij Saluja and Naveen Aggarwal	
An IoT-Based System Architecture for Environmental Monitoring	507
Binod Kumar Pattanayak, Deojeet Nohur, Sanjeev K. Cowlessur, and Rajani Kanta Mohanty	
Performance Evaluation of VM Allocation Strategies on Heterogeneous Environments in Cloud Data Center	515
Rajni Garg, Indu Arora, and Anu Gupta	
QSens: QoS-Aware Sensor Node Selection in Sensor-Cloud Architecture	527
Arijit Roy, Sudip Misra, and Aditya Kotasthane	
Introduction to Adjacent Distance Array with Huffman Principle: A New Encoding and Decoding Technique for Transliteration Based Bengali Text Compression	543
Pranta Sarker and Mir Lutfur Rahman	

BSAT: A New Tool for Analyzing Cryptographic Strength of Boolean Function and S-Box of Symmetric Cryptosystem 557
 Pratap Kumar Behera and Sugata Gangopadhyay

Sorted Galloping Prevention Mechanisms Against Denial of Service Attacks in SIP-Based Systems 571
 Sheeba Armoogum and Nawaz Mohamudally

Performance Enhancement and Reduce Energy Consumption with Load Balancing Strategy in Green Cloud Computing 585
 Hitesh A. Bheda, Chirag S. Thaker, and Darshan B. Choksi

A Framework for Secure Communication on Internet of Things (IoT) 599
 Mohammad Reza Hosenkhan and Binod Kumar Pattanayak

COVTrac: Covid-19 Tracker and Social Distancing App 607
 Subhashish Das Mohapatra, Suwendu Chandan Nayak, Sasmita Parida, Chhabi Rani Panigrahi, and Bibudhendu Pati

JOB-DCA: A Cost Minimizing Jaya Optimization-Based Data Center Allocation Policy for IaaS Cloud Model 621
 Sasmita Parida, Bibudhendu Pati, Suwendu Chandan Nayak, and Chhabi Rani Panigrahi

Sustainable Computing and Engineering

Teaching and Learning Concepts of Audio Modulation Using Tangible User Interfaces 637
 Ah-Kwet Rémi Wong Suk Hee, Hadija Ramadhani Halfani, and Girish Bekaroo

Performance Investigation of Dipole and Moxon Antennae for VHF Communication 653
 Akshay Jain, Pranay Chavan, Pratik Maradiya, and Kiran Rathod

Cloud Computing Load Balancing Using Amazon Web Service Technology 661
 Nagarjuna Hota and Binod Kumar Pattanayak

New Management Algorithms for Smart Electricity Network: Designing and Working Principles 671
 Ando Ny Aina Randriantsoa, Ali Hamada Damien Fakra, Manitra Pierrot Ranjaranimaro, Mohamed Nasrouline Mohamed Rachadi, and Jean Claude Gatina

Analyzing Key Barriers for Adoption of Digitalization in Indian Construction Industry: A Case Study 683
 Avirag Bajpai and Subhas Chandra Misra

Improved Peak-to-Average Power Ratio Reduction Method for OFDM/OQAM System	695
V. Sandeep Kumar	
Implementation of a Smart Parking System for a University Campus	707
M. Saheer Jhugroo, M. Shahil Kataully, and Soulakshmee D. Nagowah	
High-Resolution Wind Speed Mapping for the Island of Mauritius Using Mesoscale Modelling	719
Tyagaraja S. M. Cunden and Michel R. Lollchund	
Analysis of Wind Energy Resources for the Island of Mauritius Using Concepts of Thermodynamics	735
Tyagaraja S. M. Cunden, Naafeera B. R. Abdel Hassan, and Michel R. Lollchund	
Design and FPGA Implementation of an Efficient 8×8 Multiplier Using the Principle of Vedic Mathematics	749
Smitha Bhat Kaje and Jagadish Nayak	
Fault Detection and Isolation in a Leaky Water Distribution Network Using Fuzzy Logic Control Based on Residual Pressure Analysis	763
Lekhramsingh Latchoomun and Tsiatsipy Durand Brunel	
Robo-Friend: Can a Social Robot Empathize with Your Feelings Effectively?	777
Eshtiaq Ahmed, Ashraful Islam, Atiqul Islam Chowdhury, Mohammad Masudur Rahman, Shahnaj Chowdhury, and Md Imran Hosen	
Management and Banking Applications	
Using Stakeholder Expectations and Perceptions to Guide the Brand Refresh of a Tropical Airline	791
Vimi Neeroo Lockmun-Bissessur, Swaleha Peeroo, and David Savy	
An Analysis of Communication Strategies of Fast Food Outlets on Social Media in Mauritius	805
Swaleha Peeroo and A. Mooznah Auleear Owodally	
Factors Affecting Task Allocation and Coordination in Distributed Agile Software Development	817
Chitra Nundlall and Soulakshmee D. Nagowah	
Optimizing Recruitment Process Within Businesses: Predicting Interview Attendance Using C4.5 Algorithm	831
Shivianee Sandhip Laldjee, Chiamaka Ann Marie Ajufo, and Girish Bekaroo	

Training Engineers as Drivers of e-Learning in a University 839
Nirmal Kumar Betchoo

Using Scratch Software as a Teaching-Learning Tool in French Language Classes: A Case Study at Université Des Mascareignes (Mauritius) 851
Neelam Pirbhai-Jetha

Improving Fraud Detection Mechanism in Financial Banking Sectors Using Data Mining Techniques 861
Hanan Hamdan AL-Abri, Basant Kumar, and Joseph Mani

Digital Learning for Millenials: IT in Eduation and/or IT for Education 871
Neelam Pirbhai-Jetha, Pascal Boncoeur, and Normada Bheekharry

Supply Chain Management—Marketing Integration a Key Element in the Digital Era 881
Normada Devi Bheekharry

Bank Customer’s Credit Score Prediction Using Feature Selection and Data Mining Algorithm 889
Durgesh Kumar Singh and Noopur Goel

The Use of the WhatsApp Platform as an Educational Tool During the Confinement Period of the Outbreak of Covid-19 899
Nundini Devi Akaloo

Author Index 913

About the Editors

Dr. Chhabi Rani Panigrahi is Assistant Professor in the Department of Computer Science at the Rama Devi Women's University, Bhubaneswar, India. She received her Ph.D. degree in Computer Science & Engineering from IIT Kharagpur, India. Her current research interests include mobile cloud computing, machine learning, and software testing. She has more than 19 years of teaching and research experience. She has published several scholarly articles in journals and conferences of international repute. She is a life member of ISTE and member of IEEE.

Dr. Bibudhendu Pati is Head in the Department of Computer Science at Rama Devi Women's University (Only Govt. Women's University in the state of Odisha, India). He received his Ph.D. degree from Indian Institute of Technology Kharagpur, India. He has around 22 years of experience in teaching and research. His current research interests include cloud computing, big data, IoT and advanced network technologies. He has got several papers published in reputed journals, conference proceedings and books of international repute. He was invited to chair several international conferences. He is a life member of Indian Society for Technical Education, Computer Society of India, and a senior member of IEEE.

Dr. Binod Kumar Pattanayak completed M.S. in Computer Engineering in the year 1992 from Kharkov Polytechnical Institute, USSR and Ph.D. in Computer Science and Engineering from S'O'A (Deemed to be University), Bhubaneswar, India. Currently, he is working as Professor in the Department of Computer Science and Engineering at the Institute of Technical Education and Research under S'O'A (Deemed to be University), Bhubaneswar, India. He has to his credit around 100 research publications in journals and conferences of international repute. He has acted as Visiting Professor to Build Bright University, Phnom Penh, Cambodia, and Université des Mascareignes, Mauritius. His research interests include computer networking, IoT, cloud computing, compiler design and ad hoc networks.

Mr. Seeven Amic is Dean of the Faculty of Information and Communication Technology at Université des Mascareignes. He received his M.Sc. (Hons.) degree in Computer Engineering at the Kharkov Polytechnic Institute, Ukraine, in 1992 through a scholarship. In 1995, he joined the Swami Dayanand Institute of Management, now Université des Mascareignes, where he holds office as Senior Lecturer in the Department of Software Engineering at the Swami Dayanand Campus. His research areas are cryptology, cyber-security, metaheuristics, artificial intelligence and emerging technologies. He has a strong interest in experimental and collaborative research.

Prof. Kuan-Ching Li is currently Professor in the Department of Computer Science and Information Engineering at the Providence University, Taiwan. He was Vice-Dean for Office of International and Cross-Strait Affairs (OIA) in this same University since 2014. Prof. Li is Recipient of awards from Nvidia, Ministry of Education (MOE)/Taiwan and Ministry of Science and Technology (MOST)/Taiwan, as also guest professorship from different Universities in China. He got his Ph.D. from the University of Sao Paulo, Sao Paulo, Brazil, in 2001. His areas of research are networked and GPU computing, parallel software design and performance evaluation and benchmarking. He is a Fellow of the IET, a senior member of the IEEE and a member of TACC.

AI and Machine Learning Applications

Forecasting Price of Indian Stock Market Using Supervised Machine Learning Technique



Mohit Iyer and Ritika Mehra

Abstract There has been increasing enthusiasm for displaying and determining stock costs over recent decades. The artificial neural network is not acceptable as it is a grouping of both fictitious and experimental disciplines, which can be a successful way to enhance the performance of the mix of different models if the model is very unusual. In this paper, the different strategies like linear regression, kNN, Naïve, MA, AR, ARMA, ARIMA, and autoARIMA are used for forecasting the stock markets. This paper suggests which method is the best to use for predicting the stock market. This paper endeavors to address the determining of stock costs. On this unique circumstance, we gathered information on a month to month shutting stock indices of SENSEX. Subsequently, it very well may be utilized these models for estimating a task, particularly when higher anticipating precision is required.

Keywords Stock market forecasting · SVM · LR · kNN · AR · MA · ARMA · ARIMA · Machine learning

1 Introduction

The common tendency of society toward the securities exchange is that it is exceedingly hazardous for investment or not appropriate for business. The prediction of the share price is a daunting problem because the market prediction is quite hard. The stock prices vary so rapidly which makes the forecasts progressively hard. Many of them trust that stock prices are arbitrary and cannot be anticipated [1]. Some of the research has been made which challenge the above statement. These researches concluded that the current price of the stock is affected by the historical stock data. Therefore, stock market predictions based on stock data are an important topic for

M. Iyer (✉) · R. Mehra
DIT University, Dehradun, India
e-mail: mohitiyer007@gmail.com

R. Mehra
e-mail: hod.mca@dituniversity.edu.in

financial studies and the primary purpose of prediction research is to focus primarily on the progress of smart systems. It incorporates the presumption that the past companies stock data are openly accessible which has various prescient relations to the returns of the future stock which clarify that past stock data provides an insight into the behavior of the future[2]. The stock market, commonly known as an open market for an organization, gathers monetary assets to list its stock with a satisfactory price [3].

Consistently, massive data are generated inside the different applications which are hard to oversee. A large database was produced to store this massive data [4]. The users can extract this data as per their requirement. The basic data mining technology searches or extracts useful data from huge databases on the basis of similarity among them [5]. In this area, over time a variety of neural networks, machine learning, pattern recognition, and different algorithms have been introduced. In case when the existing platforms are updated and new products are developed, these techniques can be imposed on new systems. When the tools for data mining are applied on concurrent processing systems with maximum performance, it is possible to study large databases in a few minutes [6–8]. Different types of techniques and mechanisms that help in analyzing historical data, such as predictions about future unknown events can be collected, are included in the predictive analysis study [9].

1.1 Stock Market Prediction

Stock market predictions includes some of the major tasks like investigating market patterns, planning investment strategies, distinguishing the best time to purchase/sell stocks, and furthermore recognizing which stocks to purchase/sell. The budgetary items like bonds, shares, treasury bills, future and option, mutual funds, and others are purchased and sold on the capital market, which is additionally separated into the primary as well as a secondary market. The primary market implies that stocks are purchased as well as sold legitimately from the organization, whereas now these securities are further traded in the secondary market by the investors. The involvement trading between two investors is done in the stock market that is why it is also called a secondary market. The prices are fixed by share market based on interest and supply [10]. Highly sold stock will decrease in price, and high price will increase the demand of the stock. It is complicated to model stock exchanges with high accuracy because it is a nonparametric framework [1]. It relies on the merchants who need to purchase to sell or to hold shares at a specific time. The prediction has always been and will remain an enormous research area for the researchers and data scientist as they are always keen to increase the current forecast model.

Many of the traders or investors are still seeking for a new technology that can predict the stock trends and which can maximize the returns and also reduces the risk of investment as well. This is the purpose behind which a scientist gets inspirations to work in this area [11].

The rest of the paper is composed as pursues. Section 2 examines the literature survey of recently distributed research papers which identifies the stock patterns. Section 3 talks about the methods and materials. Section 4 talks about the outcomes and implementations which are used to foresee the stock patterns. Section 5 incorporates a portion of the screenshots of anticipated outcomes. Conclusion of the paper along with the future scope is in Sect. 6. At last, references have been incorporated.

2 Literature Survey

The most proficient approach to figure what is to come is to comprehend the present situations. The author [12] endeavored to build up a proper model that helps the concealed estimations of the Indian stock market, in light of the data gathered on the month to month closing price of the stock. They foresee the future stock price based on the ARIMA model which has the well-built performance of the Indian economy. For some financial experts, investors, and analysts, the Indian stock exchange is the center of interest, so it is important to comprehend the current situation of the market on this basis. It has been hypothesized that the execution of the Indian Stock Exchange symbolizes to a reasonable time arrangement, i.e. Arima (1, 0, 1) model, which makes reasonable predictions for what's to come.

The authors [13] have built up a comprehensive process of building a forecast model of a stock price using the ARIMA model after getting information from NYSE and NSE. The artificial neural network (ANN) is exceptionally well known because it has the ability to learn data patterns and surmised solutions from ambiguous data. Hybrid approaches are taking advantage of the unique strengths of each of them to improve the stock price, forecast model. The outcomes, acquired from gathering genuine information, which exhibited the quality of the ARIMA model, to present short-term forecasts to the shareholders that may perhaps assist in the decision-making for investment.

3 Methods and Material

3.1 Dataset Description

The extremely large volume of data is available on numerous websites where users share and exchange their perception and beliefs. Yahoo finance has been used for data collection with stock prices for various companies (Fig. 1).

Fig. 1 Dataset type

Data columns (total 7 columns):

Date	2472	non-null	datetime64[ns]
Open	2465	non-null	float64
High	2465	non-null	float64
Low	2465	non-null	float64
Close	2465	non-null	float64
Adj Close	2465	non-null	float64
Volume	2465	non-null	float64

3.1.1 Some Stock Market Terminologies

OPEN—is the price of the stock at the beginning of the trading day (it need not be the closing price of the previous day).

HIGH—is the highest price of the stock at closing time.

LOW—is the lowest price of the stock on that trading day.

CLOSE—is the price of the stock at closing time.

VOLUME—indicates how many stocks were traded.

ADJUSTED CLOSE—is the closing price of the stock that adjusts the price of the stock for corporate actions.

3.2 Methodology

Securities exchange forecast is partitioned into two sections which are as per the following:

Trend prediction model: It is utilized to anticipate the financial exchange by securing connections among an assortment of specialized variables and development of the offer cost.

Time series forecasting: It predicts by breaking down the stock’s past return, and it is utilized to anticipate the arrival of future stock costs.

Sentiment analysis process: Sentiment analysis—the problem of understanding the emotional tone of a text has been solved with very high accuracy (Fig. 2).

Machine learning approach: Machine learning methods are powerful approaches that have revolutionized many fields. These methods are able to take advantage

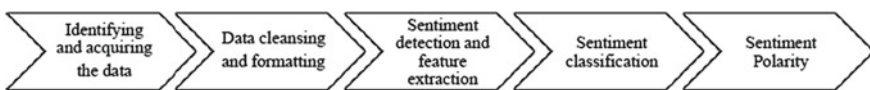


Fig. 2 Sentiment analysis process

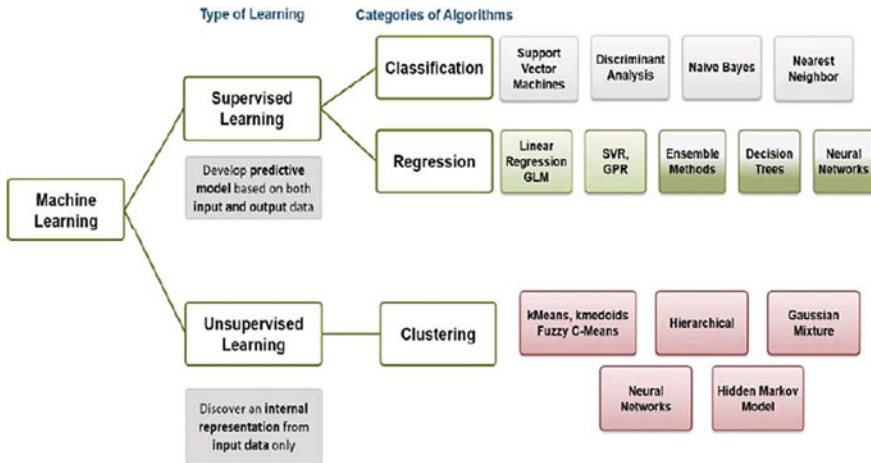


Fig. 3 Machine learning techniques

of high-dimensional input—an important asset for sentiment analysis. A few other trendy strategies are additionally used to foresee the securities exchange are SVM, liner regression, k-nearest neighbor (kNN) algorithm, time series forecasting (Fig. 3).

3.2.1 Support Vector Machine

The support vector machine is based on the theory of supervised learning, and it is considered as the best-recognized theory at present as it is used for both regression and classification analysis. The other name of SVM is probabilistic binary linear classifier. Using different types of kernels, it can also perform no-linear classification. categorization of hyper text, image classifications, permutation test, prediction of weather, prediction of sales, etc., are the solutions for several real-world problems. This theory covers the systematic study of the principle of minimizing the experiential risk, the association between the experiential risk and the estimated risk under some degree of a sample, and how to use these assumptions to locate principles and methods of new learning.

The support vector machine provides some of the prolific advantages as mentioned below:

- Works very well when we have unstructured or semi-structured data such as images and text
- Risk of overfitting is very less, and it also minimizes the computational cost
- With the help of kernel function, we can solve any complex problems.

3.2.2 Linear Regression

The most essential algorithm of machine learning that can be effectively applied to this type of data is nothing but the linear regression. The linear regression model provides a condition that is useful in deciding the relationship between the independent and dependent variables.

The equation of linear regression is written as:

$$Y = \theta_1 X_1 + \theta_2 X_2 + \dots \theta_n X_n$$

where independent variables are represented by $X_1, X_2, \dots X_n$, and the weights are represented by coefficients $\theta_1, \theta_2, \dots \theta_n$.

3.2.3 K-Nearest Neighbor (kNN) Algorithm

The model for kNN is the entire training dataset. When a prediction is required for an unseen data instance, the kNN algorithm will search through the training dataset for the k-most similar instances. The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instances.

The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. For other types of data, such as categorical or binary data, Hamming distance can be used.

- In the case of regression problems, the average of the predicted attribute may be returned
- In the case of classification, the most prevalent class may be returned.

3.2.4 Time Series Forecasting

There is one logically essential field to think about in the machine learning, i.e., time series forecasting, which is for each situation regularly overlooked. Time series forecasting has frequently overlooked because of time component which makes it increasingly hard to deal with the information. Data are having the time component and to predict such type of data, time series analysis is found to be the finest tool [14], 15. Latest examinations have shown that the strategy of time series data uses data confirmed as an explanation behind the survey of future results. Time series data could be illustrated as numerical data accumulated in a particular organization over a normal interim of time span. Time series data can be collected consistently on a weekly, monthly, quarterly or yearly basis. In order to analyze the time series data, the useful statistical information is extracted to identify the features of the data. The main motive is to analyze and find the possible link between the collected data and how the data be changed in respect of time. Time series forecasting is also useful for finding and able to analyze the trend of a stock market that helps investors to find

patterns. The stock prices vary in time, forming a time series [12, 16]. Regression is a method usually used for prediction. It has a dependent variable which relies upon at least one variable that is independent. The price of the stock can be taken as a dependent variable. The factors affecting it, like current demand, government policies, buyers' assessments, etc., are independent variables. In the case of the stock market, all the factors affecting it are not known. In such situations, a time series analysis is possible. In this method, the current price is the dependent variable and previous prices are independent variables. The source of time series analysis is the dependency of the current value on the previous price. But sometimes, tracking the flow of the market can be difficult because it depends on the season. Traders are very much engrossed in knowing the trend or the flow of past behavior, seasonal-wise growth or stock variations [17].

Arima Model

Autoregressive integrated moving average (ARIMA) is a well-known statistical method for time series forecasting, also called a Box–Jenkins approach. Box and Jenkins coined that it is possible to make a non-stationary data into stationary by means of differencing the series, X_t . The general model for X_t is given as,

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} \dots \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots \theta_q \epsilon_{t-q}$$

where

- X_t differenced time series value,
- ϕ , and θ unknown parameters, and
- ϵ independent identically distributed error terms with zero mean.

Here, X_t is shown which is based on the current and previous values of its old values and error conditions.

ARIMA is a technique which predicts the values for of a series based on its own activity. ARIMA models work on the following assumptions—

- If the mean and variance are not varying with time, it defines that the data is stable or stationary. By means of log transformation or differencing the series, a sequence can be prepared stable or stationary.
- The input data provided must have one variable series as the previous value is used by ARIMA to predict future values.

ARIMA consists of three components, i.e., autoregressive term (AR), I differencing term, and moving average term (MA) mentioned in brief below:

- AutoRegression (AR)—The meaning of AR is to use the previous values to estimate the next values. It is defined by parameter ‘p’ in the ARIMA model, and its value is determined by means of PACF plot. In a particular context, the relationship between the observation and some lagged observations used by it.

- **Moving Average (MA)**—The meaning of MA is used to predict future values by defining the number of previous forecast errors. It is defined by parameter ‘ q ’ in the ARIMA model, and its value is determined by the ACF plot. In other words, it uses the dependency from a moving average model among an observation and a residual error applied to lagged observations.

Order of differencing (Integrated (I))—Order of differencing (I) is used to eliminate the trend from a time series data and convert the time series of non-stationary into stationary data. It is defined by “ d ” value in the ARIMA model. Test like ADF and KPSS can be used to determine whether the series is stationary and help in identifying the d value. In other words, differencing of unprocessed observation is utilized by it.

4 Outcomes and Implementation

There are many time series analysis methods like linear regression, kNN, naïve, moving average (MA), autoregression (AR), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) and auto-ARIMA[9, 18]. MA tries to model the current price as a function of the white noise. AR tries to model the current price as the function of previous prices. ARMA combines both and models the current price as a function of both the white noise and previous prices. The time series data of a stock price is non-stationary. Before modeling, make it stationary. ARIMA does this by differentiating the series until the series becomes stationary. We used all the above-mentioned algorithms to model and to make predictions using the Python programming language in Jupyter Notebook.

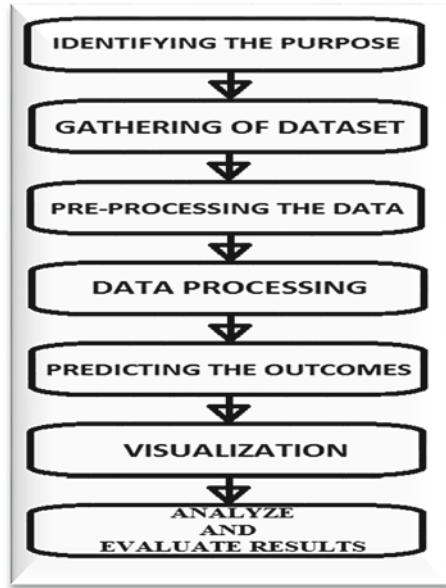
It prescribes that ARIMA is an algorithmic way to deal with change the series is better to forecast justifiably, and moreover, it gives gradually accurate outcomes [19]. For model estimation, this approach has not included any criteria for testing. Nifty Market futures have been planned for a short run in order to analyze and predicting of price causality [20]. This short run examination would not convey any remarkable results among the naïve shareholder as well as market penetrator [17].

Here, the main focus is to look over the current problems in the stock exchange. The seasonal trend and flow is the key attraction of the stock exchange. Ultimately, investors and stock broking companies would also examine the continuation, diversification and steady growth of the index. This will assist new investor and existing people in order to make a strategic decision. Investors could be attained by experience and constant observations. ARIMA algorithm is being recommended in order to conquer the above-said problems, in three steps,

1. Step 1: Identifying the model
2. Step 2: Estimating the model
3. Step 3: Prediction.

The entire architecture of the proposed model is shown in Fig. 4

Fig. 4 Architecture of the proposed method



5 View and Analyze Results

Once implementing the proposed model, prediction can be done in terms of visualization. Some of the screenshots of different models are being followed in the next section which can help the shareholders to analyze and can track the trend of the price of future stocks at a specific period of time. This can be used as assistance to the investors to decide the right time to buy, sell, or hold the stocks.

5.1 Screenshots of the Outcome

X-axis illustrates the time period in terms of years, and Y-axis illustrates the predicted price values over actual data in the above figures. All the above figures are of different models, viz. AR model, MA model, naïve model, ARMA model, ARIMA model, kNN model, and LR model are shown in Figs. 5, 6, 7, 8, 9, 10 and 11 respectively, that are showing the approximate prices based on the actual prices of shares. Compared to the prediction of all models, the ARIMA proved to be much better than others, as clearly shown in the above set of figures. Estimated prices have been shown in Tables 1 and 2 using various techniques of HDFC Bank Stock Data, which have been collected from Yahoo Finance.



Fig. 5 Prediction using AR model

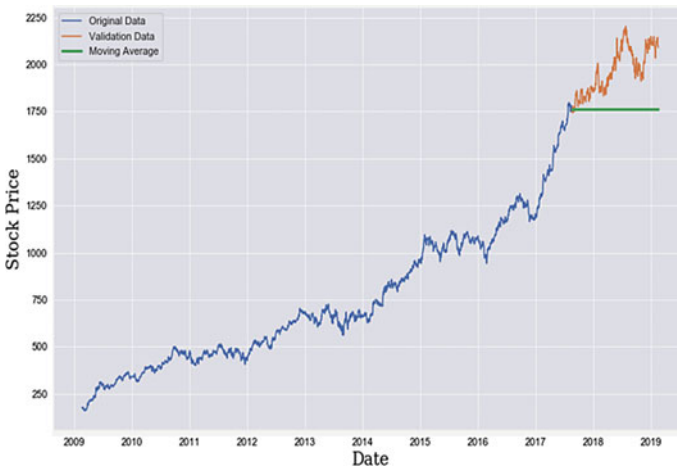


Fig. 6 Prediction using moving average

6 Conclusion

The prediction analysis is the approach which is applied to predict future instance from the current information. The prediction analysis technique is applied in the two phases which are feature extraction and classification. The various classification techniques are applied so far for the prediction analysis like SVM, kNN, and other useful methods. In this research work, a novel approach is proposed for the prediction analysis in order to obtain a better outcome. The proposed model enhances accuracy and reduces the execution time to predict the stock market, as well as helps the



Fig. 7 Prediction using ARIMA



Fig. 8 Prediction using Naive

investors make accurate and reliable estimates to when to buy and sell shares to make profitable investments. Outcomes can be improvised by integrating technical and fundamental analysis techniques. Evaluation of social media analysis can be included to get better results, especially by using fundamental analysis techniques based on public analysis. In this way, investors can get better outcomes in choosing the right time to enter the stock market and leverage beneficial investment decisions.



Fig. 9 Prediction using ARMA

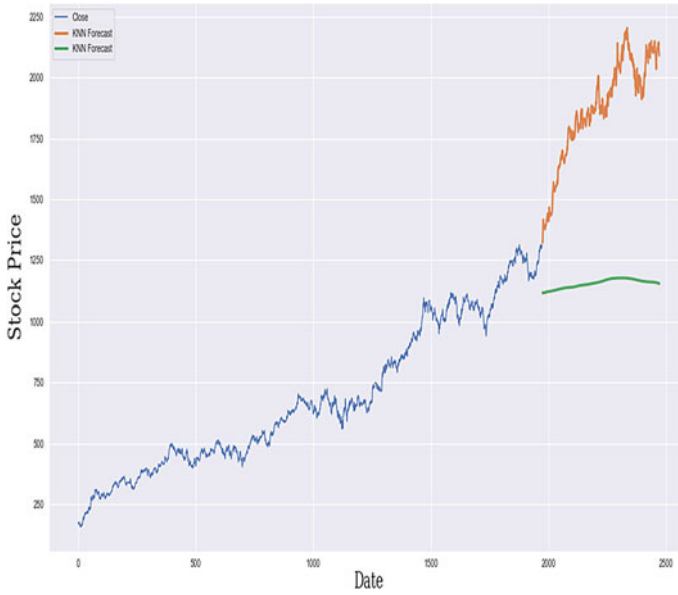


Fig. 10 Prediction using KNN model

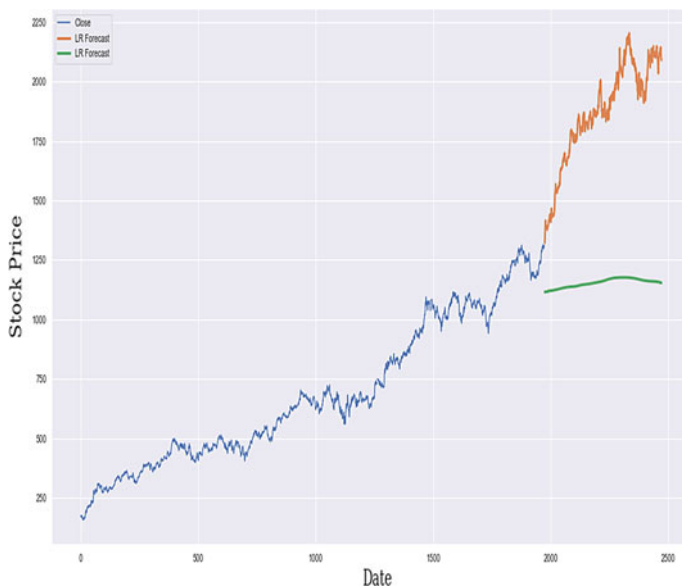


Fig. 11 Prediction using LR model

Table 1 HDFC bank stock prediction using naive, linear regression, and kNN models

DATE	Actual	Naïve forecast	LR model	kNN model
		prediction	prediction	prediction
2/5/2019	2114.05	1752.15	1157.047	1157.047
2/6/2019	2122.65	1752.15	1156.748	1156.748
2/7/2019	2117.25	1752.15	1156.463	1156.463
2/8/2019	2122.65	1752.15	1156.153	1156.153
2/11/2019	2139.65	1752.15	1155.835	1155.835
2/12/2019	2129.7	1752.15	1155.532	1155.532
2/13/2019	2143.45	1752.15	1155.23	1155.23
2/14/2019	2110.2	1752.15	1154.944	1154.944
2/15/2019	2100.65	1752.15	1154.646	1154.646
2/18/2019	2089.9	1752.15	1154.333	1154.333

Table 2 HDFC bank stock prediction using AR, MA, ARMA, and ARIMA models

DATE	Actual	AR model	Moving average model	ARMA model	ARIMA model
		Prediction	Prediction	Prediction	Prediction
2/5/2019	2114.05	2574.35	1760.917	997.48805	2105.1543
2/6/2019	2122.65	2577.048	1760.917	997.50076	2113.9294
2/7/2019	2117.25	2579.748	1760.917	997.51346	2122.9333
2/8/2019	2122.65	2582.451	1760.917	997.52616	2117.3461
2/11/2019	2139.65	2585.157	1760.917	997.38866	2123.9857
2/12/2019	2129.7	2587.866	1760.917	997.55157	2140.3587
2/13/2019	2143.45	2590.578	1760.917	997.56427	2129.0871
2/14/2019	2110.2	2593.292	1760.917	997.57697	2145.2595
2/15/2019	2100.65	2596.01	1760.917	997.58967	2109.5504
2/18/2019	2089.9	2598.73	1760.917	997.60237	2103.8299

References

- Liao, S.-H., Ho, H.-H., Lin, H.-W.: Mining stock category association and cluster on Taiwan stock market. *Expert Syst. Appl.* **35**, 1929 (2008)
- Padhiary, P.K., Mishra, A.P.: Development of improved artificial neural network model for stock market prediction. *Int. J. Eng. Sci. Technol. (IJEST)* **3**(2), 2 (2011)
- Pegah, F.: Stock trend prediction using news articles a text mining approach. In: 2007, Master Thesis, Luleå University of Technology
- Suresh, R., Harshni, S R.: Data mining and text mining—a survey. In: 2017 International Conference on Computation of Power, Energy Information and Communication (ICCPEIC)
- Verma, M., Srivastava, M., Chack, N., Diswar, A.K., Gupta, N.: A comparative study of various clustering algorithms in data mining. *Int. J. Eng. Res. Appl. (IJERA)* **2**(3), 1379–1384 (2012)
- Gharehchopogh, F.S.: Approach and developing data mining method for spatial applications. In: *Proceedings of International Conference on Intelligent Systems & Data Processing (ICISD)*, India, pp. 342–345 (2011)
- Gharehchopogh, F.S., Khaze, S.R.: Data mining application for cyber space users tendency in blog writing: a case study. *Int. J. Comput. Appl.* **47**(18), 40–46 (2012)
- Han, J., Kamber, M., Pei, J.: *Data Mining Concepts and Techniques*, 3rd edn. Morgan Kaufmann Publishing (2012)
- Rai, P., Singh, S.: A survey of clustering techniques. *Int. J. Comput. Appl.* (2010)
- Brockwell, P.J., Davis R.A.: *Time Series: Theory and Method*. Springer (1987)
- Gharehchopogh, F.S., Mohammadi, P., Hakimi, P.: Application of decision tree algorithm for data mining in healthcare operations: a case study. *Int. J. Comput. Appl.* **52**(6), 21–26 (2012)
- Banerjee, D.: Forecasting of Indian stock market using time-series ARIMA model. In: 2nd IEEE International Conference on Business and Information Management (ICBIM), Jan 2014, pp. 131–135
- Adebiyi, A.A., Adewumi, A.O., Ayo, C.K.: Stock price prediction using the ARIMA model. In: 16th IEEE International Conference on Computer Modelling and Simulation (UKSim), March 2014, pp. 106–112
- Sureshkumar, K.K., Elango, N.M.: Exploiting data mining techniques for improving the efficiency of time series data using spss-clementine. *J. Arts Sci. Commerce*. E-ISSN 2229–4686, ISSN 2231–4172
- Faisal, F.: Forecasting Bangladesh's inflation using time series ARIMA models. *World Rev. Bus. Res.* **2**(3), 100–117 (2012)

16. Pentaho Data Mining Community Documentation. Time Series Analysis and Forecasting with Weka (2014)
17. Wadia, A., Mohd Tahir Ismail, S.: Selecting wavelet transforms model in forecasting financialtime series data based on ARIMA model. *Appl. Math. Sci.* **5**(7), c315–326 (2011)
18. Sneha, S.: Applications of ANNs in stock market prediction: a survey. *Int. J. Comput. Sci. Eng. Technol.* **2**(3) (2011) (A review of stock market prediction with Artificial neural network (ANN))
19. Pradhan, K.C., Sham Bhat, K.: An empirical analysis of price discovery, causality, and forecasting in the nifty futures markets. *Int. Res. J. Fin. Econ.* Issue 26 (2009). ISSN 1450–2887
20. Khan, Z.H., Alin, T.S., Hussain, M.A.: Department of CSE, SUST, Sylhet, Bangladesh, Price prediction of share market using artificial neural network (ANN). *Int. J. Comput. Appl.* (0975–8887), vol. 22(2) (2011)

Speaker Recognition Using Noise Robust Features and LSTM-RNN



Mohit Dua, Pawandeep Singh Sethi, Vinam Agrawal, and Raghav Chawla

Abstract A tremendous growth has been observed in terms of active research in the field of speaker recognition. This has been mainly due to the increasing need of zero-touch interfaces in devices and mobile biometric authentication systems. This paper discusses implementation of text-independent speaker verification system using long short-term memory (LSTM)-based neural network for speaker modeling by using various approaches for the front-end feature extraction including Mel Frequency Spectral Coefficients (MFSC), Mel Frequency Cepstral Coefficients (MFCC), Gammatone Filter Spectra (GTF), and Gammatone Filter Cepstral Coefficients (GFCC). Additionally, to determine the best-suited speaker verification system for given noisy conditions of environment, all the combinational systems are tested under induced noisy conditions with white noise at -20 and -40 dB, as well as under clean environmental condition. The results show that the MFSC-based LSTM-RNN combination tends to perform better than all the other combinations regardless of the noise added in the dataset.

Keywords Speaker recognition · LSTM · MFSC · GTF · RNN

1 Introduction

The human hearing system is unique and becomes effective in about 25 weeks of gestation [1], even before our brains develop enough to learn the language. We first

M. Dua (✉) · P. S. Sethi · V. Agrawal · R. Chawla
Computer Engineering Department, National Institute of Technology, Kurukshetra, Kurukshetra,
India
e-mail: er.mohitdua@nitkkr.ac.in

P. S. Sethi
e-mail: pawandeep2007@gmail.com

V. Agrawal
e-mail: vinam.agrawal@gmail.com

R. Chawla
e-mail: raghavchawla7@gmail.com

learn to recognize our mother's voice [2]. Thus, from a sample of sounds, humans can identify which sound belongs to whom. However, to make a computer achieve a similar feat poses many problems. One major problem in recognizing voice of a person in noisy environments. Also, for reliable speaker recognition, sufficiently long speech utterance by a user is needed. The purpose of the speaker recognition system is to correctly verify (or identify) the speaker from a particular set of speakers, thereby making a speaker recognition system a biometric authentication task. The task of general biometric authentication is a way of verifying a person's identity based on his or her physiological behavior [3]. Speaker recognition (or verification) has many applications including speaker recognition on-the-go in videos (as in case of Prime Video), mobile device unlocking (as in case of Google Assistant, Apple's Siri, etc.), biometric authentication systems using voice, etc. Some major current applications include Scobey-Coronach Border Crossing (Canada-US border) which uses a speaker recognition system in order to allow registered local citizens to cross the border without having to declare anything, Barclays Wealth in 2013 announced plans about using speaker recognition passively in order to verify telephone customer's identity within as less as 30 s of normal telephonic conversation, use of speaker recognition in criminal investigation as in the case of 2014 execution of James Foley, and HSBC offering their customers biometric authentication-based banking software for them to access their account online using either their fingerprint or voice, among others.

1.1 Related Work

The development of automatic speaker recognition systems started in the early 1960s. The researchers at Bell Labs (Bell Telephone Laboratories) at that time worked on creating a similarity measurement that compared two signals using filter bank arrays. One of the biggest achievements includes determination of similarity of two signal spectrograms by cross-correlating them [4]. At around the same time, researchers at IBM corporation used linear discriminant analysis (LDA) to achieve an accuracy of more than 90% for the task of telling a known speaker from an imposter [5]. Texas Instruments, in 1976, was successful in creating the first fully automated text-dependent speaker verification system using digital filter banks for spectral analysis which was then tested by the US Air Force [6]. The Speech Group was set up at National Institute of Standards and Technology (NIST) in mid-1980s. The purpose of this Speech Group was to study and develop new techniques for speech processing [7]. The speech input used for speaker recognition systems generally contains noise, which causes distortions and information loss [8]. The National Security Agency (NSA) funded NIST Speech Group has been responsible for driving the technology forward and finding the most promising approaches. [9]

In 2000s, many modeling techniques utilized Gaussian mixture models (GMMs) and support vector machine (SVM) for classification. The GMM mean super vector SVM system represents acoustic observations as a series of GMM vectors with

discriminative SVM classification [10]. In 2002, Reynolds gave an overview of the various front-end and back-end technologies which can be exploited to implement a robust speaker recognition [11]. In 2011, Krishnamoorthy created a text-independent speaker recognition system [12] using TIMIT dataset. In 2015, Richardson applied DNN approaches for recognition of speaker and language and significant gains were observed [13]. In the same year, Weninger used LSTM-based RNNs to further improve the performance of the system [14]. Snyder, in 2017, used DNN embeddings to create text-independent speaker verification system to improve the performance [15]. In 2019, Guan et al. [16] developed methods and systems for automatic discovery of fraudulent calls using speaker recognition.

In this paper, a text-independent speaker recognition system with various combinations of front-end techniques with neural network-based back-end technique has been developed. In order to make the proposed system noise robust, the developed combinations are tested in clean and as well as noisy speech data, prepared by inducing white noise in TIMIT dataset.

2 Preliminaries

The performance of any speaker recognition system largely depends on the extraction technique used for the speech sample as well as processing technology. The process of speaker recognition begins by first collecting the audio from which the speaker needs to be verified or identified. This is usually done by recording the voice of the speaker using a device such as a microphone. Now the audio needs to be processed to extract useful information using which the speaker can be verified. The pre-processing consists of windowing, framing, and filtering. The features are then extracted using feature extraction techniques such as Mel Frequency Cepstral Coefficients (MFCC), Mel Frequency Spectral Coefficients (MFSC), Gammatone Filter Spectra (GTF), and Gammatone Filter Cepstral Coefficients (GFCC). These features are then used to obtain speaker embeddings and these embeddings are matched against embeddings of speaker model to verify the identity of the speaker.

2.1 Feature Extraction Techniques

Mel Frequency Cepstral Coefficients (MFCC): The most notable feature extraction technique used today is Mel Frequency Cepstral Coefficients (MFCC) [16]. Human auditory system is unable to well-perceive linear scale of speech. As a result, MFCC bases itself more on the critical variations observed in human auditory scale, spacing filters on low frequencies linearly, while filters on higher frequencies are spaced logarithmically. In order to achieve this, MFCC uses the Mel frequency scale. Mel scale is used as it is linearly spaced below 1000 Hz and has logarithmic spacing above

1000 Hz. MFCC uses Fourier transform and Mel filter banks to extract features. Mel frequency (f_{mel}) can be calculated using sound frequency (f_c) as follows:

$$f_{\text{mel}} = \log_{10} \left(1 + \frac{f_c}{700} \right) * 2595 \quad (1)$$

where f_{mel} is frequency on Mel scale and f_c is the given frequency in hertz.

Mel Frequency Spectral Coefficients (MFSC): In Mel Frequency Spectral Coefficients (MFSC), the cepstral representation of the audio stream which is used to derive the MFCCs is directly used as speech feature representations. The drawback of doing this is that the features that are derived have non-local characteristics. This can be explained as the discrete cosine transform (DCT) operation for generating MFCC features is meant to eliminate the correlations among energy coefficients. The order of filter bank energies changes. This disturbs the local property. Instead of this, another approach that can be employed is to use the log energies directly derived from filter bank energies. This is known as MFSC feature extraction. The relation between MFCC and MFSC features can be summed up as follows: MFSC are MFCC features but before the application of the last DCT layer.

Gammatone Filter Cepstral Coefficients (GFCC): Gammatone Filter Cepstral Coefficients (GFCC) [17] is another one of the popular perceptually inspired front-end feature extraction technique used in speech recognition. Here, gammatone filter banks are used instead of Mel filter banks as in case of MFCC or MFSC. Other steps remain almost the same in GFCC as in MFCC. Fourier transform is applied on the pre-processed speech signal before it is passed through gammatone filter banks. Another difference in GFCC is the use of equal loudness pre-emphasis on gammatone filtered signal. Finally, DCT is to this signal to obtain GFCC features.

2.2 Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) model [18] is a type of recurrent neural network (RNN) architecture used in the field of deep learning. Commonly, an LSTM unit is made up of various components, namely input gate, output gate, cell, and forget gate. Broadly, the input gate specifies the magnitude or degree of effect that a new value flowing in has on cell. Similarly, forget gate specifies the extent to which the cell retains a value and output gate denotes the extent to which the cell's value is used for computing activation of output for the LSTM unit. The cell holds the responsibility of keeping record of the dependencies among elements in input sequence. Intuitively, the cell stores values for random time periods, while the input, output, and forget gates control information flowing into and out from the cell. Most prominent applications of LSTM networks include classification, processing, and prediction based on time-sequenced data. Traditional RNNs, unlike LSTMs, are prone to vanishing and

exploding gradient problems. LSTMs were developed to specifically deal with these problems. Theoretically, plain RNNs are capable to track long-term dependencies in input data sequences. The problem is that when training such a plain RNN using back-propagation, the back-propagated gradients may tend to vanish or explode. Vanishing gradients occur when gradients tend to zero, while gradients explode when they tend toward infinity. RNNs made up of LSTM units solve the vanishing gradient problem by allowing gradients to flow unchanged. Although vanishing gradient problem is solved, LSTM networks are still prone to exploding gradient problem.

3 Proposed Architecture

The major factors that affect the performance of a speaker recognition system are: the technique which we use to extract features from a speech utterance, i.e., front-end technique, and the back-end technique which is used for speaker modeling. In this work, we create speaker recognition systems with a variety of front-end techniques, namely MFCC, MFSC, GTF, and GFCC. These techniques are used in conjunction with various neural network-based back-end speaker modeling techniques including LSTM networks. The performance of these combinational systems is then compared to determine the best performing system. For the purpose of measuring system performance in all of the above cases, we use the heuristic called as equal error rate (EER). Additionally, environmental factors such as noise signal attenuation, etc. might have significant effect the performance of the speaker recognition system. In order to make our system more noise robust and withstand real-life noisy speech utterances accurately, we train our system not only under clean, but also under noisy environmental conditions induced by manually adding white noise to our training and testing data. The dataset that we use for our speaker recognition task is the Texas Instruments-Massachusetts Institute of Technology (TIMIT) Dataset for English speech recognition. We modify the dataset for suiting our speaker verification task. For training on noisy data, we add white noise to the aforementioned TIMIT dataset using the Linux SoX (Sound eXtraction) utility. We add the noise at two varying signal to noise ratios (SNRs), i.e., at -10 , at -20 , and at -40 dB. This is done in order to determine the system best suited to specific noise conditions in the environment.

Figure 1 shows the pipeline of the speaker recognition system. Speech signal which is in .wav format is first pre-processed. Then features are extracted from the pre-processed signal using front-end techniques such as MFCC, MFSC, GFCC, GTF. The extracted features are then fed into the LSTM stack which creates speaker representation. Then cosine similarity is obtained for speaker representation which is obtained from the LSTM stack and the average speaker representation. Then the claimed speaker is verified based on the cosine similarity.

Speech signal first needs to be pre-processed (Fig. 2) before feature extraction. The speech signal pre-processing consists of first sampling the signal by using a sampling rate. In our case, we use 16 KHz as sampling rate. Filtering of the signal is done based on a threshold value. In our case, we take the threshold value to be 30 dB. Then the

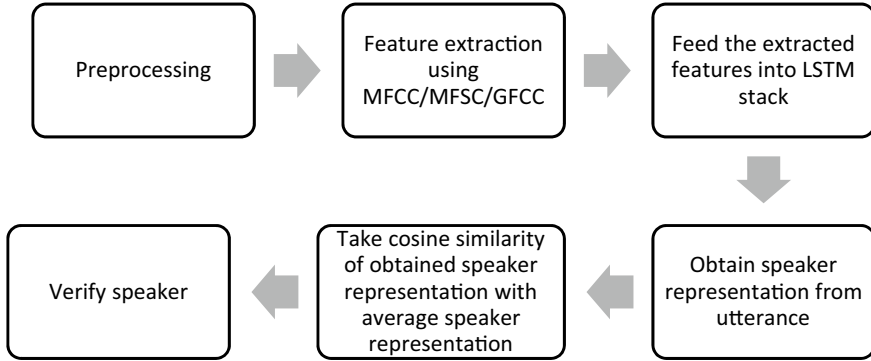


Fig. 1 High-level architecture of the proposed speaker recognition system

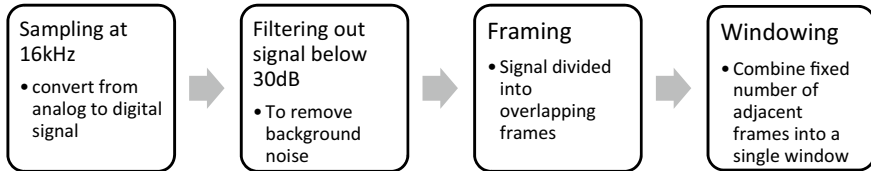


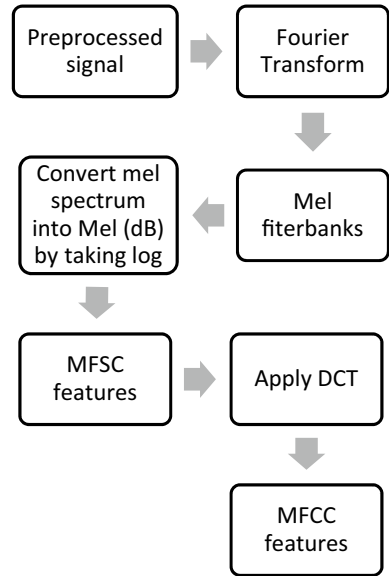
Fig. 2 Pre-processing of speech signal

frames are created for the filtered sample. The frames are overlapped to some extent to prevent data loss. Then, constant number of adjacent frames is combined to form a single window. This process of combining adjacent frames is called windowing. We take a frame size of 0.25 ms with a frame shift of 0.1 ms.

MFSC and MFCC feature extraction involves a similar process (Fig. 3). To extract MFCC features, we need to apply DCT to the extracted MFSC features. In case of MFSC feature extraction, we start with applying Fourier transform and we obtain the magnitude spectrum after this step. Then Mel filter banks are used to obtain Mel spectrum out of magnitude spectrum. Then Mel spectrum is converted into Mel dB by taking log. This yields MFSC features. After this step, we apply DCT to obtain MFCC features. For obtaining GFCC features, step of applying Fourier transform is same as in MFSC and MFCC. Then instead of Mel filter banks, gammatone filter banks have used. Gammatone filter banks have similar impulse response as human auditory filers.

After this equal loudness step is performed. Then log of the obtained amplitude is taken to convert into decibel. Then DCT is applied to finally obtain the GFCC features. If the DCT step is not performed, the features that we get are known as GTF features. For obtaining GFCC features, step of applying Fourier Transform is same as in MFSC and MFCC. Then instead of Mel filter banks, gammatone filter banks are used.

Fig. 3 Calculating MFCC and MFSC features



Gammatone filter banks have similar impulse response as human auditory filers. After this equal loudness step is performed. Then log of the obtained amplitude is taken to convert into decibel. Then DCT is applied to finally obtain the GFCC features (Fig. 4). If the DCT step is not performed, the features that we get are known as GTF features (Fig. 4). The extracted features serve as input to the LSTM stack (Fig. 5). We use three LSTM layers to form the LSTM stack. Each LSTM layer consists of 768 LSTM units. The output layer is linear, consisting of 256 projection units. The output vector from this layer is used as speaker representation. The optimizer which we use is stochastic gradient descent (SGD) and the learning rate is 0.01 initially.

Fig. 4 Calculating GFCC and GTF features

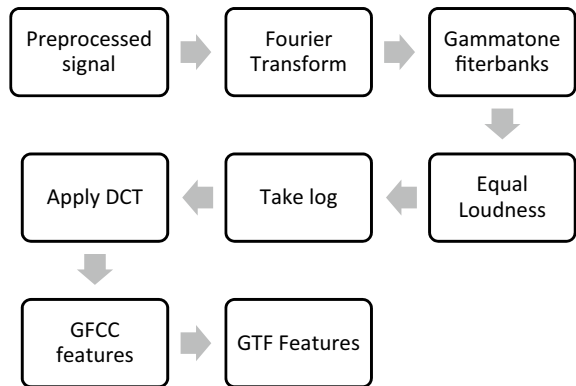
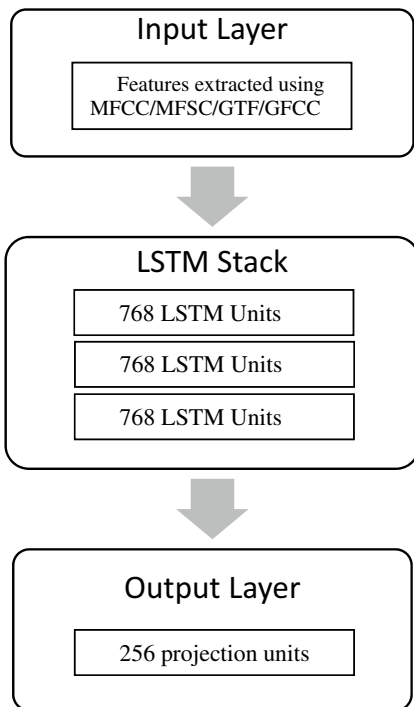


Fig. 5 LSTM-based RNN network



4 Implementation Details and Results

For training and testing purposes, TIMIT dataset consisting of 6300 utterances by 630 speakers of 8 dialects of the USA, with every speaker having 10 utterances each, has been used. The models have been trained on 950 epochs. For creation of noisy data, we use SoX (Sound eXtraction) utility in Linux to add white noise having -20 and -40 dB SNR. Learning rate is initially kept at 0.01 in configuration. With each 300th epoch, the learning rate is halved. For training, a batch size of 16 speakers with each speaker's 10 utterances have been picked. For testing purpose, the same batch size of 16 speakers with 10 utterances per speaker has been used. The models are tested on 10 epochs over the test set. The order of occurrence speakers is shuffled in each epoch. This is done so that concentration of certain test cases in a particular batch of an epoch does not affect our results. The average EER for all the 10 epochs is used to measure the overall accuracy.

The implementation of this system is done using PyTorch library. PyTorch is a Torch-based machine learning library for Python, which was created at Facebook for open-source use. It has been used due to intuitive and readable nature, along with widely available and up-to-date documentation. The training and testing of our model, along with feature extraction, are carried out using the PARAM Shavak

Table 1 Comparative analysis for different combinations

Noise type	Equal error rate (%)			
	MFSC	GTF	MFCC	GFCC
Noiseless data	3.18	12.57	13.97	13.59
White noise (−40 dB)	3.99	13.87	18.83	15.11
White noise (−20 dB)	11.36	14.36	21.78	16.15

supercomputer which has two GPUs: Nvidia Quadro P5000 (16 GB DDR5X) and Nvidia Quadro P400 (2 GB GDDR5) for deep learning tasks.

4.1 Results

As described by Table 1, the results have been observed for different combinations made by varying front-end feature extraction techniques and extent of noise in data. The front-end techniques are MFCC, MFSC, GFCC, and GTF. The noise levels are −20 and −40 dB. The tests have been carried out for noiseless data also.

It can be analyzed that MFSC tends to perform better than all the other feature extraction techniques, regardless of the noise added in the dataset. Also, with an LSTM back-end, MFCC feature extraction tends to be outperformed by all other feature extraction techniques. In case of GFCC and GTF, only a marginal difference in performance is observed. The steps needed to compute log Mel energies are inspired by the properties of speech signal and human perception of such signals. Due to the limitations of some machine learning algorithms, the DCT step is needed to compute MFCC. DCT de-correlates filter bank coefficients.

5 Conclusion and Future Work

With the introduction of deep learning systems in the field of speech and speaker recognition, the choice of MFCCs gets debatable since DNNs are less vulnerable to highly correlated inputs. Speech signals are usually highly nonlinear, while DCT is a linear transformation. Due to this, some information in the speech signals might have been lost as DCT was applied over filter banks. Therefore, DCT becomes no longer necessary. This also explains why MFSC features outperform MFCC features in our work.

References

1. Graven, S.N., Browne, J.V.: Auditory development in the fetus and infant. *Newborn Infant Nurs. Rev.* **8**(4), 187–193 (2008)
2. Kisilevsky, B.S., Hains, S.M., Lee, K., Xie, X., Huang, H., Ye, H.H., Wang, Z.: Effects of experience on fetal voice recognition. *Psychol. Sci.* **14**(3), 220–224 (2003)
3. Wayman, J.L., Jain, A.K., Maltoni, D., Maio, D. (eds.): *Biometric systems: technology, design and performance evaluation*. In: Springer Science & Business Media (2005)
4. Pruzansky, S.: Pattern-matching procedure for automatic talker recognition. *J. Acoust. Soc. Am.* **35**(3), 354–358 (1963)
5. Li, K.P., Dammann, J.E., Chapman, W.D.: Experimental studies in speaker verification, using an adaptive system. *J. Acoust. Soc. Am.* **40**(5), 966–978 (1966)
6. Haberman, W., Fejfar, A.: Automatic identification of personnel through speaker and signature verification—system description and testing. In: *Proceedings of Carnahan Conference on Crime Countermeasures*, pp. 23–30 (1976)
7. NSTC Biometrics: “Speaker Recognition,” 7 August 2006. <https://www.biometrics.gov/Documents/speakerrec.pdf>. Accessed on March 2014
8. De La Torre, A., Segura, J. C., Benitez, C., Ramirez, J., Garcia, L., Rubio, A.J.: Speech recognition under noise conditions: compensation methods. In: *Robust Speech Recognition and Understanding*, 439 (2007)
9. Speaker Recognition Evaluation, 5 March 2012. Available <https://www.nist.gov/itl/iad/mig/sre.cfm>
10. McLaren, M., Vogt, R., Baker, B., Sridharan, S.: A comparison of session variability compensation techniques for SVM-based speaker recognition. In: *Eighth Annual Conference of the International Speech Communication Association* (2007)
11. Reynolds, D.A.: An overview of automatic speaker recognition technology. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. IV-4072. IEEE (2002)
12. Krishnamoorthy, P., Jayanna, H.S., Prasanna, S.M.: Speaker recognition under limited data condition by noise addition. *Expert Syst. Appl.* **38**(10), 13487–13490 (2011)
13. Richardson, F., Reynolds, D., Dehak, N.: Deep neural network approaches to speaker and language recognition. *IEEE Signal Process. Lett.* **22**(10), 1671–1675 (2015)
14. Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J.R., Schuller, B.: Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: *International Conference on Latent Variable Analysis and Signal Separation*, pp. 91–99. Springer, Cham (2015)
15. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: *Interspeech*, pp. 999–1003 (2017)
16. Guan, Z., Ashby, C.S., Moulinier, I.A.Y., Dickison, M.E.: U.S. Patent No. 10,659,588. U.S. Patent and Trademark Office, Washington, DC (2020)
17. Wanli, Z., Guoxin, L.: The research of feature extraction based on MFCC for speaker recognition. In: *Proceedings of 2013 3rd International Conference on Computer Science and Network Technology*, pp. 1074–1077 (2013)
18. Shi, X., Yang, H., Zhou, P.: Robust speaker recognition based on improved GFCC. In: *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pp. 1927–1931 (2016)

Brain Tumor Segmentation Using Random Walks from MRI Images



Shiv Naresh Shivhare and Nitin Kumar

Abstract Magnetic resonance imaging (MRI) is a vital and universally recognized medium to assess brain neoplasms. This paper presents a study on brain tumor segmentation based on the random walk algorithm which is a graph-based method in which pixels of a brain MR image are treated as nodes. Segmentation is performed by interactively labeling certain nodes as foreground and background seeds, followed by computing the probability of each unlabeled node to reach all the labeled nodes using random paths. The method is applied on two different MR modalities viz. T2-weighted MRI with fluid attenuated inversion recovery (FLAIR), and T2 MRI to segment complete tumor, and tumor core regions, respectively, by utilizing visual traits of MRI images and identifying local and global brain tissues information. Efficacy is validated quantitatively as well as qualitatively through performing the experiments on publicly available brain tumor segmentation challenge (BRATS-2013) dataset. Results demonstrate that the proposed method performs favorable as compared to several existing methods.

Keywords Brain tumor segmentation · Magnetic resonance imaging · Random walks

S. N. Shivhare (✉) · N. Kumar

Department of Computer Science and Engineering, National Institute of Technology Uttarakhand, Srinagar Garhwal, Uttarakhand, India
e-mail: shiv827@gmail.com

N. Kumar

e-mail: nitin@nituk.ac.in

S. N. Shivhare

School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_3

1 Introduction

Automated MRI image analysis plays a significant role in the diagnosis and assessment of various brain tumor types and for sophisticated treatment planning. Glioblastoma multiforme (GBM), a malignant grade-IV tumor type, is one of the most affected and dangerous brain diseases in the world [3, 17]. Automation, accuracy, and time efficiency for a method of brain tumor segmentation are the three major needs of the hour because of the following reasons, (i) manual brain image analysis is quite time taking and depends upon the expertise of the radiologist, (ii) variable and amorphous structure of tumor affects the robustness of any tumor segmentation method. Except to the healthy or normal brain tissues, i.e., white matter, gray matter, and cerebro-spinal fluid (CSF), brain lesion is subdivided into four major components, edema, non-enhancing solid core, enhancing tumor, and necrosis. MRI in the form of its various modalities expresses extensive global and local level information of brain's healthy and neoplastic tissues. Community of worldwide researchers involved in BRATS challenge is playing a significant role by creating and maintaining benchmarks specially in terms of providing a dataset [17] and defining evaluation measures. BRATS dataset contains MR images in four MRI modalities, i.e., FLAIR, T1-weighted, T2-weighted, and post-contrast T1-weighted MRI as displayed in Fig. 1. Annotation of dataset is performed by different experts from various institutions based on the visual characteristics of MR image modalities. T2-weighted MR modality images are helpful to extract active tumor also known as tumor core without edema. Similarly, FLAIR images are suitable to identify whole tumor region by observing the hyper-intense lesion [8].

State-of-the-art methods in the literature of brain lesion detection and segmentation can be categorized in four major ways as follows: atlas-based methods, region-based methods, gradient and edge-based methods, clustering and classification-based methods. Atlas-based methods utilize the anatomical structure of healthy brain tissues followed by registering the same with tumorous brain image in order to capture and analyze the affected brain region [1, 17, 23]. As the anatomy of human brain tissues possesses significant variation case by case, atlas-based methods are only useful for the rough estimation of lesion area, and hence, they are generally used as a preprocessing step in any lesion segmentation method. Region-based methods

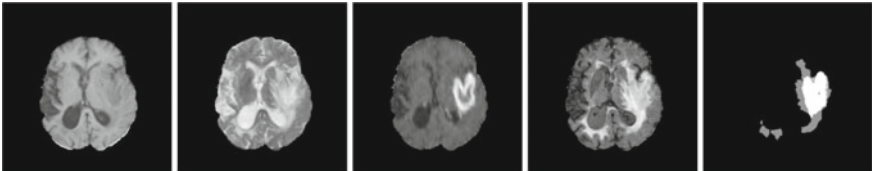


Fig. 1 Axial view of MRI modalities of a selected subject from BRATS dataset [14]. From left to right: T1-weighted, T2-weighted, T1c, FLAIR followed by the corresponding annotated ground truth

perform segmentation by first investigating features of different pixels (or voxels) and further finding the affinity among them [15, 22]. Feature extraction is the major step in region-based segmentation followed by a similarity measurement criteria which often leads to oversegmentation problem. Edges and corner regions are the group of pixels which represent the significant information whenever there is an abrupt change in the intensity or appearance in an image. Edges and corners are the joint gradient information in both horizontal and vertical directions which can be exploited to recognize salient objects specially abnormal lesion in brain MR images [11, 20]. However, disadvantages of such approach include user interaction for initial region selection and appearance of lesion in low contrast images. Combining the hypo-intense and hyper-intense appearance of multiple MR modalities together has been a solution of such problems. Moreover, clustering and classification-based methods are currently the most popular due to the significant growth of machine learning paradigm. Classification-based methods usually rely on certain features of different MR modalities such as local intensity of pixels, texture information around the neighborhood, statistical information, and spatial distribution of pixels. Thus, extracted features are exploited to segment the image using either supervised or unsupervised (also known as clustering) classification algorithms [2, 3, 12, 13, 18, 24].

Agar et al. [1] proposed an atlas-based lesion segmentation method in which the initial lesion area is detected by registering the input brain MR images with the healthy atlas-based brain tumor prior. Tumor sub-compartments have been further segmented by implementing a Convolutional Restricted Boltzmann Machines classifier (RBMs). Another generative method is suggested by Menze et al. [17] to segment brain tumor in multi-modality MR images by deriving a tumor estimation algorithm. However, atlas-based segmentation methods are prone to false segmentation since the segmentation quality depends on accurate registration procedure. Working toward the similar objective, Letterboer et al. [15] proposed a region-based method which gives radiologists an initial estimation of brain abnormality towards providing a better treatment planning. Recently, Tong et al. [22] presented a method using texture feature extraction and kernel dictionary learning. Dictionary coding for normal and abnormal brain voxels is performed which in order to further classify the tumor region using linear discrimination function. Sachdeva et al. [19] proposed a whole tumor region segmentation method based on statistical and texture features and active contour model. Content-based active contour is a deformable model which is employed to evolve a user-initialized curve around the specified object boundary under the influence of internal and external forces. Exploiting the classification-based method, Bauer et al. [2] suggested joint architecture of hierarchical conditional random field (CRF) and support vector machine (SVM) classifiers for efficient segmentation of whole tumor volume. Corso et al. [3] proposed a method to segment tumor and edema regions based on Bayesian classification in which weighted aggregation is used to compute the model-aware affinities. Moreover, advancements of classification-based methods have been carried out using convolution operation-based machine learning methods in order to improve the performance of brain lesion segmentation. Havaei et al. [12] proposed a two-path way convolutional neural network (CNN)-based segmentation

method which utilizes local as well as global features of brain tissues. Likewise, Kamnitsas et al. [13] presented a three-dimensional version of CNN architecture with 11 layers in its pipeline. A three-dimensional CRF is also exploited in the end to refine the performance of lesion segmentation. Pei et al. [18] suggested another classification-based method utilizing tumor cell density and textures features of brain tissues in association with random forest (RF) classifier. The method also implements joint-label fusion approach with another segmentation method to improve the performance. In a recent study, Zhao et al. [24] trained the RF classifier by extracting gradient and circular context-sensitive features. The total extracted features are reduced by using minimum redundancy maximum relevance (MRMR) approach.

We propose a semi-automated tumor segmentation method with following major contributions:

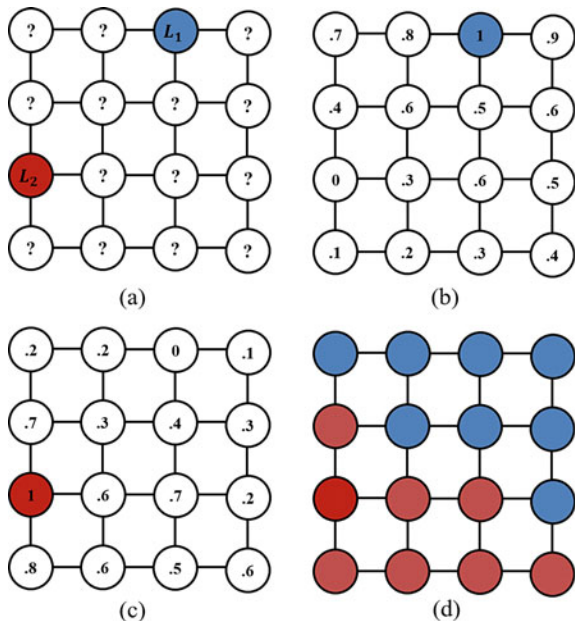
1. Random walks algorithm is implemented in semi-automated manner to detect and segment complete tumor and tumor core regions from FLAIR and T2-weighted MR images, respectively.
2. By utilizing the prior visual constraints of different MR modalities, multiple seed points representing foreground and background are chosen interactively in order to significantly improve the segmentation performance.
3. The proposed method segments complete tumor and tumor core regions in 0.2 s for each set of T2-weighted and FLAIR MR images which illustrates the computational efficiency of the method.

Remaining of the paper is structured in the following way: Sect. 2 describes the proposed brain lesion segmentation method in detail. Section 3 demonstrates the experimental setup and results implemented on the benchmark datasets along with the comparison with other existing methods. Lastly, conclusion presents the significant outcomes of this work with future works in Sect. 4.

2 Proposed Method

In this study, a semi-automated method is proposed to recognize and segment two different affected regions of brain tumor by exploiting random walks method. The segmentation procedure of random walks algorithm is shown in Fig. 2 which illustrates that how a node (pixel in an image) is labeled based on the interactively selected foreground and background seed points. Probability shown at each node represents that how easy or difficult it is to reach a selected seed point starting from that particular node by following random walks. The proposed method is implemented on FLAIR and T2 MR images to obtain complete tumor and tumor core regions, respectively. Segmentation in various classes can be done based on the chosen seed points. However, we have chosen seeds corresponding to two labels (for foreground and background) by utilizing the visual constraints and hyper-intense regions of FLAIR and T2-weighted MRI modalities.

Fig. 2 Example showing segmentation using random walks. **a** Two seeds points L_1 and L_2 as shown in blue and red colors. **b** Calculation of probabilities to reach seed L_1 starting from each node through random walks. **c** Calculation of probabilities to reach seed L_2 starting from each node through random walks. **d** Assigning labels to each node of the graph based on the calculated probabilities corresponding to initially selected seed points



2.1 Random Walks Segmentation

Image segmentation using random walks is a graph-based semi-automated method in which a 2-dim input image represents a graph and each pixel is treated as a node (vertex) of the graph [10]. A few nodes (seed points) of the graph are interactively chosen as labels. In our case, seeds points are chosen corresponding to two labels, i.e., foreground and background for each image. A probability corresponding to each unmarked node is computed which shows that how easy or how difficult it is to reach a particular label (background or foreground) starting from each unmarked node through a randomly chosen walk. Thus, two probabilities are calculated for each unmarked node based on which the node is segmented in either of the class.

The mathematical formulation and description of the random walks method is given as follows [10]: Let $G = (V, E)$ be the graphical representation of an image where V and E represent the set of vertices (pixels) and set of edges between two vertices, respectively. Each edge $e \in E$ between vertices v_p and v_q has a weight w_{pq} . Let G be an undirected graph and $d_p = \sum_q w_{pq}$ be the degree of a vertex v_p for the edges e_{pq} which incident on the vertex v_p . Value of the weight w_{pq} is defined based on the Gaussian function applied over the difference between the local level features (pixel's intensity) of the two vertices v_p and v_q as given below:

$$w_{pq} = \exp(-\beta(f_p - f_q)^2) \quad (1)$$

where f_p and f_q represent the local features of pixels v_p and v_q respectively. β is the additional parameter whose value can be set empirically depending on the application and image type. The random walks segmentation method is inspired by the flow of electrons through random paths in an electrical circuit network. β indicates the conductance phenomenon over the network by guiding how easy or difficult it is to reach a marked node from each unmarked node following through random paths. It becomes very easy to move from one node v_p to another node v_q if the value of w_{pq} is very close to 1. Similarly, it is very difficult to move from one node to another in case the value of w_{pq} is close to 0.

The solution of random walks segmentation problem can be solved through electrical current and voltage laws. However, another alternate solution of the same is given in [10] by computing the probabilities of random walks analytically in order to improve the efficiency of the method. Computing the probabilities of random walks can be represented in terms of combinatorial Dirichlet problem which is solved over a Laplacian equation derived in context of random walks problem. The functional of combinatorial Dirichlet integral [10] is given as below:

$$D[\mu] = \frac{1}{2} \int_{\Psi} |\nabla \mu|^2 d\Psi \quad (2)$$

where μ and Ψ represent field and regions, respectively [4]. The Dirichlet problem is defined on the basis of a harmonic function with its corresponding boundary constraints. A harmonic function compatible to the boundary constraints also optimizes the value of Dirichlet integral due to the Euler-Lagrange property of the Laplacian equation [4]. The combinatorial Laplacian matrix [7] is defined as follows:

$$L_{pq} = \begin{cases} d_p & \text{if } p = q, \\ -w_{pq} & \text{if } v_p \text{ and } v_q \text{ are adjacent vertices,} \\ 0 & \text{Otherwise,} \end{cases} \quad (3)$$

where L_{pq} is indexed by nodes v_p and v_q .

Let V_M be the set of marked nodes (seed points) corresponding to the foreground and background and V_U be the set of unmarked nodes (except to the chosen seed points); the probabilities of the set of nodes V_U can be calculated by considering it as a type of Laplacian function which is compatible to the predefined boundary conditions of the marked nodes V_M . Based on this consideration, the probability estimation of each node of image graph is equivalent to the minimization criterion of the above stated Dirichlet integral:

$$D[x] = \frac{1}{2} x^T L x = \frac{1}{2} \sum_{e_{pq} \in E} w_{pq} (x_p - x_q)^2 \quad (4)$$

where x_p and x_q represent the probabilities at vertices v_p and v_q , respectively. The critical points of Dirichlet integral $D[x]$ become minima due to the semi-definite and positive property of the Laplacian matrix L . The above stated equation can be rewritten in terms of set of marked and unmarked node V_M and V_U as follows:

$$D[x_U] = \frac{1}{2} [x_M^T \ x_U^T] \begin{bmatrix} L_M & B \\ B^T & L_U \end{bmatrix} \begin{bmatrix} x_M \\ x_U \end{bmatrix} \quad (5)$$

where x_M and x_U represent the potential or the computed probabilities of marked seed points and unmarked vertices, respectively. The following equation is obtained by performing differentiation of $D[x_U]$ with respect to x_U as the solution of random walks problem in order to calculate the potential or probabilities of unmarked vertices:

$$L_U x_U = -B^T x_M \quad (6)$$

Practically, number of unmarked nodes in an image graph is much greater than that of marked seed nodes which makes this solution computationally complex due to the complex calculation of L_U^{-1} while solving the equation $x_U = -L_U^{-1} B^T x_M$. However, user interaction for choosing initial foreground and background seed points makes this method efficient in terms of producing accurate and fast segmentation results.

3 Experimental Results

Efficacy of the method is validated qualitatively as well as quantitatively through performing the experiments on real glioma images from publicly available benchmark dataset¹ (BRATS-2013) as shown in Fig. 3 and Table 1. Experimental work has been carried out on the Dell desktop having 64-bit Windows 10 OS, with Intel(R) i-7 processor and 8 GB of random access memory (RAM).

3.1 Dataset

BRATS-2013 dataset contains numerous real Glioma images of low-grade (LG) and high-grade (HG) brain tumor type. Each 2-dim slice of dataset is skull stripped in order to maintain the anonymity of patients and available in the form of four different modalities such as T1-weighted images, T1c images, T2 weighted images, and FLAIR images. However, only T2-weighted and FLAIR MR image are utilized in this work. Real image data consists of the images of 20 HG and 10 LG Glioma subjects.

¹<https://www.smir.ch/BRATS/Start2013>.

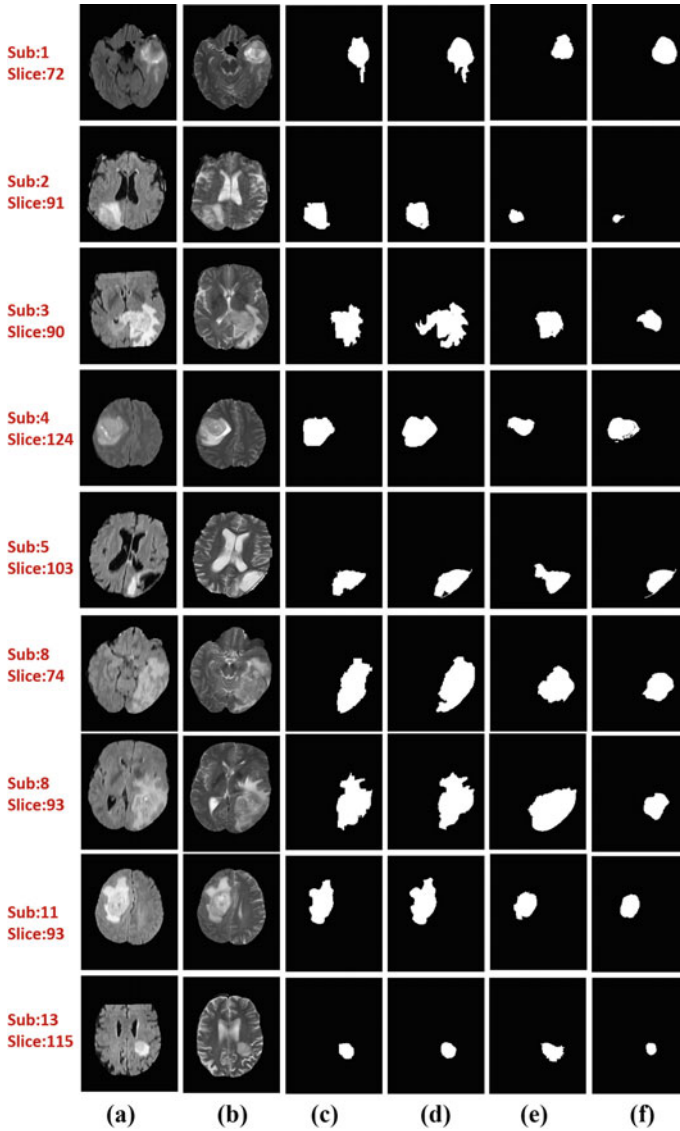


Fig. 3 Segmented complete tumor and tumor core regions on the selected subjects of BRATS2013 high-grade (HG) dataset. Each row delineates the subject id slice sequence number. **a** FLAIR MR image, **b** T2 MR image, **c** segmented complete tumor region, **d** ground truth corresponding to complete tumor region, **e** segmented tumor core region, **f** ground truth corresponding to tumor core region

Table 1 Results obtained on BRATS-2013 real image dataset for the prediction of complete tumor and tumor core regions and comparison with other existing methods

Methods	Complete tumor			Tumor core		
	DSC (HG/LG)	Sensitivity (HG/LG)	PPV (HG/LG)	DSC (HG/LG)	Sensitivity (HG/LG)	PPV (HG/LG)
Demirhan et al. [5]	0.64/0.62	0.69/0.72	0.74/0.69	0.48/0.50	0.63/0.67	0.58/0.62
Geremia et al. [9]	0.65/0.53	0.73/0.55	0.68/0.53	0.52/0.26	0.51/0.28	0.79/0.32
Taylor et al. [21]	0.74/0.40	0.28/0.23	0.71/0.51	0.44/0.05	0.25/0.57	0.67/0.92
Proposed	0.67/0.66	0.73/0.75	0.79/0.78	0.49/0.52	0.69/0.76	0.61/0.60

As per the BRATS challenge guidelines, Glioma brain tumor type can be classified into edema, non-enhanced solid core, enhancing tumor, and necrosis [16]. Edema can be identified as the swelling around other brain tumor tissues. Therefore, edema tissues are predominantly curable as compared to other tumor tissue types. Performance of a method is evaluated by effective segmentation of the following regions:

1. Complete tumor region as the fusion of edema, non-enhanced solid core, enhancing tumor, and necrotic core.
2. Tumor core region also known as complete tumor region without edema as the fusion of non-enhanced solid core, enhancing tumor, and necrotic core.

3.2 Qualitative Evaluation

Segmented complete tumor region as well as tumor core region from few of the selected 2-dim slices of HG subjects is shown in Fig. 3. Identification of subject and its corresponding slice is shown at the extreme left of each row. Each row shows the input MR images as well as the segmented brain tumor region along with the ground truth. The visual appearance of the shown images illustrates that the proposed approach produces favorable results for brain tumor segmentation. Moreover, the method lags while detecting tumor core regions as it produces many false positive cases due to the similar range of intensity local feature of lesion and CSF.

3.3 Quantitative Evaluation

The segmentation performance is also validated quantitatively with the available ground truth using the relevant evaluation measures such as dice similarity coefficient

(DSC), sensitivity or true positive rate, and positive predictive value (PPV). DSC [6] also known as the dice score represents an overall agreement between the predicted result (output) and the actual result (ground truth) by measuring the overlapped region. Sensitivity represents the rate with which our prediction is true positive (TP) means the predicted result is positive (for tumorous brain tissues) and actual result is also positive. Positive predictive value (PPV) refers to the number of true positive predictions over all the positive predictions (either positive or negative) predicted by the proposed method.

$$\text{DSC} = \frac{2 * \text{TP}}{(\text{TP} + \text{FP}) + (\text{TP} + \text{FN})} \quad (7)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. Average quantitative results for all 20 HG as well as 10 LG subjects in terms of aforementioned performance measures are given in Table 1. Each subject's 3-dim data size is $216 \times 176 \times 176$ which means total 176, 2-dim slices are available of 216×176 size each. Out of the total 176 slices, tumorous brain tissues information is present only in few of the slices only. Typically, most of the brain tissues information is present in the slices which belong to the middle position. Due to the same fact, 20 middle slices of each subject's brain slices are considered for our experiments.

The segmentation results of the proposed method are also compared with the works of Demirhan et al. [5], Geremia et al. [9], and Taylor et al. [21] after reproducing the segmentation results on the same settings. Segmentation results of Demirhan et al. [5] are obtained by training the corresponding model on BRATS-2013 training dataset. As BRATS dataset provides us the real Glioma images which are skull stripped and registered with T1c MRI, these preprocessing steps were skipped while implementing this model. While reproducing the results of Taylor et al. [21], a Hidden Markov Model (HMM) is trained on 80% of BRATS 2013 training dataset. Pixel-wise classification is performed for both low-grade and high-grade Glioma dataset images. As the dataset is highly unbalanced, down-sampling is done for pixels representing healthy brain tissues before training the model. The detailed segmentation scores corresponding to the selected performance measures are shown in Table 1. We achieved the DSC of 0.67 for HG subjects and 0.66 for LG subjects for segmenting complete tumor region. The DSC for tumor core region is achieved as 0.49 and 0.52 for HG and LG subjects, respectively. Performance in segmenting tumor core region is due to the complex appearance of T2-weighted MR images where not only abnormal brain tissues but also CSF appear hyper-intensive. Still the result of tumor core segmentation is competitive as reported in Table 1. The performance of our method

in terms of sensitivity and PPV is also favorable with the compared methods for both complete tumor and tumor core regions. However, our method outperforms in all the measures for segmenting LG subjects data for both complete tumor and tumor core regions. The performance score obtained indicates that the proposed brain tumor segmentation approach gives significant contribution.

3.4 Discussion

The quantitative and qualitative results shown above suggest that the proposed brain tumor segmentation approach is suitable and efficient for real time brain tumor segmentation. The proposed approach detects and segments the most significant tumor regions, i.e., complete tumor and tumor core regions in order to assist radiologists and provide better treatment planning. However, enhancing brain tumor region cannot be detected separately due to the visual constraints of FLAIR and T2-weighted MRI. All the experiments are conducted on the middle axial slices of each subject of the dataset to ensure that the significant amount of tumorous and healthy tissues present in each slice. Pixels were marked for both background and foreground interactively based on which probability of all unmarked pixels was computed in order to segment each node in the specified labels. The proposed segmentation approach is based on random walks algorithm in which β is the only free parameter in Eq. 1 in order to perform accurate segmentation. For all the experiments, value of parameter β is kept constant as 350. The proposed method is also time efficient as it segments both complete tumor and tumor core regions for each set of T2-weighted and FLAIR MR images in 0.2 s only.

4 Conclusion

In this paper, an efficient segmentation approach is proposed to detect and segment tumor in two significant complete tumor and tumor core regions. A probability-based random walks segmentation approach is exploited slice-by-slice manner to divide each MR image in two segments based on the pre-selected (marked nodes) foreground and background seed points interactively. Each MR image is treated as a graph in which probability of each unmarked node is computed by finding its accessibility against marked nodes through considering random walks in its four-connected neighborhood. All four tumor labels, i.e., edema, non-enhanced solid core, enhancing tumor, and necrotic core are extracted and combined as complete tumor region by implementing the proposed method in FLAIR MR image slices of each subject. Similarly, tumor core region (all tumor labels without edema) is extracted from T2-weighted MR image slices. The evaluation results demonstrated the high performance of the proposed brain tumor segmentation approach when compared to other existing methods. The proposed approach can assist radiologists because of its

computational efficiency. In future, the proposed brain tumor segmentation approach can be extended by selecting foreground and background seed points automatically using feature descriptors like scale invariant feature transform (SIFT) and Harris edge and corner detector.

References

1. Agn, M., Puonti, O., Law, I., af Rosenschöld, P., van Leemput, K.: Brain tumor segmentation by a generative model with a prior on tumor shape. In: *Proceeding of the Multimodal Brain Tumor Image Segmentation Challenge*, pp. 1–4. Springer (2015)
2. Bauer, S., Nolte, L.P., Reyes, M.: Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 354–361. Springer (2011)
3. Corso, J.J., Sharon, E., Dube, S., El-Saden, S., Sinha, U., Yuille, A.: Efficient multilevel brain tumor segmentation with integrated Bayesian model classification. *IEEE Trans. Med. Imag.* **27**(5), 629–640 (2008)
4. Courant, R., Hilbert, D.: *Methods of mathematical physics*, vol. 2. Wiley (1989)
5. Demirhan, A., Törü, M., Güler, I.: Segmentation of tumor and edema along with healthy tissues of brain using wavelets and neural networks. *IEEE J. Biomed. Health Inform.* **19**(4), 1451–1458 (2014)
6. Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology* **26**(3), 297–302 (1945)
7. Dodziuk, J.: Difference equations, isoperimetric inequality and transience of certain random walks. *Trans. Am. Math. Soc.* **284**(2), 787–794 (1984)
8. Drevelegas, A.: *Imaging of brain tumors with histological correlations*. Springer Science & Business Media (2010)
9. Geremia, E., Menze, B.H., Ayache, N., et al.: Spatial decision forests for glioma segmentation in multi-channel mr images. In: *MICCAI Challenge on Multimodal Brain Tumor Segmentation*, vol. 34, pp. 14–18. Citeseer (2012)
10. Grady, L.: Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 1768–1783 (2006)
11. Harris, C., Stephens, M.: A combined corner and edge detector. In: *Alvey Vision Conference*, vol. 15, pp. 147–151. Citeseer (1988)
12. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. *Med. Image Anal.* **35**, 18–31 (2017)
13. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Med. Image Anal.* **36**, 61–78 (2017)
14. Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P.: The virtual skeleton database: an open access repository for biomedical research and collaboration. *J. Med. Internet Res.* **15**(11), e245 (2013)
15. Letteboer, M.M., Olsen, O.F., Dam, E.B., Willems, P.W., Viergever, M.A., Niessen, W.J.: Segmentation of tumors in magnetic resonance brain images using an interactive multiscale watershed algorithm. *Acad. Radiol.* **11**(10), 1125–1138 (2004)
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imag.* **34**(10), 1993 (2015)

17. Menze, B.H., Van Leemput, K., Lashkari, D., Weber, M.A., Ayache, N., Golland, P.: A generative model for brain tumor segmentation in multi-modal images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 151–159. Springer (2010)
18. Pei, L., Bakas, S., Vossough, A., Reza, S.M., Davatzikos, C., Iftekharuddin, K.M.: Longitudinal brain tumor segmentation prediction in mri using feature and label fusion. *Biomed. Signal Process. Control* **55**, 101,648 (2020)
19. Sachdeva, J., Kumar, V., Gupta, I., Khandelwal, N., Ahuja, C.K.: A novel content-based active contour model for brain tumor segmentation. *Magn. Resonance Imag.* **30**(5), 694–715 (2012)
20. Shivhare, S.N., Kumar, N., Singh, N.: A hybrid of active contour model and convex hull for automated brain tumor segmentation in multimodal mri. In: *Multimedia Tools and Applications*, pp. 1–23 (2019)
21. Taylor, T., John, N., Buendia, P., Ryan, M.: Map-reduce enabled hidden Markov models for high throughput multimodal brain tumor segmentation. In: *Multimodal Brain Tumor Segmentation*, vol. 43 (2013)
22. Tong, J., Zhao, Y., Zhang, P., Chen, L., Jiang, L.: Mri brain tumor segmentation based on texture features and kernel sparse coding. *Biomed. Signal Process. Control* **47**, 387–392 (2019)
23. Ullmann, J.F., Janke, A.L., Reutens, D., Watson, C.: Development of mri-based atlases of non-human brains. *J. Comp. Neurol.* **523**(3), 391–405 (2015)
24. Zhao, J., Meng, Z., Wei, L., Sun, C., Zou, Q., Su, R.: Supervised brain tumor segmentation based on gradient and context-sensitive features. *Front. Neurosci.* **13** (2019)

Factors Accountable for Diabetes Using Artificial Intelligence in Medico-Care



Karuna Babber and Shruti Wadhwa

Abstract The new advancements in technology have given us heaps of data. The healthcare industry is one such industry which has accumulated mounds of data from varied sources. To manage this large volume of heterogeneous data which is usually referred as ‘big data’, data analytic tools have been developed. With the changing demands of healthcare management like from disease-centric to patient-centric and volume-based to value-based, healthcare delivery models need to be built using artificial intelligent analytical tools. This paper presents factors accountable for diabetes using artificial intelligence tools. Python programming language has been used to produce the results.

Keywords Blood pressure · Body mass index · Insulin · Diabetes · Big data analytics · Artificial intelligence

1 Introduction

The healthcare industry is world’s largest growing industry. With the peculiar demands of healthcare management, a paradigm shift from conventional methods to new innovative analytical methods is the need of the hour. According to recent study, on a global map, big data expenditure in healthcare has to grow towards Compound Annual Growth Rate of at least 42% during 2015–2020 [1, 2]. The tons of health data like societal, clinical, scientific or diagnostics are adding up on a day-to-day basis. In the technological world, the term ‘big data’ has been coined to describe such huge volumes of data. Douglas Laney [3] described big data by three V’s that is volume, velocity and variety wherein ‘big’ part of big data indicates its large volume, velocity represents the rate and speed of data collection and variety stands for different types

K. Babber
Post Graduate Government College, Chandigarh, India
e-mail: karunababber.kb@gmail.com

S. Wadhwa (✉)
Nidus Technologies Pvt. Ltd., Chandigarh, India
e-mail: shrutiwadhwa99@gmail.com

of organized and unorganized data. Over the time, few other authors [4] added two more V's—veracity and variability into the big data definition. In this big data, a large chunk that is around 90% comprises of unstructured data. To manage such unstructured data and to transform it into subject-oriented information is nothing short of a herculean task. Moreover, it is almost impossible to process such unstructured big data with conventional softwares. We need technically advanced softwares with high-end computational power to make sense of huge data. The implementation of artificial intelligence [5–7] and machine learning algorithms [8] can help us to generate decision-making information.

1.1 Big Data Analytics in Healthcare

The healthcare system has a multidimensional model wherein the treatment, prevention and diagnosis of all health-related issues have to be dealt with extra caution and care. The major stakeholders of healthcare system are health professionals [9–11] (doctors, nurses and other support staff), health facilities (hospitals, clinics, diagnostic and other treatment technologies) and finance houses (governmental/private and insurance organizations). The data coming from all these levels comprises of patients' medical data (prescription and diagnosis reports), clinical and medical data (images, wearable device data and laboratory examinations) and other private/public medical data [12, 13]. Previously, the common practice to store such medical data was hand written or typed reports but with the advent of computer systems, digitization of clinical and medical reports [14–16] has become a widely accepted norm in healthcare. In 2003, Institute of Medicine came up with the idea of maintaining 'Electronic Health Records'. In [17, 18], authors defined Electronic Health Records (EHR) as computerized medical records of patients which provides any information related to past, present or future physical/mental health of an individual. Figure 1 provides general view of big data health analytics.

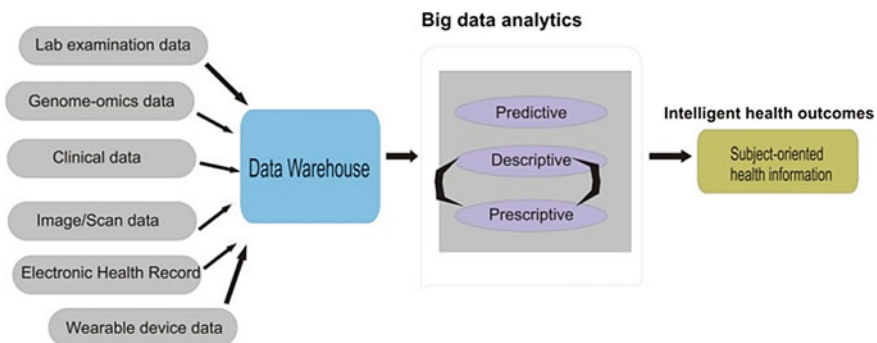


Fig. 1 Big data analytics in healthcare

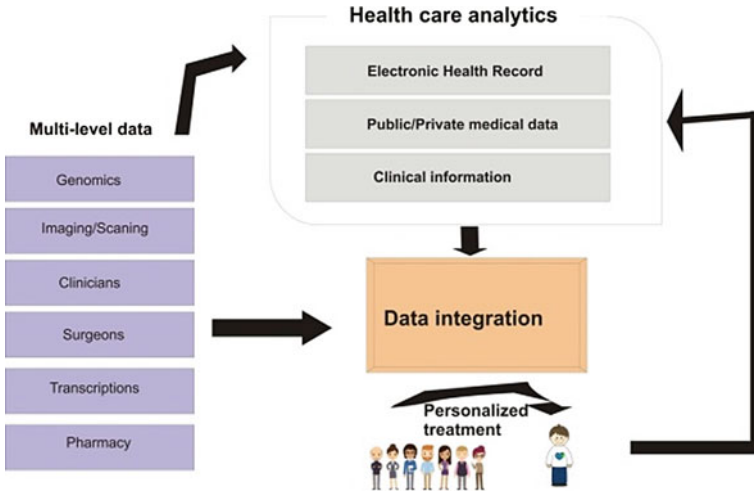


Fig. 2 Framework of healthcare data analytics

1.2 Electronic Health Records

The National Institutes of Health (NIH) of USA has taken up the ‘All of Us’ initiative (<https://allofus.nih.gov>) [19] to gather all types of patients’ data like clinical diagnosis, medical imaging or environmental and socio-behavioural data. The purpose of gathering huge amount of patients’ data is to streamline the healthcare services and to provide timely and coordinated medical help to the persons in need [20, 21]. Moreover, this initiative will be going to reduce redundant additional examination costs, ambiguities caused by illegible handwritten prescriptions [22, 23] and above all it is surely going to cut the ever growing costs of health insurances. By collecting and processing all type of heterogeneous patients’ data, automatic reminders, periodic checkups as well as one stop screenings can be made possible. Figure 2 depicts the framework for integrating multi-level information for personalized medico-care.

2 Data

We have taken dataset for our observations from the Pima Indians Diabetes Database collected by National Institute of Diabetes and Digestive Kidney Diseases [24]. The observational data comprises of different parameters like blood pressure, pregnancy status, insulin, age of people and body mass index (BMI) of individuals. Python and its data science-related packages have been used to carry on the analysis.

In Python, firstly we use ‘Pandas’ library file to read our dataset which is in csv format. For data scrubbing ‘Numpy’ library and few extract, transform, load (ETL)

tools have been used (see Fig. 3). To plot the graphs and to construct inferences, the ‘Matplotlib’ and ‘Sklearn’ libraries of Python are used [25–28].

To begin with, firstly we import ‘Pandas’ library file to read our dataset and after reading the data the normalization of data is carried out. By applying statistical techniques, we split the data into two outcomes, i.e. an outcome of zero (0) received for non-diabetic people and an outcome of one (1) received for diabetic people. The following bar graph (Fig. 4) shows the output.

From Fig. 4, it is clear that more than 250 people with ‘0’ outcomes are non-diabetic, whereas around 125 people with ‘1’ outcome are diabetic. Now to make inferences of different parameters on diabetes, we have taken number of features into consideration, the details of which are provided below.

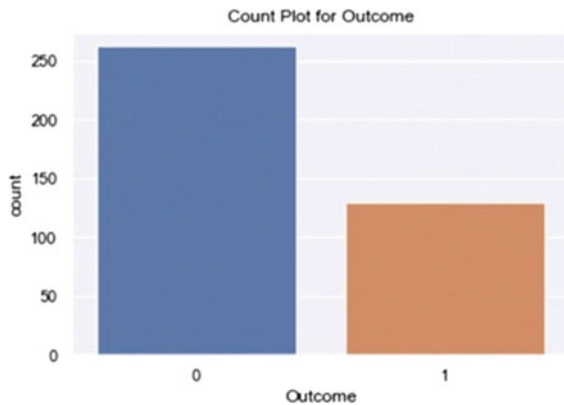
```
4]: diab.head()
4]:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin  BMI  DiabetesPedigreeFunction  Age  Outcome
0            6     148             72           35         0   33.6                0.627   50         1
1            1      85             66           29         0   26.6                0.351   31         0
2            8     183             64           0         0   23.3                0.672   32         1
3            1      89             66           23         94  28.1                0.167   29         0
4            0     137             40           35        168  43.1                2.288   33         1

[5]: diab.isnull().values.any()
## To check if data contains null values
[5]: False

[6]: diab.describe()
## To run numerical descriptive stats for the data set
[6]:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin  BMI  DiabetesPedigreeFunction
count  768.000000  768.000000  768.000000  768.000000  768.000000  768.000000  768.000000  768.000000  768.000000
```

Fig. 3 Pima Indians diabetes dataset

Fig. 4 Bar graph visualization of non-diabetic (0) and diabetic (1) people



3 Results

Blood pressure (BP): The measure of pressure of the blood in the circulatory system is defined as blood pressure [29, 30]. To find a correlation between BP and diabetes, the bar graph (Fig. 5) and box-plot visualization (Fig. 6) are provided below.

In graph I, the BP value is about normal with mean around 65–85 mmHg as against the ideal value of 80 mmHg. In graph II, for diabetic patients, the BP value plot is little skewed towards right, i.e. it is around 75–85 mmHg, whereas for non-diabetic people, it is around 60–75 mmHg. The box-plot in Fig. 6 clearly shows that the ‘whiskers’

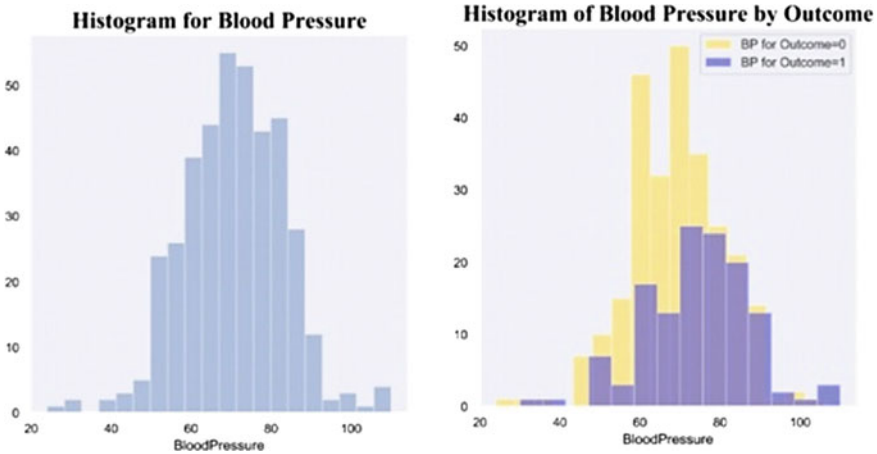
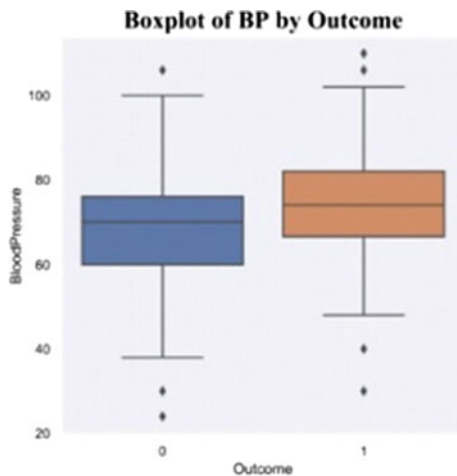


Fig. 5 Bar graph visualization of non-diabetic (0) and diabetic (1) people with respect to blood pressure

Fig. 6 Box-plot visualization of non-diabetic (0) and diabetic (1) people with respect to blood pressure



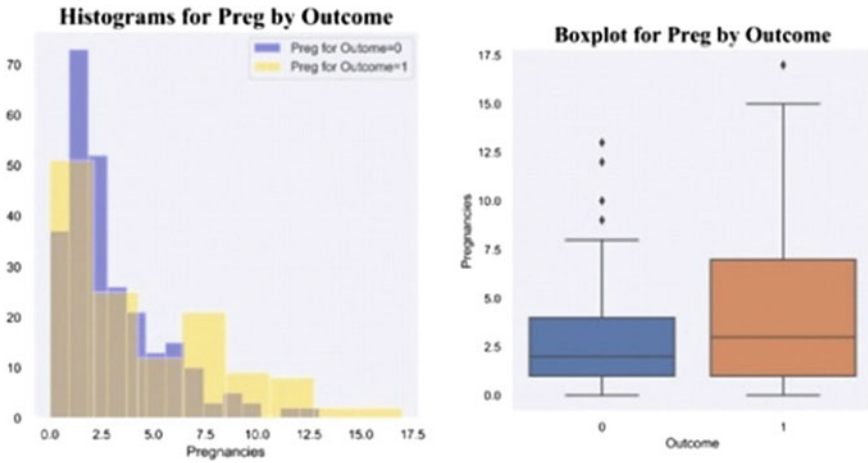


Fig. 7 Bar graph visualization of non-diabetic (0) and diabetic (1) people with respect to pregnancy status

that is the maximum value of BP for non-diabetic people is 100 but it is at 110 with few outliers in case of diabetic people. Secondly around 75% diabetic people have BP in the range of 75–85 mmHg, whereas in case of non-diabetic people, around 75% have a BP range of 70–75 mmHg. Both the plots indicate strong association between the two parameters.

Pregnancy: In the observational data, we found pregnancy status ‘1’ and to calculate correlation between pregnancy and our outcome variables (0 and 1), statistical tools were applied. In graph I (Fig. 7), we observe a peak at average of 7.5 pregnancies among the diabetic patients which indicates low correlation between the two. Again in graph II, both the plots (non-diabetic and diabetic) have medians in lower values, thereby signalling weak correlation among the two features.

Age of people: The age is more correlated to the outcome variables (0 or 1). In the below graph I (Fig. 8), it is clear that people within the age group of 20 or late 20s are non-diabetic, whereas after the age of 30 and above, people are prone to diabetes. Again in graph II, it is depicted that for non-diabetic people median is at 25 years but for diabetic people median is at 32 years. Both the graphs signalled substantial relation between the two features.

Body mass index (BMI): The body mass index [31] is a person’s weight in kilograms (kg) divided by his/her height in metres (m). The National Institutes of Health (NIH) has adopted BMI as a parameter to define normal weight, overweight and obesity of people rather than the traditional height/weight charts. The histograms in Fig. 9 show the BMI level of all the people irrespective of their age groups (graph I). While in graph II, for non-diabetic people, BMI level peak is from 28 to 32 but for diabetic people, it is skewed with peak at 36. Again in box-plot of BMI (Fig. 10), we observe that more than 50% diabetic people have high BMI level as against non-diabetic

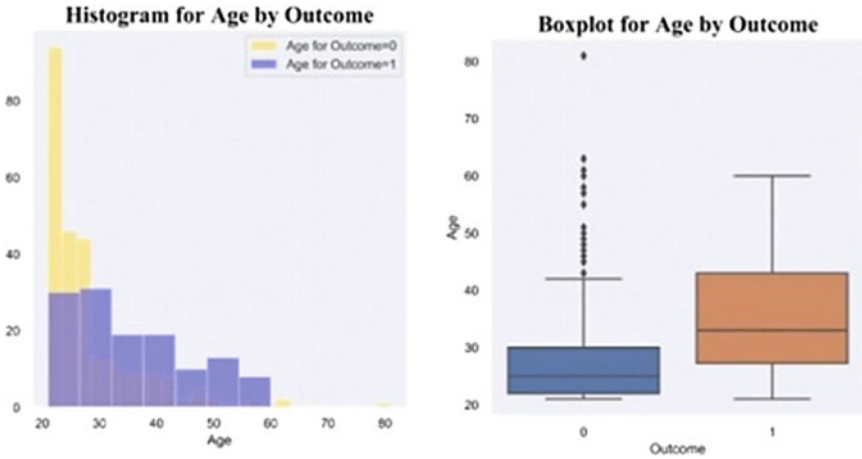


Fig. 8 Bar graph visualization of non-diabetic (0) and diabetic (1) people with respect to age of people

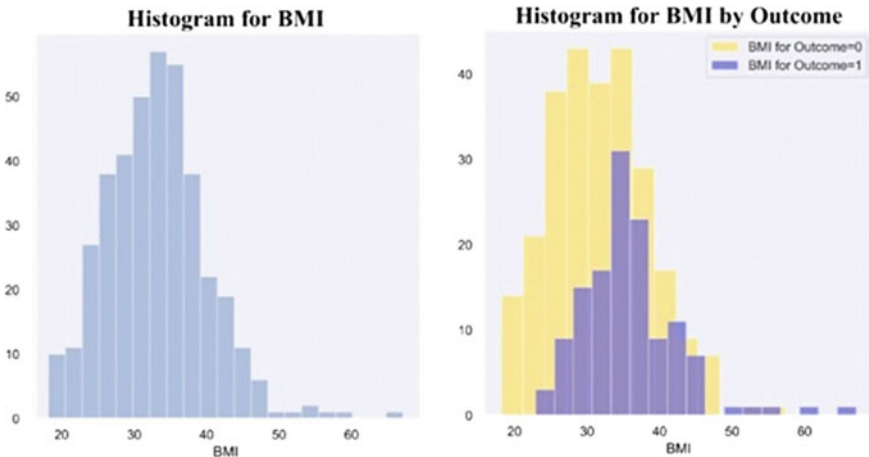


Fig. 9 Bar graph visualization of non-diabetic (0) and diabetic (1) people with respect to BMI

people. Moreover for diabetic people, some outliers are beyond maxima whereas it is not the case with non-diabetic people. Thus, a strong correlation between BMI and diabetes is indicated.

Insulin: The insulin is the very important factor to regulate plasma glucose and within moderate levels it needs to be present in human body [32]. The graph I (Fig. 11) clearly shows peak at 100 for non-diabetic people, whereas the peak shifts to 200 for diabetic patients. The box-plot (Fig. 12) shows the median for diabetic people at 180, whereas it is at 100 in case of non-diabetic people. Secondly, the maximum value

Fig. 10 Box-plot visualization of non-diabetic (0) and diabetic (1) people with respect to BMI

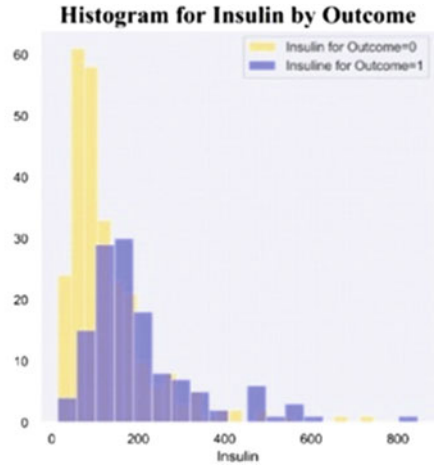
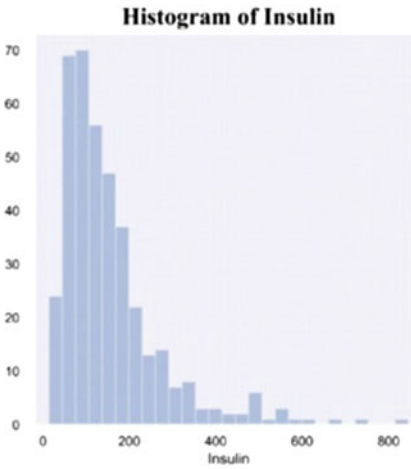
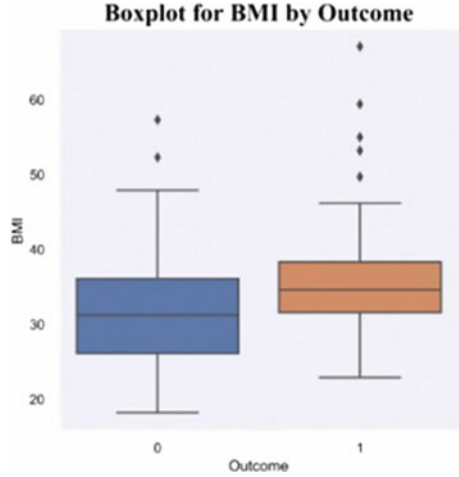
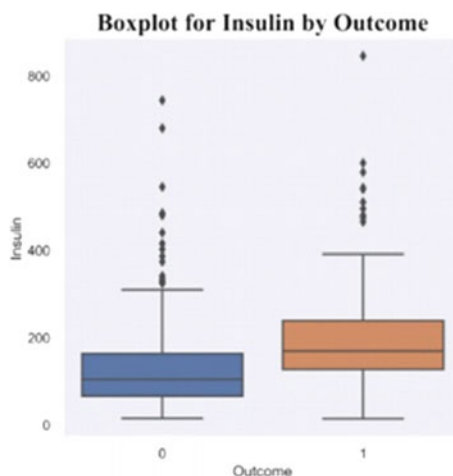


Fig. 11 Bar graphs of non-diabetic (0) and diabetic (1) people with respect to insulin

with many outliers is at 400 for diabetic people but it lies at 300 with few outliers for non-diabetic people. A strong correlation between the two parameters cannot be ruled out.

Fig. 12 Box-plot visualization of non-diabetic (0) and diabetic (1) people with respect to insulin



4 Result Analysis and Discussion

From the above results and by applying the thumb rule of correlation, we can infer that BMI, BP, age and insulin levels of people showed strong correlation with diabetes, whereas pregnancy status have signalled moderate correlation with diabetes.

5 Conclusion

The healthcare sector has already witnessed many revolutionary changes in the recent years and especially big data health analytics has opened doors for new predictive systems. Working on the same lines, this paper tries to explore the usage of artificial intelligence in health data. We have found strong correlation of diabetes with BP, BMI, age factor and insulin levels of humans. Although we may have heard of influence of these different features on diabetes from our doctors in our lives but getting these inferences matched through analytical tools is the sole purpose of this paper. Through this paper, we successfully visualize the correlations of features accountable for diabetes.

References

1. Fred, D.: Healthcare artificial intelligence market CAGR of 47–50%. Networking news. <https://hitinfrastructure.com/news/reports-healthcare-artificial-intelligence-market-cagr-of-47-50> (2019)
2. Clare: The Global virtual healthcare market. Reportlinker. <https://www.prnewswire.com/news-releases> (2020)

3. Laney, D.: 3D data management: controlling data volume, velocity and variety, Application delivery strategies. META Group Inc., Stanford (2001)
4. Mauro, A.D., Greco, M., Grimaldi, M.: A formal definition of big data based on its essential features. *Library Rev.* **65**(3), 122–135 (2016)
5. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. *Commun. ACM* **51**(1), 107–113 (2008)
6. Nguyen, T.L., Thi Thu, H.D.: Artificial intelligence in healthcare: a new technology benefit for both patients and doctors. In: *The Proceedings of 2019 Portland International Conference on Management of Engineering and Technology (PICMET)*, IEEE Xplore, No. 19185164 (2019). <https://doi.org/10.23919/picmet.2019.8893884>
7. Buch, V.H., Ahmed, I., Maruthappu, M.: Artificial intelligence in medicine: current trends and future possibilities. *Brit. J. General Pract.* **68**(668), 143–144 (2018). <https://doi.org/10.3399/bjgp18X695213>
8. Rong, G., Mendez, A., Assi, E.B., Zhao, B., Sawan, M.: Artificial intelligence in healthcare: review and prediction case studies. *J. Eng. Elsevier* **6**(3), 291–301 (2020). <https://doi.org/10.1016/j.eng.2019.08.015>
9. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. *Health Inf. Syst.* **2**(3) (2014)
10. Waring, J., Lindvall, C., Umeton, R., Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *J. Artif. Intell. Med. Elsevier* **104** (2020). <https://doi.org/10.1016/j.artmed.2020.101822>
11. Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., Seroussi, B.: Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. *J. Artif. Intell. Med. Elsevier* **94** (2019). <https://doi.org/10.1016/j.artmed.2019.01.001>
12. Shameer, K.: Traditional bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Brief Bioinform.* **18**(1), 105–124 (2017)
13. Nasi, G., Cucciniello, M., Guerrazzi, C.: The role of mobile technologies in health care processes: the case of cancer supportive care. *J. Med. Internet Res.* **17**(2) (2015)
14. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Hao, L., Sufeng, M., Wang, Y., Dong, Q., Shen, H., Wang, Y.: Artificial intelligence in healthcare: past, present and future. *Stroke Vascular Neurol.* E:000101 (2017). <https://doi.org/10.1136/svn-2017-000101>
15. Jha, S., Topol, E.J.: Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA* **316**, 2353–2354 (2016)
16. Neil, D.B.: Using artificial intelligence to improve hospital in-patient care. *IEEE Intell. Syst.* **28**, 92–95 (2013)
17. Reisman, M.: EHRs: the challenge of making electronic data usable and interoperable. *J. Pharma Theory* **42**(9), 572–575 (2017)
18. Doyle-Lindrud, S.: The evolution of electronic health record. *Clin. J. Nurs.* **19**(2), 153–154 (2015)
19. <https://allofus.nih.gov/>
20. Lucci, S., Kopec, D.: *Artificial Intelligence in the 21st Century*. Sterling Stylus Publishing, LLC (2015)
21. Pan, Y.: Artificial intelligence 2.0: theories and applications. Special issue of *Front. Inf. Techno Electron. Eng.* **19**(1), 1–2 (2018)
22. Murdoch, T.B., Detsky, A.S.: The inevitable application of big data to health care. *J. Appl. Med. Appl. (JAMA)*. **309**, 1351–1352 (2013)
23. Dilsizian, S.E., Siegel, E.L.: Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current Cardio Reports* **16**, 441–446 (2014)
24. <https://www.kaggle.com/uciml/pima-indians-diabetes-database/>
25. Zaharia, M.: Apache Spark: a unified engine for big data processing. *Commun. ACM* **59**(11), 56–65 (2016)
26. Gerard, B.: Analysis of random forests model. *J. Mach. Learn. Res.* **13**, 1063–1095 (2012)

27. Pataky, T.C.: Power1D: a Python toolbox for numerical power estimates in experiments involving one-dimensional continua. *J. Comput. Sci.* **3**(1), e125 (2017). <https://doi.org/10.7717/peerj-cs.125>
28. Lutz, M.: *Programming Python*. O'Reilly Media, 4th edn. (2011)
29. Dianjianyi, S., Zhou, T., Heianza, Y., Xiang, L., Mengyu, F., Vivian, A.F., Qi, L.: Type 2 diabetes and hypertension: a study of bidirectional causality. *Circul. Res.* **124**, 930–937 (2019). <https://doi.org/10.1161/CIRCRESAHA.118.314487>
30. Tsimihodimos, V., Gonzalez-Villalpando, C., Meigs James, B., Ferrannini, E.: Hypertension and diabetes mellitus: co-prediction and time trajectories. *J. Hypertension* **71**, 422–428 (2018). <https://doi.org/10.1161/HYPERTENSIONAHA.117.10546>
31. Nuttall, F.Q.: Body mass index: obesity and health—a critical review. *J. Nutrition Today* **50**(3), 117–128 (2015)
32. Accilli, D.: Insulin action research and the future of diabetes treatment. At American diabetes association's 77th scientific sessions in San Diego. **67**(9), 1701–1709 (2018). <https://doi.org/10.2337/dbi18-0025>

An Evaluation of Deep Learning Networks to Extract Emotions from Yelp Reviews



Yasser Chuttur and Leevesh Pokhun

Abstract User reviews are increasing exponentially in the online world and they play a crucial role in the decision-making process of buying products or services. A lot of thoughts go by a customer mind when it comes to making online purchases. While reviews are important for customers, businesses can also derive various advantages from the data generated by its user base. They can leverage user opinions, sentiment and emotion to gain customer insights. While several successful studies have already been conducted in analysing sentiments expressed in documents, we find that latest trends in machine learning, specifically deep learning, are still not well studied when it comes to emotion detection. We therefore, present here a study to evaluate the application of two deep learning networks namely CNN and LSTM in identifying emotions in yelp reviews.

1 Introduction

In this ubiquitous and technological evolved world, it is a necessity to check for a product or service online before making the decision to purchase it [1]. Each manufacturer makes it a must to have their products full specifications available online for their customers and these customers usually post reviews after having made the purchase to express their satisfaction, dissatisfaction, or their experience with others, who are in the search of similar product or service. As customers are always in the search of the optimum experience, these reviews are some sort conviction whether it is worth investing their time and money [2]. In the online world filled with scam, reviews are among the product attributes that can reassure a potential customer that they are stepping into the right direction.

Y. Chuttur (✉) · L. Pokhun
Software and Information Systems Department, University of Mauritius, Reduit, Mauritius
e-mail: y.chuttur@uom.ac.mu

L. Pokhun
e-mail: leevesh.pokhun1@umail.uom.ac.mu

These reviews are not only helpful for online customers, but also for the business. For an enterprise, it is important to know what their customer base are expecting from them or how they feel about the purchase. Despite having customers posting thousand reviews, it is not feasible to read each one of them, and as time goes by, the review count keeps increasing. So, to that end, it is important for enterprises to only extract meaningful information such as opinions, sentiments, and emotions out of those reviews. Sentiment analysis, also known as opinion mining, focuses on classifying text into three main categories, namely positive, negative, and neutral, whereas emotion analysis seeks to capture the emotion expressed in a message [3].

Yelp¹ is a worldwide business directory service and review site with social networking features. It allows users and overall public to give their ratings and review businesses that they have recently experienced. In this paper, a deep learning approach is taken towards emotion classification in the Yelp dataset. Novel techniques such as convolutional neural network and long short-term memory are used to assess their effectiveness in detecting emotions. The rest of the paper is organized as follows. In section two, a brief overview of CNN and LSTM as deep learning classes is presented along with evaluation metrics used for deep learning applications. Related works at emotion extraction are discussed in Sect. 3 followed by a description of the emotion classification model built for this study in section four. Section 5 details the implementation of the classification model and presents the performance metrics obtained when applying different optimized versions of CNN and LSTM. We then present our conclusions following a discussion of the results obtained.

2 Deep Learning

Deep learning is a subfield of artificial intelligence, which mimics the thinking capabilities of the brain with layers in neurons in the neocortex. Deep learning studies hierarchical structures and levels of representation and abstraction to comprehend the patterns of data that come from diverse source, for instance images, videos, sound, and text.

In computer science, deep learning is most known for its application in areas such as facial recognition, optical character recognition, text to speech or image caption and also in natural language processing and for which several classes of data processing networks have been proposed [4]. Each network consists of one input and one output layer with one or more hidden layers. Each hidden layer acts as a container of *neurons*, whose purpose is to calculate the weighted sum of inputs and weights, add any bias, and execute an activation function so as to fulfil a learning process. The two deep learning classes evaluated in this proposal are convolutional neural network (CNN) and long short-term memory (LSTM).

¹<https://www.yelp.com/>.

2.1 Convolutional Neural Network (CNN)

CNNs are a kind of neural network that is used to address problems related to computer vision. It is designed to benefit from the two-dimensional format of images and that surrounding pixels are related. Its architecture depends on feature sharing, that is, each feature map is created from convolution across the same filter. Max pooling is another step introduced in CNN where the network can see larger input [5]. CNN for text processing uses the same idea. The textual data are transformed into one-dimensional vector and is fed as input to the network.

2.2 Long Short-Term Memory (LSTM)

LSTM is an extension of the recurrent neural network (RNN) [5]. LSTM has a memory function that permits it to read and write information. It is managed by three gates, namely *input*, *forget*, and *output* gates. The input gate is influenced by the current input and previous state. The forget gate is tasked to address the issues related to vanishing and exploding gradient. Lastly, the output gate receives the information from the input and forget gate, which then determines the memory cell state. LSTM is known for its performance in natural language processing (NLP) because it addresses issues with time series, that is, it can remember the sequence of texts [6] and its capacity to retain information in its memory [7].

2.3 Word Embeddings

Word embeddings are a method used to obtain words that are semantically related. Words are represented in an n-dimensional vector space and the distance between the word vectors indicates the semantic relatedness of similar words. In NLP, word embeddings are used in tasks such as named entity recognition (NER) and machine translation [8]. GloVe from Stanford University and Word2Vec from Google are the two most recognized embedding models used by researchers. Word embeddings are particularly important in deep learning when processing texts to address problems of synonyms, polysemy, and disambiguation.

2.4 Evaluating Performance of Deep Learning Methods

The performance of emotion analysis techniques is usually evaluated according to accuracy, precision, recall, and F1-score. The formulae for each metric, as used in this study, are as follows:

Accuracy refers to the number correctly classified data on the entire dataset [9].

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TruePositive} + \text{FalsePositive} + \text{FalseNegative} + \text{TrueNegative}} \quad (1)$$

Precision, also referred as confidence in data mining, represents the amount of predicted positive data that are correctly true positive [10].

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}, \quad (2)$$

Recall, also known as sensitivity, represents the amount of true positive data that are correctly predicted as positive [10].

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad (3)$$

F-measure, also known as F1-score, is the weighted combination of precision and recall [11] for the purpose of measuring the classification effectiveness [9].

$$\text{F1} = \frac{2 \times \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

3 Related Works

Emotion is a complex state of mind, which represents a human feeling that can influence corresponding physical and psychological behaviour [12]. Happiness, sadness, love, and hatred are some examples of emotions expressed by human beings. Emotion models, also known as models of emotion, set forth the different criteria to make various emotions expressed by an individual measurable and distinguishable [13]. In particular, five emotion models, namely discrete, appraisal, dimensional, circuit, and componential model have been used in previous studies to determine the state of feeling of an individual [14]. However, a look at studies conducted on emotion analysis reveals that researchers do not necessarily agree on a universal emotion model. It is common to find a study using a combination of emotion model or part of an emotion model, to suit the need for the study conducted.

Su et al. [15], for example, have attempted to detect seven emotion classes: *anger*, *disgust*, *happiness*, *sadness*, *surprise*, *boredom*, and *anxiety*, from text using long short-term memory (LSTM). They extracted semantic word vectors user input for each inputted word using Word2vec. Then they proceed to the extraction of affect bearing words using the Chinese Valence-Arousal Words (CVAW) affective lexicon. An autoencoder is also used to capture bottleneck features of emotional words which is then concatenated with semantic word vector features to generate textual features.

Finally, the seven emotions extraction is performed by LSTM. The LSTM model is built on top a recurrent neural network (RNN), which has 128 hidden nodes. The model was trained on NLPCC-MHMC-TE dataset, which consists of *natural language processing, and Chinese computing* dataset and their manually collected data. They reported an accuracy of 70.66%.

Zhang et al. [16] employed a multi-task convolutional neural network (CNN) in an attempt towards emotion prediction. Compared to other studies where texts are categorized into a single label, the authors attempted multi-class emotion prediction. They first apply label distribution learning which based on a single label text returns the truthiness of that label, that is, based on a lexical database, it detects the affect bearing words in a sentence and calculates the density of each emotion found. Then, the multi-task CNN predicts the different emotions present within a sentence by employing cross-entropy and Kullback–Leibler as optimization function. The CNN model consists of a convolutional layer followed by max pooling and the network is fully connected with dropout. Stochastic gradient descent (SGD) is chosen as backpropagation algorithm. The model was tested on a series of datasets, namely *ISEAR, Fairy Tales, TEC, CBET*. They reported their performance metrics *precision, recall, F-score, and accuracy*. *Fairy Tales* obtained the highest performance across all metrics, *precision* = 78.2%, *recall* = 78.21%, *F1* = 78.72%, *accuracy* = 79.21%.

Chatterjee et al. [17] identified emotion classification as a multi-class problem in real conversations, that is, there not a single sentence involves, but rather multiple. They classified a user input into four emotion classes, happy, sad, angry, and others. They used two LSTM layers to leverage the power of sequential processing. Layer one uses semantic word embedding and the other uses semantic word embedding where the end result of the layers is concatenated and passes to a fully connected network to classify a conversation emotion. They built their own dataset from tweets. They reported using *precision, recall* and *F-score* performance metric for each emotion class. *Angry* obtained the highest *precision* = 87.69%, *sad* obtained the highest *recall* = 76.63%, and *F-score* = 80.79%.

Kratzwald et al. [18] attempted emotion recognition using long short-term memory (LSTM) and transfer learning to detect four emotion classes, anger, fear, joy, sadness. Their algorithm consists of the use of transfer learning, word embeddings, and finally, the LSTM layer. They named their transfer learning method Sent2affect. It basically means that they applied sentiment analysis and the output is then transferred to LSTM and they used pretrained GloVe embedding. The model was experimented on *natural language processing and Chinese computing* (NLPCC) dataset where performance was reported using accuracy where it obtained 70.66%.

Abdullah et al. [19] proposed a framework which uses CNN and LSTM to detect and predict sentiments and emotions in Arabic tweets. The Arabic language faces shortage of research on it due to its complex nature of its dialects. The authors attempted the first emotion and classification by using deep learning to contribute towards Arabic language processing. Four emotion classes are used, *anger, joy, fear* and *sadness*. The emotions are also classified into intensities which range from 0: low emotion to 3: high emotion. Their model first layer consists of a CNN layer, which is followed by an LSTM layer. They used AraVec as embedding model. Evaluation

reported using Spearman correlation where it obtained an overall result of 56.9% across all emotions. Sad obtained the highest score of 61.1%.

Li et al. [7] attempted emotion classification on WeChat articles comments. They propose a basic LSTM model which it used Word2Vec embeddings and followed by the LSTM layers. They reported an accuracy of 25.12%.

He et al. [6] proposed a bi-directional LSTM-CNN framework to detect emotion intensity in texts. Compared to other works, where n emotions are identified for a related text, they claim to give more insights on each emotion intensity being expressed. CNN is used to capture textual features, ignoring the text global information and LSTM is used to obtain the long-distance dependencies by sequentially modelling texts across words. The model also uses rectified linear unit as activation function mean squared error loss function. They used EmoInt-2017 tweet dataset to test their model, where emotion intensities are classified into classes, *anger*, *fear*, *joy* and *sadness*. Performance is reported using Pearson and Spearman correlation coefficient. On average, their model obtained Pearson = 67.7% and Spearman = 66.5% and it was able to classify more sad tweets with Pearson = 70.9% and Spearman = 71.3%.

4 Implementing the Emotion Classification Model for Yelp

Figure 1 provides a general overview of the classification model developed for this study. Yelp reviews are passed as sentences to a pre-processing phase, which reduces sentence dimensionality while keeping useful words only. Two deep learning models are fed with the pre-processed data where each model projects a sentence into an embedding vector matrix, then the features generated are fed to different hidden layers for each of CNN and LSTM until a classification output is obtained for any emotion detected. The main reason why CNN and LSTM were chosen for this study was because previous studies have reported good performance results with those two networks.

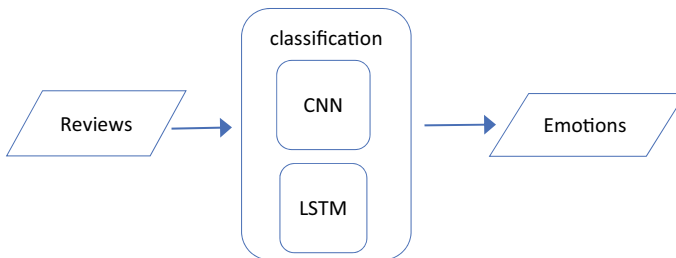


Fig. 1 Emotion classification model overview

Table 1 Number of reviews per emotion class

Emotion class	Review count
Anger	90
Concerned	1
Disappointment	98
DISGUST	20
Fear	8
Happy	281
Neutral	56
Sad	9
Surprise	30
Total	593

4.1 Dataset Used for the Study

The dataset used for this study was obtained from the Yelp dataset challenge.² Yelp encouraged researchers to use their data for creative and breakthrough innovations such as in the field of natural language processing and affective analysis among others. The dataset consists of one million reviews compiled in a CSV file with twelve fields marked as: Business_id, Categories, Stars, User_id, Text, attributes_Alcohol, attributes_Noise.Level, attributes_Parking, and attributes_Price.Range. The “Text” field contained all reviews of customers.

To train our classification model, a new column, “Emotion_class”, was added to the CSV file after manually reviewing 593 reviews. A total of nine emotion classes: *anger*, *concerned*, *disappointment*, *disgust*, *fear*, *happy*, *neutral*, *sad*, and *surprise* were assigned to the reviews based on the text contents [4]. Table 1 lists the number of reviews per emotion class. Emotions classes used for this study were taken from the discrete emotion category [20] as the focus is on core emotions, which are usually relatively easy to be interpreted from text data.

4.2 Data Pre-processing

The NLTK library³ in Python is used to clean the dataset. First the reviews are tokenized, and word level analysis is done. If a word forms part in the name entity relationship (NER) list “*GEO; ORG; PER; GPE; TIM; ART; EVE; NAT*”, they are removed from the sentence.

Then, each word is lemmatized, and the lemmatized output is then passed to the part of speech (POS) tagging to check if after the word level analysis process, the

²<https://www.yelp.com/dataset/challenge>.

³<https://www.nltk.org/>.

Fig. 2 NLTK output

```

>>> nltk.pos_tag(nltk.word_tokenize('loved'))
[('loved', 'VBN')]
>>> nltk.pos_tag(nltk.word_tokenize('love'))
[('love', 'NN')]
>>> nltk.pos_tag(nltk.word_tokenize('feared'))
[('feared', 'VBN')]
>>> nltk.pos_tag(nltk.word_tokenize('fear'))
[('fear', 'NN')]
>>> |

```

word is still a verb. The motivation behind this step is that emotion words are not tagged as verbs in their root form; they are tagged as either nouns or adverbs as shown in Fig. 2. So, the lemmatization process returns the word to its initial form and this can be used to determine if the word is a verb or an emotion word.

An example showing a review before and after being pre-processed is shown below:

Before pre-processing

Mr Hoagie is an institution. Walking in, it does seem like a throwback to 30 years ago, old fashioned menu board, booths out of the 70s, and a large selection of food. Their speciality is the Italian Hoagie, and it is voted the best in the area year after year. I usually order the burger, while the patties are obviously cooked from frozen, all of the other ingredients are very fresh. Overall, its a good alternative to Subway, which is down the road.

After pre-processing

Mr institution walking seem throwback year ago old fashion menu board booths large selection food speciality vote best area year usually order burger patty obviously cook frozen ingredient fresh overall good alternative road.

As seen in the example above, pre-processing aims at keeping only meaningful words. It can also be seen that the sentence dimensionality is considerably reduced (here from 83 to 34 words).

4.3 Deep Learning Models Tested

The effectiveness of emotion extraction in Yelp reviews was tested using convolutional neural networks (CNN) and long short-term memory (LSTM) with GloVe embeddings [21].

The CNN model first represents the sentence into an embedding vector matrix and then employs multiple widths of filters and max-pooling operation on it. The model combines cross-entropy loss for classification with an optimization function.

The features are then concatenated and passed to a hidden layer with dropout and softmax activation.

The LSTM model first layer is an embedding layer followed by a one-dimension convolutional layer. A max-pooling layer is applied and then followed by LSTM layer with hundred RNN cells and two dense layers. Each layer uses rectified linear unit (ReLU) activation function and softmax is used for the last layer.

To obtain the best performance, the model was trained with various combinations of hyper parameters, number of hidden layers and number of neurons per layer.

4.4 Model Training

The model was trained with 90% of the dataset and the remaining 10% records were used as evaluation set. The trained samples were randomly shuffled and divided into mini batches. During the training phase, three optimizers were used: *AdaDelta*, *Adam*, and *stochastic gradient descent (SGD)* for gradient descent calculation.

AdaDelta short for adaptive learning rate can dynamically adapt over time using only first-order information and has minimal computational overhead beyond vanilla stochastic gradient descent [22]. Adam (adaptive momentum) is an algorithm like AdaDelta but it can calculate adaptive learning rates of various parameters from estimates of first and second moments of gradients [23]. Stochastic gradient descent algorithm divides the dataset into mini batches and then computes the average gradient descent over a mini batch [24]. Categorical cross-entropy is used as loss function for all models and trained upon fifty epochs.

5 Model Evaluation

The classification model built for this study was implemented and tested on a PC running Windows 10 with specifications listed in Table 2.

Results obtained for accuracy, F1, precision, and recall are further listed in Tables 3, 4, 5 and 6.

Table 2 PC specifications used in this study

Hardware/Software	Specifications
Processor	Intel(R) Core(TM) i7-4710HQ CPU @ 2.50 GHz (8 CPUs)
Memory	16,384 MB RAM (16 GB) DDR3
GPU	Nvidia GeForce GTX 860 m DDR5 with CUDA version 10.0
Python	3.7 (as development tool)
TensorFlow	2.0 (as development tool)

Table 3 Model evaluation result for accuracy

Models	Accuracy	
	Training	Evaluation
CNN—Adadelata	0.4841	0.3667
CNN—Adam	0.4841	0.3667
CNN—SGD	0.4841	0.3667
LSTM—Adam	1	0.3667
LSTM—Adadelata	0.4841	0.3667
LSTM—SGD	0.8161	0.35

Table 4 Model evaluation result for F1

Models	F1	
	Training	Evaluation
CNN—Adadelata	0.2509	0.2254
CNN—Adam	0.2498	0.2254
CNN—SGD	0.2324	0.2183
LSTM—Adam	1	0.3638
LSTM—Adadelata	0.1314	0.1843
LSTM—SGD	0.809	0.3128

Table 5 Model evaluation result for precision

Models	Precision	
	Training	Evaluation
CNN—Adadelata	0.1436	0.1288
CNN—Adam	0.1428	0.1288
CNN—SGD	0.1315	0.1228
LSTM—Adam	1	0.3638
LSTM—Adadelata	0.3728	0.4487
LSTM—SGD	0.9229	0.3852

Table 6 Model evaluation result for recall

Models	Recall	
	Training	Evaluation
CNN—Adadelata	0.9926	0.9018
CNN—Adam	0.9945	0.9018
CNN—SGD	0.9982	0.9821
LSTM—Adam	1	0.3638
LSTM—Adadelata	0.081	0.1183
LSTM—SGD	0.723	0.2634

Regardless of the optimization approach used, CNN models gave out an accuracy of 48.41% during training and 36.67% during evaluation. LSTM with Adam, *AdaDelta* and *SGD* optimizer obtained an accuracy of 100%, 48.1% and 81.61%, respectively, during training but we observed a relatively low accuracy of around 35–36% almost similar to CNN models during evaluation.

CNN models gave F1 scores between 23 and 25% during training and similarly; F1 score ranging from 21 to 22% was obtained during evaluation. In contrast, during training, LSTM, with Adam optimization gave out an F-score of 100% closely followed by LSTM SGD with an F-score of 80.9% while LSTM Adadelata was far behind with a very low F-score of 13.14%. Surprisingly, a very poor F-score was observed when testing the LSTM models. F-scores obtained were 36.38%, 18.43%, and 31.28% for LSTM Adam, Adadelata, and SGD optimization, respectively.

Precision values between 13 and 14% were obtained for CNN models during training. During evaluation, all CNN gave the same precision values of 12.28%. LSTM with Adam optimization gave a score of 100% during training but gave out a precision value of 36.38%. LSTM SGD gave out a score of 92.29% during training and similar to LSTM Adam gave poor results during evaluation with a precision value of 38.52 value. LSTM Adadelata gave lower precision value (37.28%) during training but a higher precision value during testing (44.87%).

Finally, for recall calculations, we observe that CNN models all gave almost similar recall values of approximately 99% during training and high recall values of around 90% for CNN Adadelata and CNN Adam. The best recall result was obtained for CNN SGD with a recall value of 98.21%.

LSTM Adadelata gave a very low recall value of 8.1%; LSTM SGD a higher recall value of 72.3%, while LSTM Adam gave a recall value of 100%. During evaluation, LSTM Adadelata also gave a low recall value of 11.83%; LSTM SGD a relatively higher recall value of 26.34% and LSTM Adam gave a better recall value of 36.38%.

6 Discussion

Emotion classification is not an easy task as evidenced by results obtained in this study. Other than for LSTM SGD model and CNN models for recall values, in general, training and evaluation results obtained using the two popular deep learning models, CNN and LSTM, did not provide good performance as already reported in previous studies. Such observations raised multiple questions on our side, which we list below.

- Is CNN or LSTM really effective in detecting emotions in texts?
- Are individual Yelp reviews too short for processing by deep learning models?
- Could the low performance values come from the size of Yelp reviews used for training and evaluation?
- How can we modify CNN and LSTM such that they may be optimized for processing small data size?

- Given that each review was manually assigned emotions, could the subjectivity involved be a factor in the poor performance results obtained?

In the future, we intend to address the above questions by increasing the sample size used in our classification models. We also recommend that any effort in building a corpus for emotion classification must take into account the existence of multiple emotions in a dataset and to label the underlying data accordingly. Our dataset also suffers from imbalance of emotions as shown in Table 1. With few labelled emotions, the classification model eventually could not learn each emotion class from the dataset used.

7 Conclusion

LSTM and CNN are two widely used deep learning networks in computer science. Successful applications of LSTM and CNN have been reported in previous studies. When applied to classification of emotions in user reviews, we, however, did not obtain similar performances as reported in previous studies. We are aware that deep learning networks like CNN and LSTM do not have good performance with small size datasets—but how small is small? We are tempted to believe that the larger the dataset, the better the performance will be. This, in summary, highlights a major drawback of deep learning models. In the future, we may expect to find better performance learning models, which can learn quicker and with smaller dataset.

References

1. Sun, X., Han, M., Feng, J.: Helpfulness of online reviews: examining review informativeness and classification thresholds by search products and experience products. *Decis. Support Syst.* **124**, 113099 (2019). <https://doi.org/10.1016/j.dss.2019.113099>
2. Nikolay, A., Anindya, G., Panagiotis, G.L.: Deriving the pricing power of product features by mining consumer reviews. *Manage. Sci.* **57**, 1485–1509 (2011). <https://doi.org/10.1287/mnsc.1110.1370>
3. Yam, C.: Emotion Detection and Recognition from Text Using Deep Learning, <https://www.microsoft.com/developerblog/2015/11/29/emotion-detection-and-recognition-from-text-using-deep-learning/>. Accessed on 31 Jan 2019
4. Di, W., Bhardwaj, A., Wei, J.: *Deep learning essentials: your hands-on guide to the fundamentals of deep learning and neural network modeling*. Packt Publishing Ltd (2018)
5. Hadji, I., Wildes, R.P.: *What Do We Understand About Convolutional Networks?* (2018)
6. He, Y., Yu, L., Lai, K.R., Liu, W.: YZU-NLP at EmoInt-2017: Determining Emotion Intensity Using a Bi-directional LSTM-CNN Model, pp. 238–242 (2017)
7. Li, P., Wang, P.: Short Text Emotion Analysis Based on Recurrent Neural Network, pp. 1–5 (2017). <https://doi.org/10.1145/3078564.3078569>
8. Seyeditabari, A., Tabari, N., Zadrozny, W.: *Emotion Detection in Text: A Review* (2018)
9. Powers, D.M.W.: What the F-measure doesn't measure (2014). <https://doi.org/10.13140/RG.2.1.1571.5369>

10. Powers, D.M.W.: Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation (2007)
11. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction BT. In: Proceedings of DARPA Broadcast News Workshop, pp. 249–252 (1999)
12. Bull, N.: The attitude theory of emotion. Johnson Reprint (1951)
13. Scherer, K.R.: What are emotions? and how can they be measured? *Soc. Sci. Inf.* **44**, 695–729 (2005). <https://doi.org/10.1177/0539018405058216>
14. Kim, K.: Emotion Modeling and Machine Learning in Affective Computing, pp. 1–12 (2014)
15. Su, M.H., Wu, C.H., Huang, K.Y., Hong, Q.B.: LSTM-based text emotion recognition using semantic and emotional word vectors. In: 2018 1st Asian Conference and Affective Computing Intelligent Interaction, pp. 1–6. ACII, Asia (2018). <https://doi.org/10.1109/ACIIAsia.2018.8470378>
16. Zhang, Y., Fu, J., She, D., Zhang, Y., Wang, S., Yang, J.: Text emotion distribution learning via multi-task convolutional neural network. *IJCAI International Joint Conference on Artificial Intelligence*, July 2018, pp. 4595–4601 (2018)
17. Chatterjee, A., Gupta, U., Chinnakotla, M.K., Srikanth, R., Galley, M., Agrawal, P.: Understanding emotions in text using deep learning and big data. *Comput. Human Behav.* **93**, 309–317 (2019). <https://doi.org/10.1016/j.chb.2018.12.029>
18. Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., Prendinger, H.: Deep learning for affective computing: text-based emotion recognition in decision support. *Decis. Support Syst.* **115**, 24–35 (2018). <https://doi.org/10.1016/j.dss.2018.09.002>
19. Abdullah, M., Hadzikadicy, M., Shaikhz, S.: SEDAT: sentiment and emotion detection in Arabic text using CNN-LSTM deep learning. In: Proceedings of 17th IEEE International Conference on Machine Learning Applications, ICMLA 2018, pp. 835–840 (2019). <https://doi.org/10.1109/ICMLA.2018.00134>
20. Barrett, L.F.: To Understanding variability in emotion. *Cogn. Emot.* **23**, 1284–1306 (2009). <https://doi.org/10.1080/02699930902985894>. *Variety*
21. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. *EMNLP 2014—2014 Conference on Empirical Methods in Natural Language Processing, Proceedings Conference*, pp. 1532–1543 (2014). <https://doi.org/10.3115/v1/d14-1162>
22. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method (2012)
23. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representation, ICLR 2015—Conference Track Proceedings, pp. 1–15 (2015)
24. Chaudhari, P., Soatto, S.: Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In: 2018 Information Theory and Applications Work. ITA 2018, pp. 1–20 (2018). <https://doi.org/10.1109/ITA.2018.8503224>

Classification of Brain Tumor MRIs Using Deep Learning and Data Augmentation



Gulshansingh Bhagbut and Zahra Mungloo-Dilmohamud

Abstract Brain tumor identification and classification is crucial in everyday life. This paper focuses on a four-class classification problem to differentiate between three prominent types of brain tumor namely glioma, meningioma, pituitary tumors and no tumor. The proposed system uses deep transfer learning and two pre-trained and one custom model to classify these brain MRI images. The empirical work is performed using a custom dataset made from existing public datasets. The proposed system registers a classification accuracy of up to 99%. Performance measures such as precision, recall and F-score have also been calculated. Moreover, since the dataset is not so big, the results show that transfer learning is a useful technique when the availability of medical images is limited.

Keywords Deep learning · CNN · Brain tumor · Classification · Data augmentation · Pre-processing

1 Introduction

Accurate diagnosis and identification of brain tumors, in a timely manner, is crucial in the selection of the best treatment plan for patients. Brain tumors are abnormal cell or tissue growth in the brain and have more than 120 types [35]. A brain scan is necessary to identify the presence of a brain tumor and to know its exact location. The proper detection greatly relies on the image, the knowledge, experience and availability of radiologists and neurologists. The choice of a treatment methodology also relies upon a number of factors. Computer aided diagnosis (CAD) procedures have been helping medical practitioners for various tasks, including tumor detection, classification and severity detection.

G. Bhagbut · Z. Mungloo-Dilmohamud (✉)
FoICDT, University of Mauritius, Moka, Mauritius
e-mail: z.mungloo@uom.ac.mu

G. Bhagbut
e-mail: gulshansingh.bhagbut@umail.uom.ac.mu

Deep learning, a subcategory of machine learning, has been used extensively in the medical field. It learns by using multiple layers to gradually extract higher level features from the raw input [16]. One of the areas where deep learning excels is image classification [19]. Hence, deep learning strategies have been used in medical image analysis in oncology for oral cancer detection [13], breast cancer studies [31], lung cancer diagnosis [8], classifying liver cancer as malignant or benign [15], brain metastases and brain tumor detection [3, 6], skin cancer classification between different types [7], diagnosis outcome of esophageal cancer [10] and identification of renal tumors [32]. Deep transfer learning, a more recent technique, is also used in image analysis as instead of starting the learning process from scratch, the system starts from patterns that have been learned when solving a different problem and is now applied to a new field.

In this work, three different deep transfer learning models have been tested to classify brain tumor MRI images. The proposed solution makes use of several data augmentation and image pre-processing techniques to reinforce the data and therefore generate more robust models. The project makes use of a custom dataset, built from publicly available datasets as described in Sect. 2.4.

This paper proceeds as follows. Section 2 presents the relevant topics, existing brain tumor datasets and the experimental framework. Section 3 presents the results obtained using the different DL models tested as well as an evaluation of the results. Finally, Sect. 4 concludes the paper and presents the future work.

2 Materials and Methods

2.1 Brain Tumor Datasets

A number of brain tumor datasets are available for download online. These datasets differ in the number of images, the resolution of images, the size of images, the different types of brain tumors that are represented by the images, the number in each class as well as the views in which the images have been acquired. The details of some of the most popular datasets are provided in Table 1, where a stands for axial, c for coronal and s for sagittal.

2.2 CNN Models and Brain Tumor

Convolutional neural network (CNN) is a type of deep neural network mainly used for computer vision tasks. Deep learning has streamlined the workflow by incorporating an automated feature engineering approach unlike the traditional machine learning approach where features had to be extracted manually.

Table 1 Existing brain tumor datasets

Dataset name	Number of images	Size of image	Number of classes	Classes	Views	Source
BraTS	Variable	256 × 256	2	High-grade and low-grade glioma	a, c and s	[2]
Figshare brain tumor	3064	512 × 512	3	Meningioma, glioma and pituitary tumors	a, c and s	[6]
Kaggle brain tumor dataset	253	Variable	2	Tumor and non-tumor	a	[34]
The cancer imaging archive (TCIA)	Variable	Variable		Depends on the selected dataset	Variable	[33]
Brain tumor MRI image data collection	96,115	64 × 64	5	Astrocytoma, glioblastoma, oligodendroglioma Unidentified tumors Healthy brain	all	[1]
MIRIAD	708	124 × 256	1	Healthy brain	s	[37]
IXI	600	150 × 256	1	Healthy brain	c	[36]
NFBS		192 × 256	1	Healthy brain	c	[38]
MIDAS	110	170 × 256	1	Healthy brain	a	[38]

CNNs are made up of three main types of layers: Convolutional, pooling and fully connected (dense) layers as given in Fig. 1. Convolutional layers are used to achieve the mentioned automatic feature extraction. During this operation, a set of filters

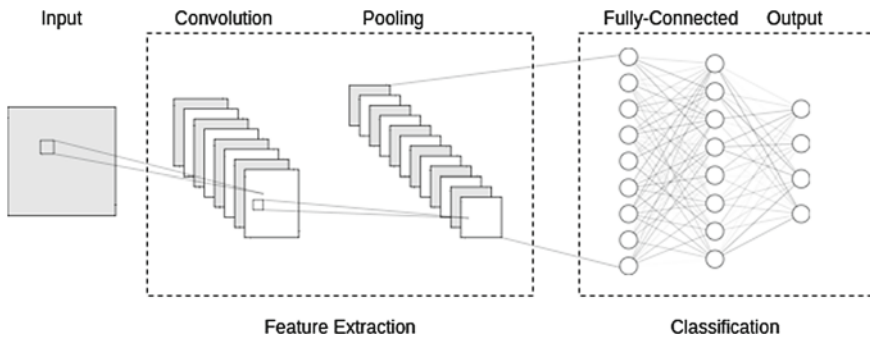


Fig. 1 Different layers in a CNN

Table 2 Some CNN models

Model	Number of parameters (in Millions)	Layers	Top-5 error rate
VGG16 [24]	138	16	8.0
VGG19 [25]	138	19	7.5
ResNet-50 v2 [9]	23	50	5.7
ResNet-152 [9]	60	152	3.6
InceptionV1/GoogleNet [27]	6.7	27	6.7
InceptionV3 [28]	23.6	48	3.5
MobileNet [11]	4.2	28	10.5
MobileNetV2 [22]	3.4	53	9.9
DenseNet-121 [12]	8	121	5.17
Xception [5]	22.8	71	0.06/3.5
AlexNet [14]	60	8	15.3

are convoluted on the image, while performing a dot product between the filter and the area covered to detect and learn patterns such as edges, lines and shapes. The principal task of pooling layers is to reduce the spatial dimension of feature maps, thus reducing the computational power needed. Pooling also reduces the number of network parameters. Dense layers are responsible for making classifications based on the previous learned features [1].

A survey of literature shows that there are many CNN models that are available for use for image recognition when implementing a CAD system and Table 2 provides a comparison of some of these.

As mentioned above, CNN models have been used extensively in the field of image classification in the medical field. Some prior work has also already been carried for brain tumor detection as given in Table 3.

2.3 Deep Transfer Learning

Deep transfer learning is a deep learning method where a model, built for a specific task, is reused as the base for another model performing a different task [29]. Many state-of-the-art models currently exist and may be repurposed for other tasks, in other domains, such as in the medical field. Transfer learning is quite popular since it allows us to build models quicker and with good accuracy. Moreover, hardware requirements, for training, are not as high as when training a model from scratch. A convolutional neural network (CNN) architecture can be categorized into two parts: the convolutional base and the classifier as given in the diagram below. Prior to using an existing model, several factors, such as the available computational power and the size and similarity of the dataset, should be considered as these will determine the

Table 3 Some of the most recent work in the field of brain tumor detection using CNN

Author/Year	Title	Dataset	Performance
Deepak et al. [6]	Brain tumor classification using deep CNN features via transfer learning	Figshare brain tumor dataset	Accuracy: 97.8%
Swati et al. [26]	Brain tumor classification for MR images using transfer learning and fine-tuning	Figshare brain tumor dataset	Accuracy: 94.8% F1 score: 94.47
Saxena et al. [23]	Predictive modeling of brain tumor: a deep learning approach	Kaggle brain tumor dataset	Accuracy: 95% F1 score: 0.952
Toğaçar et al. [30]	BrainMRNet: brain tumor detection using magnetic resonance images with a novel convolutional neural network model	Kaggle brain tumor dataset	Accuracy: 96.1% F1 score: 94.12
Balasoorya et al. [1]	A sophisticated convolutional neural network model for brain tumor classification	REMBRANDT + MIRIAD + BRAINS	Accuracy: 97.44% F1 score: 99.46%

best approach to repurpose the model. Figure 2 illustrates the three main approaches which can be used.

According to Pattanayak [18], the first approach which is to train the entire model from scratch is suitable when a large dataset is available. The dataset is also different from the pre-trained model's dataset and computational power is not an issue. The second approach is to freeze some layers in the convolutional base and train the rest of the layers. If a large and similar dataset is used, only a few layers need to be trained. However, if a small and different dataset is used, it may be better to freeze less layers. The third approach is to freeze the entire convolutional base and train the model using a new classifier. This is suitable only if the dataset is similar and small and when computational power is limited. For example, in the context of this project, this approach will not be suitable since most existing models are trained using ImageNet's dataset and not on brain tumors specifically.

2.4 Empirical Work

A. Datasets and pre-processing

For the empirical work, the brain tumor dataset was downloaded from Figshare [4] where it is openly available. The latter consists of three classes of brain tumors: Glioma, meningioma and pituitary tumors.

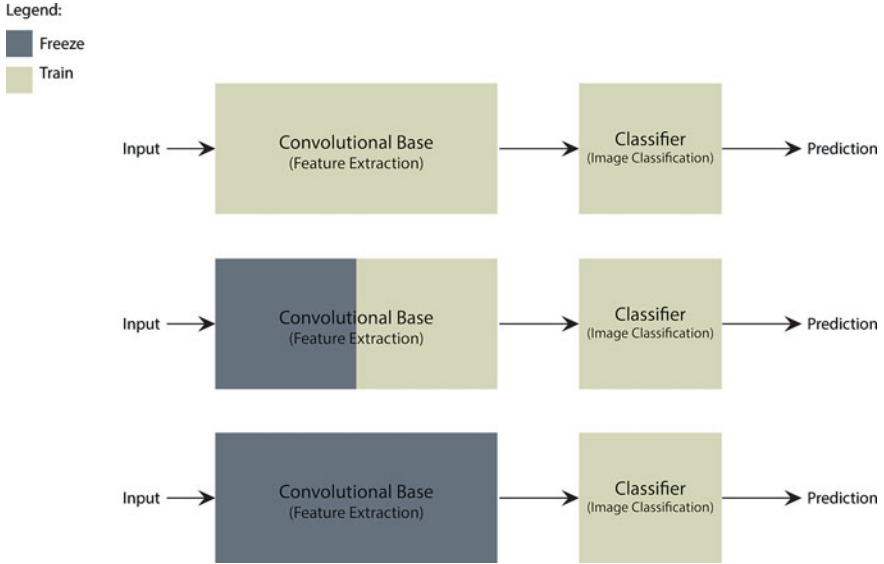


Fig. 2 Three main approaches to transfer learning

For healthy brain images, a combination of three datasets were used to make sure that all three planes (axial, sagittal and coronal) are present, thus matching the Figshare dataset. Images from IXI, MIRIAD [17] and the CASILab healthy brain MRI dataset were extracted and combined with the Figshare dataset to create a new four-class dataset.

It is also to note that only MRIs of healthy patients were taken from the MIRIAD dataset as it also consists of MRIs of patients suffering from Alzheimer’s disease. The images of the Figshare dataset are in .mat files, and as seen from Table 1, the size of each image is 512×512 . The MRI images in the new dataset were augmented and pre-processed as given in Fig. 2 and as detailed below.

B. Data augmentation

High quality and abundant data are crucial to the successful implementation of different deep learning models [20]. In the context of deep learning, the volume and quality of the data is as important as the algorithm, and therefore, data augmentation is often used. Data augmentation refers to the application of one or more deformations to annotated dataset which results in new, additional training data [21]. Data augmentation can help in resolving data imbalance and can increase the overall accuracy of a model. Therefore, an initial step is data augmentation.

The data augmentation phase involved the application of several image deformations as described in Tables 4 and 5. However, these image augmentation techniques

Table 4 Summary of applied data augmentation techniques

#	Technique	Parameters
1	Image flip	Horizontal
2	Random rotation	Range: 18°
3	Random translation	Range: -10 to 10 pixels
4	Shear: right	Range: -10 to 10 pixels
5	Shear: bottom	Range: -10 to 10 pixels

Table 5 Overview of dataset after data augmentation

Tumor class	Initial size	After augmentation
Glioma	1426	7126
Meningioma	708	7126
Pituitary tumor	930	7435
No tumor	5687	7444

were not applied individually, a random combination of the aforementioned deformations were applied on each image with the objective of generating different images each time.

C. Data pre-processing

Once the images were augmented, they were then pre-processed as given in Fig. 3. The objective of this phase is to apply several image processing algorithms to improve the overall quality and resize images to the required dimensions.

First each image is converted to grayscale, which reduces it to a single channel image. It is then resized to a square image of size 224 × 224. The resizing operation preserves the aspect ratio of the original image by adding the necessary padding. Contrast limited adaptive histogram equalization (CLAHE) is then applied to the resized image, followed by non-local means denoising and an unsharp mask. The image is then transformed to a three-channel image by replicating the image three more times, since the CNNs used in this study have been initially designed to accept RGB images.

D. Dataset splitting

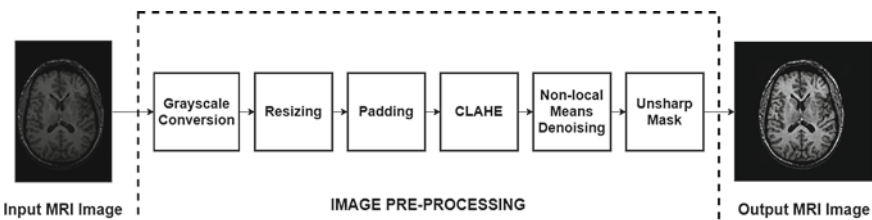


Fig. 3 Pre-processing steps of the images

The dataset splitting phase is done to split the augmented and pre-processed dataset into three parts: training, validation and testing with percentages 70%, 15% and 15%, respectively. This leaves the training size to 20,356 and the validation and testing data with 4362 samples.

E. CNN architecture and selected models

Three (3) different models have been used for this study, ResNet-50 v2, DenseNet-121 and our own model. These two existing architectures have been pre-trained on ImageNet’s dataset and have been chosen mainly for their relatively low-parameter count, compared to other models. Choosing a low-parameter model may help to counteract overfitting during training since the size of our dataset is relatively small, compared to ImageNet.

For the experimental work, we also designed a custom CNN architecture for comparison purposes. The model has 600,392 parameters, 599,510 of which are trainable (Fig. 4).

F. Classifier and training parameters

For the transfer learning approach, only the feature extractors of ResNet and DenseNet were used. A new classifier was then developed to be used on top of the mentioned models. It is composed of a Global Average Pooling layer, followed by two dense layers of 128 and 256 units, respectively, with ReLU activation function and a final four-way dense layer with a softmax activation function.

Regarding our custom CNN model, the classifier consists of a global average pooling layer, followed by two dense layers of 256 and 512 units respectively with ReLU activation function. A dropout of 0.5 was also added to help with the issue of overfitting, and finally, a dense four-way layer with softmax activation is present at the end to make predictions. Figure 5 illustrates the classifiers that were used and Table 6 provides an overview of the parameters used for training of the models.

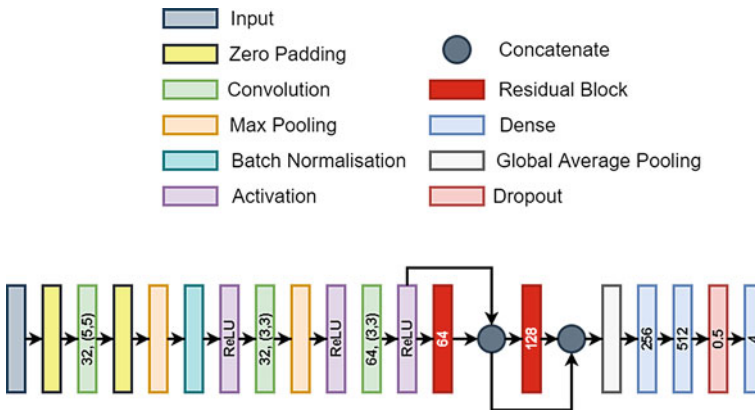


Fig. 4 Custom architecture diagram

Fig. 5 Two classifiers used
(Left: ResNet and DenseNet,
Right: custom model)

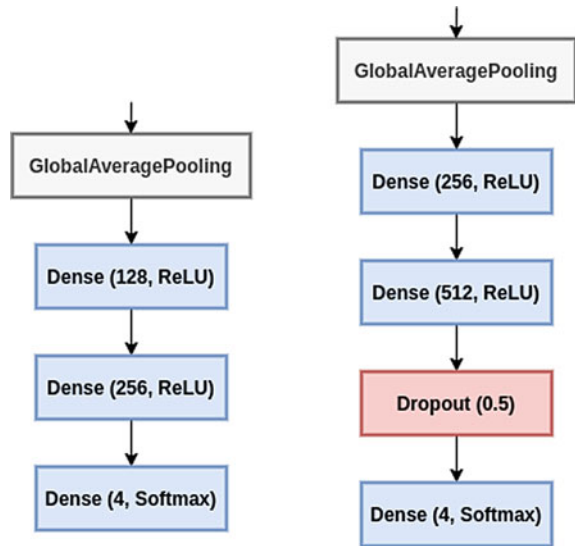


Table 6 Overview of training parameters

Model	Parameter	Value
Resnet50 v2 Densenet121 Custom model	Optimizer	SGD
	Loss function	Sparse
	Batch size	50
	Learning rate	0.0001
	Decay	1e-6
	Momentum	0.9
	Layers trained	100%

G. Implementation Environment

Python programming language along with various libraries such as Keras with Tensorflow as backend and scikit-learn has been used for the implementation of the system. Data augmentation and image pre-processing were achieved using OpenCV. Matplotlib has been used for generating graphs and confusion matrices.

The development and training process was extensively done on Google's Colab platform as the latter provides access to powerful hardware components such as an Intel Xeon CPU, NVidia P100 or K80 GPUs and 26 GB RAM.

Table 7 Overall model performance for experiment 1

Model	Macro-average			Accuracy
	Precision	Recall	F1-score	
Resnet50 v2	0.9780	0.9777	0.9778	0.9785
Densenet121	0.9893	0.9893	0.9893	0.9897
Custom model	0.8731	0.8694	0.8705	0.8721

3 Results and Discussions

3.1 Performance Metrics

It is imperative to evaluate the performance of any machine learning or deep learning project to ensure that optimal results are being delivered to the user by the model. The performance of a ML model should be evaluated using more than one metric. While the model may be observed to be performing well using a certain metric, its performance may be worse when measured using another metric. There are a number of performance metrics available for determining the success or failure of a model.

Amongst the most common metrics used are accuracy, precision, recall (or sensitivity) and F1 score. These are calculated using the Eqs. (1)–(5) below and where, TP, FP, TN and FN are the number of classified cases of true positives, false positives, true negatives and false negatives, respectively.

3.2 Evaluation of Results

Amongst the various metrics defined above, classification accuracy is the most extensively used quality index. The classification accuracies obtained in our experiments are as given in Tables 7 and 8. These results give an indication of the performance of the different models used for this study. After the training phase, each model is tested using the testing portion of our dataset and the corresponding classification reports and confusion matrix were generated as given in Table 9.

3.3 Discussion

After assessing the data from the graphs in Figs. 6, 7 and 8 as well as Tables 7, 8 and 9, it can be seen that the model trained using transfer learning yields far better classification accuracies than the custom model which was trained from scratch and has overfitted. The F1-score, precision and recall for the ‘no tumor’ easily achieve 99%, possibly due to the superior quality of images and since no tumor traces are present

Table 8 Class-specific evaluation for the different model

Class	Precision	Recall	F1-score	Accuracy
(a) ResNet model				
G	0.9797	0.9638	0.9716	0.9785
M	0.9554	0.9582	0.9568	
No T	0.9991	0.9982	0.9987	
P	0.9777	0.9904	0.9840	
(b) DenseNet model				
G	0.9801	0.9867	0.9834	0.9897
M	0.9834	0.9776	0.9805	
No T	1.0000	1.0000	1.0000	
P	0.9939	0.9930	0.9935	
(c) Custom model				
G	0.8664	0.8284	0.8470	0.8721
M	0.7348	0.7969	0.7646	
No T	0.9991	0.9383	0.9677	
P	0.8921	0.9138	0.9028	

it is more distinguishable from tumor images. Regarding the misclassifications from tumor classes, the dataset used for these experiments contains several degraded images and in some cases is impossible to identify. The ResNet and DenseNet models perform very well and this can be justified by their excellent architecture design and the fact that they were pre-trained on ImageNet.

The results obtained from our experiments also prove to be comparable to the mentioned works above, and in some cases, better performance can be observed.

4 Conclusion

In this paper, several concepts of deep learning have been applied to establish an approach to classify brain tumors. Convolutional neural networks have proved to be ideal for computer vision tasks by showing higher accuracy in classifications. However, misclassifications were still present in our tests and may have been caused by degradations in the quality of images used or even by the size of our dataset. The integration of deep learning in medical applications will require data of superior quality but it may still be used in the creation of decision support systems for radiologists.

A future iteration of this work might include more tumor classes depending on data availability. A new model could also be developed for multi-label classification, thus allowing sub-classes of tumors to be classified (e.g., the sub-types of glioma). The model might also create bounding boxes on tumor regions.

Table 9 Confusion matrix for the different models

(a) ResNet model					
		Predicted			
		G	M	No	P
Actual	Glioma	1011	34	0	4
	Meningioma	20	986	1	22
	No tumor	0	2	1133	0
	Pituitary	1	10	0	1138

(b) DenseNet model					
		Predicted			
		G	M	No	P
Actual	Glioma	1035	13	0	1
	Meningioma	17	1006	0	6
	No tumor	0	0	1135	0
	Pituitary	4	4	0	1141

(c) Custom model					
		Predicted			
		G	M	No	P
Actual	Glioma	869	156	1	23
	Meningioma	118	820	0	91
	No tumor	3	54	1065	13
	Pituitary	13	86	0	1050

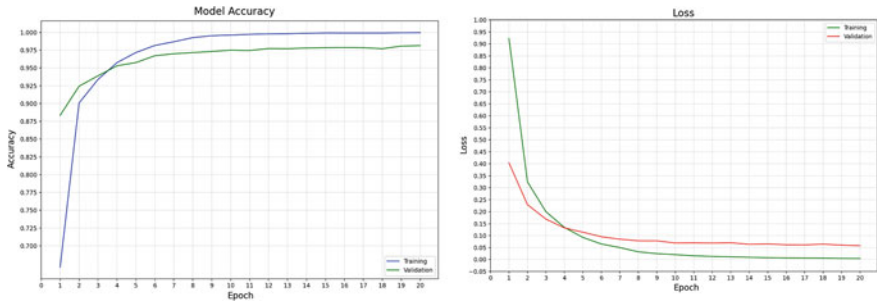


Fig. 6 Model accuracy and loss graph for the ResNet model

Generative adversarial networks (GAN) may also be considered regarding the generation of synthetic MRI images for training but would require the assistance of a domain expert to assess the validity of the generated data.

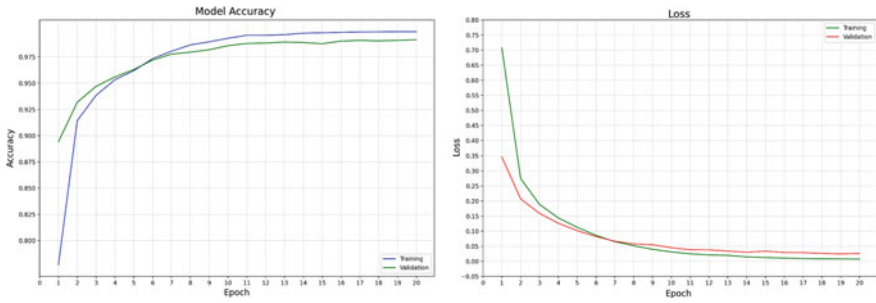


Fig. 7 Model accuracy and loss graph for the DenseNet model

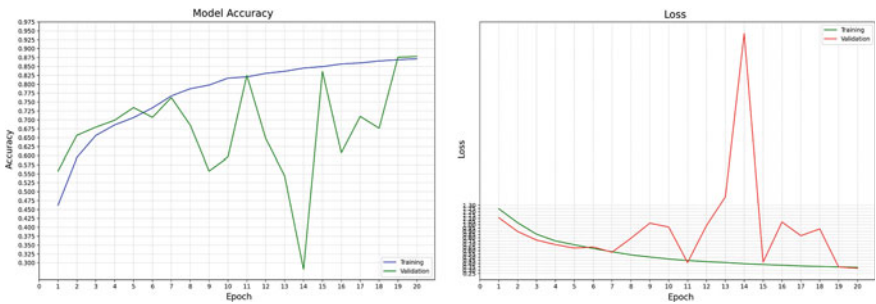


Fig. 8 Model accuracy and loss graph for the custom model

References

1. Balasooriya, N.M., Nawarathna, R.D.: A sophisticated convolutional neural network model for brain tumor classification. In: 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), pp. 1–5. IEEE (2017)
2. BraTS: Multimodal Brain Tumor Segmentation Challenge 2019, CBICA, Perelman School of Medicine at the University of Pennsylvania, <https://www.med.upenn.edu/cbica/brats2019/data.html>
3. Charron, O., et al.: Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Comput. Biol. Med.* **95**, 43–54 (2018)
4. Cheng, J.: Brain tumor dataset. Figshare (2017)
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807. IEEE (2017)
6. Deepak, S., Ameer, P.M.: Brain tumor classification using deep CNN features via transfer learning. *Comput. Biol. Med.* **111**, 103345 (2019)
7. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118 (2017)
8. Gu, Y., et al.: Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs. *Comput. Biol. Med.* **103**, 220–231 (2018)
9. He, K. et al.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE (2016)
10. Horie, Y., et al.: Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest. Endosc.* **89**(1), 25–32 (2019)

11. Howard, A.G., et al.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications (2017)
12. Huang, G., et al.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. IEEE (2017)
13. Jeyaraj, P.R., Samuel Nadar, E.R.: Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *J. Cancer Res. Clin. Oncol.* **145**(4), 829–837 (2019)
14. Krizhevsky, A., et al.: ImageNet Classification with Deep Convolutional Neural Networks (Semantic Scholar. Undefined) (2012)
15. Kutlu, H., Avci, E.: A novel method for classifying liver and brain tumors using convolutional neural networks, discrete wavelet transform and long short-term memory networks. *Sensors Basel Sensors* **19**, 9 (2019)
16. LeCun, Y., et al.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
17. Malone, I.B., et al.: MIRIAD—public release of a multiple time point Alzheimer’s MR imaging dataset. *Neuroimage* **70**, 33–36 (2013)
18. Pattanayak, S.: Pro Deep Learning with TensorFlow: A Mathematical Approach to Advanced Artificial Intelligence in Python. Apress (2017)
19. Rawat, W., Wang, Z.: Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* **29**(9), 2352–2449 (2017)
20. Sajjad, M., et al.: Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J. Comput. Sci.* **30**, 174–182 (2019)
21. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **24**(3), 279–283 (2017)
22. Sandler, M. et al.: Mobilenetv2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510–4520. IEEE (2018)
23. Saxena, P. et al.: Predictive modeling of brain tumor: a deep learning approach (2020) (Unpublished)
24. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (2014)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Presented at the 3rd International Conference on Learning Representations, ICLR (2015)
26. Swati, Z.N.K., et al.: Content-based brain tumor retrieval for MR images using transfer learning. *IEEE Access.* **7**, 17809–17822 (2019)
27. Szegedy, C. et al.: Going deeper with convolutions. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. IEEE (2015)
28. Szegedy, C. et al.: Rethinking the inception architecture for computer vision. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. IEEE (2016)
29. Tan, C., et al.: A survey on deep transfer learning. In: Kůrková, V. (ed.) Artificial Neural Networks and Machine Learning—ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, Oct 4–7, 2018, Proceedings, Part III, pp. 270–279. Springer International Publishing, Cham (2018)
30. Toğaçar, M., et al.: BrainMRNet: brain tumor detection using magnetic resonance images with a novel convolutional neural network model. *Med. Hypotheses.* **134**, 109531 (2020)
31. Yousefi, M., et al.: Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. *Comput. Biol. Med.* **96**, 283–293 (2018)
32. Zhou, L., et al.: A deep learning-based radiomics model for differentiating benign and malignant renal tumors. *Transl. Oncol.* **12**(2), 292–300 (2019)
33. Access the Data—The Cancer Imaging Archive (TCIA), <https://www.cancerimagingarchive.net/access-data/>
34. Brain MRI Images for Brain Tumor Detection, Kaggle, <https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection>

35. Brain Tumor Information, National Brain Tumor Society, <https://braintumor.org/brain-tumor-information/>
36. IXI Dataset—Brain Development, <https://brain-development.org/ixi-dataset/>
37. Minimal Interval Resonance Imaging in Alzheimer’s Disease (MIRIAD), Dementia Research Centre—UCL—London’s Global University, <https://www.ucl.ac.uk/drc/research/methods/minimal-interval-resonance-imaging-alzheimers-disease-miriad>
38. NFBS Skull-Stripped Repository, https://preprocessed-connectomes-project.org/NFB_skulls-tripped/

Deep Neural Network Optimization for Handwritten Text Recognition



Chahat Goel, Aishwarya Chaudhary, S. Indu, and Sudipta Majumdar

Abstract These days a large number of handwritten documents are available in scanned images. The aim of Handwritten Text Recognition (HTR) is to transcribe the offline document images by a computer through the use of deep neural networks. Even though hybrid architectures, designed for this purpose are gaining popularity, we aim to further optimize the prescribed models by changing the hyperparameters associated with the number of convolutional neural network layers on the overall accuracy. We explore the implementation of hybrid architectures and show that mainly the depth of the network plays a noteworthy role in improving the accuracy of the trained model.

Keywords Handwritten text recognition (HTR) · Convolutional neural network (CNN) · Recurrent neural networks (RNN) · Multidimensional recurrent neural network (MDRNN) · Long short-term memory (LSTM) · Connectionist temporal classification (CTC) · Multi-directional long short-term memory (MDLSTM)

1 Introduction

HTR is the method used to transcribe the handwritten text into digital text. There are two approaches to HTR: online and offline. The online approach involves writing with a special pen on an electronic device, thereby enabling the availability of geometric and temporal information. On the other hand, offline approach involves scanning the handwritten documents and processing the resulting images. The offline approach

C. Goel (✉) · A. Chaudhary · S. Indu · S. Majumdar
Electronics and Communication Department, Delhi Technological University, Delhi, India
e-mail: chahatgoel98@gmail.com

A. Chaudhary
e-mail: chaudharyaishwarya96@gmail.com

S. Indu
e-mail: s.indu@dce.ac.in

S. Majumdar
e-mail: korsudipta@rediffmail.com

involves many challenges like the cursive nature of handwriting, the variety of each character in shape and size and large vocabularies. Thus, the offline approach is a relatively difficult approach as compared to online. However, its successful implementation will help in various areas such as automated evaluation of handwritten documents, word spotting, transcribing historical manuscripts, performing data analytics on medical transcripts.

Earlier, standard Optical Character Recognition (OCR) techniques were used for recognition work. OCR technique used for word recognition fails due to the following reasons: (1) For individual characters, it recognizes at a significantly higher accuracy. However, poses a challenge for cursive handwriting recognition. (2) It requires manually hard coding of individual properties of symbols like aspect ratio, pixel distribution, number of strokes, etc.

Neural networks overcome the limitations of OCR as they are able to learn parameters during the training phase from the large dataset, thereby making the manual hard coding redundant. Recently, the hybrid architecture of Convolutional Neural Network-Recurrent Neural Network (CNN-RNN) is gaining popularity as the classifier makes use of CNN layers to extract features from the input image and recurrent neural network layers to propagate information through the image. The recurrent neural network outputs a matrix which contains a probability distribution over the characters at each image position. Decoding this matrix yields the final text and is done by the connectionist temporal classification operation. Even though this hybrid architecture gives significantly accurate results, however, further optimization is achieved by changing the hyperparameters of the trained model.

2 Related Work

Cun et al. [1] demonstrated how CNNs eliminate the need for hand-crafted feature extractors. CNNs work with data that is organized in a grid-like structure, such as images which have a 2D structure or plain audio data which have a 1D structure [2]. A Rectified Linear Unit (RELU) is taken as a non-linear activation function in CNNs [2]. The RELU essentially zeros out the negative part of its argument while it keeps the positive part untouched. Pooling [2] is added on top of the non-linearity: It replaces the output at a certain location by a summary statistic of nearby outputs.

An RNN is a special kind of ANN which is used to model sequential data such as handwritten text or speech [3]. Sequential data can be of arbitrary length along the time axis, where in the context of handwritten text, it means the axis along which the writing happens. Additionally, to the input and output sequence x_t and y_t , RNNs also have an internal state sequence h_t , which memorizes past events. The internal state h_t is a function of the current input x_t and the last internal state h_{t-1} that means $h_t = f(h_{t-1}, x_t)$ [2]. Graves et al. [4] proposed MDRNNs, where the number of connections equals the spatiotemporal dimensions found in the data, which substitute the single recurrent connection of conventional recurrent networks.

MDRNN connections enable the network to form an elastic internal depiction of neighboring context, which is immune to localized variations and distortions.

Hochreiter [5] discussed the vanishing gradient problem of RNNs. To avoid them, he introduced Long Short-Term Memory recurrent neural networks (LSTM) [6]. A LSTM block learns when to remember and when to forget past events, which are achieved by introducing gating neurons [2]. Graves et al. [7] noted that although RNNs are powerful in learning sequences, they require training data which is pre-segmented, and post-processing is required for transforming their outputs into label sequences, thereby limiting their usage. This limitation is overcome by CTC.

Most of the recent works have used convolutional layers for feature extraction and applied variations on the recurrent network to achieve transcription. Wu et al. [8] showed that separable MDLSTM-RNN not only consumes less computational time and resources but is also effective in processing the variations in handwriting in various directions for better learning. Stuner et al. [9] proposed a framework, where complimentary neural networks are stacked together in a cascading fashion and contain a rejection stage depending upon the lexicon validation.

Jain et al. [10] demonstrated on the Urdu dataset how the use of CTC layer allows segmentation-free transcription and how patterns can be captured in both forward and backward direction through the use of bidirectional RNNs. Chen et al. [11] devised a multi-task model that was capable of both identifying the script and performing handwriting recognition. They used an LSTM variant called Separable Multidimensional Long Short-Term Memory (SepMDLSTM). They also made use of CTC loss to train their model.

In this work, we optimize the efficiency of two models: (1) a CNN-RNN hybrid architecture proposed by [12] and (2) Multidimensional recurrent neural network proposed by [13] by changing the hyperparameter associated with the number of hidden layers, since hyperparameters have a significant impact on the performance of the model being trained.

3 Problem Formulation

Offline HTR is generally considered to be difficult as compared to online handwriting recognition. In the online case, the input data forms a 1D sequence which can be easily fed to an RNN. However, the offline case poses a huge challenge since the input images are not one dimensional. Presenting the image as a 1D sequence by giving the images to the network in the form of one vertical line at a time would be a very novice approach as such an approach is least likely to handle the variations and the distortions which occur on the vertical axis; for example, even if an image is moved above or below by a single pixel, it would appear entirely different from the original one. Significant amount of work has been done in this direction, with the two models described earlier being the most distinctive. We aim to optimize these models by changing the hyperparameters associated with the increase in the number of layers to achieve optimum accuracy.

4 Methodology

4.1 CNN-RNN Hybrid Architecture

Figure 1 illustrates the architecture that we use in our work which consists of CNN layers, RNN layers, and a CTC best path algorithm [14].

Neural network can be expressed as a function which maps an image of size $W \times H$ to a character vector (C_1, C_2, \dots) lying in the range of 0 and L .

$$NN : M \rightarrow (C_1, C_2, \dots, C_n) \tag{1}$$

where M is the matrix of size $W \times H$, and " n " ranges from 0 to L .

CNN: The CNN layers extract appropriate features from the input image. Each layer performs three operation: (1) the convolution operation—The tradition convolution process is achieved when the function f is set to retrieve n parameters from the patch of data x [15].

$$f(x) = (\sigma(\text{sum}(W_1 \odot x) + b_1), \dots, \sigma(\text{sum}(W_n \odot x) + b_n)) \tag{2}$$

where b_1, b_2, \dots, b_n are biases, W_1, W_2, \dots, W_n are weight matrices, all the components of a matrix are summed by the sum operator, and \odot is a component by component multiplication. This leads to n diverse parameters, and the values of a particular feature over windows give rise to a feature map (2) non-linear RELU activation function, given by $R(x)$, makes all the negative value to zero.

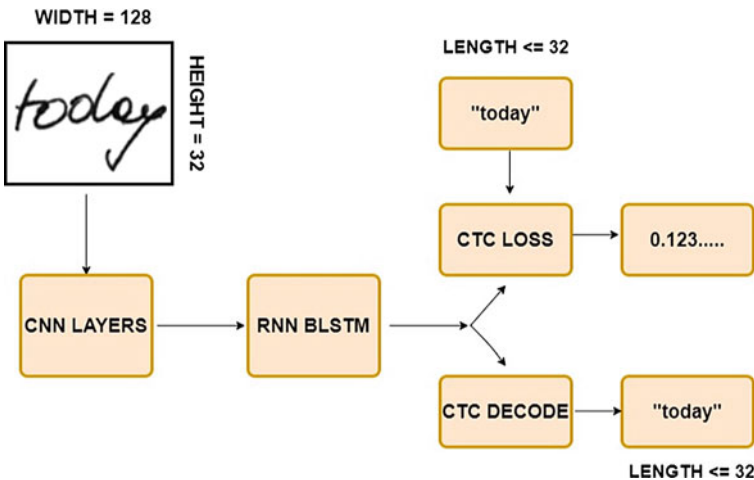


Fig. 1 Model outline of CNN-RNN hybrid architecture

$$R(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{3}$$

(3) Pooling layer down samples each feature map independently, diminishing the height and width, keeping the depth unchanged.

RNN: The information sequence propagates through RNN. The Long Short-Term Memory (LSTM) implementation of RNNs is used that consists of a gating mechanism which is capable of propagating information through distances that are much longer and offers additional vigorous training features than simple RNN. The output vector obtained from RNN is converted into a matrix of size 32×80 . There are 79 different characters in IAM dataset and one extra character, CTC blank character, which is essential for the CTC operation; hence, we have 80 characters for each time-step repeated for 32 times.

CTC: RNN output matrix is fed to the CTC algorithm which acts as an output layer of RNN designed to label the sequence. The output of CTC is the probability distribution over the labeled sequences along with the CTC loss.

4.2 *Multidimensional Recurrent Neural Network*

In this architecture, several layers of convolution and MDLSTM are alternatively stacked together. After the final layer of MDLSTM, the 2D vector is converted into a 1D sequence. The alignment between the output and the input sequence is handled by the softmax layer. We attempt to optimize the network topology by changing the number of convolution and MDLSTM blocks, accompanied by the occurrence and the absence of max pooling layer. The network architecture is described by the sequence which has a nomenclature of the form BP-B-BP-BP-B. Here, BP indicates a unit of MDLSTM and convolution with max pooling applied to it, and B indicates

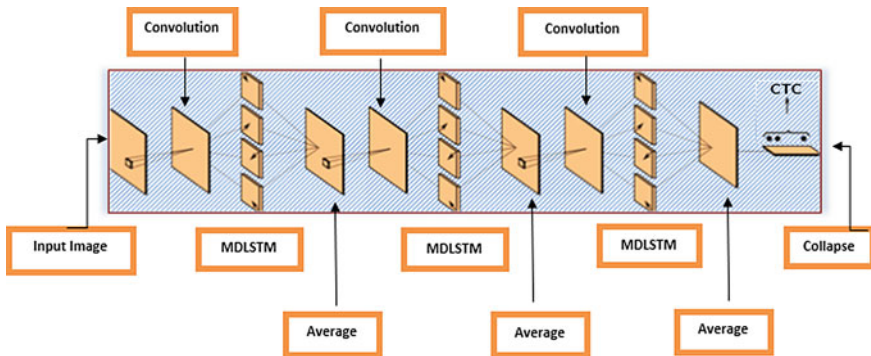


Fig. 2 Model overview of MDLSTM-based architecture

a unit of MDLSTM and convolution with no max pooling. Figure 2 depicts this architecture.

A slight variation has been applied to the network architecture proposed by [13]. In this, we don't divide the input image into blocks of size 2×2 ; instead, we directly feed the input image to a convolution layer. Next, four MDLSTM layers are applied to scan through the input in four different directions, as indicated by the arrows. The average of all the values coming from the four directions is fed to a single convolution layer. A CTC layer is applied to transcribe the input image after it is made to pass through a cascade of convolutional and MDLSTM layers.

4.3 Evaluation Metric

Character Error Rate (CER) and Word Error Rate (WER) [16] are very popular methods for calculating errors for HTR. CER is calculated by counting the number of edit operations to translate the recognized text into the ground truth text, divided by the length of the ground truth text [12]. For WER, the text is split into a sequence of words. This word sequence is then used to calculate the WER in the same way as a character sequence is used to calculate the CER. Both CER and WER can take on values greater than 100% if the number of edit operations exceeds the ground truth length [12]: If "xyz" is recognized but the ground truth is "ab," then the CER has a value of $3/2 = 150\%$.

$$\text{CER} = \frac{I + D + S}{\text{Length of Ground Text}} \quad (4)$$

where I represents character insertions, D represents character deletions, and S represents character substitutions required to convert the recognized output to the ground truth text.

$$\text{WER} = \frac{I(w) + D(w) + S(w)}{\text{Length of Ground Text}} \quad (5)$$

where $I(w)$ represents word insertions, $D(w)$ represents word deletions, and $S(w)$ represents word substitutions required to convert the recognized output to the ground truth text.

4.4 Hyperparameters

Hyperparameters are the parameters that determine the structure of the trained architecture and are set before training, for example, number of hidden layers in CNN,

kernel size, stride, padding, batch size, etc. The value of hyperparameters cannot be determined from the data, and by hit and trial method, best combination of hyperparameter is calculated in order to achieve maximum efficiency of the model.

Hyperparameters are classified into two categories:

- Those that controls the network structure:
 - Number of hidden layers.
 - Size of the kernel.
 - Values of stride.
 - Values of padding.
- Those that controls the training of the network:
 - Batch size.
 - Number of epochs.
 - Learning rate.

In this work, we analyze the impact of varying the number of hidden layers in CNN on the accuracy of the model. We used manual search hyperparameter optimization technique to tune the hyperparameter. The model was trained on different number of layers and scored on the validation data.

4.5 *Decoding Algorithms*

Hyperparameter optimization is carried out and compared for two decoding algorithms

1. **Best path decoding:**

Best path decoding [7] is a simple algorithm which finds the most probable path by selecting the most expected character per time-step. Encoding is done by removing all the duplicate characters and blank spaces in the path. Final output is the recognized text.

2. **Vanilla beam search:**

Best path decoding is only an approximation, and for advanced decoding, we use vanilla beam search. Beam search decoding [14] calculates the probability of multiple labeling candidates. In each time-step, the candidates are extended by all possible labels. This gives new candidates for the next time-step. To keep the algorithm feasible, only the most promising labels are kept.

4.6 *Calculation of CTC Loss*

CTC helps in aligning the input with the output. For a given input X , we need to train our model to maximize the probability associated with the right answer. For this

conditional probability $p(Y/X)$ is computed [17], where Y is the output sequence. To calculate CTC loss, scores in all the possible alignments of the ground truth text are summed up such that the position of appearance of the text in the image does not matter. The score for each alignment is computed by multiplying the corresponding character scores.

5 Experiments and Results

We have used IAM handwriting database 3.0 [18] for the research. It contains 115,320 words spanning across 13,353 handwritten sentences contributed by 657 writers. Table 1 shows the data splitting description for training and validation purposes. The goal is to check the performance of the model by changing the depth of the network keeping the other hyperparameters constant.

The hybrid architectures are implemented through TensorFlow framework on Nvidia GTX 1070 GPU. Training was done with the batch size of 50. The kernel size of first two layers is 5×5 , and for the rest of the layers, it is 3×3 . The choice of keeping the filter size in odd numbers is because it allows all the previous layer pixels to be symmetrically placed around the output pixel. Without this symmetry, distortions across the layers need to be accounted which happens while using an even sized kernel. In general, small odd sized kernel filters are applied eliminating the 1×1 filter as the feature extracted would be highly localized, with no information from the neighboring pixels. Performance is measured in Character Error Rate (CER) and Word Error Rate (WER).

The results for hyperparameter optimization for best path decoding algorithm and vanilla beam search decoding algorithm are tabulated in Tables 2 and 3, respectively. We see that both the CER and WER improve when the number of hidden layers increases from 3 to 5 layers to finally the 7 layers with significant average word recognition accuracy of 71.245% as against 67.675% in 3 layers and 70.3% in 5 layers.

Table 1 Data splitting description

Data section	Splitting (%)
Training	80
Validation	20

Table 2 Comparison of different layers for best path decoding

Hidden layers	CER (%)	WER (%)
3	12.57	32.68
5	10.62	30.0
7	8.77	29.07

Table 3 Comparison of different layers for vanilla beam search decoding

Hidden layers	CER (%)	WER (%)
3	12.01	31.97
5	10.46	29.36
7	8.48	28.24

Table 4 Comparison of different MDLSTM-based network topologies

Network	Hidden layers	CER (%)	WER (%)
BP-BP-BP	6	3.8	10.2
BP-BP-BP-B	8	3.7	10.4
BP-B-BP-BP	8	3.6	10.0
BP-BP-BP-B-B	10	3.5	9.3
BP-B-BP-B-BP	10	3.5	9.3

With the change in the network topology, i.e., the combination of CNN layers and MDLSTM layer, CER and WER significantly improves. The results for MDLSTM are tabulated in Table 4. We observe that by simply increasing the depth of the model, it yields a word recognition accuracy of 90.7% in 10 layers as against 89.8% in 6 layers.

6 Conclusions

We presented our comparison and concluded that the depth of the networks plays a pivotal role in determining the performance of the model. We trained the CNN-RNN hybrid architecture up to 7 layers and the MDLSTM-based hybrid architecture up to 10 layers and achieved substantial improvements in the performance in terms of word recognition accuracy. Also, we concluded that the presence or the absence of max pooling in the MDLSTM-based network doesn't have much significant effect. We believe that these experiments were limited for analyzing the effect of only one hyperparameter on the proposed architecture; however, significant improvements and further optimization can be achieved by analyzing various other hyperparameters as well.

References

1. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE (1998)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
3. Greff, K., Srivastava, R., Koutník, J., Steunebrink, B., Schmidhuber, J.: LSTM: a searchspace odyssey. IEEETrans. NeuralNetw. Learn. Syst. 2222–2232 (2017)

4. Graves, A., Fernandez, S., Schmidhuber, J.: Multidimensional recurrent neural networks. In: International Conference on Artificial Neural Networks. Porto, Portugal, Sept 2007
5. Hochreiter, S.: Untersuchungen zudynamischen neuronalen Netzen. Master's thesis, Technische-Universität. München (1991)
6. Hochreiter, S., Schmidhuber, J.: Longshort-term memory. *Neural Comput.* (1997)
7. Graves, A.: Supervised sequence labelling with recurrent neural networks. Springer (2012)
8. Wu, Y.-C., Yin, F., Chen, Z., Liu, C.-L.: Handwritten Chinese text recognition using separable multi-dimensional recurrent neural network. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto (2017)
9. Stuner, B., Chatelain, C., Paquet, T.: Handwriting recognition using cohort of LSTM and lexicon verification with extremely large lexicon (2016). CoRR, abs/1612.07528
10. Jain, M., Mathew, M., Jawahar, C.V.: Unconstrained OCR for Urdu using deep CNN-RNN hybrid networks. In: 4th IAPR Asian Conference on Pattern Recognition (ACPR), pp. 747–752. Nanjing (2017)
11. Chen, Z., Wu, Y., Yin, F., Liu, C.L.: Simultaneous script identification and handwriting recognition via multi-task learning of recurrent neural networks. ICDAR (2017)
12. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intel.*
13. Pham, V., Bluche, T., Kermorvant, C., Louradour, J.: Dropout improves recurrent neural networks for handwriting recognition. In: International Conference on Frontiers in Handwriting Recognition (2014)
14. Scheidl, H., Fiel, S., Sablatnig, R.: Word Beam Search: A Connectionist Temporal Classification. ICFHR (2018)
15. Keren, G., Schuller, B.: Convolutional RNN: an enhanced model for extracting features from sequential data. In: International Joint Conference Neural Network, pp. 3412–3419 (2016)
16. Bluche, T.: Deep neural networks for large vocabulary handwritten text recognition. Ph.D. Thesis Université Paris Sud-Paris XI (2015)
17. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intel.* **31**(5), 855–868 (2009)
18. Marti, U.-V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. *IJDAR* (2002)

Consolidating Online Real Estate Data Using Image Analysis and Text Processing



Yasser Chuttur and Haydar Mahadooda

Abstract Property buyers often access online real estate websites to search for properties which they might be interested in. However, property sellers may in turn have their properties listed on several real estate agency websites to increase their chances of reaching a large audience. Consequently, same properties are often found on multiple websites. Problems arise when same properties are listed with different prices and different details on different websites. To users, such a situation impedes on the decision-making process. Ideally, having a single portal, where all data pertaining to the same property from different real estate agency websites are consolidated that would be useful for potential buyers prior for making a decision. Identifying similar real estate properties posts, however, is not a straightforward exercise. With different information available from different websites, additional processing is required. In this paper, we propose an approach that makes use of image analysis and text processing to detect similar real estate properties which advert from different websites posted for the Mauritius market.

Keywords Information processing · Image analysis · SURF

1 Introduction

The real estate industry comprises of various economic activities that include property development, appraisal, marketing, selling, leasing, and management of commercial, industrial, residential properties. In Mauritius, the real estate industry is an important economic pillar, which generates foreign direct investment, revenues and creates jobs through marketing and sales activities. Today, the Internet has simplified these activities. For instance, a single click on Google, can allow property buyers to access

Y. Chuttur (✉) · H. Mahadooda

Software and Information Systems Department, University of Mauritius, Reduit, Mauritius
e-mail: y.chuttur@uom.ac.mu

H. Mahadooda

e-mail: muhammad.mahadooda@umail.uom.ac.mu

a variety of real estate offers. The greater number of platforms owners list their properties, the greater the diffusion of information.

While having large numbers of websites do encourage buyers to search more intensively, discover and visit more properties, it also incurs excessive time and energy spent. Therefore, the use of Internet does not benefit homebuyers in terms of search time, flexibility, and intuitive results. Another example of a frequently encountered problem is related to a property being listed on more than one real estate websites with inconsistency in the type and level of information, namely price, size, amenities, and features amongst others. Lack of full cost disclosure may be a huge issue if not stated accurately. This inconsistency in result display may negatively influence the decision-making process of buyers, thus leading to loss of business opportunities for the property owners.

To address these issues, we propose a federated search system to consolidate real estate contents on a single platform. The main feature of the platform is related to merging the same property listings identified from different websites with high information density and also provides the links to the source websites. Speeded Up Robust Features (SURF), an image recognition technique has been employed to match identical properties. To further improve the matching accuracy, the addresses of the listings were compared using the Levenshtein distance similarity ratio.

The paper is divided into the following five sections as follows: (i) a brief overview of online federated search systems, image identification techniques, and text processing techniques, (ii) general description of the proposed system architecture, (iii) implementation of the system and result output, (iv) evaluation of the federated search platform, and (v) concluding section that also discusses the limitations of the proposed system.

2 Online Federated Search System

A federated search system is one that provides access to information from multiple sources using a single user interface [1]. Data from multiple existing information sources are fetched, processed, and consolidated for display on the federated system interface. According to Shokuhi and Si [2], an information source can be a database on a server, a web portal, a government website, or a general-purpose search engine. The data retrieved may undergo transformation, refining, and other post processing techniques depending on the purpose and architecture of the system [3].

In general, Cahoon et al. [5] explain that a federated search system should consists of three general components; (1) a resource selector, an application which performs the task of fetching information from the sources, (2) a broker that can analyze, fetch, and process relevant information from the imported data and merge them into a single list, and (3) a user interface that displays the result.

Furthermore, a federated search system may vary in the underlying architecture depending on the purpose of the system [4]. Two well-known architectures used

in federated search system are: the single database architecture and the Distributed Information Retrieval (DIR) architecture as described below.

2.1 Single Database Architecture

In single database architecture, information needed by the system is made available in a unified database [4]. A single database architecture does not provide access to real-time data. Data sources consist mostly of websites. Data can be obtained through different information extraction techniques, and among the most used technique for this architecture is web scraping [6]. The single database architecture provides the system using it with a highly desired feature which is the ability to do post-processing on the data before making it available to its end user.

2.2 Distributed Information Retrieval Architecture

Thomas, [3], explains the concept of Distributed Information Retrieval (DIR) as one which queries multiple collections or search engines in real time, aggregates the data, and displays them in one view. Callan, [7], describes the architecture in three phases, (1) resource description, (2) resource selection, and (3) result merging.

2.3 Popular Federated Search System

Kayak¹ is a popular online travel agency that provides its users facilities for flight bookings with options to book flights from the airline carrier website or third-party booking agent. The prices for each agent providing the booking service for a particular flight is given on the website. Kayak uses federated search technologies to obtain its data. In order to have price comparison ability and presenting the cheapest option, Kayak connects to its source websites through custom built API (Application Programming Interface) for each source website [7].

Other websites providing similar services such as Expedia,² Travelocity³ and Skyscanner⁴ uses a range of techniques for information extraction based on the source websites type and the services it offers. Aside from an API connection to the source website, other techniques include web scraping, XML form feed from the source website, Rich Site Summary (RSS) feeds, and Iframe (Inline Frame) [8].

¹<https://www.kayak.com/>.

²<https://www.expedia.com/>.

³<https://www.travelocity.com/>.

⁴<https://www.skyscanner.net/>.

3 Detecting Similar Images

To be able to find similar real estate properties listed on multiple websites, an effective image matching techniques are essential. Meharban and Priya [9] describe Content-Based Image Retrieval (CBIR) as an image retrieval technique using visual features such as color, shape, and texture. These visual features which are low-level features of an image content are made up of: (1) an extraction algorithm that detects and extracts features producing a feature vector, also known as feature descriptors, and (2) a similarity measure to compare two images [10]. We present here a brief review of several image matching algorithms in terms of their feature detection capabilities, feature matching time, and computing resource needed.

3.1 *Scale-Invariant Feature Transform (SIFT)*

The framework proposed by Lowe [11] makes use of the scale-invariant feature transform (algorithm), which is among the most widely used algorithm in applications for object recognition, image reconstruction, and object detection nowadays. SIFT algorithm performs image matching in four stages [12], namely:

1. Scale space extrema detection
2. Keypoint localization
3. Orientation assignment
4. Keypoint description.

3.2 *Speeded Up Robust Features (SURF)*

The Speeded Up Robust Features (SURF) algorithm, proposed by Bay et al. [13], operates for image analysis on the Gaussian scale space domain. SURF detectors use intrinsic images to increase the speed of Hessian matrix-based feature detection. SURF features are rotational and scale-invariant. SURF's major benefit over SIFT is that it requires a minimal amount of computing resources.

3.3 *KAZE*

KAZE uses non-linear scale space by filtering on non-linear diffusion. This approach makes blurring locally adaptable to feature-points in an image, thereby minimizing distortion and maintaining the outlines of regions in the target images at the same time. It is based on the Hessian matrix normalized scale determinant, measured at multiple scale stages [14]. Using a moving frame, the detector response maxima is collected as

feature-points. Feature description incorporates the rotation invariance attribute by identifying dominant orientation around each observed object in a spherical region. KAZE features are asymptotic to rotation, size, affinity and have more distinctive traits at different levels with the expense of moderate resource increase.

3.4 AKAZE

Built upon the non-linear diffusion filter of KAZE, Accelerated-KAZE (AKAZE) non-linear scale domains are built using a simulated annealing system known as Quick Explicit Diffusion (FED). The AKAZE detector is founded upon the Hessian matrix determinant. The efficiency of rotation invariance is enhanced through Scharr filtering. Maxima for detector results are obtained as feature-points in contextual regions. AKAZE descriptor is centered on the Modified Local Difference Binary (MLDB) algorithm which is efficient. Due to non-linear scaling space, the AKAZE characteristics are asymptotic to size, rotation, minimal affinity and have more distinctive features in different scales [15].

3.5 Oriented FAST and Rotated BRIEF (ORB)

Oriented FAST and Rotated BRIEF (ORB) technique, developed by Rublee et al. [16], uses a combination of modified Features from Accelerated Segment Test (FAST) [17] identification and path normalized definition method of Binary Robust Independent Elementary Features (BRIEF) [18]. Across each level of the scale pyramid, FAST points are identified and are evaluated using the Harris corner score to select only the highest quality scores. Since the BRIEF approach with rotation is unstable, a revamped version of the BRIEF descriptor was used. ORB characteristics are invariant to change in rotation and scale. ORB is a highly resourced technique in terms of computational capacity [15].

4 Finding Similar Text Contents

When the same real estate properties are listed on different websites, any text description of those real estate properties should be similar as well. In other words, although properties may be described in different wordings, key features (number of rooms, region, number of levels, size, etc.) of similar properties should remain invariant. Hence, a proper text matching algorithm would be useful in detecting similar properties.

In this section, we cover two popular string matching algorithms: the Levenshtein distance similarity ratio and the Rabin–Karp algorithm.

4.1 Levenshtein Distance Similarity Ratio

When evaluating the similarity between two strings of characters, Levenshtein distance similarity ratio measures the number of events required to convert one string into the other [19]. More events amount to less similarity for the two strings being compared. Every index letter of the string is given an equal significance in this method [20]. The permissible transformations are insertion—inserting a new character, deletion—removing a character, and substituting—replacing one character. By running these three operations, Levenshtein distance similarity ratio attempts to change the first string corresponding to that of the second. Ultimately, an edit distance between 0 and 1 is obtained that can be used to obtain the percentage similarity between the two strings [21].

4.2 Robin Karp Algorithm

Rabin Karp algorithm is a string matching algorithm that takes advantage of a hash function to perform its job. This technique employs hashing to search within a text for one of many series of string patterns [22]. Rabin Karp algorithm is widely used to detect copyright infringement activities on platforms that contain text, such as papers or blogs [23]. The text is converted to a token which is then converted to an integer containing the hash value. The Rabin Karp algorithm is based on the assumption that if two strings identical, then the hash value of both strings must be the same, so the algorithm can only search on strings under the same hash value as the searched string value [22].

5 Proposed Real Estate Federated System

To simplify the search for real estate properties, we implemented an online portal based on the concept of federated search, which employs image and text matching algorithms. More specifically, we used the Speeded Up Robust Features (SURF) and the Levenshtein distance similarity ratio to find similar properties. SURF has the best feature matching capabilities but does not have the best performance in terms of image matching time. ORB, BRISK, and AKAZE have shorter matching time compared to SURF [13, 26].

We make use of web scraping technologies to extract information from different real estate properties websites, and those are processed for image and text similarity. All data are stored in a consolidated real estate property database, which allows users to quickly obtain for properties from different websites in one place.

To further address the problem of inconsistent information for identical properties listed on different websites, the portal displays all properties as a single listing with

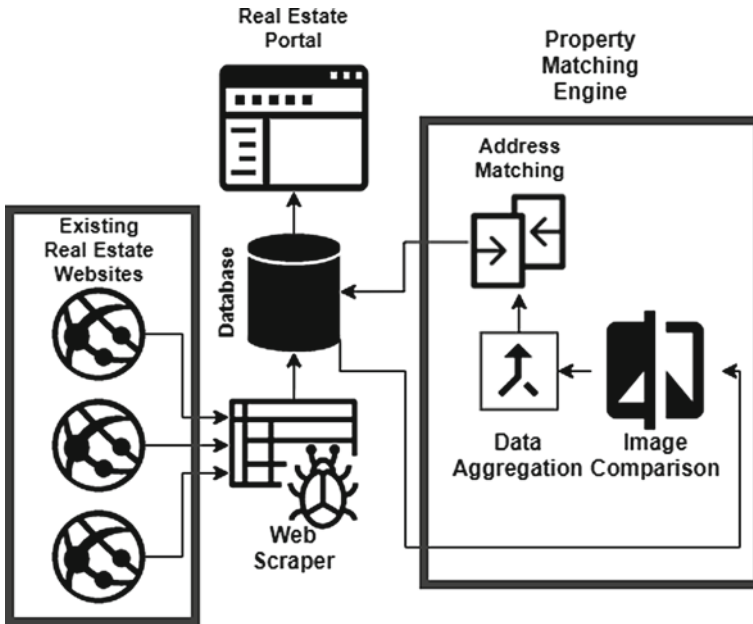


Fig. 1 Federated real estate search system architecture

maximum information density and price comparison. The architecture diagram of our proposed system is shown in Fig. 1. The system consists of two core processes: information extraction part and a property matching part. Description for each part of our system follows.

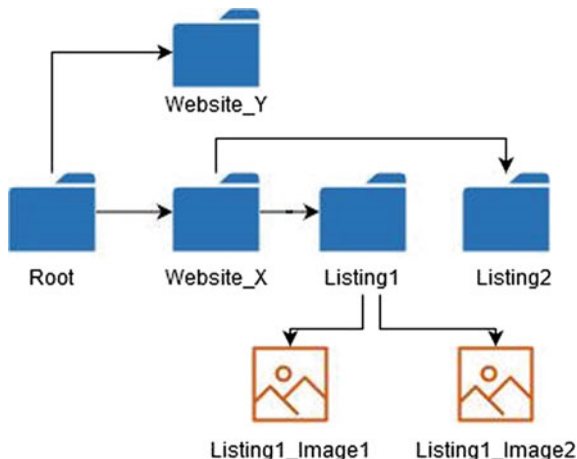
5.1 Information Extraction

This process makes use a web crawler, responsible to visit and extract real estate property information from a pool of real estate websites. Scraped data are made available to the property matching engine as a CSV file containing property contents from different websites. Images of listed properties are also downloaded in this phase.

5.2 Property Matching Engine

This process is responsible for image comparison, data aggregation, and address matching. SURF image identification algorithm with Fast Library for Approximate Nearest Neighbors (FLANN) based matching is used to identify similarities in images. FLANN is a feature matching algorithm built upon Nearest Neighbor

Fig. 2 Classification of images' folder structure. The images are named based on their listing ID and image sequence number



Distance Ratio (NNDR). For this matching process, each feature descriptor (from the first feature set) is checked for the nearest neighbor as well as second nearest neighbor [24]. Subsequently, for each feature descriptor, the ratio of the nearest neighbor to the next closest neighbor is determined, and a certain default ratio is used to screen out the preferred matches [25]. The default ratio is maintained as 0.7. The Levenshtein distance similarity ratio is also applied in this process to detect similar text contents.

6 Implementation, Testing, and Results

To extract information from different real estate websites, we used a desktop-based web scraping software freely available from [import.io](https://www.import.io/).⁵ A separate sitemap for each visited website had to be customized for the web scraper. The sitemap serves as a guideline for data extraction by the crawler. The scraped data was then outputted to a CSV file format. The data went through a data cleansing process in which string manipulations techniques was applied. The cleansed data is then imported to the system's database. In this phase, images for listed properties are downloaded using the URLs from the scraped data and stored in a classified structure (Fig. 2), which will help in the image identification process.

Following information extraction, the property matching engine is triggered so as to start property comparison process. The batch of images for a property pertaining to a website (batch A) is compared to the batch of images (batch B) of another website, and this process is repeated for all data collected by the web scraper. The SURF algorithm with FLANN-based matcher is applied to identify similar images. Once a batch of images is deemed as similar to another, a unique key is generated which is

⁵<https://www.import.io/>.

inserted into a similar records table in our consolidated database. Every unique key generated is stored in the consolidated database and will contain details of properties that are alike.

Once the matching process is completed, the system re-iterates another set for batch B until batch A is compared to all other batch of images pertaining to listings from other websites. SURF and FLANN were deployed into the system using python's OpenCV package, which has state of the art image processing capabilities.

For every unique key generated, the system performs a results merging process. Data from all websites for a property, which is found similar, are aggregated using a data frame in Python. The data frame undergoes forward and backward fills operations, a process, which ensure that the final output for an identical property has maximum information density. The data frame was implemented using the *pandas* package in python.

When merging results for similar properties, the data frame retains the original physical address or location (district, region, city, etc.) for each property as listed from different websites. Those addresses are subjected to a similarity comparison using the Levenshtein distance similarity ratio. The Levenshtein distance similarity ratio was implemented using the *fuzzywuzzy* package in Python. Addresses found similar indicate that the listings originally found similar by the image processing phase which can be qualified as identical. The merged result is then inserted into our consolidated database for display on the portal.

Figure 3 shows two identical listings on two different real estate websites but with different information. Result obtained by scraping data from those two websites and processed in our system is shown in Fig. 4.

7 Evaluation and Discussion

Real estate properties listings from three different real estate property websites were used to evaluate our federated system. Our focus is on real estate data available on different websites for the Mauritius market. The challenge with real estate data is that information available on different websites is unstructured and of different density. Our system relies on two attributes, image and address description, of real estate posts to determine similarity of two properties. The algorithm used for image matching is SURF, and for text similarity is the Levenshtein distance similarity ratio.

The total number of real estate listing processed was 700 from the websites: *lexpressproperties*, *davylandproperties*, and *mwproperties*.⁸ The dataset was manually analyzed, and of 185 identical listings were identified on at least two websites. The developed system was used to perform property matching on the 700 listings. Our approach, i.e., combined image and text analysis, gave us a total of 172 listings qualified as identical by the system. In other words, 92% of the listings could be correctly matched.

⁸<https://mwproperty.mu/>.

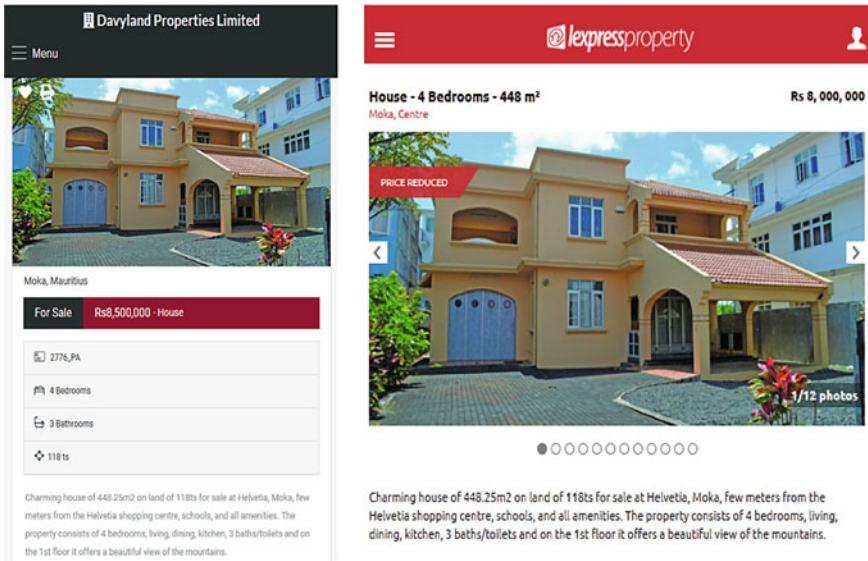
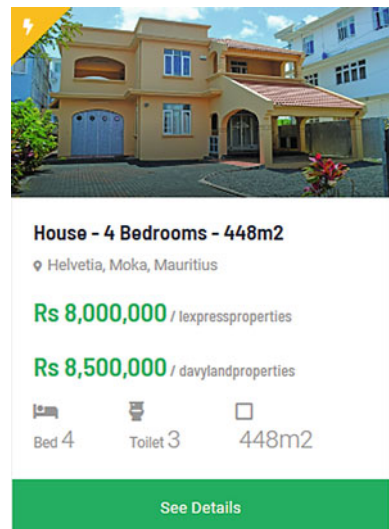


Fig. 3 Property on sale on [davylandproperties.com](https://www.davylandproperties.com/) ⁽⁶⁾ (left) and [l'expressproperty.com](https://www.lexpressproperty.com/en/) ⁽⁷⁾ (right)

⁶<https://www.davylandproperties.com/>.

⁷<https://www.lexpressproperty.com/en/>.

Fig. 4 Consolidated property information on our portal showing identical property listed on two different websites



Sample results obtained on our federated search portal are shown in Figs. 3, 4, 5, and 6. It is seen that even when images posted are from different angles (Fig. 5), the SURF algorithm is able to detect close similarity between the two listed properties demonstrating the effectiveness of our system.

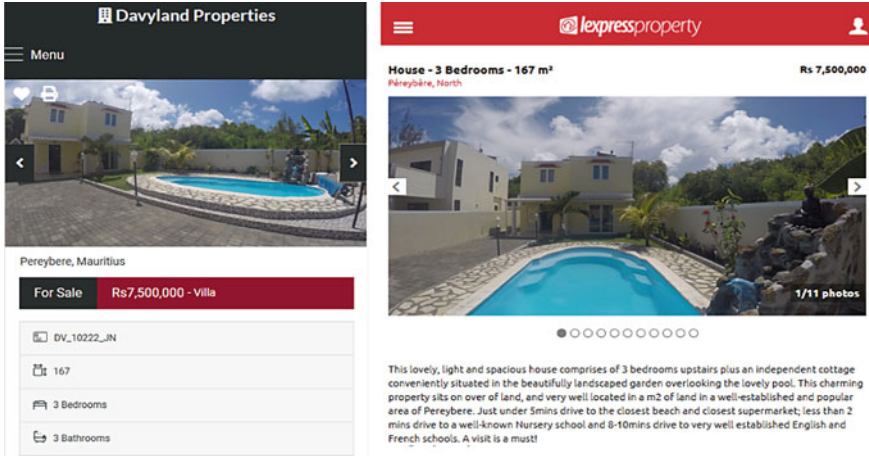
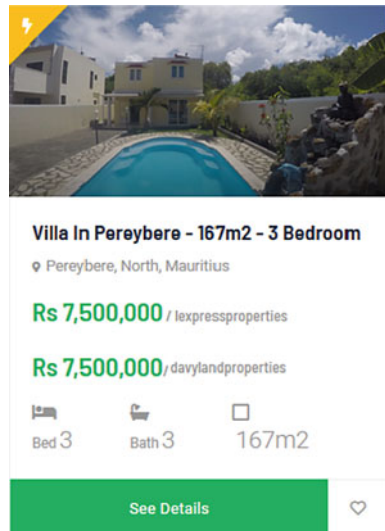


Fig. 5 Identical properties listed on two different websites with different pictures

Fig. 6 Consolidated property information on our portal showing identical property listed on two different websites



8 Conclusion

Having properties listed on different websites with different level of information is detrimental to the decision process of buyers. This paper has described the implementation of a federated real estate portal, which consolidates data from multiple real estate websites. The portal easily shows identical properties from multiple real estate websites as a single property listing enriched with reliable, consistent and real time information. The SURF algorithm and Levenshtein distance similarity ratio techniques used in our implementation gave fairly accurate results for image and text matching. In future, we intend to improve on the processing time of our system by applying the use of online GPUs such as Google Colab.

References

1. Avrahami, T., Yau, L., Si, L., Callan, J.: The FedLemur project: federated search in the real world. *J. Am. Soc. Inform. Sci. Technol.* **57**(3), 347–358 (2006)
2. Shokouhi, M., Si, L.: Federated search. *Found. Trends® Inf. Retrieval* **5**(1), 1–102 (2011)
3. Thomas, P.: To what problem is distributed information retrieval the solution? *J. Am. Soc. Inform. Sci. Technol.* **63**(7), 1471–1476 (2012)
4. Gibson, I., Goddard, L., Gordon, S.: One box to search them all: implementing federated search at an academic library. *Library Hi Tech* **27**(1), 118–133 (2009)
5. Cahoon, B., McKinley, K.S., Lu, Z.: Evaluating the performance of distributed architectures for information retrieval using a variety of workloads. *ACM Trans. Inf. Syst. (TOIS)* **18**(1), 1–43 (2000)
6. Chang, C.-H., Kayed, M., Girgis, M., Shaalan, K.: A survey of web information extraction systems. *IEEE Trans. Knowl. Data Eng.* **18**(10), 1411–1428 (2006)
7. Callan, J.: Distributed information retrieval. In: *Advances in Information Retrieval*, pp. 127–150. Springer, Boston, MA (2002)
8. Golgher, P.B., Laender, A.H., Da Silva, A.S., Ribeiro-Neto, B.: An example-based environment for wrapper generation. In: *International Conference on Conceptual Modeling*, pp. 152–164. Springer, Berlin, Heidelberg (2000)
9. Meharban, M.S., Priya, S.: A review on image retrieval techniques. *Bonfring Int. J. Adv. Image Process.* **6**(2), 7 (2016)
10. da Silva Torres, R., Falcao, A.X.: Content-based image retrieval: theory and applications. *RITA* **13**(2), 161–185 (2006)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* **60**(2), 91–110 (2004)
12. Almeida, J., Torres, R.D.S., Goldenstein, S.: SIFT applied to CBIR. *Revista De Sistemas De Informacao Da FSMA* **4**, 41–48 (2009)
13. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **110**(3), 346–359 (2008)
14. Alcantarilla, P.F., Bartoli, A., Davison, A.J.: KAZE features. In: *European Conference on Computer Vision*, pp. 214–227. Springer, Berlin, Heidelberg (2012)
15. Tareen, S.A.K., Saleem, Z.: A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In: *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1–10. IEEE (2018)
16. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: *2011 International Conference on Computer Vision*, pp. 2564–2571. IEEE (2011)

17. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European Conference on Computer Vision, pp. 430–443. Springer, Berlin, Heidelberg (2006)
18. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: Brief: binary robust independent elementary features. In: European Conference on Computer Vision, pp. 778–792. Springer, Berlin, Heidelberg (2010)
19. Rahim, R., et al.: Searching process with Raita algorithm and its application. *J. Phys. Conf. Ser.* **1007**(1), 1–7 (2018)
20. Khairul, M., Simaremare, Siahaan, A.P.U.: Decision support system in selecting the appropriate laptop using simple additive weighting. *Int. J. Recent Trends Eng. Res.* **2**(12), 215–222 (2016)
21. Ho, T., Oh, S.R., Kim, H.: A parallel approximate string matching under Levenshtein distance on graphics processing units using warp-shuffle operations. *PloS One* **12**(10) (2017)
22. Siahaan, A.P.U.: Rabin-Karp elaboration in comparing pattern based on hash data. *Int. J. Secur. Appl.* **12**(2), 59–66 (2018)
23. Lubis, A.H., Ikhwan, A., Kan, P.L.E.: Combination of Levenshtein distance and Rabin-Karp to improve the accuracy of document equivalence level. *Int. J. Eng. Technol.* **7**(2.27), 17–21 (2018)
24. Li, J.F., Wang, G., Li, Q.: Improved SURF detection combined with dual FLANN matching and clustering analysis. In: *Applied Mechanics and Materials*, vol. 556, pp. 2792–2796. Trans Tech Publications Ltd. (2014)
25. Vijayan, V., Kp, P.: FLANN based matching with SIFT descriptors for drowsy features extraction. In: 2019 Fifth International Conference on Image Information Processing (ICIIP), pp. 600–605. IEEE (2019)
26. Karami, E., Prasad, S., Shehata, M.: Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images (2017). [arXiv:1710.02726](https://arxiv.org/abs/1710.02726)

Time Series Visualization of Customer Emotions Using Artificial Neural Network



Yasser Chuttur and Nandishta Rawoteea

Abstract Online customer reviews have been recognized as having a high influence on both customers and business managers' decisions. While current approach to display general customer rating score is useful for customers to obtain a quick indication of overall customers experience with a given business, a fixed score does not provide useful information for managers on different customer service strategies adopted over a given period of time. In this study, we propose a method to analyze customer reviews over a period of time, so that emotions expressed by customers can be visualized across a timeline. We make use of Long Short-Term Memory (LSTM) as an artificial neural network technique on reviews posted on marideal.mu, a hotel booking website in Mauritius. Emotions such as 'love', 'sadness', 'joy', 'surprise', 'anger' and 'fear' were identified from 4270 reviews for 110 hotels found on marideal.mu website. We demonstrate how hotel managers can drill through the classified comments so as to identify various degree of satisfaction from their customers' experience and have a better insight about the performance of their businesses over a desired period of time. Findings of this proposal can help managers develop means to better identify business strategies to improve both customer satisfaction and business performance.

Keywords Emotion detection · Customer reviews · LSTM

1 Introduction

Customer reviews are common in most online business platforms. In this study, we shall focus on customer reviews for hotels. Traditionally, people had to call in hotels or hotel agencies in order to book a stay at a particular hotel. With recent development in technology and especially online hotel reservation platforms, customers have now

Y. Chuttur (✉) · N. Rawoteea

Software and Information Systems Department, University of Mauritius, Reduit, Mauritius
e-mail: y.chuttur@uom.ac.mu

N. Rawoteea

e-mail: nandishta.rawoteea@umail.uom.ac.mu

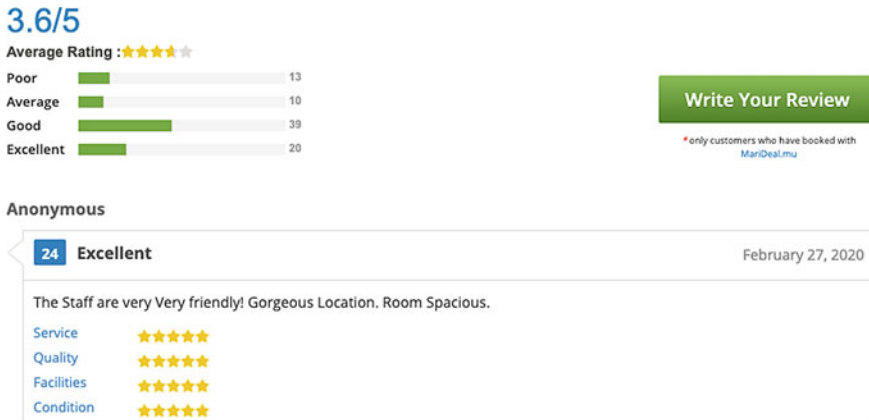


Fig. 1 Typical customer review page as seen on marideal.mu

the privilege of browsing and selecting the best package that suits their preferences and makes a hotel reservation within a couple of minutes anywhere and anytime they want. Online hotel reservation platforms have also become a popular means for customers to express their thoughts and experiences about their stay. Rightly so because as found by a recent study [1], around 81% of customers consult reviews of former customers' experience as a determining factor while making a booking decision.

In this context, this study has been conducted to analyze customer reviews using a machine learning technique known as Long Short-Term Memory (LSTM). The goal is to classify reviews into emotions to allow for a better understanding of customer experiences. Compared to sentiment analysis, which normally takes only two directions, namely positive and negative, we believe that emotional analysis may help detect from texts reviews the type of feelings such as love, joy, surprise, sadness, anger, and fear expressed by a customer. Those types of emotions, when recognized, may better serve the purpose of management in identifying the strengths and weaknesses of customer satisfaction strategies adopted.

We also note that online customer reviews consists of two main parts: (1) a general score in the form of star review (1 star to 5 stars) and/or experience category (good, excellent, average or poor) and (2) a comment section to let customers describe their experiences. An example from the website marideal.mu¹ is shown in Fig. 1.

Such an approach although useful to some extent does not necessarily provide the most accurate description of a customer's experience. For instance, while a customer may describe a service as *poor*, the same customer may provide a general description of his/her experience, which when analyzed may reveal that the experience was not in overall 'poor'. In other words, while a customer may have mixed experience, the latter cannot accurately express his/her experience with one score or experience category only. One way to address this problem is for decision makers to go over

¹<https://www.marideal.mu/>.

all text reviews in the comment section to analyze the actual experience for each customer. But such task is very tedious and time consuming. Instead, an automated approach that can read the emotions expressed within each comment would be useful.

Another problem we notice with current approach of collecting customer reviews is that decision makers are unable to keep track and monitor the trend of customer satisfaction, issue status and performances ranking over time. As shown in Fig. 1, decision makers have access to an aggregated score (here 3.6/5), which indicates the overall rating of the organization concerned. Such a score is useful for customers but not for managers and other stakeholders. This is because, it is not possible for managers to understand how their organizations have been doing over time in terms of strategies adopted to address customer complaints or other customer service issues. For instance, given the need to evaluate the percentage of positive/negative feedback received for a particular period of time, all the corresponding reviews for that period of time will have to be manually searched and analyzed. Moreover, it is difficult to analyze the frequency of complaints and investigate whether positive reviews have increased or decreased over previous months and years. Thus, with current approach of aggregating customer reviews (see Fig. 1), it is difficult to keep track of customer ranking over time.

Thus, with current approach of collecting customer reviews, it is difficult to acquire reliable and meaningful insights to support effective decision making. Addressing the problems mentioned earlier using emotion analysis to extract useful data would benefit both customers and hotels administrators. In the following sections, we present a background on online customer reviews and emotion analysis. We then present the architecture of our proposed emotion classification system for emotion analysis of online reviews. Along with a discussion of the implementation details of our system, we then present some results obtained from our classification model. We conclude the paper with future directions after an evaluation of our proposal performance.

2 Online Customer Reviews

With the advent of the world wide web, social networking site and online shopping, more and more people are sharing their opinions, reviews and comments about their experience relative to a product or a service they had access to [2]. In line with this, the ultimate success of a company, a product or a service is highly dependent on the ability for the firm to provide customer satisfaction [3]. In this section, we provide a background on the concept of customer reviews.

2.1 *Electronic Word of Mouth*

Customer reviews are a form of electronic Word of Mouth (e-WOM). In contrast to conventional Word of Mouth (WOM), eWOM has a stronger impact and reach as anyone having access to the Internet can access a review of another person anywhere, anytime. It is stated that there is approximately 2.4 billion active Internet users, who write and/or read on the Internet all around the world [4], and more and more people are gaining access to the Internet. This implies that when consumers are sharing their experiences through e-WOM, billions of users have access to these experiences and therefore are being influenced by the shared experience as opposed to traditional WOM, where the number of people being influenced is limited. We also find that one of the most impacted industries by the eWOM is the hospitality industry. Potential customers and travellers regularly look for information about a destination, hotel, restaurants and other services related to the hospitality industry before they make a decision [5].

2.2 *Reviews Classification*

Review contents can be classified in two categories [6], namely the *attribute-centric* type and the *benefit-centric* type. The difference between the two types of reviews lies in the way that the product is evaluated to support recommendations. For instance, in attribute-centric reviews, the review text will contain a description, which is based on the technical attributes of the product or service, such as capacity of hard drive, buffet contents [6]. Therefore, while the evaluations are subjective, their data are supported by objective data and descriptions. On the other hand, the benefit-centric reviews do not contain arguments, which support subjective evaluations about such technical attributes but rather are focussed on subjective experience only. For instance, a reviewer may describe how a product helped him/her in storing all his favourite movies on a hard drive, but does not actually provide the technical details of the hard drive capacity to support his/her review.

2.3 *Reviews Characteristics*

Reviews have two main types of characteristics, quantitative and qualitative. The quantitative characteristics of a review refer to 'star ratings' attributed and the length of the review made or review elaborateness [7, 8]. Researchers [7] view star ratings as numeric summary statistics or overall ratings, depicted as five-point star recommendations for the general assessment of the product. In fact, researchers [9, 10] have observed that there is a linear relationship between star ratings and consumer behaviour. In terms of review elaborateness, researchers [11] found that reviews differ

in length and that long reviews usually contain in-depth information about a product in terms of how and where the product was purchased and of its usage over time. The benefit that elaborate reviews bring is that they tend to alleviate the consumer's uncertainty regarding a product and its quality and therefore help the consumer to be confident in their purchasing decision process.

Qualitative characteristics of reviews focus on customer perceived enjoyment and review readability. With respect to perceived enjoyment, the term has been coined to define the extent to which reviews are perceived to be enjoyable [12]. It has the purpose of representing the consumer's emotion, that is, feeling of *pleasure, depression, disgust, rage, anger, happiness* or *hate* [13]. In terms of review readability; it is important that people who read the reviews are able to understand the written information about the product [14]. Understandability therefore is an important qualitative aspect, which depicts the extent to which customers will accept online review on review sites and other platforms.

2.4 Motivations for Providing Online Reviews

Studies have shown that the main factors which motivate consumers to write reviews through eWOM are social benefits, economic rewards, concerns for others as well as self-enhancement [15]. Social benefit involves the willingness of the customer for social incorporation and identification while economic rewards boil down to the remuneration received from online review sites or platforms, where the consumers share their experience. Concern for others expresses the wish of the consumer to recommend a positive experience or even to inform consumers of the bad experiences. Self-enhancement, on the other hand, refers to the need for recognition required by consumer and is often linked to economic rewards. Factors that motivate people to look for reviews include finding the best offers and value for money in terms of quality and convenience, seeking social reassurance and information about a hotel or destination and reducing the risk and uncertainty [16].

2.5 Impacts of Online Reviews for Companies

As reported in [17], there are seven main impact factors of eWOM from the perspective of an organization, namely quality control and new procedures, revenue management, customer interactions response and recovery, specific marketing strategies, focus on target communication, online reputation comparison and generating loyalty. The authors classified these seven impact factors into two categories: informational and revenue interest. The informational perspective involves the procedures that are required for managers to influence the information about the destination and quality control [18]. The data acquired can be utilized to obtain satisfaction from customers,

generate loyalty, solve their issues and monitor the organization's image and reputation. The engagement in positive eWOM is likely to result in more customers and therefore more business activity. In addition, eWOM plays a significant role in building the organization's reputation; that is, positive comments will elevate the organization's reputation and will have a positive effect on the image of the company while negative comments will have a bad effect on the company and may impact on the price competitiveness and profits.

Previous studies [18] found that eWOM can influence loyalty and cause customers to repeat their purchases as well as to recommend the product or the service to other potential customers. According to Ye et al. [19], there is a significant positive relationship between business performance of companies and positive online reviews. Therefore, managers can use these reviews as a marketing tool to interact with the customers in order to improve their relationship with their customers and achieve improved customer satisfaction, sales and revenue management. The impacts of eWOM can therefore be seen as opportunities for companies to review their strategies to obtain a competitive edge in their business, such as in the areas of quality service, branding and product development [17].

3 Emotion Analysis

As defined by the Oxford Dictionary, 'Emotion' is a 'strong feeling deriving from one's circumstances, mood or relationships with others'. Emotion is considered to be an extremely complicated and represents a multidimensional characteristic, which is embodied by humans in their daily life [2]. However, with the advent of technology and the World Wide Web (WWW), people express their emotions through text using different social networking systems. It is estimated that there are approximately 2.4 billion active Internet users, who write and/or read on the Internet all around the world [4]. Researchers from different fields such as psychology, business, computer science, affective computing and artificial intelligence have worked on the detection and analysis of emotions from text [2]. However, it is acknowledged that the complexity of human emotions and the ability to detect the correct emotion from text present complex challenges, which are yet to be addressed.

Indeed, emotion recognition is even more challenging when different emotions are expressed within a single portion of text. In addition to this, it is often seen that emotion in a text is sometimes so implicit that the ability to detect emotion automatically becomes compromised. Furthermore, sarcastic texts are sometimes difficult to understand by human themselves, making it more difficult for a computer to detect it [2].

Emotion analysis, also known as emotion mining, is widely recognized as being in its infancy and has a very long way to proceed [20]. Indeed, there are not many research works done in this area. One of the first studies related to affective computing brought about the concept of training computers to detect human emotions and explained why computers would be useful to recognize emotions [21] while various

applications of textual emotional-mining methods can be found in [20]. For instance, emotion mining can help gain information about customer satisfaction, whereby the owner can improve or revise their products or services in order to build a stronger relationship with the end user.

3.1 *Emotion Models*

Emotion models set forth the different criteria to make various emotions expressed by an individual measurable and distinguishable [22]. In particular, five emotion models, namely *discrete*, *dimensional*, *componential*, *circuit* and *appraisal model* have been reported in previous studies [23]. Prior to conducting emotion detection, it is important to choose a suitable emotion model, which will enable the researcher understand how emotions are explained and described and also provides the researcher with the knowledge required to appraise the textual data [24, 25]. Along with the appropriate emotion model, an appropriate emotion detection tool or technique should be selected for text analysis.

3.2 *Analytical Approaches of Emotion Detection from Text*

As noted by several researchers [26], emotion detection from text is usually done by a learning-based method. The learning-based approach makes use of a trained classifier in order to classify the target text into different emotion categories through the usage of keywords. This approach is easier and quicker to align to domain changes as it can learn new features quickly by inputting a large training set to a machine learning algorithm for building the classification model [27]. This approach, however, has the drawback of leading to not-so-clear boundaries between the emotion categories.

3.3 *Emotion Detection Algorithms*

The algorithms usually used to detect emotions from text that can be classified into three broad classes, namely rule-based approaches, non-neural machine learning approaches and deep learning approaches. With regards to deep learning approaches, the deep neural networks have had significant success in text, speech and image domains [28]. The variations of recurrent neural networks include long short-term memory networks LSTM and BiLSTM, and these have shown effectiveness in the modelling of sequential information [29, 30].

In addition, convolution neural networks have also been a successful method in the image domain. Introduction of convolution neural networks to text domain has shown that they are capable of deciphering abstract concepts from raw signals [31].

One of the approaches that CNNs use is the classification of emotional features. The limitations of the CNNs pertain to the fact that this approach is designed to analyze only a single dimension, and therefore, it is not so clear as to how it is able to generalize across multi-class predictions or regression tasks that form part of dimensional emotion models [32]. While this approach makes use of a 'deep' network, the network architecture handles texts of predefined size, which is similar to traditional machine learning. As opposed to the approach of CNNs, recurrent networks are able to iterate over sequences and can therefore handle textual data of arbitrary size. Some researchers have also used the novel deep learning approach known as Sentiment and Semantic LSTM (SS-LSTM) to detect emotions in text domains [33]. Yet another study [34] also used a similar way to resolve issues of emotion detection in English texts. The authors fed the input user utterance in the model into two LSTM layers through the usage of two different word embedding matrices. The first layers made use of a semantic word embedding while the second layer made use of a sentiment word embedding. Together, the two layers learn semantic and sentiment feature representation and encode sequential patterns in the user utterance.

4 Proposed Approach to Analyze Online Hotel Reviews

Current approach for collecting reviews online gives only a cumulative view of customers experience with a given organization's product or service. A single aggregate score or star rating does not give enough information about customers experience over a period of time. Given that decision makers can benefit from capturing emotions expressed by customers over any given period of time, we propose here to implement a system that can extract emotions from online reviews and graphically display those emotions as a time series. Our focus will be on hotel reviews available on the website *marideal.mu*, which is a popular site for booking hotels in Mauritius.

The architecture of our proposed system is shown in Fig. 2. Four main information processing stages are involved in our system, namely: data extraction, training of data, processing of data and finally classification of data. We classify each review under the emotions *love*, *sadness*, *joy*, *surprise*, *anger* and *fear* [23] and further include a cloud platform, whereby hotel decision makers may view and drill through the classified reviews to have a better insight about the performance of their hotels over a desired period of time. In this way, a better understanding of the experiences of customers during their stay can be obtained. Decision makers will also have the ability to view emotions reports in order to make improvements accordingly.

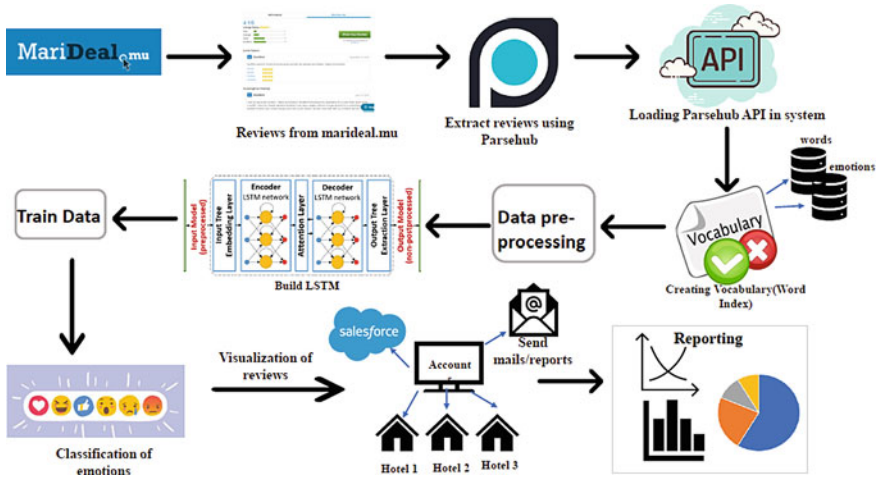


Fig. 2 System architecture of the proposed emotion classification and plotting system

5 Implementation, Results and Interpretations

We used ParseHub² to extract reviews from the source website *marideal.mu*. ParseHub is free and offers the necessary tools to easily select data needed for our analysis. To ensure that we have access to latest reviews online, we also include an implementation of ParseHub API, which is an extension for ParseHub implementation. To train our LSTM classifier, an emotion dataset containing 422,202 reviews labelled into six categories of emotions: ‘love’, ‘joy’, ‘surprise’, ‘sadness’, ‘fear’ and ‘anger’ were used [23]. We built the relevant LSTM model using keras³ library, which makes use of TensorFlow⁴ in the back end. TensorFlow acts as an abstraction layer hiding all the complexity behind the current implementation.

In the data pre-processing phase, extracted reviews were cleansed so as to remove unwanted or unnecessary information. More specifically, reviews were tokenized, and all brackets, punctuations and stop words were removed. Reviews were converted to lowercase, and dates of collected reviews were transformed to a consistent date-time format. Finally, classification of each review is done through our trained LSTM model. A total of 4270 reviews was classified for this study. Because a review may express various emotions, our system is able display all emotions detected in a set of reviews for a particular hotel.

For each emotion detected, a score is generated and is used to plot a chart or a graph. For the visualization part, we used the cloud platform offered by salesforce⁵

²<https://www.parsehub.com/>.

³<https://keras.io/>.

⁴<https://www.tensorflow.org/>.

⁵<https://www.salesforce.com/>.

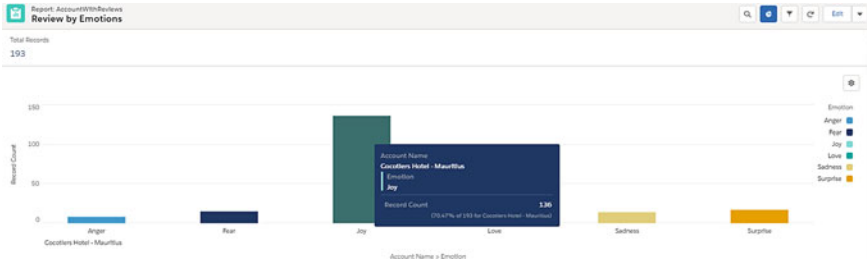


Fig. 3 Aggregate total emotions detected for one hotel

customer relationship management service. Results obtained for a particular hotel is shown in Figs. 3, 4 and 5. In Fig. 3, the latest aggregate emotions are displayed. It is seen that customers have expressed all emotions used to train our LSTM model. Among those emotions, however, we notice that most customers expressed joy as demonstrated by highest score obtained for that emotion class.

In Fig. 4, we demonstrate how a user can decide to view the emotion trend for a stated period of time as a time series. Here, we plot the emotions trend from the year 2015 to 2020 for the same hotel. Again, it is observed that customer reviews have mostly expressed the emotion ‘joy’, which surprisingly is seen to decrease from 2015 to 2020 (*graph reads 2020–2015 from left to right*). In other words, lesser customers have had the feeling of joy expressed in their reviews. Such kind of information could be useful to a decision maker to identify any issues with customer service and to take necessary remedial actions.

Lastly, as shown in Fig. 5, users may further drill down a set of classified emotions to obtain more details on the emotion trends detected for a given period of time. Here, the chart displays emotions for all the months for the year 2015–2020. It is easy to see that the emotion ‘joy’ has been the dominant emotion expressed by customers, but others such as ‘fear’, ‘anger’ and ‘sadness’ are also present throughout the same period. Decision makers may wish to dig further into those reviews to better understand how to improve customer service.



Fig. 4 Emotions variations per year for one hotel from 2015 to 2020

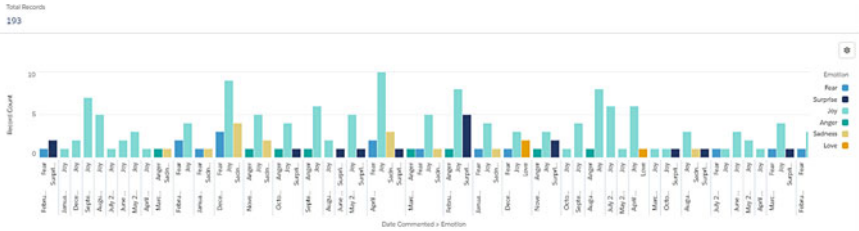


Fig. 5 Monthly breakdown of emotions observed for one hotel

6 Evaluation and Discussion

To evaluate the performance of our system, we calculated the *accuracy* and *F-Score* of our classification model. A total of 4270 reviews was considered for the evaluation. For each review, an overall emotion was manually assigned as actual emotion, and the emotion for which our classifier detected the highest score was considered as the classified emotion. The confusion matrix Table 1 shows the results obtained when plotting the classified emotions versus the actual emotions of the reviews used for evaluation.

Our classifier is found to have an overall accuracy of 76.8%. However, since the model used for our system is conducting multiple class classification on an imbalanced data, accuracy remains somewhat an approximate metric to evaluate the performance of our system. Instead, we measure the system’s performance using the F-score metric. Using the above confusion matrix, we obtain a recall value of 60.2% and precision value of 80.5%. Following which, we calculate an overall F-score of 68.9% for our system.

With a performance of 68.9%, we infer that our approach to classify emotions from online reviews is satisfactory but for which there is room for improvement. As raised by Wang [35], one avenue that can be explored would be to collect more data for training purpose so as to increase the performance of our classifier.

Table 1 Confusion matrix for classified emotion versus actual emotions

		Actual emotion					
		Love	Joy	Surprise	Anger	Fear	Sadness
Classified emotion	Love	231	4	5	2	1	3
	Joy	542	2316	202	25	16	35
	Surprise	6	4	145	22	21	13
	Anger	8	0	2	445	24	11
	Fear	12	0	17	0	81	1
	Sadness	1	2	0	3	7	63

7 Conclusion

In this paper, we presented the implementation of a system that can display customer emotions from online reviews over a given period of time. Our proposal makes of LSTM, which is part of the artificial neural network family. Our approach demonstrates how useful information about customer feelings can be extracted from rich source of online reviews. At the same time, we demonstrate that plotting emotions as a time series can uncover hidden information not available with conventional average scores used to identify an organization customer rating. We also proposed a novel approach to handle reviews at various aggregate levels, i.e. a decision maker has the flexibility of drilling through different period of time to obtain further insights into customer feelings. Our artificial neural network model, however, could have better performance. We expect that combining the LSTM model used in this study with other models such as CNN may likely improve the performance of our emotion classifier.

References

1. Snehasish, B., Alton, Y.K.: StatisticBrain: trust in online hotel reviews across review polarity and hotel category. *J. Comput. Human Beh.* **90**, 265–275 (2019)
2. Sailunaz, K., Alhaji, R.: Emotion and sentiment analysis from twitter text. *J. Comput. Sci.* **36**, 1–18 (2019)
3. Patterson, P.G., Cowley, E., Prasongsukarn, K.: Service failure recovery: the moderating impact of individual-level cultural value orientation on perceptions of justice. *Int. J. Res. Mark.* **23**(3), 263–277 (2006)
4. Louis, F., Wojciech, C.: Advances in the human side of service engineering. In: 5th International Conference on Applied Human Factors and Ergonomics, Volume Set, Proceedings of the 5th AHEE Conference 19–23 July 2014
5. Lee, J., Park, D., Han, I.: The effect of negative online consumer reviews on product attitude: an information processing view. *Electron. Commer. Res. Appl.* **7**(3), 341–352 (2008)
6. Park, D.H., Kim, S.: The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. *Electron. Commerce Res. Appl.* **7**(4) (2008)
7. Willemsen, L.M., Neijens, P.C., Bronner, F., de Ridder, A.J.: Highly recommended! the content characteristics and perceived usefulness of online consumer review. *J. Comput. Mediated Commun.* **17**(1), 19–38 (2011)
8. Racherla, P., Friske, W.: Perceived usefulness of online consumer reviews: an exploratory investigation across three services categories. *Electron. Commer. Res. Appl.* **11**(6), 548–559 (2012)
9. Clemons, E., Gao, G., Hitt, L.: When online reviews meet hyper differentiation: a study of the craft beer industry. *J. Manage. Inf. Syst.* **23**(2), 149–171 (2006)
10. Park, D.H., Lee, J., Han, I.: The effect of on-line consumer reviews on consumer purchasing intention: the moderating role of involvement. *Int. J. Electron. Commerce* **11**(4), 125–148 (2007)
11. Mudambi, S.M., Schuff, D.: What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quart.* **34**(1), 185–200 (2010)
12. Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: Extrinsic and intrinsic motivation to use computers in the workplace. *J. Appl. Soc. Psychol.* **22**(14), 1111–1132 (1992)

13. Triandis, H.C.: Values, attitudes, and interpersonal behavior. In: Howe, H., Page, M. (eds.) Nebraska Symposium on Motivation, pp. 195–295. University of Nebraska Press, Lincoln, NE (1980)
14. Zakaluk, B.L., Samuels, S.J.: Readability: Its Past, Present, and Future. International Reading Association, Newark (1988)
15. Hennig-Thurau, T., Walsh, G.: Electronic word-of-mouth: motives for and consequences for reading customer articulations on the Internet. *Int. J. Electron. Commer.* **8**(2), 51–74 (2004)
16. Cantalops, A.S., Salvi, F.: New consumer behavior: a review of research on eWOM and hotels. *Int. J. Hospital. Manage.* **36**, 41–51 (2014)
17. Zhang, Z., Ye, Q., Law, R., Li, Y.: The impact of e-word-of-mouth on the online popularity of restaurants: a comparison of consumer reviews and editor reviews. *Int. J. Hospital. Manage.* **29**(4), 694–700 (2010)
18. Loureiro, S.M.C., Kastenholz, E.: Corporate reputation, satisfaction, delight, and loyalty towards rural lodging units in Portugal. *Int. J. Hospital. Manage.* **30**(3), 575–583 (2011)
19. Ye, Q., Law, R., Gu, B., Chen, W.: The influence of user-generated content on traveler behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Comput. Hum. Behav.* **27**(2), 634–639 (2011)
20. Yadollahi, A.G., Shahraiki, O.R., Zaiane, A.: Current state of text sentiment analysis from opinion to emotion mining. *ACM Comput. Surv.* **50**(2), 25 (2017)
21. Picard, R.W.: Affective Computing. Massachusetts Institute of Technology (1997)
22. Scherer, K.R., Wallbott, H.G.: Evidence for Universality and Cultural Variation **66**, 310–328 (1994)
23. Kim, K.: Emotion Modeling and Machine Learning in Affective Computing, pp. 1–12 (2014)
24. Canales, L., Martínez-Barco, P.: Emotion detection from text: a survey. In: 11th International Workshop on Natural Language Processing and Cognitive Science—NAACL, pp. 1–8 (2014)
25. Binali, H., Chen, W., Vidyasagar, P.: Computational approaches for emotion detection in text. In: 4th IEEE International Conference on Digital Ecosystems and Technologies, pp. 172–177 (2010)
26. Haggag, M., Samar, F., Nahla, E.: Ontology-based textual emotion detection. *Int. J. Adv. Comput. Sci. Appl.* **6**(9), 239–246 (2015)
27. Ivica, J.: Interdisciplinary approach to emotion detection from text. *XA Proc.* **1**(1), 34–72 (2018)
28. Eiman, K., Eman, M.G.Y., Chee, S.A.: Deep learning analysis of mobile physiological, environmental and local sensor data from emotion detection. *Int. J. Inf. Fusion* (2018)
29. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
30. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *Signal Process. IEEE Trans.* **45**(11), 2673–2681 (1997)
31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, Lake Tahoe, 3–6 Dec 2012, pp. 1097–1105
32. Kratzwald, B., Feuerriegel, S.: Putting Question-Answering Systems into Practice: Transfer Learning for Efficient Domain Customization (2012)
33. Gupta, U., Chatterjee, A., Srikanth, R., Agrawal, P.: Sentiment-and-Semantics-Based Approach for Emotion Detection in Textual Conversations. Nei-IR, Japan (2017)
34. Chatterjee, A., Narahari, K.N., Joshi, M., Agrawal, P.: EmoContext contextual emotion detection in text. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 39–48 (2019)
35. Wang, L., Xia-kang, W., Juan-juan, P., Jian-qiang, W.: The differences in hotel selection among various types of travellers: a comparative analysis with a useful bounded rationality behavioural decision support model. *L. Tour. Manage.* **76**, 103961 (2020)

Sentiment Analysis Using Deep Learning for Recommendation in E-Learning Domain



Rawaa Alatrash, Hadi Ezaldeen, Rachita Misra, and Rojalina Priyadarshini

Abstract Sentiment analysis (SA) is one of the methods that can assist in extracting information from a large amount of data. It is considered one of the research fields in text mining, which has become vital to employ within recommendation systems, as well as in e-learning environments. In the current work, we present a new method of recommendation model utilizing sentiment analysis based on convolutional neural network (SABCNN) and natural language processing (NLP) techniques. Starting from collecting and analyzing the learners' sentiments of reviews for the e-content with their corresponding rating within e-platforms, a sentence or a specific text is classified to multi-levels by determining what semantics of feelings it holds. Our research aims towards recommending learning resources that are relevant to the learners' preferences with the aid of the previous reviews of other learners, sharing him/her the top preferences.

Keywords Recommender system · E-learning · Sentiment analysis · Deep learning · Natural language processing (NLP) · Word Embedding

R. Alatrash (✉) · H. Ezaldeen
Department of Computer Science and Engineering, C.V. Raman Global University, Bidyanagar,
Mahura, Janla, Bhubaneswar, Odisha 752054, India
e-mail: rawaa.alatrash@gmail.com

H. Ezaldeen
e-mail: hadi.talal@gmail.com

R. Misra · R. Priyadarshini
Department of Computer Science and Information Technology, C.V. Raman Global University,
Bidyanagar, Mahura, Janla, Bhubaneswar, Odisha 752054, India
e-mail: rachita.dhunu.misra@gmail.com

R. Priyadarshini
e-mail: rojalinapriyadarshini@cvrce.edu.in

1 Introduction

Humans are emotional in nature and opinions matter to them. The analysis of sentiments (emotions) aims to build a system that analyzes what the individual wants of a product, subject or event, in that people express their opinions as a text in form of reviews or comments. Exploration of these opinions available to the user is a difficult process, but in practice, it is very useful [1]. With the aid of current studies of sentiment analysis, it is divided into three levels, i.e., attribute level, document-level, and sentence level [2]. The aim of sentence-level is to classify sentiments in each sentence; the first task is to identify the expression behind the sentence whether it contains the entity under scrutinizing (object), or contains explicit words representing the opinion of that entity (subjective), as exploration of opinions and the sentiment analysis fell into scope of NLP [3]. What matters in this research is exploration of opinions for the purpose of deducing an idea about a particular subject (book) that positively affects learners and is reflected in their opinions and reviews. In our model, the ratings of reviews about the books are predicted by sentiment analysis methods that we have used, books with higher ratings will be then recommended to the learner according to his/her education domain.

1.1 *Incorporating Emotions to Support Adaptive E-Learning*

Many institutions aspire flexibility in education through combination intelligent techniques and approaches to develop adaptive e-learning systems [4, 5].

Content analyzer system (edX-CAS) has been developed in [6] to enhance the adaptability of students learning experience, a tool that uses natural language processing techniques to detect the feelings, opinions, and emotions reflected in the online learning material courses, such as “Massive Open Online Courses (MOOCs) and Small Private Online Courses (SPOCs).” The objective is to detect factors such as subjectivity and polarity (positivity or negativity) of the content of the courses and the emotional opinion related, then to propose, where appropriate, recommendations to the corresponding teaching-staff for improvement of the content.

On the other hand, the authors in [7] have implemented a hybrid approach in integration with topic revelation with student interest. It combines lexical-based machine learning models to infer the emotional state of users from the comments accessible to SentBuk application in Facebook, which can be used to support personalized e-learning. So that the sentiment of the student across a particular course can be used as feedback for the tutors. This approach, based on the analysis of feelings, allows to extract information on the polarity of the feelings transmitted by the users in the messages they write (negative, neutral or positive), and to model the polarity of the habitual feeling of each user and detect significant emotional changes. Thus, the adaptable information can be involved in e-learning system for recommending. The most convenient actions to be handled each time.

Knowing the emotional state of learner is useful to expand the possibilities of adaptive e-learning. It contributes to know his/her potential needs at certain time, selection of content based on the emotional state of the learner at that time.

1.2 *Sentiment Analysis Techniques*

One of the most important techniques that has achieved brilliant results on the task of classifying the sentence is convolutional neural networks (CNN) [8, 9], which was invented for computer vision, then used on the field of search query retrieval [10], sentence modeling [11], semantic parsing [12], and other conventional NLP missions [13].

Furthermore, these NLP models need to determine model's parameters such as the style of the network, hyper-parameters, features size, regularization, and so on. Here, it is very important to take into account the sensitivity analysis which is affected by the architecture components; also, the performance of the model relies on the changes in the configurations to classify the sentence. Many design decisions of the deep networks are to identify what is important or relatively inconsequential. Because of the relative simplicity and powerful performance, the authors [14] concentrated on the CNNs with single-layer (excluding the sophisticated one) which forms a baseline model that is new norm similar to support vector machine (SVM) and logistic regression.

The authors in [15] have compared three different machine learning (ML) techniques such as multinomial naive bayes (MNB), linear support vector machine (LSVM), and long short-term memory network (LSTM) in order to analyze the sentiment of the sentences so that they can classify the sentence into 2 classes (Positive and Negative). They have chosen randomly samples for training and testing out of a big dataset of reviews, which is verified and publicly available from Amazon that included a different user's reviews from Amazon.com. In addition, they created their own dataset from user's reviews of Amazon products of different categories with their rating. They showed that the classification process of sentiment reviews doesn't take into account where this review comes from or its species.

There can be cases of mismatch between the user's review and its corresponding rating given by the user. A model has been developed [16] using deep learning on "Amazon.com product review data" to reveal the reviews with the ratings mismatched, where a web service provides suitable feedback. The sentiment analysis process determines whether the review has positive or negative emotions. They train a "Recurrent Neural Network with gated recurrent unit," using reviews' text that has been translated into vectors by using a paragraph vector. In addition, the product information has been incorporated with the semantic relationship of review.

The motivation has been derived from [8], to achieve the most benefit of CNNs for the task of sentence classification in real-world settings, and the findings of our experiments in this work have been relied on to derive practical advice.

In Sect. 2, outline of SABCNN is presented. Section 3 discussed the proposed model. Experiments, results, and discussion are given in Sects. 4 and 5, respectively. Section 6 concludes the paper along with some future research directions.

2 Outline of SABCNN

SABCNN can be addressed with two processes, which are: the training process and the recommendation process, as shown in (Fig. 1), where the training process can provide with CNN parameters learned to be the stepping stone for the recommendation process, as follows.

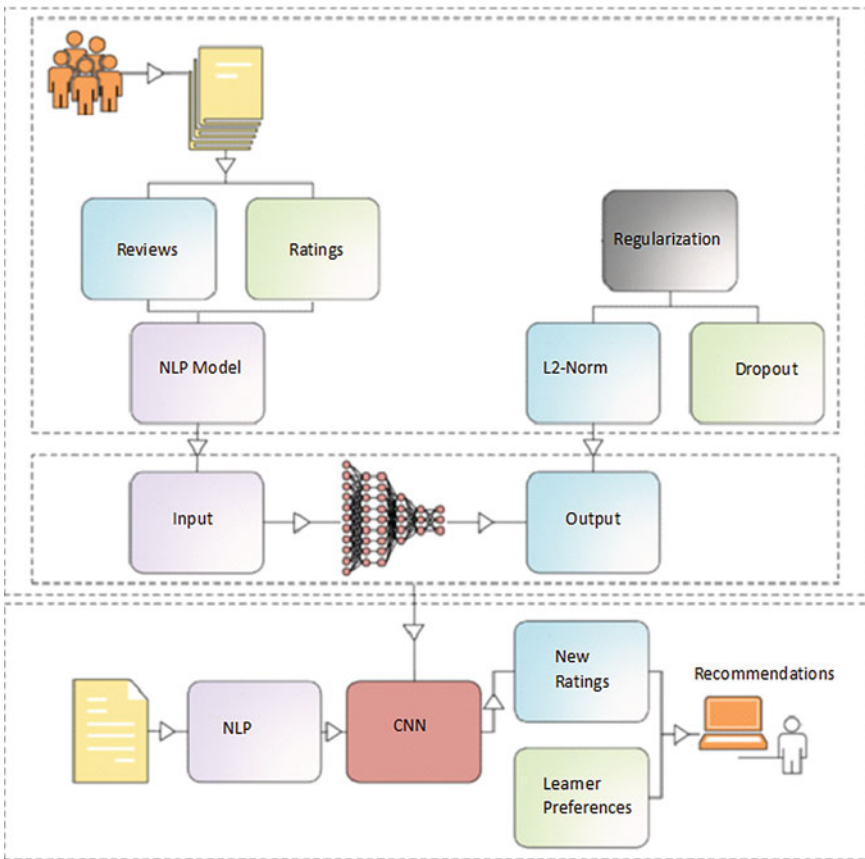


Fig. 1 Architecture of our proposed recommendation model (SABCNN)

2.1 Training Process

The training data for CNN are the review's sentences labeled by the rating scores corresponded to the learning resources, which are made by the learners. The input data for the language model are the reviews' text information of the existing learning resources where the target is the rating score. The output is regularized by L2-norm and dropout technique [17].

2.2 Recommendation Process

The predicted rating score can help to classify the obtained reviews of the learning resource, which indicates whether it achieves a good rating and is needed by the learner or not. To obtain the final predicted rating of the learning resource linked to the scope of learner's interest, we apply the mean of all predicted ratings extracted of the reviews regarding of this learning resource. Then, we select {Q} learning resources with the top ratings computed for all the learning resources linked to the scope of learner's interest to be then recommended to the learner. The proposed recommendation algorithm can be used as a new recommender system, and it can also be utilized to enhance an existing recommendation approach within the e-learning system.

3 The Proposed Model

In the following section, the construction of SABCNN is described, and shows how one can use CNN to predict the ratings from input data provided by learners, which depict their emotions about learning resource in terms of text entries. The CNN input is made using Word Embedding techniques involved in NLP according to the text entries, the output deduces the ratings of the e-content which are then validated whereon to be the base of the recommendation phase. Figure 1 elucidates this architecture, and the details are in Sect. 3.3.

3.1 Data Sets

We have created our own dataset using scraping of different books' reviews from Amazon.com using python script. We collected around 15,000 records which includes the customers' reviews to books—with varying terms such as {"Machine Learning," "Deep learning" ..., etc.}—with their corresponding ratings. Our dataset (DS) comprises of the reviews for books based on book's rating score from "5" to

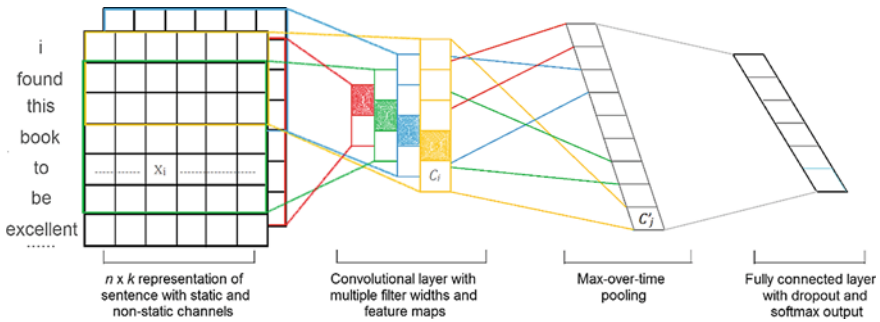


Fig. 2 CNN model construction with two channel for an example sentence

In the model, the input data is from the representation of the words, which is to be going from Embedding space. Each sentence in the corpus was represented where $x_i \in \mathfrak{R}^k$ is the k -dimensional word vector, corresponding to the i th word in a sentence of length n , which is represented by concatenating as:

$$x_{1:n} = [x_1, x_2, \dots, x_n], \quad \text{Where } x \in \mathfrak{R}^{nk} \tag{1}$$

The process of convolving a filter $w \in \mathfrak{R}^{hk}$ with a kernel size of v words is applied to generate a new feature through a convolution operation is involved in each convolutional layer. Formally, Let c_i a feature is extracted from a window of words $x_{i:i+v-1}$ as follow:

$$C_i = f(w \cdot x_i + b) \tag{2}$$

where $b \in \mathfrak{R}$ denotes to a bias expression, and f refers to non-linear activation function in which we used the “ReLU function.” This filter is convolved to each potential window of words $\{x_{1:v}, x_{2:v+1}, \dots, x_{n-v+1:n}\}$ to produce a feature map:

$$C = [C_1, C_2, \dots, C_{n-v+1}], \quad \text{where } C \in \mathfrak{R}^{n-v+1} \tag{3}$$

The later layer is a pooling process in that max-pooling over time takes the highest value $C' = \text{Max}\{C\}$ from each feature map. This pooling process usually treats with variable lengths of the sentence which has been used to generate one feature from one filter. Various filters are applied to obtain different features, using verities of kernel sizes. This convolutional scheme produces a vector $z = [C'_1, C'_2 \dots, C'_j]$, here j refers to the window size number. These features obtained from the penultimate layer which then fed to a fully connected Softmax layer. Finally, the probability of distributed classes over the output.

The output set corresponds to a probable class rotatory wherein the categories number is five classes as mentioned in Sect. 3.1. Back-propagation algorithm is then used to further tune the CNN parameters through the “gradient of the cross-entropy

loss function.” To deal with overfitting problem or co-adaption through L2-norm and dropout technique [18], during forward propagation in training process, we applied a ratio of dropping out “ p ” randomly in the penultimate layer of the hidden units, wherein “Element-wise multiplication operator” is applied on the output unit with using, a vector with binary values that is Bernoulli random variables which includes the probability to be 1, and the dropout rate applied is set to 0.5. While through testing process, the learnable vectors of weight are reformed by “ p ” wherein $\hat{w} = pw$, then \hat{w} is applied to estimate the testing sentence.

4 Experiments

Our focus is on predicting the books with the highest rating. We have done several experiments on our customized dataset mentioned in Sect. 3.1. Two ways were applied which are involving under Word Embedding; either to train the word vectors during the training process from scratch that is presented in the random model wherein each word maps into an integer, or to use the pre-computed vectors (GloVe) [19] and here the word maps into vectors. Similar terms are grouped closely, and dissimilar are grouped far away by using “Euclidean or Cosine distance” in their Embedding space. Also, two options are there during the pre-computed vectors procedure in which the word vectors are; either kept frozen without training that is presented in the static model, or left to fine-tuning that is presented in non-static model.

Practically, we adopt the dataset involving pre-trained “GloVe” word vectors that are obtained by using “aggregated global word-word co-occurrence statistics from a corpus” to train our CNN model. This method was vital to improve the performance, and we also adopted randomly initializing way to generate word vectors for the words not presented in the series of pre-computed. The dimensionality of pre-trained vectors have is “200” and that has been trained using “log-bilinear model with a weighted least-squares objective” [19]. In the experiments, we employed several variants of the CNN models, differing regarding their construction, as follow:

- CNN-random: In this baseline model, we randomly initialized the word vectors as an integer. During training, these vectors were got slight modification.
- CNN-static: “GloVe” word vectors were adopted to pre-train this model. All words vectors were kept static whereas the model’s other parameters were tuned.
- CNN-non-static: “GloVe” word vectors were adopted and then fine-tuned for each task of this model.
- CNN-two-channel: We customized two groups of word vectors initialized by “GloVe,” where each group was considered as a “channel,” then we applied each filter on these two channels. The gradients are done for only one channel through back-propagation, and fine-tune operation is conducted for one group, while another group is left static.

We used four convolutional filters (3, 4, 5, 6) with 250 feature maps each, L2-norm constraint of 3, and mini-batch size of 55 is used to extract the reviews' categorization during the training process. From the training set, 10% is randomly chosen for the "dev set." The training experiments is conducted by using "Stochastic Gradient Descent" over "Shuffled Mini-Batches" with the "Adam Optimization" [20]. And these parameters were optimized utilizing "random search technique." These models have been performed using python programming that utilized "Keras Package," as well as the parameters that have been depicted.

5 Results and Discussion

Our experimental results are obtained from implementing the four models described in Sect. 4. The random model shows accuracy of 0.72, while the non-static model performed a higher accuracy value of 0.74, but both models have given overfitting. Whereas the static model shown lowest accuracy but with less overfitting. In order to get better accuracy and acquire lower overfitting, we combined the two methods; pre-computed vector (Static) and fine-tune (Non-Static) in a two-channel model which resulted in low overfitting and accuracy of 0.77 (Fig. 3). The accuracy obtained through the four models is presented in Table 1.

From the empirical results, improved performance is we observed with promising results despite implementing the sentiment analysis on the text reviews labeled with 5 sentiment classes. This provides a more forward-looking and enhanced process of sentiment analysis, whereas most current researches were conducted to determine positive and negative polarity of the sentiment analysis of text reviews.

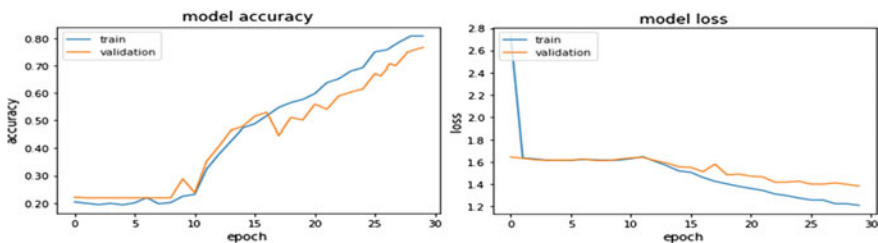


Fig. 3 CNN-two-channel model accuracy and loss

Table 1 Performance comparison

The model	CNN-random	CNN-static	CNN-non-static	CNN-two-channel
Accuracy	0.72	0.47	0.74	0.77

6 Conclusion

Within e-learning platforms, it becomes vital to adopt an efficient recommender system to achieve a learner's efficient learning experience. Our proposed model analyzes the users' reviews of the books, and when there is new data, the model must predict its usefulness, which leads to recommending better books for the learner based on top ratings predicted by the model. To enrich the e-learning platform, we used our customized dataset comprising of users' reviews labeled with 5 classes. Various CNN models with respect to the basic architecture were implemented to classify the reviews in a range of 5 classes, along with Word Embedding for learning representation of text. Remarkable results were obtained regarding multiple classifications performance, practically through the multichannel model, wherein most of the researches focuses on the two-polarity in the classification of sentiment analysis. Future work will be conducted in multichannel with "POS tag" with more efficient deep networks, and to improve the datasets. A hybrid e-learning recommender system with implicit learner profile built using intelligent algorithms can be one of the outcomes.

References

1. Patil, P., Yalagi, P.:Sentiment analysis levels and techniques: a survey. *Space* **1**, 6 (2016)
2. Bhati, R.: A survey on sentiment analysis algorithms and datasets.*Rev. Comput. Eng. Res.* **6**, 84–91 (2019). <https://doi.org/10.18488/journal.76.2019.62.84.91>
3. Bhati, R.G.:A deep literature survey on sentiment analysis.*CLIO Annual Interdisc. J. Hist.* **6**(5), 349–353 (2020)
4. Wang, K., Zhang, Y.:Topic sentiment analysis in online learning community from college students.*J. Data Inf. Sci.* **1**.ahead-of-print (2020)
5. Mite-Baidal, K., Delgado-Vera, C., Solís-Avilés, E., Espinoza, A.H., Ortiz-Zambrano, J., Varela-Tapia, E.:Sentiment analysis in education domain: a systematic literature review.In: *International Conference on Technologies and Innovation*, pp. 285–297. Springer, Cham (2018)
6. Cobos, R., Jurado, F., Blázquez-Herranz, A.: A content analysis system that supports sentiment analysis for subjectivity and polarity detection in online courses. *IEEE Revista Iberoamericana De Tecnologías Del Aprendizaje* **14**(4), 177–187 (2019)
7. Ortigosa, A., Martin, J., Carro, R.M.:Sentiment analysis in Facebook and its application to e-learning.*Comput. Human Behav.* **31**, 527–541 (2014)
8. Kim, Y.:Convolutional neural networks for sentence classification (2014). [arXiv:1408.5882](https://arxiv.org/abs/1408.5882)
9. Johnson, R., Zhang, T.:Semi-supervised convolutional neural networks for text categorization via region embedding.In: *Advances in Neural Information Processing Systems*, pp. 919–927 (2015)
10. Shen, S., Yelong, X.H., Gao, J., Deng, L., Mesnil, G.:Learning semantic representations using convolutional neural networks for web search.In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 373–374 (2014)
11. Kalchbrenner, N., Grefenstette, E., Blunsom, P.:A convolutional neural network for modelling sentences (2014).[arXiv:1404.2188](https://arxiv.org/abs/1404.2188)
12. Yih, W.-T., He, X., Meek, C.:Semantic parsing for single-relation question answering.In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, Short Papers, pp. 643–648 (2014)
13. Ronan C., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch.*J. Mach. Learn. Res.* **12**, 2493–2537 (2011)

14. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification (2015). [arXiv:1510.03820](https://arxiv.org/abs/1510.03820)
15. Güner, L., Coyne, E., Smit, J.: Sentiment analysis for Amazon.com reviews. *Big Data in Media Technology (DM2583)*, KTH Royal Institute of Technology, Stockholm. <https://doi.org/10.13140/RG.2.2.13939.37920.2019>
16. Shrestha, N., Nasoz, F.: Deep Learning Sentiment Analysis of Amazon.com Reviews and Ratings (2019). [arXiv:1904.04096](https://arxiv.org/abs/1904.04096)
17. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors (2012). [arXiv:1207.0580](https://arxiv.org/abs/1207.0580)
18. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from over fitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
19. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014)
20. Kingma, Diederik, P., Ba, J.: Adam: a method for stochastic optimization (2014). [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).

Selection of Best K of K -Nearest Neighbors Classifier for Enhancement of Performance for the Prediction of Diabetes



Subhash Chandra Gupta and Noopur Goel

Abstract Today, getting a meaningful information from an ocean of data obtained from numerous sources becomes very tedious work. A number of methods are applied to analyses various types of results obtained from them. Classification is one of them. Machine learning algorithms are used to train classifiers and enhance the capability of different classification methods. In this research paper, focus is given on K —nearest neighbor algorithm and has been analyzed its performance to classify patients in diabetic and non-diabetic class using Pima Indian Diabetes Dataset obtained from UCI repository. Diabetes mellitus, which is a metabolic disorder of human body, is one of the major health threats in world. Due to it, body becomes unable to consume insulin properly released by pancreas gland. Experimental analysis has been performed on PIMA dataset using python. A detail study is made on it using KNN algorithm with different values of K and identified the best value of K on which KNN returns best result. KNN provide a better accuracy 81.17% after applying feature selection method.

Keywords Diabetes mellitus · K -nearest neighbors · Feature selection · ANOVA · F-measures

1 Introduction

Whenever a human body becomes unable to manage the proper secretion of hormone “Insulin,” a metabolic disorder is arised, is known as diabetes mellitus. Diabetes mellitus is the result of insufficient production of insulin from pancreas in human body. In normal situation, pancreas secretes a hormone called insulin which controlled the blood glucose level. When blood glucose levels is increased after eating food or some other reasons, pancreas secretes insulin to make glucose level normal

S. C. Gupta (✉) · N. Goel
VBS Purvanchal University, Jaunpur, India
e-mail: csubhashgupta@gmail.com

N. Goel
e-mail: noopurt11@gmail.com

in blood. The inefficient production or absence of insulin causes hyperglycemia in diabetic patients. If diabetes is not diagnosed and treated properly in early stage, it causes major implications like cardio vascular disease, visual impairments, renal failure, and leg amputation delay in wound healing, numbness or tingling in hands or legs are some other symptoms of seriousness of diabetes. The types of diabetes are

In Type-1 diabetes, which is also known as Juvenile-onset diabetes or insulin-dependent diabetes (IDDM), pancreas is damaged due to the attacks on beta cells. Glucose level is increased in the blood due to the less secretion of insulin by damage pancreas. To control type-1 diabetes, patient has to rely on external insulin. Type-2 diabetes is also known as adult-onset diabetes or non-insulin-dependent Diabetes Mellitus [1]. Type-2 diabetes is the results of the resistance of body cells against insulin. In it, body cells becomes unable to consume produced insulin and it causes the pancreas to produce more insulin [2]. The over- functioning of pancreas is badly affected by it, and the secretion of insulin is stopped and glucose level is increased in body.

The National Diabetes and Diabetic Retinopathy survey Report 2019 [3] released by the Health and Family Welfare Ministry says that the prevalence of diabetes in India is at 11.8% in all age groups. This report has been prepared by Rajendra Prasad Centre for Ophthalmic Sciences, All India Institute of Medical Sciences, New Delhi during 2015–2019. This report has made a detail study about prevalence of diabetes in different age groups. A summarized view is given in following: (Fig. 1).

The objective of this paper is to analyze the performance of KNN classifier with different number of neighbors (K) and also find the optimal value of K for KNN classifier. The paper is divided into following sections: The first and current section

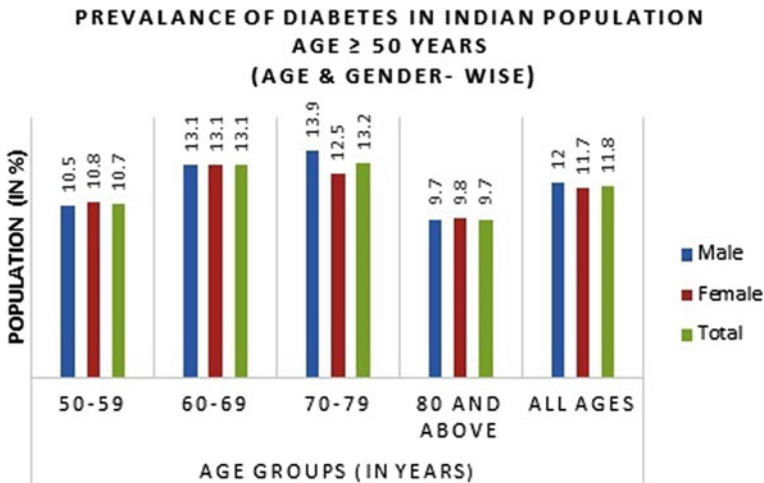


Fig. 1 Statistical facts about diabetes in India (*Source* National diabetes and diabetic retinopathy survey report 2019 [4])

are introduction, next section is of literature review, third section is about classifier, fourth section is the experimental work conducted on PIMA dataset, and last section is the conclusion.

2 Literature Review

Sneha et al. [5] has worked on a prediction model and used SVM, random forest, Naïve Bayes, decision tree, and KNN classifier for early detection of Diabetes Miletus disease. Among these classifiers, KNN shows 63.04% accuracy. Aiswarya Iyer et al. [6] applied J48 decision tree and Naive Bayes classification method on PIMA Indian Diabetes Dataset and find Naive Bayes as best scorer for their model. Implementation was done by WEKA tools. Aishwarya Jakka et al. [7] made a comparative analysis of performance analysis of KNN, decision tree, Naïve Bayes, support vector machine, logistic regression, and random forest classifiers using PIMA Indian Diabetes Dataset. They find the best accuracy level of 77.6% from logistic regression while KNN classifiers shown 73.43% accuracy.

Sengamuthu et al. [8] proposed a disease prediction model which was implemented by WEKA and MATLAB tools. They used a number of classification algorithms and got best accuracy from modified J48 classifier. Kavita Mittal et al. [9] has used cancer dataset obtained from UCI machine learning repository to make prediction about positive patient using K -nearest neighbors classifier. She further made a comparison of KNN with K -means algorithm. In [10], Amina et al. used WEKA tools on diabetes dataset to do their research work. They converted numerical value in categorical values during data preprocessing and apply KNN, decision tree and Naïve Bayes classifiers and got best accuracy by decision tree. Suresh Kumar et al. [11] divided a diabetes dataset in three clusters after preprocessing of original dataset. KNN approach was used to making the clusters. Further, they applied random forest, C 4.5 and Naive Bayes classification methods for prediction and got best result from C 4.5 classifier.

3 K-Nearest Neighbors Algorithm

K -nearest neighbors is a machine learning algorithm and is applied to solve classification and regression problem. It is a supervised learning algorithm and non-parametric since it has no information about the distribution of data. It is one of the simplest classification methods in comparison to other machine learning classification methods. It is simple, easy to implement and still produces better results in some manner. The classification model is built on the entire dataset in KNN because it uses whole dataset during testing phase [6].

K -nearest neighbor techniques are based on the concept of similarity measures (also called distance or closeness) within data items of dataset to classify new data

objects [5]. The similarity measures assign some weights to the neighbor's contribution. The weight given by the "K" neighbors plays important role for the classification of data in new class label. The value of "K" defines the number of selected data items from the nearest neighbors. The output class label is the class / group of neighbors from which the data item is the most closest. The distance metrics between two data points of a dataset are calculated by either of these methods [12]. The main advantage of KNN is that model building is not required for it. The performance of KNN method is affected by a number of factors such as—dimensionality of data points, the value of K, and data distribution. The performance of KNN is less reliable for a dataset having higher number of attributes. The reason behind it is the calculation of distance metrics. Generally, distance metrics is calculated by Euclidean distance measure which is affected by the curse of dimensionality. Choosing the right value for K is tricky but a right value makes a balance between under fitting and overfitting. But there is some problem with KNN. As the size of dataset increases, KNN becomes slow and makes it impractical choice when rapid prediction is required.

4 Experimental Process on Diabetes Dataset

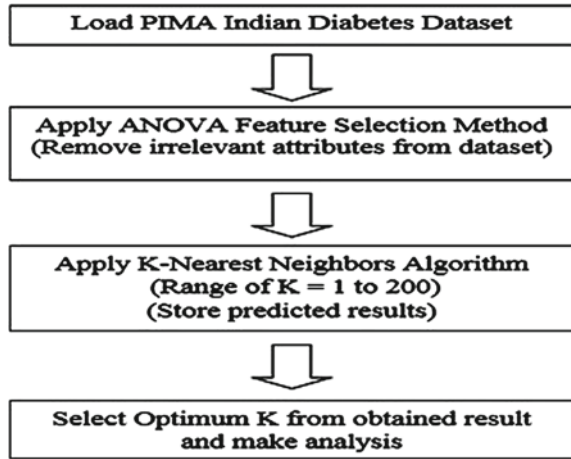
4.1 Work Flow of Proposed Model

The objective of research is to apply KNN machine learning algorithm on PIMA dataset and analyzes its performance on different numbers of neighbors. The proposed model is performed in following steps:

Step 1: Load PIMA Diabetes Dataset—The diabetes dataset is obtained from the source UCI machine learning repository [13] and available free in public domain repository. The dataset contains records of 768 patients and each record stores patient's data on 9 different test parameters. The attributes of dataset are pregnancy, glucose, blood pressure, skin thickness insulin, BMI, DPF, age, and outcome. The dataset contains one extra attribute called "Outcome" which store binary (0, 1) values. The positive diagnosis of patient has been stored by "1" and the negative diagnosis is by "0." The PIMA dataset contains data about 268 positive and 500 negative diagnosed samples.

Step 2: Apply Optimal Feature Selection Method—Irrelevant attributes of dataset are identified by optimal feature selection method and remove them from dataset. The dataset which having a number of attributes can slow the process of formation of training model and it may require a large chunk of memory. The performance of prediction model may also down. In this research article, ANOVA is used for feature selection method. ANOVA is a filter-based statistical selection method which is based on correlation statistics. For ANOVA implementation, the Python SciPy library provides `f_classif()` method along with `SelectKBest()` filtration method. The dataset contains 768 records with 9 attributes. The ANOVA f-measures score of

Fig. 2 Work flow of proposed model



eight attributes of are 213.1617, 71.77, 46.14, 39.67, 23.88, 13.28, 4.00 and 3.25, respectively. ANOVA F measure score of attribute skin thickness and blood pressure are less than 10. So, these irrelevant attributes are removed and dataset is reduced in 768×7 shape (Fig. 2).

Step 3: Apply KNN Algorithm—Classification of positive and negative class level is performed by K nearest neighbors techniques taking different value of K (range—1–200).

Step 4: A comparative analysis is performed on results obtained from different value of “ K ” and declare the best K for KNN on dataset.

5 Result and Discussion

For a prediction model, the performance of a classifier is important due to the associated cost attached to it. An incorrect diagnosis of a disease might have to pay a heavy cost to a patient and it might be even of his life. Performance metrics are calculated from these values.

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

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

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

$$\text{Precision} = TP/(TP + FP)$$

$$\text{F1 Score} = (2 * \text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

The performance metrics of KNN are represented by table given (Table 1)

Table 1 Performance metrics of KNN on its optimum value of K (72, 73 or 74)

Metrics	Score (%)
Accuracy	81.17
Precision	84.21
Recall	58.18
Specificity	93.94
F1 Score	68.82
Error Rate	18.83

For a good classifier, its F1- Score is expected to be high. The data analysis has been performed on PIMA diabetes dataset using KNN classification algorithm. The programming language used for this task is Python. The observation of graph has been drawn between number of neighbors (K) and accuracy/ F1 score.

For diabetes dataset, it is found that the performance of KNN improves with the increasing value of K until it achieves its peak value. After it KNN's performance moves down, although value of K is increases. It means after touching the peak performance, the performance is not improving when value of k is increased. However, the selection of optimal value for “ K ” may be tricky and required large size dataset for accuracy. It may vary from one dataset to another dataset. In this research work, the optimal value of “ K ” for KNN classifier is either 72, 73, or 74 when it shows its best performance. The optimum value of K and its performance are shown in following (Fig. 3).

Sensitivity (Recall) shows the correct diagnosis capability of classifiers. It is the ratio of positive results made by the classifier against total positive samples in dataset. It is also known as true positive rate. The classifier whose TPR is high shows that it

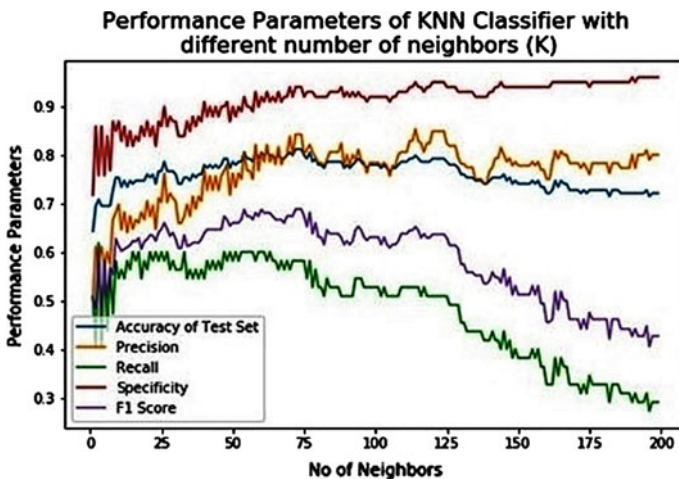


Fig. 3 Performance score of KNN with different value of K

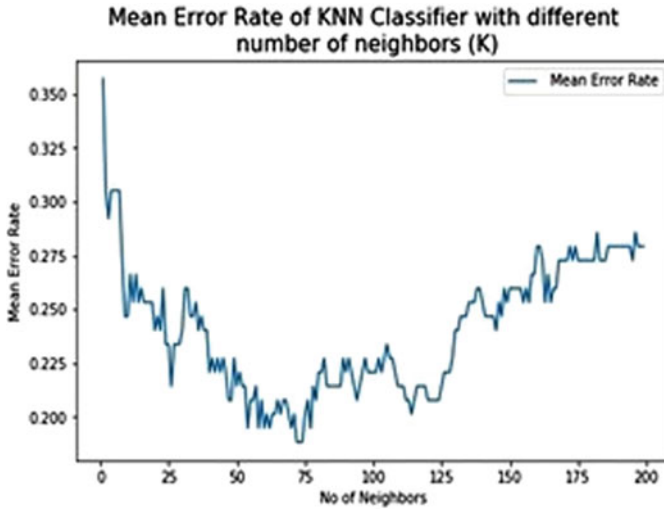


Fig. 4 Mean error rate of KNN with different value of *K*

correctly identify most of the positive cases. Here, sensitivity on optimal value of *K* is 58.18% which is neither good nor bad. Precision shows the percentage of correct positive diagnosis made by classifier in total positive diagnosis made by the classifier. A high precision shows that classifier makes less error in his positive declaration. In other words, the number of correct positive results declared by classifier in whole declarations is high. For this experiments, KNN shows high precision of 84.21% which is much better. For a disease prediction system where the cost of an incorrect negative prediction (FALSE NEAGTIVE) is more than the cost of incorrect positive prediction (FALSE POSITIVE). In such type of systems, precision and sensitivity should be high for a good classifiers. At the optimum value of *K*, precision and sensitivity of KNN classifier are 84.21% and 58.18% respectively (Fig. 4).

The minimum error rate for KNN classifier is 18.83% on optimum value of *K*. A plot is drawn between *K* and error rate. Observation shows that when the value of *K* is 1, the error rate is 35.71% due to the overfitting of boundaries. After it, the error rate decreases with the increasing value of *k* and reaches to minimum level of 18.83% when optimum value of *K* is achieved.

6 Conclusion and Future Scope

The detection of a disease in its early stage plays a vital role for the cure of patients. A disease prediction system may use various techniques to improve its performance for disease prediction. The system must be robust so that it reduces the cost of false-negative diagnosis. The false-negative case means a person which is actually a victim

of disease but put him in non-disease category. In this paper, a study is made on KNN machine learning algorithm for the diabetes prediction system and also observes the behavior of it on different value of neighbors K . As noise affects the performance of classifiers, KNN also affected by it. To get better performance of KNN, the dataset is preprocessed and irrelevant attributes are eliminated from dataset using optimal feature selection methods. The best performance of KNN is achieved when K is either 72, 73, or 74. The observations obtained from the experimental analysis performed by model on PIMA diabetes dataset is only for this dataset and can be applied for validation on other diabetes dataset. In future, this model can be used to validate the observations on large size diabetes dataset.

References

1. World Health Organization: Global Report on Diabetes. WHO Library, Geneva (2016)
2. Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., Yang, G., Chen, Y.: A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int. J. Med. Inform.* **97**, 120–127 (2017). <https://doi.org/10.1016/j.ijmedinf.2016.09.014>
3. Sushmi, D.: Nearly 12% of Indians above 50 have diabetes, finds new survey | India News—Times of India. <https://timesofindia.indiatimes.com/india/nearly-12-of-indians-above-50-have-diabetes-finds-new-survey/articleshow/71531884.cms>. Last Accessed 17 Jun 2020
4. Sharma, N.C.: Government survey found 11.8% prevalence of diabetes in India. <https://www.livemint.com/science/health/government-survey-found-11-8-prevalence-of-diabetes-in-india-11570702665713.html>
5. Sneha, N., Gangil, T.: Analysis of diabetes mellitus for early prediction using optimal features selection. *J. Big Data.* **6**, (2019). <https://doi.org/10.1186/s40537-019-0175-6>
6. Iyer, A., Jeyalatha, S., Sumbaly, R.: Diagnosis of diabetes using classification mining techniques. *Int. J. Data Min. Knowl. Manag. Process.* **5**, 01–14 (2015). <https://doi.org/10.5121/ijdkp.2015.5101>
7. Jakka, A., Vakula Rani, J.: Performance evaluation of machine learning models for diabetes prediction. *Int. J. Innov. Technol. Exploring Eng.* **8**, 1976–1980 (2019). <https://doi.org/10.35940/ijitee.K2155.0981119>
8. Sengamuthu, R., Abirami, R., Karthik, D.: In: Various Data Mining Techniques Analysis To Predict (2018)
9. Mittal, K., Aggarwal, G., Mahajan, P.: Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. *Int. J. Inf. Technol.* **11**, 535–540 (2019). <https://doi.org/10.1007/s41870-018-0233-x>
10. Azrar, A., Awais, M., Ali, Y., Zaheer, K.: Data mining models comparison for diabetes prediction (2018). <https://doi.org/10.14569/ijacsa.2018.090841>
11. Kumar, P.S., Umatejaswi, V.: Diagnosing diabetes using data mining techniques. *Int. J. Sci. Res. Publ.* **7**, 705–709 (2017)
12. Saxena, K., Khan, Z., Singh, S.: Diagnosis of diabetes mellitus using K nearest neighbor algorithm. *Int. J. Comput. Sci. Trends Technol.* **2**, 36–43 (2014)
13. Dua, D., Graff, C.: UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>

Phishing Website Prediction: A Machine Learning Approach



Anjaneya Awasthi  and Noopur Goel 

Abstract Phishing is an act of stealing our precious, personal, sensitive data such as credentials that which we use for accessing the resources and services, available across the cyberspace. Seeing these rapidly growing phishing attacks and their adverse effect on the businesses including individual users, it has now become a need for the organizations and individuals worldwide to be able to effectively predict the phishing website and differentiate them from legitimate ones. The aim of this research paper is to efficiently predict the phishing websites so that users may be benefitted from this study and prevent them from getting trapped. In this paper, machine learning techniques are used for prediction. Data mining is used worldwide by almost every face of the society viz. business organizations, govt. organizations, and other kind of data collectors to extract knowledge from the collected data. On the other hand, machine learning is a data mining technique that is used to analyze, classify the data, and efficiently predict the results for the estimation and planning by all of the organizations all around the globe. Classification algorithms, namely logistic regression, decision tree, and random forest classification, are used to predict the fake websites and presented their comparison of their predictions achieved. The results have been presented in numeric format as well as graphically with the help of chart. The data used is taken from UCI machine learning online repository. The seed value is changed and analyzed, and results achieved are least accuracy of 95.93, 97.96, and 98.78% of accuracy as the highest as well. Some future study and applying some good practices may help in designing a better and more accurate solution for the prediction of the phishing website, just by examining the URL and its features.

Keywords Phishing · Logistic regression · Decision tree · Random forest

A. Awasthi (✉) · N. Goel

Department of Computer Applications, VBS Purvanchal University, Jaunpur, UP, India
e-mail: anjaneyaawasthi@gmail.com

N. Goel

e-mail: noopurt11@gmail.com

1 Introduction

As we continuously keep drowning deeper and deeper in this technological race of computers, and computing devices, data, and technology, we are also revealing ourselves to the cyber space. As we know very little about this vast cyber space, we may be exposing ourselves to many known or unknown security threats. Phishing is not a new name nowadays for any person working with or using data computing and communicating devices as well as the carrier called the Internet. There exist many kinds of cyber threats on the public Internet thus making it more vulnerable place for novice users and newbies, as they use to fall in the trap set by these attackers by giving their personally sensitive information unknowingly.

This paper proposes a machine learning-based approach that predicts that a URL is a phishing or legitimate one. In this approach, 2456 websites data, 30 features are extracted for correct and accurate prediction of the phishing website. The features such as frequency of http or https, in the URL, redirection used, long URLs. These examples become opportunity for the cyber criminals.

According to the fourth quarter report of year 2019 published on February 24th 2020 by Anti Phishing Working Group (APWG) [1], the no. of unique phishing websites detected in the months of October, November, and December months of 2019 are 76,804, 39,580, and 45,771, respectively. Likewise, number of unique phishing emails received by the group is 45,057, 42,424, and 45,072, respectively, while number of brands targeted by phishing attack creators were 333, 325, and 341, respectively. These attacks affected almost every sector world wide including IT and IT enabled services.

This research study is conducted based on machine learning techniques. The present paper is divided into five sections. Section 2 presents the literature review that contains previously conducted researches by various authors time to time. Section 3 describes the methodology used in this study in conducting the research work, i.e., machine learning techniques to predict the phishing websites. Section 4 contains the results and discussion, where the results produced by proposed approach are given and compared the results with many previous researches. The paper is concluded in the last section, i.e., Sect. 5.

2 Literature Review

This section discusses the previous studies and researches done by different authors at different times. Some of them worked on email addresses or web URLs, while some of them worked on the online data such as: Phishtank or UCI.

Kiruthiga and Akila [2] have used many machine learning classifiers such as: Support Vector Machine (SVM), Natural Language Processing (NLP), Decision tree (DT).

To improve accuracy of detection of phishing website URLs, Sahingoz et al. [3] have introduced a machine learning oriented approach to detect phishing using various machine learning algorithms, namely Decision Tree (DT), Adaboost, KNN, naive Bayes, etc., on Ebbu 2017 dataset with Weka and tenfold cross-validation with default parameters.

Mao [4] have presented a phishing detection approach using machine learning that uses page layout features and classifiers. They used data from Phishtank.com, and four classifiers Decision Tree (DT), Support Vector Machine (SVM), AdaBoost, and Random Forest (RFC) CSS data are used for page layout and trained the classifiers on vector-based data. They predicted with the top accuracy of 97.3%.

Kang et al. [5] have used an ensemble feature selection framework that was based on machine learning. Using tools available in Weka such as: SVM, RFC, Naïve Bayes, along with cumulative distribution function for feature selection and classified the web URL data from phishtank.com with the top accuracy score of 96.17%.

Wang et al. [6] presented a phishing detection approach that involves use of convolutional neural networks. They have used statistical knowledge of nine URL character level features and put these data into their experiment on the dataset of phishtank.com and tested and compared many algorithm results out of which the top accuracy score was 95.6%.

Gupta et al. [7] have discussed dynamic classification mining techniques to detect phishing URLs. They used UCI ML repository dataset for the experiment. They used random forest, random tree, J48, naive Bayes, and LMT.

Jain et al. [8] have presented an approach of detecting any web URL by examining the source code (HTML) via CSS from the website. They used at least eight of the available classifiers, SVM, naive Bayes, neural network, random forest, etc.

Gupta [9] have also presented a comparative analysis of three algorithms of machine learning, namely Bayesian classifier, nearest neighbor classifier, and random forest, using data from APWG and phishtank.com and performed the comparison.

Many authors have compared, tested by using many tools; this study also uses three machine learning algorithms and presented the achieved results.

3 Methodology

In proposed paper, three machine learning classifiers have been used, namely: logistic regression [2, 8, 10, 11], decision tree [2, 8], random forest [2, 11–13], for experimenting on the dataset which is taken from UCI website [14]. Then classified this dataset and compared the predictions achieved by these algorithms. Thereafter displayed the results in tabular format as produced by these algorithms and plotted these results pictorially as well for the ease of understanding of the readers.

3.1 Dataset Analysis

As mentioned earlier, the dataset taken and used is the pre-processed dataset from UCI machine learning data repository, the attributes, and their column names available in.csv file are given in Table 1.

Table 1 Description of the data used

S. no.	Column (Attribute) name	Column description	Data type	Possible values
1	“has_ip”	If URL contains IP address	Integer	{ 1, 0}
2	“long_url”	Length of URL too long	Integer	{1, 0, -1}
3	“short_service”	Shortening service used for URL	Integer	{ 0, 1}
4	“has_at”	URL is having @ symbol	Integer	{ 0, 1}
5	“double_slash_redirect”	double slashes used for redirection	Integer	{ 1, 0}
6	“pref_suf”	Unwanted prefix and suffix	Integer	{-1, 0, 1}
7	“has_sub_domain”	If URL have sub domain	Integer	{-1, 0, 1}
8	“ssl_state”	URL contain SSLfinal state	Integer	{-1, 1, 0}
9	“long_domain”	Domain registration timeline	Integer	{0, 1, -1}
10	“favicon”	Favicon present	Integer	{ 0, 1}
11	“port”	If port is enabled	Integer	{ 0, 1}
12	“https_token”	If URL contains HTTPS token	Integer	{ 1, 0}
13	“req_url”	If Requests another URL	Integer	{ 1, -1}
14	“url_of_anchor”	If contain URL of anchor	Integer	{-1, 0, 1}
15	“tag_links”	tags having Links	Integer	{1, -1, 0}
16	“SFH”	Incorrectly used server form handler	Integer	{-1, 1}
17	“submit_to_email”	Submitting data to an email	Integer	{1, 0}
18	“abnormal_url”	Abnormal looking URL	Integer	{1, 0}
19	“redirect”	Redirection present	Integer	{0, 1}
20	“mouseover”	If on_mouseover event is used	Integer	{0, 1}
21	“right_click”	If RightClick is disabled	Integer	{0, 1}
22	“popup”	If it calls for a Pop Up window	Integer	{0, 1}
23	“iframe”	Iframe ambiguity	Integer	{0, 1}
24	“domain_Age”	If age of domain is too short	Integer	{-1, 0, 1}
25	“dns_record”	If DNS record not present	Integer	{ 1, 0}

(continued)

Table 1 (continued)

S. no.	Column (Attribute) name	Column description	Data type	Possible values
26	“traffic”	If web traffic is too less	Integer	{ -1, 0, 1 }
27	“page_rank”	If page_rank is not appropriate	Integer	{ -1, 0, 1 }
28	“google_index”	If not found in Google index	Integer	{ 0, 1 }
29	“links_to_page”	If links are pointing to external page	Integer	{ 1, 0, -1 }
30	“stats_report”	If statistical report is not appropriate	Integer	{ 1, 0 }
31	“target”	Result	Integer	{ 1, -1 }

This study uses all 2456 instances and 30 attributes in this dataset. First, the dataset is divided in two distinct parts; one part (80%) used for training of our algorithm and another part (20%) of dataset for testing purpose. These two parts are fed as input to all the three classifiers one by one for classification of data, and to make the predictions, the process is elaborated more clearly in Fig. 1.

3.2 Machine Learning Classifiers Used

1. **Logistic regression classifier:** In logistic regression, the prediction of probability is in two values only, while in linear regression the predicted values are outside the range of (0–1).
2. **Decision tree classifier:** It is a supervised learning-based predictive modeling tool; it splits the dataset based on multiple conditions that help us describe the condition based on multiple influences. A decision tree is generated from root following top-down approach that involves partitioning of data.
3. **Random forest classifier:** Random forest or random decision trees is unsupervised learning technique based on ensemble learning. It is used for classification and regression. It works by using multiple decision trees for training and giving output prediction in form of a class.

4 Results and Discussion

In this section, we will provide the details of data and the proposed approach in which we have applied three machine learning algorithms in python to classify and predict the phishing websites from the dataset; we have utilized this pre-processed data available on UCI website [14]. The detailed description of data, i.e., how much instances of phishing websites and how much is of non-phishing websites. First of all, we have displayed the no. of data instances and number of attributes in the data,

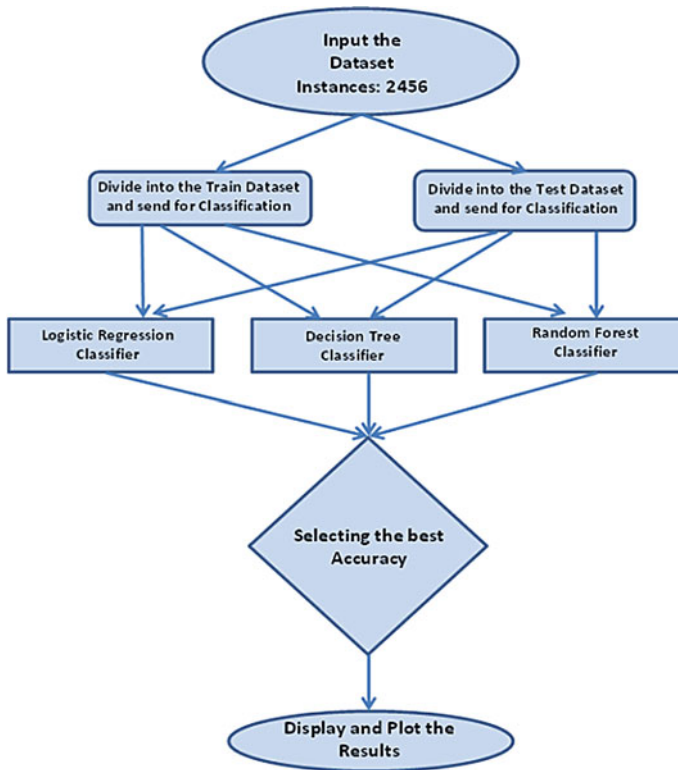


Fig. 1 Phishing website detection process

where no. of instances is: 2456, and no. of attributes: are 31 in the dataset; out of these 31 attributes, 30 different attributes are present as we explained in the previous section, and the last (31st) attribute is result attribute which contains the details of the distribution values: (-1) value that is 1362 indicates that the corresponding website is phishing website and (1) value that is 1094 indicates that the corresponding website is a non-phishing website. The number of phishing website data instances is 1362, while the number of non-phishing website instances is 1094 that sums up to our total data instances, i.e., 2456. Dataset is presented using Exploratory Data Analysis (EDA) for data analysis that incorporates many graphical techniques to examine underlying structure of data and extraction of important variables and to develop a model and determine optimal factors. To represent the statistical summary of the data, box and whisker plots are used and shown in Fig. 2.

The confusion matrix presents the summarized prediction results of a problem, where the classifiers are used for the problem. Incorrect and correct prediction summary are broken into classes, and their values are given. It shows the confusion of the classification model used while making predictions. It also gives errors and their types as well. The classification report has many constituent variable parameters that

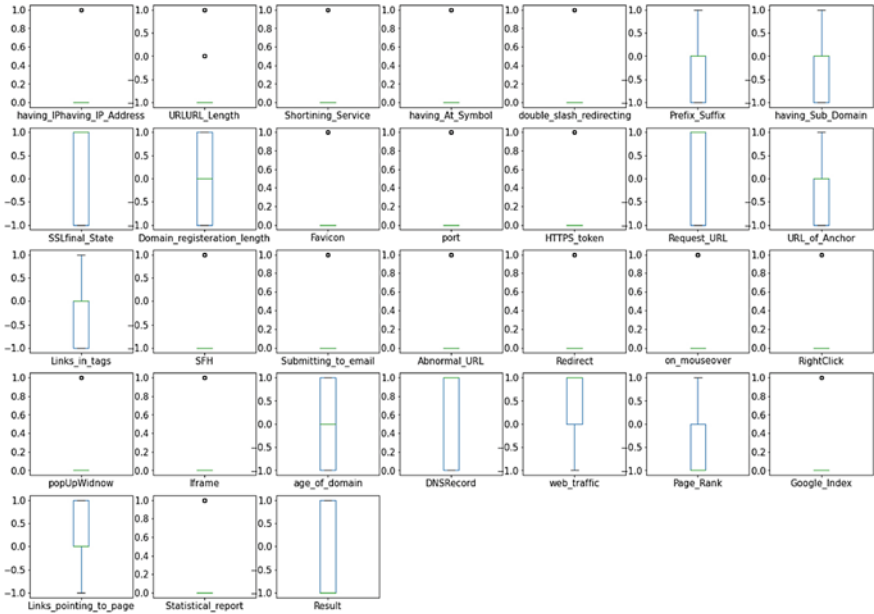


Fig. 2 Box and whisker plots showing the statistics

show the different parameters used for the calculations of an accuracy score, recall value or sensitivity, selectivity or specificity, precision, F1-score formulas are given in Table 2. The calculated values of these metrics using all the three classifiers are given in Table 4.

Seed value at the base, from which a system starts. This value/feature is required to move forward toward modeling a system. This value may be random that varies from model to model or system to system or application to application required triggering the whole system. Multiple seed values are used in the proposed approach ranging from 1 to 10, and best results are picked out of these values, and their results are described in tabular form (Table 3) as under:

At first all these three algorithms are applied; for this, we have split the data into two parts in the 80:20 ratios, one part for training, i.e., 80% and the other for

Table 2 Confusion matrix metric formula table

Type of metric	Formula	
Accuracy	$ACC = \frac{tp+tn}{tp+fp+tn+fn}$	(1)
Recall	$Recall = \frac{tp}{tp+fn}$	(2)
Precision	$Precision = \frac{tp}{tp+fp}$	(3)
F1-score	$F = 2 \cdot \frac{precision \cdot recall}{precision+recall}$	(4)
Specificity	$Specificity = \frac{tn}{tn+fp}$	(5)

Table 3 Accuracy scores achieved using different seed values

Seed value used	Accuracy achieved by logistic regression	Accuracy achieved by decision tree	Accuracy achieved by random forest
2	95.93	97.96	98.78
4	95.12	98.17	97.56
6	93.49	96.54	97.35
8	93.08	94.30	97.15
10	94.91	96.74	97.96

Table 4 Computational table for all three classifiers

Classifier used	Confusion matrix	Phishing and non-phishing	Precision	Recall	F1-score	Support	Accuracy achieved (%)
Logistic regression	[[270 10] [10 202]]	-1 1	0.96 0.95	0.96 0.95	0.96 0.95	280 212	95.93
Decision tree	[[272 8] [2 210]]	-1 1	0.99 0.96	0.97 0.99	0.98 0.98	280 212	97.96
Random forest	[[272 8] [2 210]]	-1 1	0.99 0.96	0.97 0.99	0.98 0.98	280 212	98.78

testing, i.e., 20%. And the results we obtained in form of accuracy, confusion matrix, classification report are given below in Tables 4 and 3 types of errors that have been calculated are displayed in Table 5.

The visual representation of results in form of line chart is given below in Fig. 3.

The results we achieved using these algorithms are encouraging for us to work further with more machine learning techniques as they may give us more accurate and refined results; if we try different combinations of the algorithms and techniques, then the results are compared with previous researches in Table 6.

Table 5 Errors produced for all three classifiers

Classifier used	Mean absolute error	Mean squared error	Root mean squared error
Logistic regression	0.0813	0.1626	0.4032
Decision tree	0.0406	0.0813	0.2851
Random forest	0.0243	0.0487	0.2208

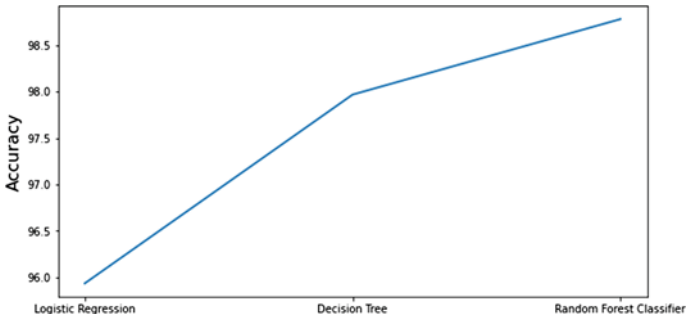


Fig. 3 Visual representation of results

Table 6 Comparison of results (Accuracy)

Author	Year	Technique(s)	Accuracy %
This paper (Proposed)	2020	LR, DT, RF (With Seed Value 2)	98.78
Mao et al.	2019	SVM, DT AB, RF	97.31
Kang et al.	2019	RF, JRIP, PART, SVM, NAÏVE BAYES	96.17
Wang et al.	2019	Convolutional neural networks, RNN, CANTINA + etc	95.6
Gupta	2016	RF, NN, BAYESIAN	88
Abdelhamid et al.	2017	C4.5, EDRI, RIDOR, ONERULE, etc.	96
Jeeva et al.	2016	APRIORI, PREDICTIVE APRIORI	93

5 Conclusion

We have conducted this experimental research on a very important security aspect of phishing websites detection, knowing the importance of this, and also it is much needed by the people. The results as we mentioned earlier are encouraging, and they also motivate us to go ahead with more research on this topic. We have used three classifiers and machine learning algorithms logistic regression, decision tree classifier, and random forest classifier in python; out of which the third algorithm (RF) has given the best accuracy score; we will try to progress in this area of research to contribute more in making the systems more accurate and help the community fight this threat more effectively.

References

1. APWG | Phishing Activity Trends Reports (2020). <https://apwg.org/trendsreports/>. Last Accessed 05 May 2020

2. Kiruthiga, R., Akila, D.: Phishing websites detection using machine learning. **2**, 111–114 (2019). <https://doi.org/10.35940/ijrte.B1018.0982S1119>
3. Koray, O., et al.: Machine learning based phishing detection from URLs. *Expert Syst. Appl.* **117**, 345–357 (2019). <https://doi.org/10.1016/j.eswa.2018.09.029>
4. Mao, J.: Phishing page detection via learning classifiers from page layout feature (2019)
5. Leng, K., et al.: A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Inf. Sci. (Ny)* **484**, 153–166 (2019). <https://doi.org/10.1016/j.ins.2019.01.064>
6. Wang, W. et al.: PDRCNN : precise phishing detection with recurrent convolutional neural networks, (2019)
7. Surbhi Gupta, A.S.: Dynamic classification mining techniques for predicting phishing. URL. Presented at the (2018)
8. Kumar, A., Gupta, J.B.B.: A machine learning based approach for phishing detection using hyperlinks information number of unique phishing sites detected. *J. Ambient Intell. Humaniz. Comput.* (2018). <https://doi.org/10.1007/s12652-018-0798-z>
9. Comparison of Classification Algorithms to Detect Phishing Web Pages Using Feature Selection and. **4**, (2016). <https://doi.org/10.5281/zenodo.61181>
10. Chiew, K.L., et al.: Utilisation of website logo for phishing detection. *Comput. Secur.* (2015). <https://doi.org/10.1016/j.cose.2015.07.006>
11. Abdelhamid, N., Abdel-jaber, H.: Learning comparison based on models content and features. 72–77 (2017)
12. Ali, W.: Phishing website detection based on supervised machine learning with wrapper features selection. **8**(9), 72–78 (2017)
13. Ho, T.K.: Random decision forests. *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR.* **1**, 278–282 (1995). <https://doi.org/10.1109/ICDAR.1995.598994>
14. UCI Machine Learning Repository: Phishing Websites Data Set (2019). <https://archive.ics.uci.edu/ml/datasets/Phishing+Websites>, Last Accessed 19 Apr 2020

COVID-19 Sentimental Analysis Using Machine Learning Techniques



Chhinder Kaur and Anand Sharma

Abstract With the rise in patients affected by the coronavirus, the World Health Organization declared it a pandemic. Globally, people are forced to stay at home to maintain social distancing. People are sharing their feelings through social media platform like Twitter. Twitter data helps to understand grief and pain due to coronavirus. In this paper, a Twitter dataset has been used for sentiment analysis of people's opinions related to coronavirus (COVID-19) that is a vital issue these days all over the world, and various countries are affected by this pandemic. So analyze the people sentiment regarding this pandemic using machine learning techniques and sentiment analysis model such as Naive Bayes, Support Vector Machine, Logistic Regression, and Random Forest Classifier has been proposed to analyze sentiment more effectively. Further, a comparison of these models has been done to prove extremely effective and accurate based on the analysis of feelings and opinions regarding coronavirus (COVID-19).

Keywords Twitter · Machine learning model · Sentiment analysis · Covid-19

1 Introduction

These days various microblogging sites are available on the internet but Twitter is the most popular site for the users to share their opinions, ideas, and messages on twitter called "tweets", with a limit of 280 characters. Twitter provides a platform to spread information around the globe using limited characters. Millions of tweets appear on Twitter every day from all over the world. Each tweet expresses opinions and ideas on various topics. These tweets help to carry out marketing campaigns, brands promotion of products, predictions in politics, sports, social issues, health issues. Twitter helps to share sentiments. A lot of sentiment analysis is carried out

C. Kaur (✉) · A. Sharma

Department of Computer Applications, UCCA, Guru Kashi University, Bathinda, India

e-mail: chhinderkaur87@gmail.com

A. Sharma

e-mail: andz24@gmail.com

using tweets. An application domain like corona issues deals with large text corpora and "formal language". The two specific issues addressed in computer-based tweet analysis are a) misspellings and slang in tweets are much higher than in other domains, users post their opinions on different topics on twitter, besides blogs, news, and other microblogging sites. This pose a challenge for sentiment analysis of tweets. The most common tweets are neutral tweets than positive or negative. Also, linguistic representational challenges for sentiment analysis and a limitation on tweet characters and limited sentimental cues.

In this research study, around 90,921 tweets from February 2, 2020, to May 15, 2020, have been collected from Twitter using Twitter scraper to analyze the people's opinions towards coronavirus pandemic from all over the world using machine learning techniques. The data set of about 73,121 tweets are used for training classifiers. Tweets have collected from Twitter by using Twitter Scraper. The collected dataset of tweets are saved as a CSV file and then pre-processed the dataset using NLTK library in python and labeled these tweets dataset as positive, negative, and neutral tweets. After that different models are applied to analyze the accuracy of models and then, summarize tweets and plot into a visual graph using Matplotlib library which is very popular in python programming to effective visualization.

TfidfVectorizer is applied for the training of the dataset. For applying the model on the dataset, the dataset is split into train and test as an 80–20 ratio. The models applied on the dataset are compared for the accuracy of the model. TfidfVectorizer methods used for extracting a new feature set in the dataset. A sentiment analysis model is built based on supervised learning such as Support Vector Machine, Logistic Regression Naive Bayes, and RandomForestClassifier to compare the performance of different models and enhancing the effective classification of models.

2 Related Works

There have been many studies in the area of sentiment analysis but almost focused on a part of texts. People can share their views in limited to 280 characters in a tweet, so it is very different to analyze sentiment clue. Liu [1], Tang et al. [2] analyzed the strong and weak points of sentiment analysis. Pang and Lee [3, 4] compared many classifiers in sentiment analysis and opinion mining for movie reviews. Most of the authors used star rating as a feature for analysis and classification. Go et al. [5] studied POS and Bigram methods. Emotions are removed from their training data for classification and compared the performance with Support Vector Machine (SVM), Naive Bayes using parameter Maximum Entropy. Barbosa and Feng [6] expressed that N-gram is very slow, and they used Microblogging features in their research. Agarwal et al. [7] used Microblogging, POS, and Lexicon feature in their research. Also, tree kernel is used to classify tweets and applied on POS and N-Gram is built by them. Kumar and Sebastian [8] analyzed the sentiment polarity of tweets using dictionary method. Mullen and Collier [9] used the SVM to analyze the sentiment in diverse information sources. Saif et al. [10] used Unigram for best accuracy. Rizzo

and Troncy [11] used various API like expose APIs: Alche-myAP in their research for better performance. Stanford University used Maximum Entropy classifiers to perform a twitter sentiment and also built a Recursive Deep Model.

3 Proposed Approach

To categorize the sentiment of each tweet as positive, negative, and neutral, an effective feature extractor is applied for enhancing the accuracy of the model using different classifiers. These classifiers are used to build a model to effectively classify such as Logistic Regression(LR), Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest Classifier(RF). Feature extraction, feature selection, BOW (bag of words) and TFIDF methods are used to increase the performance of the model.

3.1 Pre-processing of Dataset

The users can post their tweets on different languages on Twitter. Tweets have been very noisy features, so pre-processing data is carried out in the paper by using the NLTK library and also using lamda function to enhance the accuracy of tweets.

To pre-process the data following steps are carried out:

- (a) Convert each tweet upper case into lower case and splitted into sentences.
- (b) Remove stopwords from tweets like a, an, the.
- (c) Remove spaces and remove those words which do not start with the alphabet letters.
- (d) Remove Characters which are repeat such as “coorona”, converted into “corona”.
- (e) Remove usernames that are attached in tweets that start with”@” like @ck_chandan. Usernames replaced by “AT USER”.
- (f) Remove URLs that are attached in tweets like “<https://twitter.com>”.
- (g) Remove #hashtags from tweets like #socialdistancing, #coronavirus. These hashtags are provided useful information, so only removed “#” at hashtags from tweets. example, “ #CoronaVirusUpdate “to” coronavirus update”.
- (h) Word count, character count, stopword count, and remove the punctuation marks and lemmatization from the tweets.

3.2 *Feature Extraction*

In this paper, the polarity and subjectivity of each tweet is identified using TextBlob in Python. On the other hand, extracted tweets into a feature set by using count vectorizer and Tfidfvectorizer for enhancing the performance of the model also used n-gram, bio-gram, and tri-gram Following features are used for enhancing the accuracy of classifiers:

- (a) Expression of the moral in sentiment analysis of an entity or a keyword that can be extracted from each tweet as Positive, Negative, and Neutral. NLTK is used for extracting the feature for enhancing the efficiency and accuracy of classifiers. It also made the word cloud of tweets to identify which keywords are used in tweets by users to post tweets related to social distancing. Keywords that are mostly used by a user in their tweets are shown in word cloud Fig. 2.
- (b) For analyzing sentiments of Covid-19 data, the dataset of tweets was evaluated by different #hashtags like #socialdistancing, #cronavirus, #stayhomeandstaysafe, #china, #covid19 data sets. All tweets of these hashtags are collected by using twitter scraper in python. This method is used to enhance the accuracy of sentiment analysis, especially those tweets which do not contain adjectives verbs and sentiment polarity for classifying tweets used. In this research paper, different hashtags are selected related to COVID -19 and combined them in one CSV data set for better performance.

3.3 *Classification of Models*

Twitter Scraper is used as a library tool to collect the tweets related to COVID-19 pandemic from the internet for sentiment analysis and built different models based on Logistic Regression(LR), Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest Classifier(RF). Firstly, the dataset is split into training and testing. To train the different classifiers and to increase the efficiency of these different classifiers steps are followed:

- (a) Naive Bayes classifier builds to categorize subjective and objective tweets. The sentences labeled as a subjective or objective training set and applied Unigram, Bigram, for training the model.
- (b) SVM classifier used to subjective tweets into the positive class, negative class, and neutral class. The sentences labeled as positive, negative, and neutral tweets and then applied Unigram, Bigram to train the model.
- (c) Logistic Regression classifier is used to predict the values as positive, negative, and neutral.
- (d) Random Forest classifiers are applied to predict the values
- (e) After that, plot a graph after the analysis of sentiments as positive negative and neutral.

Four classification models are applied to train the dataset:

Naive Bayes (NB) Classification Method. It is based on statistics using Machine Learning. Conditional probability among words, phrases, and classes to predict statistics of them belonging to a class in this approach. The main point of this method is the independent appearance of words in machine learning. Naive Bayes does not evaluate the dependence of words with any class. In this paper applied. Scikit-Learn tool is used with a Multinomial Naive Bayes model to classify and enhance the performance of the model.

Support Vector Machine (SVM) Support Vector Machine is another popular machine learning technique. To train the model, Scikit-learn tool is used with Linear kernel. Support Vector Machine has two vector sets with m size in the training set. vector describes the presence or absence of that feature. Value is 1, if the feature is present, its value is 0 on the contrary. Object-oriented features and Bigram are the same. Figure 1 describes the Logistic Regression, Naive Bayes classifier, Support Vector Machine classifier, and Random Forest Classifier to classifying tweets from Twitter.

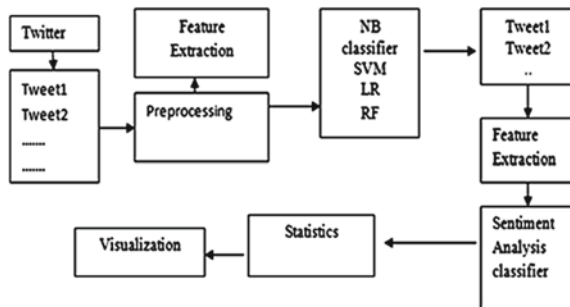
Logistic Regression(LR) Logistic Regression is a fundamental classification method in Machine Learning. LR Scikit-Learn tool is used to increase the performance of the model and to predict the values.

Random Forest Classifier(RF) Random Forest Classifier is another popular supervised learning algorithm in machine learning techniques and It can be used for both classification and regression model. It is easy to use and most flexible algorithm in machine learning techniques. This paper used a Random Forest classifier to predict the values.

The classifiers perform operations through the following steps:

1. Twitter Scraper is used to collect the tweets based on a topic COVID-19 using Python Programming Language in Jupyter notebook and then saved the tweet dataset into a database as a CSV file. In this paper, the topic is "#COVID-19", tweets have been downloaded from Twitter using Twitter Scraper related to "#socialdistancing, #coronavirus, #covid19".

Fig. 1 Proposed sentimental analysis model



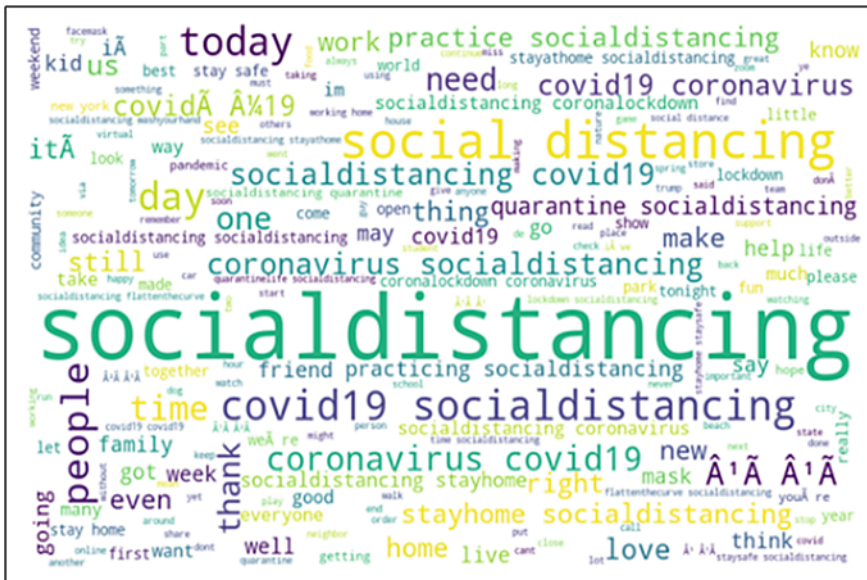


Fig. 2 Word cloud of tweets

2. After collecting the tweets, Pre-processed the tweets to remove the unnecessary noise from tweets.
3. After the completion of the pre-processing step extracts subjective features based on tweets, Bigram, unigram, using NLTK library for better performance.
4. After that train, the model through different classifiers as Support Vector Machine classifier, Naïve Bayes, Random Forest Classifiers, and Logistic Regression and then compare the performance of different models.
5. Plot different graphs by using the Matplotlib module that is most popular and easy to use the library in Python Programming based on tweets that are classified as positive, negative, and neutral.

4 Experiments and Results

4.1 Experiments

In this paper, containing 73,121 total tweets from different hashtags and labeled these tweets using TextBlob in Python in which 33,289 tweets are labeled as positive, 10,323 tweets as negative and 29,509 tweets as neutral tweets. The details of labeled tweets are shown in Table 1.

The training set consist of 73,121 tweets collected through Twitter Scraper. The labeled tweets as subjective or objective for training using Textblob and then trained

Table 1 Number of tweets as neutral, positive and negative

Class label	No. of tweets
Negative	10,323
Neutral	29,509
Positive	33,289

by using different classifiers as Naïve Bayes, Support Vector Machine, Logistic Regression, and Random Forest Classifier. Also, three types of features extracted containing Unigram, Bigram, n-gram for training, and experiments to enhance the performance of classifiers in this paper. The dataset is split into the training and testing to apply the particular model and predict the label.

4.2 Results

This paper, evaluated the accuracy of different classifiers on 73,121 tweets of the COVID-19 tweet dataset and compared them on the basis of their accuracy. The classification report method is used to test classifiers and Table 2 describes the accuracy of each model with the name of classifier and accuracy of a particular model.

The results shown in Table 2 describes the accuracy of different classifiers on a dataset of 73,121 tweets. The accuracy of Support Vector Machine is 94.16, Logistic Regression is 91.52, Random Forest classifier is 90.13 and Naive Bayes is 75.99.

The detail of Precision, Recall, and Accuracy of different classifiers on 73,121 tweets expressed in Tables 3, 4, 5 and 6:

Table 2 Accuracy of different classifiers on 73,121 tweets

Classifiers	Accuracy(%)
Support vector machine	94.16
Logistic regression	91.52
Random forest classifier	90.13
Naive bayes	75.99

Table 3 Performance evaluation of naïve bayes subjective classifier

Classes	Precision	Recall	F1-score	Support
Negative	0.70	0.90	0.79	4547
Neutral	0.30	0.92	0.46	688
Positive	0.96	0.68	0.80	9390
Avg/total	0.85	0.76	0.78	14,625
Naïve bayes accuracy			75.9931623931624	

Table 4 Performance evaluation of support vector machine subjective classifier

Classes	Precision	Recall	F1-score	Support
Negative	0.98	0.93	0.96	6194
Neutral	0.79	0.90	0.84	1828
Positive	0.95	0.96	0.96	6603
Avg/total	0.94	0.94	0.94	14,625
Support vector machine accuracy			75.9931623931624	

Table 5 Performance evaluation of logistic regression sentiment classifier

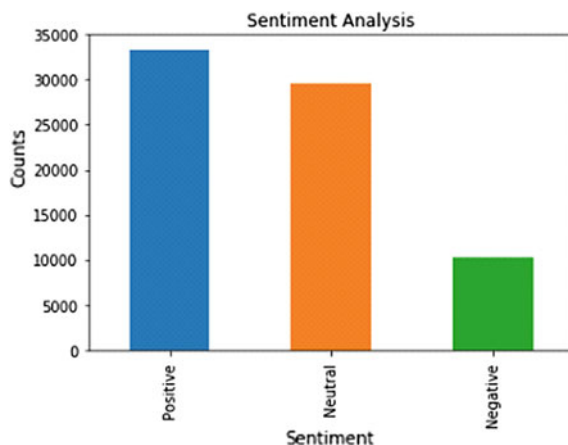
Classes	Precision	Recall	F1-score	Support
Negative	0.98	0.89	0.93	6470
Neutral	0.67	0.92	0.78	1520
Positive	0.93	0.94	0.94	6635
Avg/total	0.93	0.92	0.92	14,625
Logistic regression classifier accuracy			75.9931623931624	

Table 6 Performance evaluation of random forest classifier

Classes	Precision	Recall	F1-score	support
Negative	0.97	0.89	0.93	6429
Neutral	0.69	0.86	0.76	1661
Positive	0.91	0.93	0.92	6535
Avg/total	0.91	0.90	0.90	14,625
Random forest classifier accuracy			75.9931623931624	

The results shows the accuracy of Support Vector machine is 94.16% which is higher than other models. This research study shows better performance because in which extracted significant features for sentiment analysis using machine learning. These features help to enhance the accuracy of each classifier. Also, in Fig. 3 visualization result of tweets as positive, negative, and neutral tweets in a graph that creates for expressing analysis.

Fig. 3 Analysis of COVID-19 tweets



5 Conclusion

In conclusion, the proposed model analyzed the sentiments of Twitter using Machine Learning Techniques and Bigram, Unigram, Bow, and Tfidf features have been applied for effective feature extraction for sentiment analysis. Moreover, four different classifiers naïve Bayes, support vector machine, logistic regression, and random forest have been used for comparing the performance of different models. This research shows that peoples are aware of covid19 and still hopeful of mitigating coronavirus effects shortly.

References

1. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, 2nd edn (2010)
2. Tang, H., Tan, S., Cheng, X.: A survey on sentiment detection of reviews. *Expert Syst. Appl.* 10760–10773 (2009)
3. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the Association for Computational Linguistics (ACL), pp. 271–278. (2004)
4. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Info. Retrieval* 2(1–2), 1–135 (2008)
5. Go, A., Bhayani, R., Huang: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)
6. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of COLING(2010), pp. 36–44 (2010)
7. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau: Sentiment analysis of twitter data. In: Proceedings ACL 2011 Workshop on Languages in SocialMedia, Twitter Sentiment Analysis Using Machine Learning Techniques, vol. 289, pp. 30–38. (2011)
8. Kumar, A., Sebastian, T.M.: Sentiment analysis on twitter. *IJCSI Int. J. Comput. Sci. Issues* 9(4(3)) (2012)

9. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2004) (2004)
10. Saif, H., He, Y., Alani, H.: Alleviating: data sparsity for twitter sentiment analysis. In: Proceedings of the 2nd Workshop on Making Sense of Microposts (#MSM2012): Big Things Come in Small Packages: in Conjunction with WWW2012 (2012)
11. Rizzo, G., Troncy, R.: Nerd: Evaluating named entity recognition tools in the web of data. In: Workshop on Web-Scale Knowledge Extraction (WEKEX 2011), vol. 21 (2011)

Multimodal Music Mood Classification Framework for Kokborok Music



Sanchali Das, Sambit Satpathy, and Swapan Debbarma

Abstract This article describes one of the applications of Music information retrieval (MIR) integrated with natural language processing. The proposed work represents one of the applications of MIR that is music mood classification of one of the North-eastern regional language, which is Kokborok. It is widely spoken in the states of North East (NE) India and many other countries like Nepal, Bhutan, Myanmar and Bangladesh. The selection of the song is particular to Kokborok songs collected from the Bible, which has written in the recognized Romanized language which is accepted worldwide. We develop the multimodal corpus for audio and lyrics for Kokborok song and performed coarse-grained annotation to create mood annotated dataset and then perform classification task on both audio and lyrics separately. We projected mood taxonomy for Kokborok songs and set a mood annotated corpus with the corresponding taxonomy. Initially, we used 48 parameters for audio classification and six Text stylistic feature for lyrics based classification. The SVM classifier is used with linear kernel function for classification. Finally, Mood classification system was developed for Kokborok song consist of three different systems based on audio, lyrics and multimodal (audio and lyrics together). We also compared different classifier used to get the system performance for the above three systems. We achieved 95% accuracy for audio, 97% for lyrics and multimodal system, and the accuracy rate is about 96%.

Keywords Kokborok music · Multimodal mood classification · Music information retrieval · Natural language processing

1 Introduction

The present work is about one of the MIR research application along with Natural language processing techniques [1–9]. Maximum researchers had worked on audio and lyrics classification on western music and explore between the difference between

S. Das (✉) · S. Satpathy · S. Debbarma
Computer Science and Engineering Department, NIT Agartala, Jirania, Tripura, India
e-mail: sanchalidas1992@gmail.com; sanchalicse.sch@nita.ac.in

Hindi and English [5, 9, 10], and some of the researchers have used Indian languages like Hindi for the mood classification task [2–6]. There is a definitely less work done on any regional languages like classical music for mood classification [7, 8, 11]. In our work, we choose the music mood classification task as an application of MIR. We created dataset comprised of 300 songs of Christian Kokborok music along with their corresponding lyrics. Then we created suitable mood taxonomy for the database. There are some recent work has been carried out for kokborok music from where the subset of data has been taken [12–15]. There are some recent work has been carried out for kokborok music from where the subset of data has been taken [12–15]. But in this work, we have used only effective parameters for kokborok to create the three separate baseline model as well as the multimodal system for kokborok unlike the previous works. And we have gained much more accuracy rate for kokborok music comparatively [13–15]. Annotation is done manually to create a mood annotated dataset which is used as a ground truth set for the classification task. We then perform the mood classification task on audio files and lyrics database separately and together also (multimodal classification). It has been seen that western language and some specific language is used for MIR field whereas poor resourced language and dialects are deprived so we tried to do some original work that can be extended and help the researcher for Kokborok community. We choose a regional language which is Kokborok and generally spoken in the states of the northeast in India and other countries like Bangladesh and Myanmar, Nepal, Bhutan too.

The Christian community has intensive analysis on Christianity in Kokborok people of Tripura in the era of 1932–1988 by New Zealand Baptist community and about 50 years the Christian community in Kokborok people spread in Tripura. As of 2015, there are 840, and the total number of Kokborok Christian members is more than 98,000 in Tripura [16, 17]. We, as the researcher of Tripura, has initiated a research work towards a less-resourced language like Kokborok and incorporate natural language processing and Music Information retrieval for less-resourced language.

In Sect. 2, we have described related work in MIR field, and Sect. 3 is about proposed work, Sect. 4 has mentioned about the feature selection for audio and lyrics based classification, experimental results and comparison of the different algorithm have been shown in Sect. 5, conclusion and future works is defined in Sect. 6.

2 Related Works

In this section, we will discuss the related research by the various researchers from MIR and NLP group and also discuss the traditional mood taxonomy and the datasets.

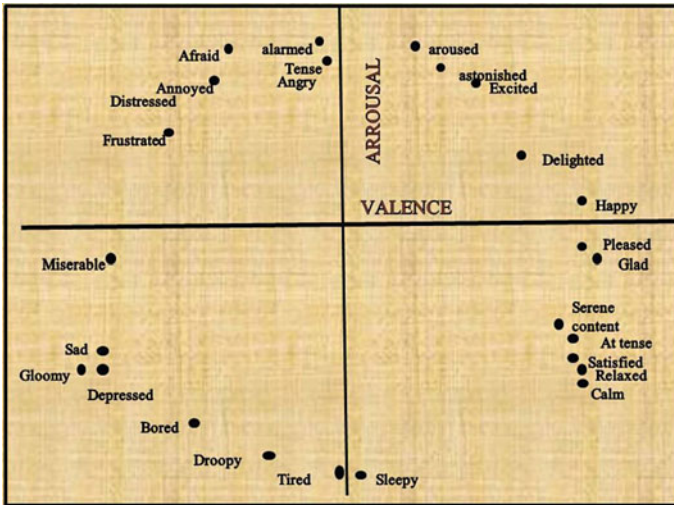


Fig. 1 Russell' taxonomy

2.1 Data Set and Taxonomy

In our work, we have collected the data from Bible as well as from the website mentioned in [18]. Mood taxonomy is the set of adjectives by which any dataset can be represented in the best way. There are several taxonomies available, i.e. Russell's taxonomy (Fig. 1.), MIREX, Havner's taxonomy (Fig. 2) [19, 20]. For Indian song, Havner's and Russell's taxonomy is found to be better fitted. For the mood classification task, prepare an extensive database of songs with similar lyrics and audio is an essential requirement. Mood annotated dataset for lyrics and audio is required to consider the mood associated with each song. For Indian music, very few works had been done in Hindi [2-6, 21-24]. In [25], find the electronic user interface that can automatically tag music mood done depending on lyrics.

2.2 Mood Classification Using Audio Feature

MIREX [19] is a mood taxonomy, an evaluation assessment of MIR algorithms to obtained the V-A score of music by different regression algorithms [10, 24, 26]. In [2-6], authors have used mood taxonomy of Russell to develop the audio classification framework in Hindi music and shows that some audio features like spectral and timbre are essential. Though, many other features like Rhythm, Pitch, Intensity are also used for audio classification.

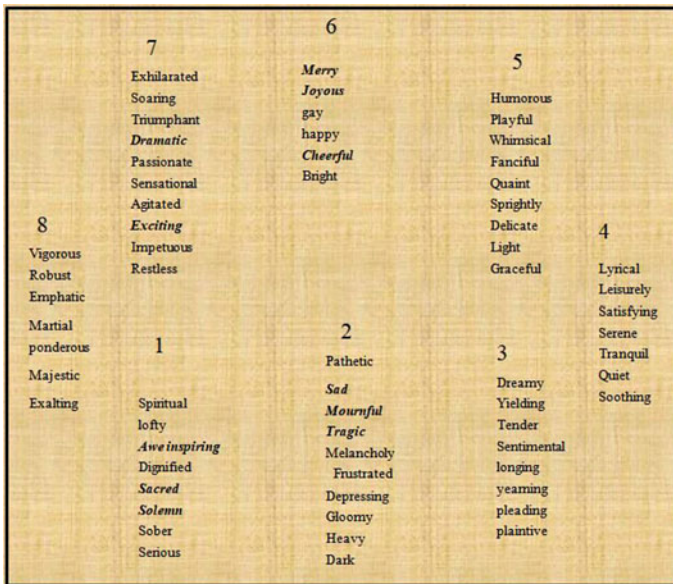


Fig. 2 Havner's taxonomy

2.3 Mood Classification from Lyric Features

There are many classification tasks conducted on western music constructed on the bag of words, sentiment lexicons (sentiwordnet) and stylistic features of a text [26]. In Hindi music [22], the author has used a combination of sentiment lexicons, stylistic and n-gram features for lyrics based classification model. In our work, only text stylistic features are used as a feature set for classification purpose because till now senti word net is not available for Kokborok. So, in future, we have to build it manually and used as a feature set for the classification task.

2.4 Multi-modal Music Mood Classification

Some researchers used a combination of audio and lyrical features to develop the automatic multimodal system of mood classification for western song. However, the multimodal model developed by a few researchers in Indian languages [2–6].

3 Proposed Work

In this segment, we have defined the preparation of required databases along with mood taxonomy generation for kokborok music.

3.1 Database Creation of Christian Kokborok Music

Our mood classification task is for one of the regional language Kokborok and our dataset confined only Kokborok music. We gathered 300 audio songs with their corresponding lyrics which are from Kokborok Christian community and related to the Holy Bible. Songs are used in this experiment are of 30-sec clip because the survey observes that first 30-sec clip of any song has the most useful information. So for the computational purpose, we remove all the noise from audio files.

3.2 Mood Annotated Dataset

For the mood classification task, it is necessary to create a ground truth set of data for audio as well as lyrics. As far as we know, mood annotated dataset is not available for Kokborok, so we have to annotate the files manually. Two annotators who know Kokborok does our annotation. Annotation is done in coarse grain method for lyrics data, and annotation is done by reading the lyrics only. For audio data, the annotation is based solely on the music not considered lyrics.

3.3 Taxonomy Generation

Mood taxonomy is used to express the feeling and the emotion regarding the song very firmly attached to it. As for our knowledge, no experiment is proposed to generate any taxonomy for Kokborok Christian song. So we adopted the subset from Havner's adjectives list. We observed in the initial observation that the adjectives in the Havner's list have fallen under the category where they can fit the database in the best way. Because songs have similar class, have to be close to each other, and songs having different cluster have to be distinct from each other in the hyperplane of v-a (Table 1).

Table 1 Proposed mood taxonomy

Sub-Class	Class			
	Happy	Sad	Calm	Excited
	Cheerful	Mournful	Sacred	Excited
	Merry	Tragic	Solemn	Dramatic
	Joyous	Pathetic	Inspiring	Aroused

4 Feature Selection

Now, we will discuss about the significant features that were used for audio and lyrics classification

4.1 Features Selection for Audio Classification

Feature extraction and selection in mood classification is an essential task for building a system by the literature survey [1, 5, 6], All the features are taken out by jAudio toolkit [27]. It is available publicly for research purpose and used by many researchers [2, 21, 24, 26, 28].

Timbre: The distinctive features of timbre have implemented for several researchers for music analysis. It is observed that MFCC features of Timber have been active features for music mood as well as a genre classification task. The spectral flux, spectral centroid, spectral shape, variability characteristics are essential for differentiation moods [2, 6, 21].

Intensity: Intensity can be considered as a prominent feature class for detection of mood of a piece of music. We consider the overall average and standard deviation of root means square, fraction of low energy which is also used by [21, 22] for calculating the values of each feature.

Rhythm: Rhythm strength, Rhythm regularity, and tempo are related to people's mood response. From the literature review, it has been seen that rhythm is steady and balanced for happy songs, sad songs generally does not show too many distinctive pattern [2, 6] (Table 2).

4.2 Future Selection for Lyrical Classification

For western music [28], text stylistic (TS) features have been considered one of the promising features of lyrics. There is some kind of TS features, i.e. number of total unique words, total repeated words etc. used by [3, 6, 21] for Hindi music. We considered some of the TS features in our experiments are shown above in Table 3.

Table 2 Features used for audio classification

Feature class	Feature description		
	Timbre	Intensity	Rhythm
Features used	Spectral roll off	Root means square	Beat histogram
	Spectral variability		Strongest beat
	MFcc's		Beat sum
	Lpc's		The strength of strongest beat
	Partial based spectral centroid		Zero crossing

Table 3 Features used for lyrics classification

Feature Name	Feature description
Number (No.) of words	Total no. of words in a lyric
No. of unique words	Total no. of unique words in a lyric
No. of repeated words	Total no. of words in a lyric whose frequency is greater than 1
No. of line	Total no. of line in a lyric
No. of the repeated line	Total no. of repeated line in a lyric
No. of unique line	Total no. of unique line in a lyric

5 Classification Result and Evaluation

For classification support, vector machine classifier (SVM) is used. SVM and decision tree are the two promising classifier that is used for audio mood classification irrespective of any language with high accuracy rate [3–6, 21].

Several open-source machine learning tools like WEKA are used for classification modelling [1, 26, 28]. We have used SVM classifier for mood classification. We have developed other models by other algorithms are also developed but with very low accuracy, so we choose LibSVM to be performed with the linear kernel to developed three particular systems as polynomial, and radial basis function do not fit fine. We faced lots of challenges, and these are described in this paper [12]. Earlier it is stated that the database is created from the Holy Bible, it is obvious to have songs with similar emotions. We observed that it creates confusion between class calm and happy and subclasses between sacred or sad. That is why we have sorted our data set up to 300 songs selectively. For western and Hindi music, lyrical classification system accuracy rates of observed a maximum of 80–90% and 50–75% [2–6, 21]. There are some work has been done recently on Kokborok audio and lyrics classification

Table 4 Classification system performance for audio

Class	Precision	Recall	<i>F</i> -measure
Calm	0.97	0.98	0.97
Excited	0.96	0.94	0.95
Happy	0.90	0.92	0.91
Sad	0.92	0.92	0.92
Average	0.95	0.92	0.95

by [13–15]. However, here we present a multimodal mood classification system for Kokborok music.

In our work, for lyrics classification, Lack of sentiment lexicons can reduce the accuracy rate comparatively. Another reason for low accuracy could be the dataset. Because in the Bible, there are fewer deviations of instruments and artist and also majority songs are religious. So, we saw that mood of the whole song is different from the mood of the first 30 s songs by annotators. We initially classify audio by 48 parameters, but it does not work well for Kokborok music. Some of the parameters do not create any impact on classification result, and we get 49% accuracy rate by LibSVM. So we select only those parameters which are significant changes in each class. We observed that only MFcc's's, Spectral Centroid, Strongest Beat, Beat Sum, Peak Based Spectral Smoothness, Zero Crossing are significant changes as classification result gets affected by those parameters only (Tables 4, 5, 6).

Table 5 Confusion matrix for the audio-based system

Predicted values					
Actual values	Class	Calm	Excited	Happy	Sad
	Calm	111	2	0	0
	Excited	3	90	2	0
	Happy	0	1	48	3
	Sad	0	0	4	37
Average accuracy rate 95%					

Table 6 Classification system performance for lyrics

Class	Precision	Recall	<i>F</i> -measure
Calm	0.98	0.99	0.97
Excited	0.95	0.97	0.96
Happy	0.97	0.90	0.94
Sad	0.95	0.97	0.96
Average	0.97	0.97	0.97

Table 7 Confusion matrix for lyrics based system

Predicted values					
Actual values	Class	Calm	Excited	Happy	Sad
	Calm	112	1	0	0
	Excited	2	93	0	0
	Happy	0	3	47	2
	Sad	0	0	1	39
Average accuracy rate 97%					

Table 8 Multimodal system performance

Class	Precision	Recall	F-measure
Calm	0.99	0.98	0.98
Excited	0.95	0.96	0.96
Happy	0.90	0.92	0.91
Sad	0.94	0.92	0.93
Average	0.96	0.96	0.96

5.1 Classification System Evaluation

It is necessary to have an enormous amount of mood annotated data for implement on a statistical model for the better results. Since this work is initially started, so the number of songs is less compared to western and other Indian languages. The mood classification has been performed using LibSVM classifier according to the features we have set. We used the WEKA API 3.8.1 for building our classification model. In Table 5, we can see the actual values and the predicted values for the audio classification system. The bold diagonal elements in each column represent the correctly predicted values. So the accuracy of the system is calculated by $286 (111 + 90 + 48 + 37)/\text{total number of the song } (300) * 100 = 95\%$. Similarly, Tables 7 and 9 shows the confusion matrix of lyrics based classification and multimodal classification, respectively. Tables 4, 6, 8 shows the precision-recall and F-measure of audio, lyrics and multimodal classification system.

5.2 Classification Based on Lyrics

See Tables 6 and 7.

Table 9 Confusion matrix for multimodal system

Predicted values						
Actual values	Class	Calm	Excited	Happy	Sad	
	Calm	111	2	0	0	
	Excited	1	92	2	0	
	Happy	0	2	48	2	
	Sad	0	0	3	37	
		Average accuracy rate 96%				

5.3 Multimodal Classifications

See Tables 8 and 9.

5.4 Comparison of Different Algorithms and System Performance

We used a different classifier to performed classification on the dataset for each of the three systems. From the Fig. 3b, we can say that support vector machine with a linear kernel and decision tree classifier that is j48 algorithm gives averagely similar and better results compared to other algorithms (Table 10).

Fig. 3 Graphical representations of system performance with different algorithms

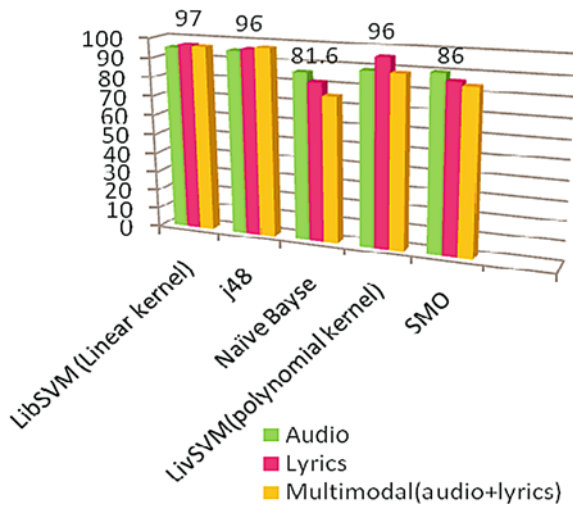


Table 10 System performance with different algorithms

Algorithms	System performance				
	LibSVM	J48	Naïve bayes	LibSVM polynomial kernel	SMO
Audio	95	95	86.3	89	90
Lyrics	97	96	81.6	96	86
Multimodal	96	97	75	88.6	84.3

6 Conclusions and Future Work

In recent work, the multimodal mood annotated database is developed for the research in the music mood classification of Kokborok. Three classification system is designed by the multimodal dataset. The audio-based system gives an accuracy rate of 0.95% and lyrics based classification system gives 0.97% accuracy rate, and for multimodal, we achieved a maximum F measure of 0.97 by LibSVM (linear kernel). We observed the mood variants during the annotation of the songs separately for audio and lyrics dataset. There are decidedly fewer variations in Kokborok Christian song found, and some of the reason may be that the same usage of instruments and the unavailability of many Kokborok singers. As audio features of a given song based on the instrumental variations too, so classification accuracy rate gets also affected. We show the comparison of system performances between three different systems and finally for the multimodal system and even using different classifier for each of the systems we can say that the LibSVM and j48 classifier both performed better on the above Christian Kokborok dataset.

We primarily considered music mood classification applications in future for Kokborok Christian music. We will work lyrics classification for various lyrical features, i.e. n-gram, bow, and sentiment lexicons. As for our knowledge, there is no sentiment lexicon available for Kokborok, so we will develop a sentiment word dictionary for Kokborok and explore all other possible features for lyrics based classification. In the multimodal system, we will study a more in-depth analysis of readers and the listener's point of view.

References

1. Tian, Y., Wu, Q., Yue, P.: A comparison study of classification algorithms on the dataset using WEKA tool. *J. Eng. Technol.* **6**(2), 329–341 (2018)
2. Patra, B.G., Das, D., Bandyopadhyay, S.: Automatic music mood classification of Hindi songs. In: *Proceedings of 3rd Workshop on Sentiment Analysis where AI meets Psychology, IJCNLP*, pp. 24–28. (2013a)
3. Patra, B.G., Das, D., Bandyopadhyay, S.: Multimodal mood classification framework for Hindi songs. *Computacin y Sistemas*, **20**(3), 515–526 (2016)
4. Patra, B.G., Das, D., Bandyopadhyay, S.: Unsupervised approach to Hindi music mood classification. *Mining intelligence and knowledge exploration*, pp. 62–69. Springer International Publishing (2013b)

5. Patra, B.G., Das, D., Bandyopadhyay, S.: Multimodal mood Classification-a case study of differences in Hindi and western songs. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1980–1989. (2016)
6. Patra, B.G., Das, D., Bandyopadhyay, S.: Labeling data and developing a supervised framework for Hindi music mood analysis. *J. Intell. Inf. Syst.* **48**(3), 633–651 (2017)
7. Banerjee, S.: A survey of prospects and problems in hindustani classical raga identification using machine learning techniques. In: Proceedings of the First International Conference on Intelligent Computing and Communication, Springer, Singapore, pp. 467–475. (2017)
8. Velankar, M.R., Sahasrabudde, H.V.: A pilot study of Hindustani music sentiments. In: Proceedings of 2nd Workshop on Sentiment Analysis where AI meets Psychology India, IIT Bombay, Mumbai. COLING-2012, pp. 91–98. (2012)
9. Yang, D., Lee, W.S.: Music emotion identification from lyrics. In: Multimedia, ISM'09. 11th IEEE International Symposium, IEEE (2009, December), pp. 624–629. (2009)
10. Malheiro, R., Panda, R., Gomes, P., Paiva, R.P.: Emotionally-relevant features for classification and regression of music lyrics. *IEEE Trans. Affect. Comput.* **2**, 240–254 (2018)
11. Degaonkar, V.N., Kulkarni, A.V.: Automatic raga identification in Indian classical music using the convolution neural network. *J. Eng. Technol.* **6**(2), 564–576 (2018)
12. Das, S., Satpathy, S., Debbarma, S.: Challenges and requirements of christian kokborok music irrespective with mood classification systems and generation of mood taxonomy. sentiment word dictionary for Kokborok. *Int. J. Comput. Intell. IoT* **2**(1), (2019)
13. Das, S., Mohan, P., Rajak, S.K., Debbarma, S.: Music mood taxonomy generation and classification of christian kokborok song: an audio-based approach. *Int. J. Adv. Intell. Paradigms* (Unpublished). (2018). [Online] Available <https://doi.org/10.1504/ijaip.2018.10020901>; <https://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijaip>
14. Das, S., Satpathy, S., Debbarma, S., Bhattacharyya, B.K.: Data analysis on music classification system and creating a sentiment word dictionary for Kokborok language. *J. Ambient Intell. Human. Comput.* 1–12 (2019)
15. Das, S., Bhattacharyya, B.K., Debbarma, S.: Building a computational model for mood classification of music by integrating an asymptotic approach with the machine learning techniques. *J. Ambient Intell. Humanized Comput.* (2020). Available <https://doi.org/10.1007/s12652-020-02145-1>
16. Detail about the Kokborok language: Available <https://en.wikipedia.org/wiki/Kokborok>
17. Detail about Christianity religion in Tripura state: Available https://en.wikipedia.org/wiki/Christianity_in_Tripura
18. Collection of Some Kokborok songs Available: https://tripuraking.com/site_0.xhtml
19. Downie, X.H., J. S., Laurier, C., Ehmann, M.B.A.F.: The 2007 MIREX audio mood classification task: lessons learned. In: Proceedings 9th International Conference Music Information Retrieval, pp. 462–467. (2008)
20. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161–1178 (1980)
21. Patra, B.G., Das, D., Bandyopadhyay, S.: Mood classification of Hindi songs based on lyrics. In: Proceedings of the 12th International Conference on Natural Language Processing, pp. 261–267. (2015)
22. Patra, B.G., Das, D., Bandyopadhyay, S.: Retrieving similar lyrics for music recommendation system. In: 14th International Conference on Natural Language Processing, December, pp. 48–52. (2017)
23. Laurier, C., Sordo, M., Serra, J., Herrera, P.: Music mood representations from social tags. In: Proceedings of the ISMIR, pp. 381–386. (2009)
24. Patra, B.G., Das, D., Maitra, P., Bandyopadhyay, S.: Feed-forward neural network based music emotion recognition. *MediaEval Workshop*, September 14–15 (2015)
25. Cano, E., Morisio, M.: Moody lyrics: A sentiment annotated lyrics dataset. In: Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics and Swarm Intelligence, ISMSI, Hong Kong, March 2017, pp. 118–124. ACM (2017)

26. Joshi, A., Balamurali, R., Bhattacharyya, P.: A fall-back strategy for sentiment analysis in Hindi: a case study. In: Proceedings of the 8th International Conference on Natural Language Processing, (ICON-2010)
27. McKay, C., Fujinaga, I., Depalle, P.: jAudio: a feature extraction library. In: Proceedings International Society for Music Information Retrieval (ISMIR), pp. 600–603. (2005)
28. Ujlambkar, A.M., Attar, V.Z.: Mood classification of Indian popular music. In: Proceedings of the CUBE International Information Technology Conference, pp. 278–283. ACM (2012)

Forecasting of Daily Demand's Order Using Gradient Boosting Regressor



Tansif Anzar

Abstract Supply chain management is an important task in terms of business process. In this task, an important terminology is forecasting. And, the order forecasting is essential for the related personals. In this paper, daily demand's forecasting is done based on the data of the Brazilian logistics company. Previously, artificial neural network (ANN) was applied in this dataset. To get the best accuracy, different settings of multi-layers perceptron neural network were used with proper optimization. But using deep neural network for forecasting orders having limited data sometimes makes a model overfitted and complex. So, a normal machine learning-based approach is used to avoid these facts, and also, there is an improvement of the model's accuracy. Gradient boosting regressor is applied in a more practical way with optimization of different parameters which decrease the error rate of previous work near about on average by 0.86% and at best by 1.44%.

Keywords Supply chain management · Order forecasting · Neural network · Machine learning · Gradient boosting regressor

1 Introduction

When the forecasting is done in a perfect way, decisions and planning will be more accurate. As the world is moving toward an expandable supply chain management system, forecasting will impact a great value for the changing demands of products and resources. As the shortage of product results in customer dissatisfaction, so the goal should be to ensure the minimum amount of inventory to satisfy the customer. On the other hand, to avoid the extra cost for over storage in inventory, another goal should be minimize the cost of buying and holding inventory [1]. None wants undersupply which can result in lost sales, so a reliable forecast is needed [2–6].

T. Anzar (✉)
Rajshai University of Engineering & Technology, Rajshahi, Bangladesh
e-mail: tansifruetce13@gmail.com

To tackle these issues, practitioners are trying to find out a better model for better accuracy [7–10]. Forecasting technique adaptation is not enough to ensure best accuracy as the process is linked up with how the forecasting process is managed and organized [11–13]. Forecasting management includes decisions on information-gathering processes and tools, organizational approaches to be adopted, inter-functional and intercompany collaboration for developing a shared forecast, and measurement of accuracy. The article is organized in the following way: In Sect. 2, the concepts and research background of supply chain forecasting are presented. In Sect. 3, gradient boosting regressor algorithm is presented. Section 4 presents the methodology of the article. Section 5 describes the experimental results. The work ends with a final conclusion and future work in Sect. 6.

2 Research Background

Supply chain can be classified into three types: direct supply chain, extended supply chain and ultimate supply chain [14]. The classification is shown in Fig. 1 [15]. The products of a manufacturer are sent to main distribution warehouses in direct supply chain management. For the increasing number of warehouses and products, accurate demand forecasting becomes more important [15].

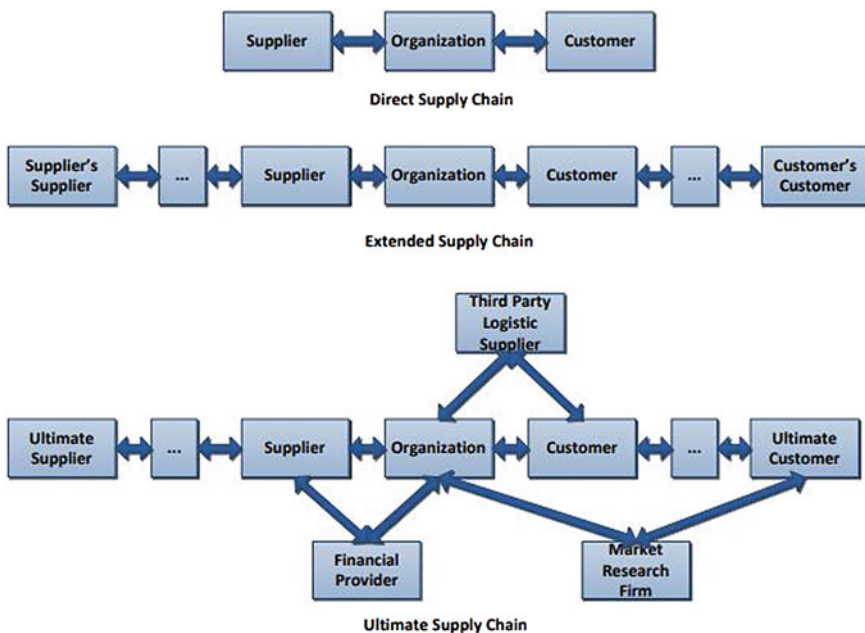


Fig. 1 Supply chain classification

In this work, demand forecasting models will be discussed. Demand forecasting has been applied to many areas like transportation, stock market, retail, etc. Traditionally, time series approaches are used like naïve method, average method, exponential smoothing, Holt’s linear trend method, exponential trend method, damped trend methods, Holt–Winters seasonal method, moving averages, autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA) models [16]. Nowadays, usage of multilayer perceptron artificial neural network [17] has been started to tackle supply chain forecasting. In 2019, two neural networks have been used by formulating price policies and forecasting retailers demand [18]. In the same year, machine learning techniques were also used for the anomaly detection in the logistic systems [19]. In 2020, a combination of genetic algorithm and multilayer feed forward network has also been applied [20]. In this paper, gradient boosting regressor is used to avoid the model’s complexity of the deep learning-based model. Different types of parameters are optimized based on the error values.

3 Gradient Boosting Algorithm

Gradient boosting is a prediction model following the mechanism of machine learning for regression and classification problems of an ensemble of weak prediction models. Model building is done in stage-wise fashion, and generalizing is done by optimizing of an arbitrary loss function [21, 22].

Like other boosting methods, gradient boosting combines weak “learners” into a single strong learner in an iterative fashion. It is easy to explain in the least-squares regression setting, where the goal is to “teach” a model F to predict value $\hat{y} = F(x)$ by minimizing the mean squared error $\frac{1}{n} \sum_{e=0}^n (\hat{y}_e - y_e)^2$, where e indexes over training set of size n of actual values of the output variable y .

- \hat{y}_e = the predicted value $F(x)$
- y_e = the real value
- n = the number of samples in y
- The model can be written as:

$$F_{m+1}(x) = F_m(x) + h_m(x) = y$$

or, equivalently,

$$h_m(x) = y - F_m(x)$$

Here,

- M = Total stages of gradient boosting algorithm
- m is different stages such $1 \leq m \leq M$
- F_m = Imperfect model for low m which could return
- h_m = New estimator to improve algorithm. [23].

Gradient boosting regressor algorithm can be written as following:

Algorithm 1 Gradient Tree Boosting Algorithm [24]

1. Initialize $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_e, \gamma)$
2. For $m = 1$ to M :
 - (a) For $e = 1, 2, \dots, N$ compute

$$r_{im} = - \left[\frac{\partial L(y_e, f(x_e))}{\partial f(x_e)} \right]_{f=f_{m-1}}$$

- (b) Fit a regression tree to the targets r_{em} giving terminal regions $R_{jm}, j = 1, 2, \dots, J_m$.
 - (c) For $j = 1, 2, \dots, J_m$ compute

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_e \in R_{jm}} L(y_e, f_{m-1}(x_e) + \gamma).$$
 - (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_m(x)$.

4 Methodology

The dataset was collected during 60 days, and this is a real database of a Brazilian logistics company. The dataset has twelve predictive attributes, and a target that is the total of orders for daily treatment [25]. The data was pre-processed. For splitting the data into the training and testing part, four fold cross-validation is done (75% for training and 25% for testing for each fold). Gradient boosting regressor function from Python's Scikit-learn machine learning library is used. From the parameters of this function, learning_rate, n_estimators, min_samples_split, min_samples_leaf, max_depth and max_features are chosen logically from different values of these parameters versus the mean squared value graph which are described in the following figures.

From Fig. 2, as the mean squared value is least for 0.05, learning_rate = 0.05 is chosen.

From Fig. 3, as the mean squared value is least for near about 150, n_estimators = 150 is chosen.

From Fig. 4, as the mean squared value is least for 6, min_samples_split = 6 is chosen.

From Fig. 5, as the mean squared value is least for 2, min_samples_leaf = 2 is chosen.

From Fig. 6, as the mean squared value is least for 2, max_depth = 2 is chosen.

From Fig. 7, as the mean squared value is least for 7, max_features = 7 is chosen.

Fig. 2 Learning rate versus mean squared error

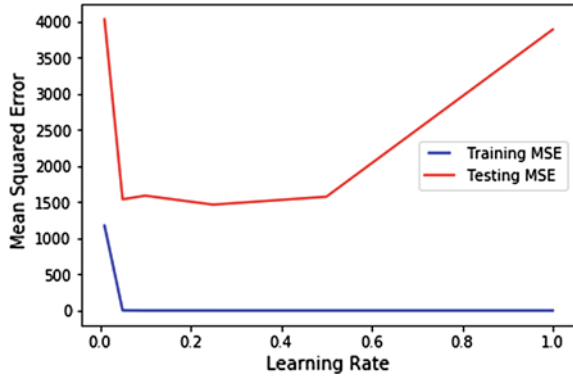


Fig. 3 Number of estimators versus mean squared error

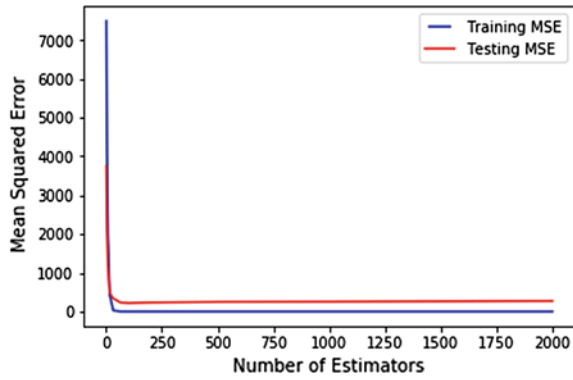
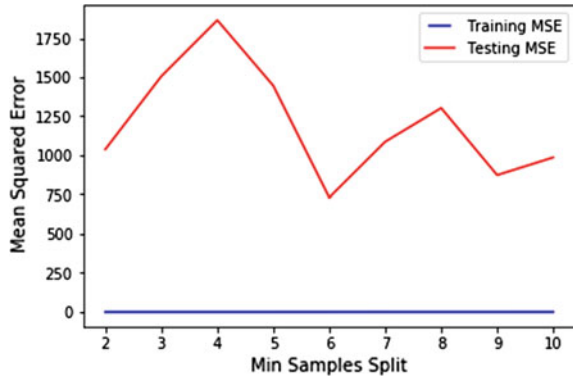


Fig. 4 Min samples split versus mean squared error



Other parameters are chosen by default of the function. In Table 1, the value of the parameters is given:

Fig. 5 Number of samples leaf versus mean squared error

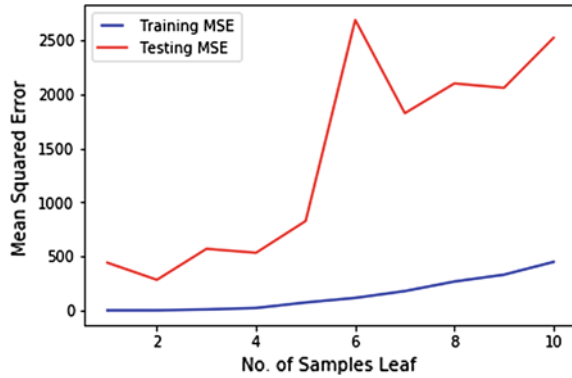


Fig. 6 Number of depth versus mean squared error

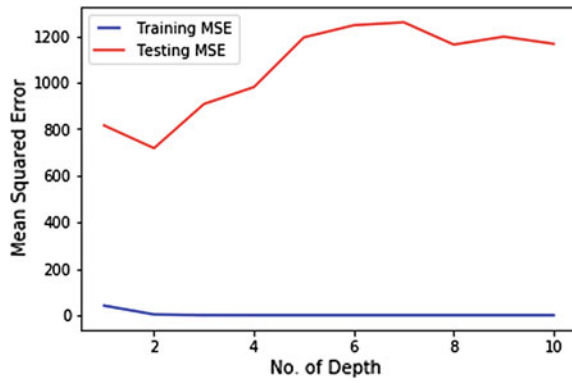


Fig. 7 Number of feature versus mean squared error

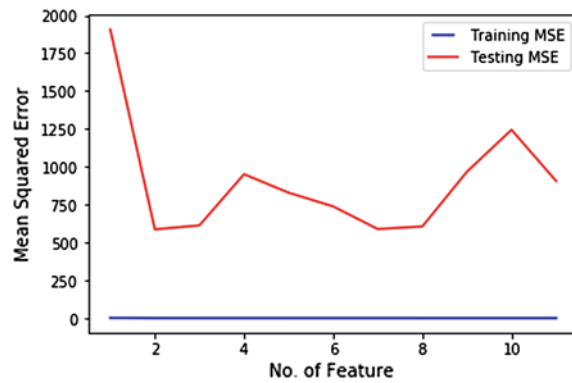


Table 1 Different parameters of gradient boosting regressor function

Parameter	Value
loss	ls
learning_rate	0.05
n_estimators	150
sub_sample	1.0
criterion	Friedman_mse
min_samples_split	6
min_samples_leaf	2
minimum_weight_fraction_leaf	0.0
max_depth	2
min_impurity_decrease	0.0
min_impurity_split	None
init	None
random_state	42
max_features	7
alpha	0.9
verbose	0
max_leaf_nodes	None
Warm_state	False
presort	'deprecated'
validation_fraction	0.1
n_iter_no_change	None
toll	0.0001
ccp_alpha	0.0

5 Experimental Results

After applying the four fold cross-validation, R square value for training and testing data and mean squared error for testing data are given in Table 2. For, the R square

Table 2 Result from four fold cross-validation

Fold No	Training data	Testing data	
	R square value	R square value	Mean squared error
1	0.99	0.94	606.34
2	0.99	0.98	58.98
3	0.99	0.93	719.03
4	0.99	0.95	316.99

Table 3 % Error of four fold cross-validation (testing data)

% Error for fold 1	% Error for fold 2	% Error for fold 3	% Error for fold 4
1.87	1.64	-3.26	-4.69
2.22	3.47	-48.25	-2.17
-10.19	1.22	0.96	0.3
0.23	2.73	3.99	-0.13
0.38	2.13	0.56	0.05
3.21	-0.05	0.34	2.72
-9.67	7.26	0.29	7.15
10.43	2.26	1.57	-6.9
3.33	1.19	13.17	7.1
-3.68	-0.15	0.52	0.18
12.39	2.07	-2.11	11.07
0.89	2.5	-0.38	-2.36
-15.69	0.63	-1.68	0.76
-11.3	-1.21	1.42	-1.45
-1.22	1.77	0.35	0.61

value, it is seen that the relationship between the predictive and target variable is high. For some specific fold, the mean squared error is also least.

In Table 3, % error for different testing data is shown.

After finding % error for all of the testing data, mean absolute % error is found for four folding, and it is clearly seen that the average % error is less than the average % error using ANN [17]. The comparison is shown in Fig. 8.

For fold-2, the error is least in gradient boosting regressor and better than least error of ANN (Fig. 9).

Fig. 8 Comparison of ANN and gradient boosting (average case)

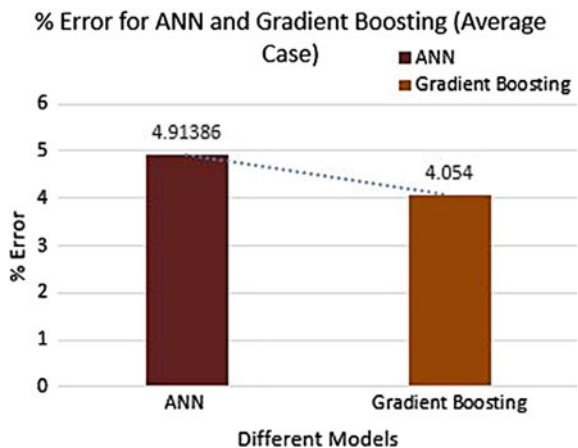
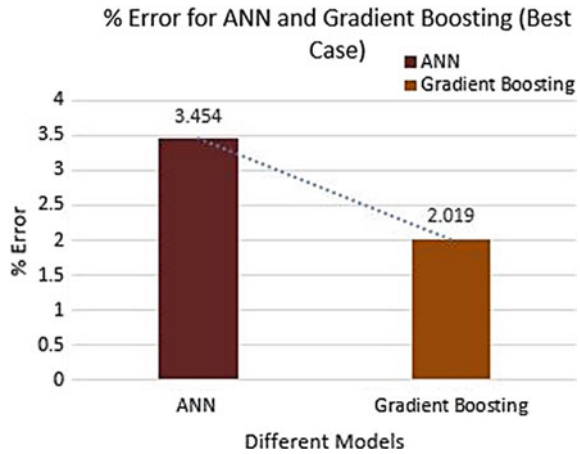


Fig. 9 Comparison of ANN and gradient boosting (best case)



6 Conclusion and Future Work

This paper tries to acquire more perfection in demand forecasting of supply chain. The contribution of the paper is to show that advanced neural network algorithm sometimes lag behind of accuracy than traditional machine learning algorithms. Another contribution is to add cross-validation in training and testing part to avoid biasness of randomness which was not done in the previous paper of this dataset. Our future work will be to add more datasets and analyze the impacts of gradient boosting and other algorithms to set state of the art in the demand forecasting.

References

1. Danese, Pamela, Kalchschmidt, Matteo: The role of the forecasting process in improving forecast accuracy and operational performance. *Int. J. Prod. Econ.* **131**(1), 204–214 (2011)
2. Vollmann, T.E., Berry, W.L., Whybark, D.C.: *Manufacturing planning and control systems*, 3rd edn. Richard D. Irwin Corp, Homewood, IL (1992)
3. Ritzman, L.P., King, B.E.: The relative significance of forecast errors in multistage manufacturing. *J. Oper. Manage.* **11**, 51–65 (1993)
4. Enns, S.T.: MRP performance effects due to forecast bias and demand uncertainty. *Europ. J. Oper. Res.* **138**(1), 87–102 (2002)
5. Zhao, X., Xie, J.: Forecasting errors and the value of information sharing in a supply chain. *Int. J. Prod. Res.* **40**(2), 311–335 (2002)
6. Kalchschmidt, M., Zotteri, G., Verganti, R.: Inventory management in a multiechelon spare parts supply chain. *Int. J. Prod. Econ.* **81**(82), 165–181 (2003)
7. Wright, D.J., Capon, G., Page, R., Quiroga, J., Taseen, A.A., Tomasini, F.: Evaluation of forecasting methods for decision support. *Int. J. Forecast.* **2**(2), 139–153 (1986)
8. Armstrong, J.S.: *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, Boston (2001)

9. Caniato, F., Kalchschmidt, M., Ronchi, S., Verganti, R., Zotteri, G.: Forecasting demand fluctuations due to promotional activities: a case in the fresh food industry. In: Proceedings of the POMS Conference, San Francisco (2002a)
10. Caniato, F., Kalchschmidt, M., Verganti, R.: A forecasting approach to manage composite demand. In: Christiansen, J.K., Boer, H. (eds.) *Operations Management and the New Economy*, pp. 227–238. Copenhagen Business School, Denmark (2002)
11. Armstrong, J.S.: The forecasting audit. In: Makridakis, S., Wheelwright, S.C. (eds.) *The Handbook of Forecasting*, pp. 584–602. Wiley, New York (1987)
12. Mentzer, J.T., Bienstock, C.: *Sales Forecasting Management*. Sage Publications, London (1998)
13. Moon, M.A., Mentzer, J.T., Smith, C.D.: Conducting a sales forecasting audit. *Int. J. Forecast.* **19**, 5–25 (2003)
14. Mentzer, J.T., DeWitt, W., Keebler, J.S., Min, S., Nix, N.W., Smith, C.D., Zacharia, Z.G.: Defining supply chain management. *J. Bus. Logistics* **22**(2) (2001)
15. Islek, I., Oguducu, S.G.: A retail demand forecasting model based on data mining techniques, *IEEE*, pp. 55–60. (2015)
16. Hyndman, R., Athanasopoulos, G.: *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia (2018)
17. Ferreira, R.P., Martiniano, A., Ferreira, A., Ferreira, A., Sassi, R.J.: Study on daily demand forecasting orders using artificial neural network. *IEEE Latin Am. Trans.* **14**(3), 1519–1525 (2016)
18. Bottani, E., Centobelli, P., Gallo, M., Kaviani, M.A., Jain, V., Murino, T.: Modelling wholesale distribution operations: an artificial intelligence framework. *Indus. Manage. Data Syst.* **119**(4), 698–718 (2019)
19. Kerdprasop, N., Chansilp, K., Kerdprasop, K., Chuaybamroong, P.: Anomaly detection with machine learning technique to support smart logistics. In: Misra, S. et al. (eds.) *Computational Science and Its Applications—ICCSA 2019*. ICCSA 2019. Lecture Notes in Computer Science, vol. 11619. Springer, Cham (2019)
20. Üstün, O., Bekiroğlu, E., Önder, M.: Design of highly effective multilayer feedforward neural network by using genetic algorithm. *Expert Syst.* **19**, e12532 (2020)
21. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Statist.* **29**(5), 1189–1232 (2001). <https://doi.org/10.1214/aos/1013203451>. <https://projecteuclid.org/euclid.aos/1013203451>
22. Friedman, J.H.: *Stochastic gradient boosting* (2002)
23. Li, C.: A gentle introduction to gradient boosting (2016). http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/slides/gradient_boosting.pdf
24. Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome: *Boosting and Additive Trees: The elements of statistical learning*, pp. 337–387. Springer, New York, NY (2009)
25. Dua, D., Graff, C.: *UCI Machine Learning Repository*, CA, University of California, School of Information and Computer Science (2019). [<http://archive.ics.uci.edu/ml/Daily+Demand+Forecasting+Orders/>]. Irvine

Improving Impulse Noise Classification Using Ensemble Learning Methods



Kunaraj Kumarasamy, S. Maria Wenisch, S. Balaji, L. J. Jenifer Suriya, A. Jerlin, and S. Robert Rajkumar

Abstract Medical image denoising is an essential pre-processing step in medical image processing which improves the performance of clinical diagnosis and prognosis. The high level medical image processing algorithms like segmentation, classification etc. works better if the image is denoised appropriately. The main objective of this research work is to find and replace only the corrupted pixels with suitable estimates of pixels in medical images. The other pixels which are not corrupted are left undisturbed, thereby preserving the image quality for proper diagnosis. For the primary task of finding the corrupted pixels, an ensemble of machine learning (EML) classifiers namely Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT or RT) and Random Forest (RF) are used by supervised learning methods. The final classification output is determined by the majority voting of the outputs of each ML classifier which works in parallel. By adopting this method, a classification accuracy of 99.87% is achieved.

Keywords Medical image processing · Denoising · Machine learning · Ensemble learning algorithms · Classifier · Estimator

K. Kumarasamy (✉) · S. Maria Wenisch · S. Balaji · L. J. Jenifer Suriya · A. Jerlin · S. Robert Rajkumar
Loyola-ICAM College of Engineering and Technology (LICET), Chennai 600034, India
e-mail: k.kunaraj@gmail.com

S. Maria Wenisch
e-mail: wenischs@gmail.com

S. Balaji
e-mail: sbalaji@licet.ac.in

L. J. Jenifer Suriya
e-mail: jenifersuriya.lj@licet.ac.in

A. Jerlin
e-mail: aruljerlin@licet.ac.in

S. Robert Rajkumar
e-mail: robertrajkumar@licet.ac.in

1 Introduction

Ensemble learning improves the results of machine learning algorithms as its overall outcome combines the results of many such learning methods. Multiple individual machine learning algorithms can be trained either sequentially or parallel with the aid of other representative methods like bagging [1], boosting [2, 3] and stacking [4] techniques. Random forest classifiers which is an ensemble of several decision trees are well known machine learning algorithms which is used for classification and other statistical problems [5]. Apart from the supervised classification problem, RF classifiers are also used for regression, estimation, manifold learning and semi-supervised learning. [6].

The pixel classification depends on the efficiency of statistical pixel parameters like Robust Outlyingness ratio (ROR) [7], Rank-Ordered Absolute Difference (ROAD) [8], Median absolute deviation (MAD) [9], Rank-ordered logarithmic difference (ROLD) [10], S-Estimate [11] and Absolute Deviation from Median (ADM) [9]. The ML classifier accuracy and the efficiency of these pixel parameters defines the efficiency of classification of corrupted pixels [20]. To further boost this, ensemble learning technique is adopted here.

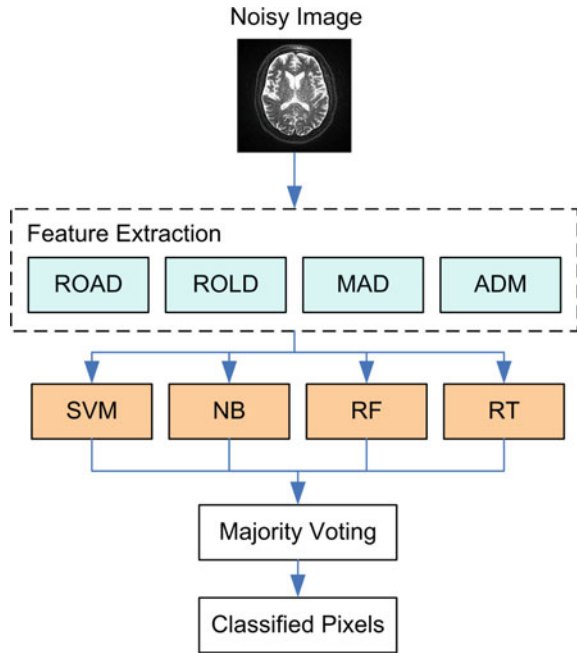
By adopting the ensemble learning methods in the machine learning classifier problem, we have improved the quality of classification of corrupted pixels. The misclassified pixels by a classifier would be rightly classified by other ML classifiers and a majority voting will put the pixels in proper groups. The paper is organized as follows: Sect. 2 introduces ensemble learning and techniques involved. Section 3 introduces statistical pixel features and various types. Section 4 gives an introduction to ML classifiers and Sect. 5 deals with the implementation methods and the experimental results obtained, and Sect. 6 provide concluding remarks.

1.1 Problem Statement

While the medical images corrupted by additive noises such as random valued impulse noise and additive white Gaussian noise is given, various statistical parameters from each pixel is computed as an initial step. Few examples are ROLD, ROAD, and ROR etc. Further these parameters are used to train an ensemble of ML classifiers like RF, SVM, and DT etc. These pre-trained classifiers can be used to detect noise pixels in real time medical images. Figure 1 represents the noise pixel classification problem statement. Majority voting is adapted to identify the corrupted pixels as classified by the ML classifiers.

Then noise model considered in our problem is the random valued impulse noise (RVIN) where the corrupted pixels takes any random value [0,255] and it has a uniform distribution. The probability distribution is given by;

Fig. 1 Pixel classification using ensemble learning



$$f(x) = \begin{cases} \frac{p}{2m} & : 0 \leq x < m \\ 1 - p & : x = y_{i,j} \\ \frac{p}{2m} & : 255 - m < x \leq 255 \end{cases}$$

2 Ensemble Learning

Many researchers use various machine learning algorithms for classification and ensemble learning to improve the accuracy of the overall classifier output for better decision making. There are other suitable applications of ensemble learning in estimation, error correction etc. The classification error is attributed to the accuracy and the precision of the classifier. The former depends on the bias and the latter depends on the variance. Normally, they have tradeoffs and a high bias (for improving accuracy) leads to low variance and hence low precision. Hence, the individual classifiers are trained with different datasets and a fixed bias. The output of these classifiers are combined together to reduce the variance without compromising accuracy and precision. Figure 2 shows the misclassified output, both false positive (FP) and false negative (FN) and it is clear that the pixels classified as FP and FN differs for each classifiers.

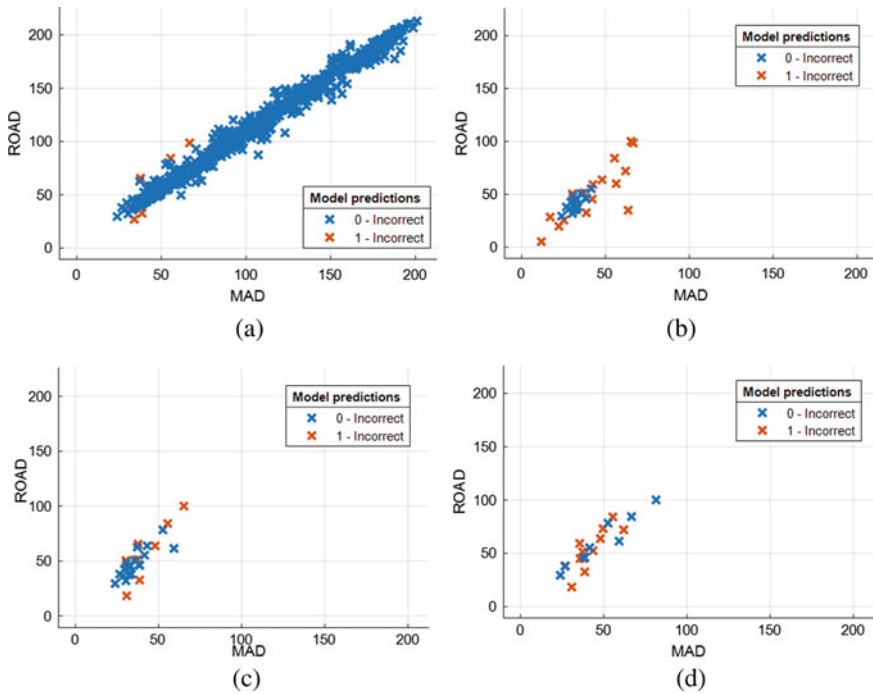


Fig. 2 Misclassified pixels of different classifiers **a** NB **b** RT **c** SVM **d** EML

Figure 2a, b and 2c corresponds to the misclassified output of the classifiers NB, RT and SVM plotted for the feature ROAD and MAD. The same plot can be done for other classifiers and features but we are restricting to this figure. Figure 2d is the misclassified output of the ensemble learner which clearly shows an improvement on classification accuracy.

The individual class labels predicted by each machine learning classifiers in parallel can be combined using several techniques. Majority Voting can be considered if more than half of the classifiers has the same predicted class. It is defined as;

$$\sum_{t=1}^T d_{t,c*} = \max_c \sum_{t=1}^T d_{t,c}$$

where d , the decision of the classifier and T is the total number of classifiers. The probability of ensemble success is given by P_{ens} .

$$p_{ens} = \sum_{k=\frac{T}{2}+1}^T \binom{T}{k} p^k (1-p)^{T-k}$$

The weighted majority voting will assume weights to individual classifiers thereby providing additional preference to whichever classifier is more suitable for the problem.

$$\sum_{t=1}^T w_t d_{t,c*} = \max_c \sum_{t=1}^T w_t d_{t,c}$$

w_t is the weight added.

Further the classifier can be trained parallel with independent data (Bagging) or sequentially by considering the output of previous stages (Boosting).

To calculate the overall probability of confidence, from the individual confidence level, sum of probabilities can be used.

$$p'_i = \frac{\sum_{j=1}^m p_{ij}}{\sum_{i=1}^n \sum_{j=1}^m p_{ij}}, \quad i = 1, \dots, n,$$

Product of individual classifier probability is given by;

$$p'_i = \frac{\prod_{j=1}^m p_{ij}}{\sum_{i=1}^n \prod_{j=1}^m p_{ij}}, \quad i = 1, \dots, n,$$

3 Statistical Pixel Parameters

We extract several statistical features which are already proven to be efficient for noise estimation and denoising problems. The following features are extracted from a surrounding window of the pixel of interest and it could be a 3X3 window or a 5X5 window, selected based on the noise intensity. The readers are recommended to see the references for the detailed discussion on the following statistical parameters.

3.1 ROR

The ROR of a POI is calculated using the following equation.

$$u(x, y) = \left| \frac{g(x, y) - \text{med}(W_5)}{\text{MADN}(W_5)} \right|$$

(x, y) is the location of the pixel $u(x, y)$. A window of size $(2N + 1) \times (2N + 1)$ is chosen around the pixel of interest (POI) to calculate the parameters. Let it be;

$$\Omega_y(N) = \{y_{i+k,j+l} : -N \leq l, k \leq N\}$$

Normally, we consider N as 1, 2 or 3 depending on the noise intensity. Median of the chosen window is given by; $\text{Med}(\Omega_y(N)) = \text{Median}(\Omega_y(N))$

From this median value the MAD, MADN and hence ROR can be calculated using the following expressions [7];

$$\begin{aligned} \text{MAD}(\Omega_y(N)) &= \text{Med}\{|\Omega_y(N) - \text{Med}(\Omega_y(N))|\} \\ \text{MADN}(\Omega_y(N)) &= \text{MAD}(\Omega_y(N))/0.6457 \\ \text{ROR}_{(y_{i,j})} &= |(y_{i,j} - \text{Med}(\Omega_y(N)))/\text{MADN}(\Omega_y(N))| \end{aligned}$$

3.2 Road

N denoting the pixel coordinate and the $(2N + 1) \times (2N + 1)$ is the current window centered at $(0, 0)$.

i.e., $\Omega_N = \{(s, t) | -N \leq s, t \leq N\}$ and let $\Omega_N^0 = \Omega_N \setminus (0, 0)$.

$d_{st}(y_{i,j}) = |y_{i+s, j+t} - y_{i,j}|$, $\forall (s, t) \in \Omega_N^0$, where d_{st} denotes the absolute difference between the centre pixel and the surrounding pixels in the taken window $(2N + 1) \times (2N + 1)$.

After sorting the elements of d_{st} in the ascending order, the sum of first k elements (r_k) is calculated to have the ROAD value of the current window.

$$\text{ROAD}_m(y_{i,j}) = \sum_{k=1}^m r_k(y_{i,j}) \text{ and } 2 \leq m \leq (2N + 1)^2 - 2.$$

3.3 Rold

To avoid the misclassification of pixel if the corrupted pixel value varies randomly, a logarithm can be considered while finding the absolute differences.

$$\tilde{D}_{st}(y_{i,j}) = \log_a |y_{i+s, j+t} - y_{i,j}|, \forall (s, t) \in \Omega_N^0$$

Hence for any $a > 1$, the number \tilde{D}_{st} is always in $(-\infty, 0]$.

To keep the dynamic range $[0, 1]$, a truncation scheme is used with a linear transformation:

$$D_{st}(y_{i,j}) \equiv 1 + \max\{\log_a |y_{i+s, j+t} - y_{i,j}|, -b\}/b; \forall (s, t) \in \Omega_N^0.$$

a and b are positive numbers which controls the shape of the logarithmic function and truncation position respectively. The efficiency of detection depends on the selection on a , b and it can be selected using the function $h_{a,b}(x)$, which is defined as;

$$h_{a,b}(x) = 1 + \max\{\log_a x, -b\}/b, (x \geq 0).$$

3.4 MAD

The noise variances can be easily calculated using MAD and at times MAD may not perform well in case of highly corrupted image. The median absolute deviation about the median is given by.

$MAD_n = b \times \text{med}_i |x_i - \text{med}_j x_j|$, where $\text{med}_j x_j$ is the median pixel value in the sub-window.

3.5 S-Estimate

This is one of the robust noise estimator which works well even for highly corrupted images with edges.

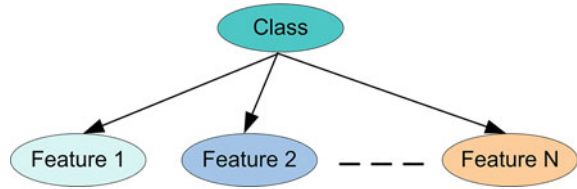
$$S = \text{med}_i \{ \text{med}_j |t_i - t_j| \}$$

$t_i, i = 1, \dots, N$ is the pixel window with N pixels and the inner median of $\{|t_i - t_j|\}$ need to be computed first followed by the outer median.

3.6 ADM

ADM (5×5) can be calculated by, $u(x, y) = \text{abs}(g(x, y) - \text{med}(W_5))$, and ADM (3×3) by, $u(x, y) = \text{abs}(g(x, y) - \text{med}(W_3))$.

Fig. 3 N-Features defining a particular class



4 Machine Learning Algorithms

4.1 Bayesian Classifier

The Bayesian classifier predominantly has a graph (directed acyclic in nature) in which each branch nodes denotes a variable X_i having a probability of $p(X_i | \prod_i)$, $p_i(i)$ is the parents of X_i .

The joint probability distribution is defined as: $p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i | \prod_i)$

The structure of the Bayesian network and the associated features determines the parameters of the network. The feature and the class are independent as shown in Fig. 3 in Naïve Bayes classifier which has good classification property even with less data [12].

4.2 Random Forest Classifiers

Decision trees are good in classification given a set of parameters and bag of random trees will further improves classification accuracy. [13, 14]. The structure of the random trees and its relevant parameters need to be fixed based on the training data which is not possible manually. The parameters extracted from the image pixels are used for this purpose. Each node of the tree is split based on the following criteria. $h(v, \theta_j) \in \{0, 1\}$

Parameters considered in supervised learning of trees are, $\theta = (\phi, \psi, \tau)$

ψ —Geometric primitive for separating data,

τ —Threshold for separating classes

ϕ —Variable meant for selecting appropriate features from the vector v .

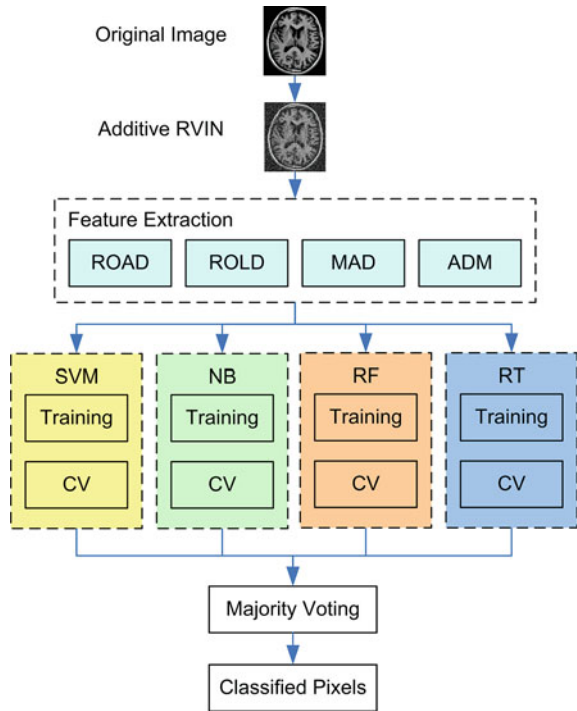
The learned class label is $c \in C$ with $C = \{c_k\}$, for the input dataset $v = (x_1, \dots, x_d) \in R^d$.

Readers are encouraged to read about various machine learning classifiers available in the literature.

5 Implementation and Results

The medical images from the MRI open dataset—Open Access Series of Imaging Studies (OASIS) [21] is collected and random value impulse noise of various degree ranging from 10 to 50% are added to the image dataset. Creating a classifier model has two stages viz. training and cross validation. Several models are chosen and they are trained in parallel with the same set of medical images. Cross validation ensures that models with acceptable accuracy are obtained and stored for real time classification. Finally the ensemble algorithm labels each pixel as corrupted or not by majority voting or using classification error probabilities. Figure 4 shows this sequence of operations schematically, where each model is trained and cross validated in parallel, after extracting the features. Figure 4 is limited, and not all the features and ML algorithms are shown. Experimentally it is found that, majority voting works well for this classification problem and it is validated by simulating with large datasets of medical images.

Fig. 4 Implementation methodology



5.1 Algorithm: Supervised Learning

```

for all the image in the database
  Read a random image
  Discard the image if already taken, else
  Change the image to fixed resolution (256 × 256)
  Convert the image of any form into binary
  for all pixels in the image

    Add impulse noise of random values and position
    Compare each pixel with original pixel
    Create a dictionary of pixel status (corrupted o original)

  end for
  for all pixels in the image

    Compute statistical parameters
    (e.g. ROAD MAD etc.)
    Store SP in each column
    Train each model

  end for

  Repeat the above steps for the next image
  Cross validate each model

end for

```

After training and cross-validation, the elite models are stored. The images in the constructed library after corrupted are randomly fed to these models for identifying corrupted pixels. The output of each model is then taken by majority voting which gives the final classified output. Table 1 provides the classification accuracy of each model and the overall ensemble output. Clearly, the classification accuracy of the proposed ensemble learner (EML) is high for low noise intensities but loses to RF classifier when noise intensity increases. It is expected to have low or very low noise pixels in the real time scenario for medical images, as high number of noise pixels

Table 1 Classification Accuracy

Noise (%)	Accuracy %				
	SVM	NB	RF	DT	EML
10	96.23	97.51	99.84	98.91	99.89
20	95.12	96.19	98.97	97.34	99.03
30	93.34	94.06	96.62	95.30	97.21
40	90.78	91.29	93.02	92.17	92.18
50	86.61	87.32	88.59	87.89	87.23

might make the image unusable because of the potential misinterpretation in the clinical context.

Figure 5 plots the classification accuracy with the experiment data available in Table 1. A steeper decrease in the classification accuracy is evident from Fig. 5, as the noise level increases over 30%. Very high noise levels are not considered here, as medical images cannot be used with high noise intensities.

Table 2 compares the average classification error of the proposed model (EML) with other standard pixel classification algorithms for randomly chosen medical images which were not used while training the model.

Figure 6 plots the classification accuracy of various algorithms for noise pixel identification. EML works well for classifying low noise intensities and works on par with ML-RF when noise intensity increases.

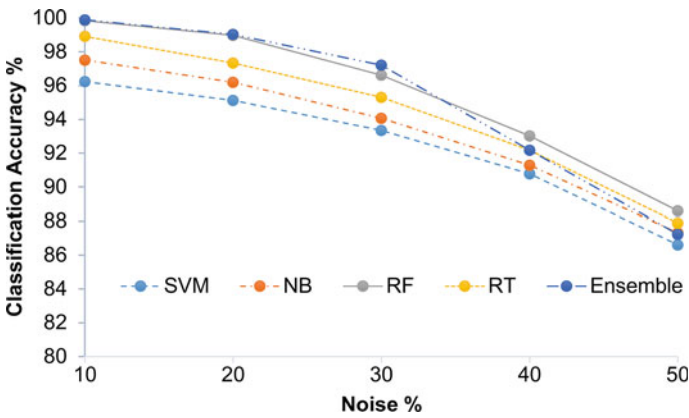


Fig. 5 Comparison of classification accuracy

Table 2 Average classification error-comparison

Noise detectors	Noise %			
	20	30	40	50
SWM [15]	4.76	7.85	9.74	12.40
TRI [16]	4.52	6.57	8.58	11.11
PWMAD [17]	6.06	8.30	10.44	12.75
ROLD-EPR [18]	5.14	7.15	8.21	9.27
ROR-NLM [19]	4.39	5.45	6.72	8.19
ML-RF [20]	1.03	3.38	5.79	7.98
EML	1.01	3.21	5.82	8.28

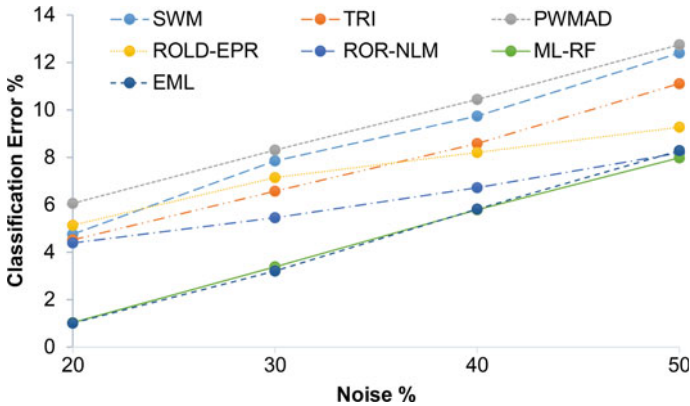


Fig. 6 Comparison of classification error

6 Conclusion

An ensemble of machine learning classifiers is proposed and it is applied for classifying the corrupted pixels in preferably medical images while it can be used for other images as well. The accuracy of classification depends on the robustness of the features (pixel parameters) and the classifiers. The advantage of ensemble learning is utilized to further improve the accuracy of classification of noise pixels. From the conducted experiments, it is clear that the proposed method outperform the conventional pixel classifiers and also provides better classification results than random forest based classifiers implemented earlier the same author. An average classification error as low as 1.01% was achieved for 20% corrupted pixels and the algorithm works well for low and mid-level of corruption while improving the accuracy for high noise level is still a challenge.

References

1. Breiman, L.: Bagging predictors. *Mach. Learn.* **24**(2), 123–1407 (1996)
2. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **51**(2), 181–2075 (2003)
3. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to Boosting. *J. Comput. Syst. Sci.* **55**(1), 119–1396 (1997)
4. Wolpert, D.H.: Stacked generalization. *Neural Netw.* **5**(2), 241–2608 (1992)
5. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
6. Criminisi, A.: Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends@ Comput. Graph. Vision* **7**(2–3), 81–227 (2011)
7. Bo, X., Zhouping, Y.: A universal denoising framework with a new impulse detector and nonlocal means. *IEEE Trans. Image Process.* **21**(4), 1663–1675 (2012)

8. Garnett, R., Huegerich, T., Chui, C., Wenjie, H.: A universal noise removal algorithm with an impulse detector. *IEEE Trans. Image Process.* **14**(11), 1747–1754 (2005). <https://doi.org/10.1109/TIP.2005.857261>
9. Petrovic, N.I., Crnojevic, X.V.: Universal impulse noise filter based on genetic program-ming. *IEEE Trans. Image Process.* **17**(7), 1109–1120 (2008)
10. Yiqiu, D., Chan, R.H., Shufang, X.: A detection statistic for random-valued impulse noise. *IEEE Trans. Image Process.* **16**(4), 1112–1120 (2007)
11. Petrovic, N.I., Crnojevic, V.: Impulse noise filtering using robust pixel-wise S-estimate of variance. In: *Proc. EURASIP J. Adv. Signal Process.* **8** (2010)
12. Sebe, N., Cohen, I., Garg, A., Huang, T.S.: In: *Machine Learning in Computer Vision*” N, vol. 25. Springer Netherlands (2005)
13. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends Comput. Graph. Vision* ISSN 1572–2740, Now Publishers, ISBN 1601985401, 9781601985408 (2012)
14. Criminisi, A., Shotton, J.: *Decision forests for computer vision and medical image analysis.* Springer, London (2013). <https://doi.org/10.1007/978-1-4471-4929-3>
15. Sun, T., Neuvo, Y.: Detail-preserving median based filters in image processing. *Pattern Recognit. Lett.* **15**(4), 341–347 (1994)
16. Chen, T., Ma, K.K., Chen, L.H.: Tri-state median filter for image denoising. *IEEE Trans. Image Process.* **8**(12), 1834–1838 (1999)
17. Crnojevic, V., Senk, V., Trpovski, Z.: Advanced impulse detection based on pixel-wise MAD. *IEEE Signal Process. Lett.* **11**(7), 589–592 (2004)
18. Dong, Y., Chan, R.H., Xu, S.: A detection statistic for random valued impulse noise. *IEEE Trans. Image Process.* **16**(4), 1112–1120 (2007)
19. Xiong, B., Yin, Z.: A universal denoising framework with a new impulse detector and nonlocal means. *IEEE Trans. Image Process.* **21**(4), 1663–1675 (2012)
20. Kunaraj K., Maria Wenisch S., Balaji S., Mahimai Don Bosco F.P.: Impulse noise classification using machine learning classifier and robust statistical features. In: Smys, S., Tavares, J., Balas, V., Iliyasa A. (eds.) *Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing*, vol. 1108. Springer, Cham (2020)
21. <http://www.oasis-brains.org/>

Image Data Preservation with Fractional Sine Transform and Dual Chaotic Sequence



Sharad Salunke, M. Venkatadri, Md Farukh Hashmi, and Bharti Ahuja

Abstract The security issue of visual data is prevalent in today's era because of the fast growth of communication technology. The cutting edge technologies like machine learning, deep learning, and cloud computing has given new heights to data communication. And in this virtual age the preservation of digital images are of prime importance because it carries much personal and sensitive information from an individual's personal data to the any nation's defence data. So, the growth in communication technology not only given heights to mankind, but also attracts the attackers for their mischievous activities of data theft. Therefore, to safeguard this, various image security methods are already proposed. So, in the chain, a hybrid method combining discrete fractional sine transform with logistic and Arnold cat map is proposed here. This hybrid combination is noteworthy in the sense that it has higher PSNR and least error in the reconstruction as compared to their ancestor algorithm.

Keywords Discrete fractional sine transform · Logistic map · Arnold cat map · Image encryption · Visual data security

S. Salunke (✉)

Department of ECE, Amity University Madhya Pradesh, Gwalior, India

e-mail: sharad.sal@gmail.com

M. Venkatadri

Department of CSE, Amity University Madhya Pradesh, Gwalior, India

e-mail: vmarriboyina@gwa.amity.edu

M. F. Hashmi

Department of ECE, NIT Warangal, Warangal, India

e-mail: mdfarukh@nitw.ac.in

B. Ahuja

Department of IT, NIT Raipur, Raipur, India

e-mail: bharti.salunke99@gmail.com

1 Introduction

Due to excellent growth in information and data transmission and reception, either text, audio or video, unapproved means of entry to information proven to be simpler and more widespread in both mobile and wired communication networks. Further, due to the continual circulation of digital information over the globe through the communication medium, protecting them from third parties has become extremely important. Information protection has now been the most critical and central issue. The bulk of the knowledge exchanged on the internet was visual pictures. Consequently, the details of the visual data are very dissimilar from the text. It has a massive collection of details, a higher redundancy, and an extreme combination across various pixels. Thus, strong safety and security while transmission of digitized images is required in a number of operations including formal and casual administration [1].

Numerous important uses, such as military visual libraries, sensitive video conferencing on various available applications used now a days, diagnostic imaging services holding sensitive patient data, cable TV, electronic personal photo archives through various social media platforms, etc., need a secure, efficient, and comprehensive storage network to save and transfer digitized photographs. Prerequisites to meet advanced picture security necessities have prompted the improvement of ground-breaking encryption strategies. Numerous security algorithms [2] have been suggested in the literature during the last decade on the basis of various standards. Among, chaotic encryption techniques are viewed as appropriate for viable use as they give a decent mix of speed, higher protection, difficulty, rational computing overheads, and computing strength, etc. Visual data have other attributes, such as: data consistency, good similarity between neighbouring pixels, less flexibility in comparison with textual information, i.e., a minor shift in the parameter of every picture element of the picture does not substantially impair the attributes of the image and the bulk capability of the data, etc.

As a result, conventional algorithms viz. AES, DES, RSA, IDEA, etc., really aren't worthy for secure image transmission as far as encryption is concerned, since these methods needs much time for computation and greater computing power to execute. Hence, only those cryptographic algorithms which completes in less time while maintaining safety as well are preferable. Therefore, as a matter of fact the encryption system, which operates very gradually, may also provide a higher degree of protection capabilities, but may be of no practical use for real-time applications.

When we talk of signal and image processing in any sort whether it is segmentation, pattern analysis, compression, or encryption, the signal processing community can't neglect the Fourier Transform for its numerous applications in this field. Besides the Fourier Transform, the sine and cosine transformations that are centered, respectively, on half-range extensions of a system along sine and cosine-dependent functions too are key instruments in signal analysis. Notwithstanding some absence of class in regards to the FT in their properties, the CT and ST have their own application territories. Fractional order of Fourier Transform is called the fractional fourier transform (FrFT). There are numerous areas where FrFT found useful such as signal

and image processing, optics, analysis of signals, encryption, and compression of visual data [3–5].

The FrFT is determined type of the classical FT in the sense that it utilizes the periodicity property by angle rotation to get the fractional order. This was published in the mathematics research a number of years ago but seems to have remained relatively obscure [6]. In view of Eigen decomposition of Discrete Fourier Transform matrices Pei [7] and Yeh he have given their explanations. In addition to above, it was also well characterized by Ozaktas and Pei [8, 9].

The fractional version of cosine and sine were primarily based on FrFT, and that was repeated by Pei. Pei and Yeh help strengthen the cosine transform with DFrCT and DFrST. Each essentially retains the DFrFT angle additive property. In addition, all are used to reduce DFrFT's computational weight [10].

The rest of the article is classified as follows: Segment 2 focuses on the preliminary concepts. Segment 3 shows the image encryption and decryption algorithm as well as segment 4, showing the performance of the proposed scheme via computational experiments in the results section. In the segment 5, at last a conclusion is made.

2 Preliminary Postulate

In this section, basic preliminary knowledge is expressed before going into the actual algorithm for better understanding. Here, utilized techniques such as logistic map, Arnold cat map, and Discrete Fractional Fourier Transform are explained.

2.1 Logistic Chaotic Map

The effectiveness of logistic chaotic map lies in its simple equation and non-linearity, therefore, extensively utilized researchers and scientists from variant areas. It is expressed in Eq. (1).

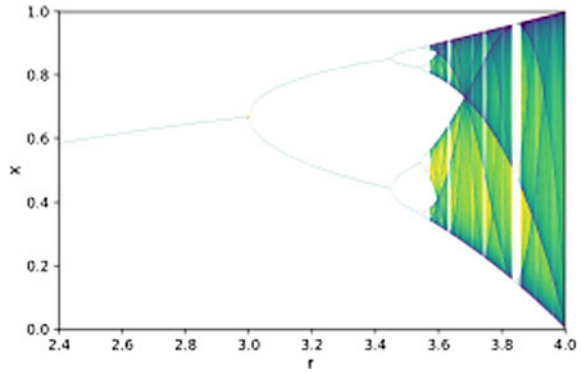
$$Y_{n+1} = p * Y_n * (1 - Y_n) \tag{1}$$

In the function stated above, two parameters are utilized to produce random sequences.

This method, therefore, is having disorderly actions, which is commonly known as chaos or chaotic behaviour. A schematic method of presenting this process is as shown in Fig. 1, where the bifurcation diagram offers knowledge regarding machine dynamics where value starts if $p = 3.75$ to $p = 4$. It acquires the values between 0 and 1 for $\{y_1, y_2, \dots, y_n\}$ [11, 15].

The method with the keys used is useful for any size of the image with the initial values as mentioned. From the related study, we noticed that several researchers have used the logistic map because of its usefulness in encryption and decryption of

Fig. 1 Bifurcation diagram of logistic chaotic map



images. The function is extremely receptive to introductory worth, with the end goal that minor varieties in beginning conditions produce broadly contrasting results. However, there are some issues with the logistic map, including stable windows, blank windows, inconsistent sequence distribution, and poor key. Double chaotic map is included in this paper to ease these challenges and create greater strategic room for disorderly behaviour.

2.2 *Arnold Cat Map*

This map was given the name after Vladimir Arnold, who proved their impacts utilizing a cat image in 1967. Arnold’s cat diagram as shown in Fig. 2 also described

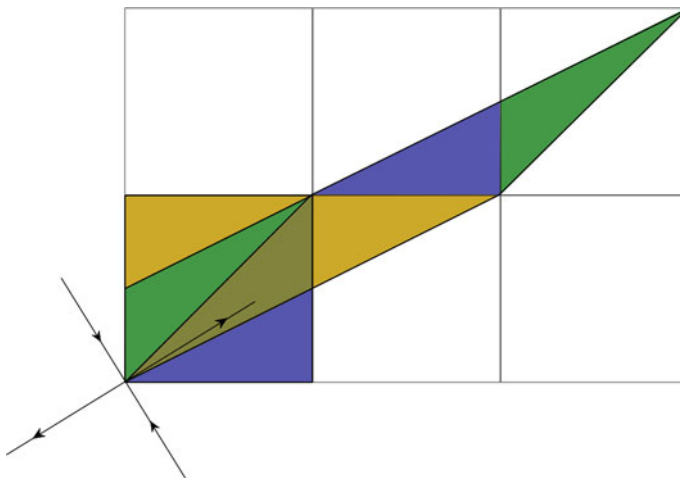


Fig. 2 Arnold cat map’s diagram

as a repetitive folding and extending cat chaotic map in a small region and widely utilized in digital chaotic encryption.

A picture is struck with the transition, which actually randomizes the initial arrangement of the picture elements, as per the Arnold’s transformation. If computed several times although, the actual picture will inevitably reoccur. The number of computations considered shall be defined as the period of the Arnold. The duration relies on the size of the picture; i.e., Arnold’s cycle would be distinct for various size images [12].

$$\begin{bmatrix} u_{n+1} \\ v_{n+1} \end{bmatrix} = A \begin{bmatrix} u_{(n)} \\ v_{(n)} \end{bmatrix} \pmod{N} = \begin{bmatrix} 1 & p \\ q & pq + 1 \end{bmatrix} \begin{bmatrix} u_{(n)} \\ v_{(n)} \end{bmatrix} \pmod{N} \tag{2}$$

In the above stated equation:

N = Image size; p, q = positive integers; $\det(A) = 1$; and (u_n, v_n) = Sample position for $N \times N$ information of image, such that, $(u_n, v_n) \in \{0, 1, 2 \dots N-1\}$ and (u_{n+1}, v_{n+1}) is the location converted from cat map [16].

There are two primary cause to Arnold map that influence disorderly development: stress (duplicate framework to extend u, v) and crease (calculated mod to get u, v in network). Equation (2) is utilized to change over the picture arranges for every single pixel. The subsequent picture is a mixed picture, when all the directions are changed. If the resulting picture meets our expected goal (i.e., up to hidden key) at a certain point of computations, we have obtained the scrambled picture that was demanded. Photo decryption is dependent on transition intervals (which is the amount of iterations to execute = cycle hidden key for Arnold) [13].

2.3 Discrete Fractional Sine Transform (DFrST)

The method extends with DFrFT. It is an ordinary DST type, whose association with DFrST and DFT is homogeneous [14]. When formulating the DFrST, the similar approach is being used to establish the matrix as used in DFrFT, i.e., $V = [v_1, v_2, \dots, v_N]$. In addition, from the above observation, DFrST’s eigen vectors are represented by,

$$V_s = [s_1, s_2, \dots, s_N] \tag{3}$$

The Kernel of DFrST is composed as,

$$R_s^\alpha = V_s D_s^\alpha V_s^t \tag{4}$$

Eigen value diagonal matrices for DFrST, i.e., D_s^α , is therefore selected as,

$$D_s^\alpha = \text{diag}[\exp(-2i\pi\alpha/M), \exp(-6i\pi\alpha/M), \dots, \exp(-2i\pi(2N - 1)\alpha/M)] \tag{5}$$

DFrST's 1-dimensional x and 2-dimensional representation y are formulated as:

$$X_s^\alpha = R_s^\alpha x \tag{6}$$

$$Y_s^\alpha = R_s^\alpha y (R_s^\alpha)^t \tag{7}$$

When $\alpha = 0$, the kernels of the DFrST is an identity matrix.

We are using orthogonal Eigen vectors to characterize the DFrST, so DFrST computational properties are equitable to DFrFT.

3 Proffered Algorithm

In this segment the proposed encryption system for digital image preservation is given advances included and clarified with the assistance of a process model as portrayed in Fig. 3.

3.1 Encryption and Decryption with DFrST and Dual Chaos

Digital images include many data but the data are strongly correlative and steady images often include plenty of redundant spatial information. In order to hide or protect the digital data or an image, we are using two-dimensional DFrST algorithm. The periodicity property of DFrST is used here which uses DFrST's proper order to decrypt the encrypted image. Further, two chaotic maps; i.e., Logistic and Arnold Map are too incorporated to enhance the security which improves the degree of randomness with two chaotic equations and also helps to increase the key space.

The steps listed below depict the proposed method as shown in Fig. 3.

1. Arnold cat map is added to actual picture I , and the resulting converted image is I_1 .

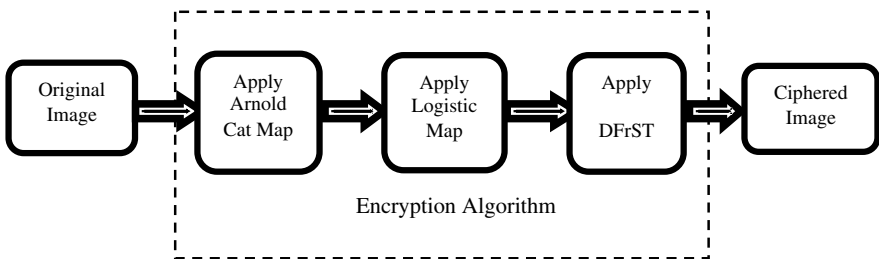


Fig. 3 Process of the image data preservation with fractional sine transform and dual chaos

2. Logistic map is applied to I_1 and which further gives output I_2 .
3. Every row element of I_2 is changed by one-dimensional DFrST, with $a = 0.77$ as order of fraction resulting in changed picture I_3 .
4. Every column element of I_3 is again changed by DFrST for column, with $b = 0.77$ resulting in I_4 . This I_4 is indicated as the ciphered picture.
5. In decryption phase, inverse DFrST utilizing $a' = (-a)$ and $b' = (-b)$ is applied to ciphered image and the resulting image is now I_1' .
6. Finally, apply inverse logistic map and inverse Arnold map to get I' which is the decrypted image.

4 Results and Discussion

This part is dedicated to the computational outcomes. The programming and simulations are done on MATLAB platform to test the effectiveness and also to validate the algorithm. For the illustration, 256×256 grayscale cameraman image is taken. Figure 4 illustrates the original, encrypted, and decrypted image using the proposed method. The correlation co-efficient is an important metric for testing the effectiveness of the encryption algorithm. The horizontal, vertical, and diagonal value comparison is shown in Table 1 and its distribution graph is shown in Fig. 5. Further, the PSNR between actual image and decrypted image is come out to be infinite and the mean square error is 0, which proves the quality and successful reconstruction of the proposed work.

5 Conclusion

We have proposed a novel hybrid combination for encrypting and decrypting the image using a discrete transformation of fractional sine and dual chaos. Due to the extreme sensitivity of the chaos function, the effect of proposed cryptosystem is better than with DFrST alone. The two chaotic maps used further enhance security in terms of key space. Finally, proposed cryptosystem is highly stable and reliable due to the nature of DFrST and the confusing properties of the chaos. For future work, this novel hybrid combination may be used with the advanced methods like machine and deep learning to trace out and classify the various kinds of attacks. Also, the method is suitable and might find applications with internet of things to fulfill real-time data needs of artificial intelligence.

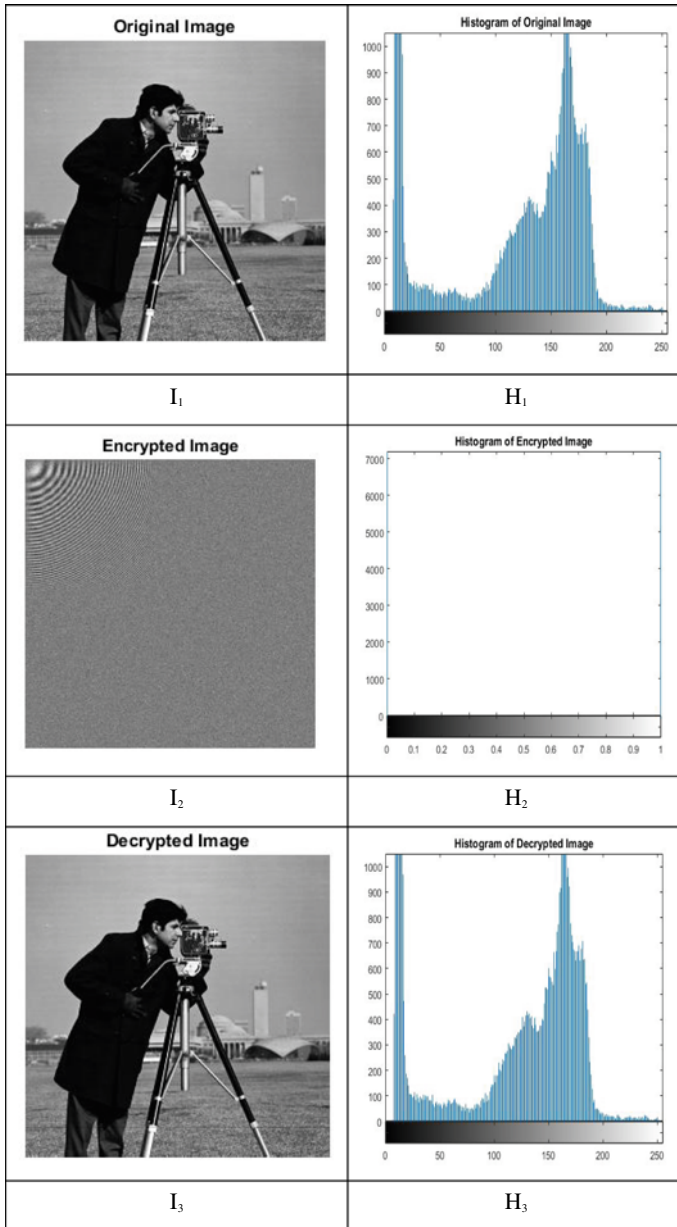


Fig. 4 Illustration of processed image cameraman: (I_1-I_3) are original, encrypted, and decrypted images. And (H_1-H_3) are corresponding histograms

Table 1 Correlation co-efficient horizontal, vertical, and diagonal values for original and encrypted image

Original image			Encrypted image		
Horizontal	Vertical	Diagonal	Horizontal	Vertical	Diagonal
0.9334	0.9592	0.9086	0.1836	0.2072	-0.2948

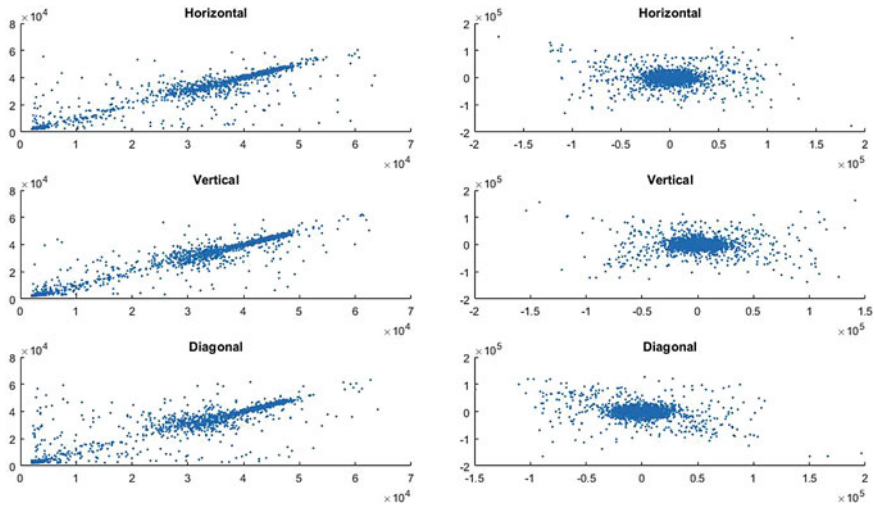


Fig. 5 Illustration of correlation co-efficient for original and encrypted image

References

1. Singh, K., Kaur, K.: Image encryption using chaotic maps and DNA addition operation and noise effects on it. *Int. J. Comput. Appl.* **23**, FCS (2011)
2. Pareek, N.K., Patidar, V., Sud, K.K.: Image encryption using chaotic logistic map. *Image Vision Comput.* **24**, 926–934, Elsevier (2006)
3. Tao, R., Deng, B., Wang, Y.: Research progress of the fractional fourier transform in signal processing. **49**, 1–25, Springer (2006)
4. Ozaktas, H.M., Zalevsky, Z., Kutay, M.A.: *The Fractional Fourier Transform with Applications in Optics and Signal Processing*. Wiley, New York (2000)
5. Namias, V.: The fractional order Fourier transform and its application to quantum mechanics. *J. Inst. Math. Appl.* **25**, 241–265, Oxford university press (1980)
6. Almeida, L. B.: The fractional fourier transform and time-frequency representations. *IEEE Trans. Signal Process.* **42**(11), 3084–3091, IEEE (1994)
7. Pei, S.C., Yeh, M.H.: Improved discrete fractional fourier transform. *Opt. Lett.* **22**, 1047–1049, OSA The optical Society (1997)
8. Candan, C., Kutay, M.A., Ozaktas, H.M.: The discrete fractional fourier transform. *IEEE Trans. Signal Process.* **48**(5), 1329–1337, IEEE (2000)
9. Pei, S.C., Hsue, W.L.: The multiple-parameter discrete fractional Fourier transform. *IEEE Signal Process Lett.* **13**(6), 329–332, IEEE (2006)
10. Pei, S.C., Yeh, M.H.: The discrete fractional cosine and sine transform. *IEEE Trans. Signal Process.* **49**, 1198–1207, IEEE (2001)

11. Sam, I.S., Devaraj, P., Bhuvaneshwaran, R.S.: Chaos based image encryption scheme based on enhanced logistic map. pp. 290–300. Springer (2011)
12. Pradhan, C., Saxena V., Bisoi, A.K.: Imperceptible watermarking technique using Arnold's transform and cross chaos map in DCT domain. *Int. J. Comput. Appl.* **55**, (2012)
13. Abbas N.A. M.: Image encryption based on independent component analysis and Arnold's Cat map. *Egypt. Info. J.* **17**, 139–146 (2016)
14. Salunke, B.A., Salunke, S.: Analysis of encrypted images using discrete fractional transforms viz. DFrFT, DFrST and DFrCT. In: International Conference on Communication and Signal Processing (ICCSP), pp. 1425–1429. IEEE (2016)
15. Yu, C., Li, H., Wang, X.: SVD-based image compression, encryption and identity authentication algorithm on cloud. *IET Image Process.* **13**, 2224–2232, IET (2019)
16. Wang, C., Ding, Q.: A new two-dimensional map with hidden attractors. *Entropy* **20**, 322, MDPI (2018)

Enhancing Deep Learning Capabilities with Genetic Algorithm for Detecting Software Defects



Kajal Tameswar, Geerish Suddul, and Kumar Dookhitram

Abstract Regardless of existing and well-defined processes, some defects are inevitable, resulting in software performance degradation. The use of traditional machine learning techniques can automate the prediction of software defects. This automated approach significantly improves the quality of the finished product and reduces the cost incurred during development and maintenance stages. The accuracy of artificial neural networks for the automatic prediction of software bugs, can be further enhanced with the use of metaheuristics algorithms. We propose a hybrid approach which combines Genetic Algorithm (GA) and Deep Neural Network (DNN) to better classify software defects. GA is used as a pre-learning phase to automatically optimize the input features for the DNN, as irrelevant variables have a substantial negative impact on the prediction accuracy. Results from experiments using the PROMISE dataset, demonstrates that a DNN consuming optimized features yields better results.

Keywords Machine learning · Software bugs · Defect prediction · Hybrid model · Genetic algorithm · Deep neural network

1 Introduction

The technological evolution of software has become an essential and pervasive part of our personal and professional life. Today's wearable technology, implanted medical devices and autonomous driving cars are just a few examples that demonstrate the beginning of a new era of software transformation. The purpose of software defect prediction is to discover major design and programmatic issues which can reduce

K. Tameswar (✉) · G. Suddul · K. Dookhitram (✉)
University of Technology, Mauritius (UTM), Pointe-Aux-Sables, Mauritius
e-mail: kajaltameswar@gmail.com

K. Dookhitram
e-mail: kdookhitram@umail.utm.ac.mu

G. Suddul
e-mail: g.suddul@umail.utm.ac.mu

the huge costs and time imperatives associated with them [3]. IEEE defined the term fault or bug as ‘inappropriate and unexpected behavior in a computer program’. However, due to exponential growth in application software, the assurance of quality in software remains largely an unnoticed subject leading to performance degradation of the industry. As a concern, testing comes into play to find defects or bugs while running a program to produce a zero-defect software (Chauhan and Singh 2014). Without proper testing, a project becomes a definite recipe for disaster that can raise its cost and affects its quality. While bugs persistently continue to worsen the performance of software, the necessity of effective and rapid methods to find software defects is high.

There have been several techniques introduced to reduce the presence of defects in software. For instance, metrics based on object-oriented, traditional and process approaches have broadly been employed in almost every defect prediction model [6]. Further applications of software metrics are demonstrated by Khoshgoftaar et al. [13] with a statistical prediction model based on function-approximation problem analysis and regression. Unfortunately, most of the presented methods fail in providing efficient results, mainly because the architecture of each software is almost unique. As such, prediction models have to take into consideration parameters which are completely different, thereby having difficulty to generalize. To overcome this complicated issue, non-parametric techniques like machine learning and computational intelligence can be considered.

Despite of multiple powerful advances in programming languages and bug detection techniques, software defects affect virtually almost all software products and services. In response to this problem, researchers have widely been studying the topic bug prediction using machine learning approaches which have the potential to leverage the prediction of software bugs [1, 17, 21–23]. Nevertheless, there still exist many uncertainties with machine learning approaches, as no single techniques have prevailed due to existing imbalanced datasets and lack of formal approaches [8]. We present a novel hybrid approach in this paper using deep neural network along with GA to build an efficient classification-based optimization system for prediction of software defects.

The rest of the paper has been prepared as follows. Section 2 provides a short view of related works that have been done in the field of software defects defection. Section 3 provides the proposed model. Experimental outcomes and results of the proposed approach are described in Sect. 4 Performance analysis and discussion of results are discussed in Sect. 5. Finally, Sect. 6 presents a brief conclusion and future work of the proposed model.

2 Literature Review

This section discusses the different software defect prediction techniques identified in the literature. Recently, machine learning approaches have become very popular techniques for defect prediction in software [6, 11]. In this context, many algorithms

have been designed each having its own data requirements and levels of complexity. Examples comprise of regression algorithms, classification techniques, clustering methods, deep learning and hybrid techniques which is a blend of optimization algorithm and machine learning.

Several supervised classification algorithms such as neural networks, naïve bayes, Support Vector Machines (SVM), linear regression, and K-nearest neighbor, as described by Perreault et al. [14] have been used for the prediction and detection of software bugs. On the other hand, regression approaches have been tackled using SVM by Elish et al. [7]. SVM has also been used for classification [10] of defects, which has a special focus on the pre-processing of the input data. Shivaji et al. [19] investigated a naïve bayes classification algorithm combined with feature selection module for efficient prediction. Each of the approaches have shown different levels of efficiency, making them difficult to implement. A more efficient deep learning neural network model is presented by Yang et al. [25, 26]. Along the same approach the work of Gondra et al. [9] demonstrates that labelled datasets with software metrics can help better train neural network models. Another model proposed by Yang et al. (2006) shows the combination of neural network with radial basis function and Bayesian method.

In unsupervised clustering algorithms, the application of ambiguous datasets has been very popular. For instance, Bishnu et al. [5] came up with a k-means clustering model for software bug prediction. Hybrid approaches based on K-means algorithms have been attempted, such as application of the Neural-Gas and Quad Tree techniques for optimum exploration and cluster labelling of real-world datasets (Rani and Rajalakshmi, 2012; Meenakshi et al. 2012).

Hybrid approaches have the advantage of combining the best of different techniques and hence further improves the accuracy of prediction models. Azar et al. [2] developed a model using ant-colony optimization technique for prediction of software bug. Another study (Rong et al. 2016) proposed a hybrid Support Vector Machine model combined with the bat search algorithm. Manjula and Florence (2018) presented a machine learning based hybrid model by combining genetic optimization algorithm with decision tree algorithm. Wahono et al. [24] build a model using neural network based on bagging technique and genetic algorithm for prediction of software bug in order to improve performance.

In regards to the above work, we noticed that software defect prediction models have a high cost associated with it. While some approaches have high processing time other are intricately complex. Genetic Algorithm (GA) has been extensively used in neural network optimization and is known to be successful in achieving optimal solutions. While substantial work has been done regarding neural network parameter optimization using GA in several applications, there has not been sufficient research performed on investigating them in the field of defect prediction. To overcome this problem, we present a hybrid-based model using GA to optimize deep neural network for software defect prediction.

3 Proposed Model

The proposed software defect prediction model comprises of a Deep Neural Network (DNN) and Genetic Algorithm (GA). It therefore follows a two-fold approach, as below:

- (i) Application of GA for feature optimization.
- (ii) Application of DNN for classification purpose.

3.1 Genetic Algorithm

Genetic Algorithm is a metaheuristic evolutionary algorithm based on the principle of selection and mutation. In our context, GA is applied for the purpose of searching the parameter space, finding the global optimum solution and optimizing the weight and threshold of the neural network effectively [20]. The parameters that has been used for the implementation of GA were set as: size of population = 100; number of generations = 50; probability of crossover = 0.5; and mutation probability = 0.2.

3.1.1 Deep Neural Network

This section defines the Deep Neural Network used for this study related to prediction of software defects. Deep neural network is useful for the learning of effective features and discriminative patterns in nature, especially for software bug prediction [25, 26]. DNN can also be applied to unlabeled datasets. In our model, we used one input layer and 10 hidden layers to produce the output.

3.2 Hybrid Intelligence of Genetic Algorithm with Deep Neural Network

The fundamental ideologies of GA are to generate an initial population of chromosomes followed by selection and crossover in order to achieve effective population having the fittest chromosome (optimal value) among them. Figure 1 below shows the proposed architecture, involving the steps required to build the hybrid predictive model using GA together with DNN.

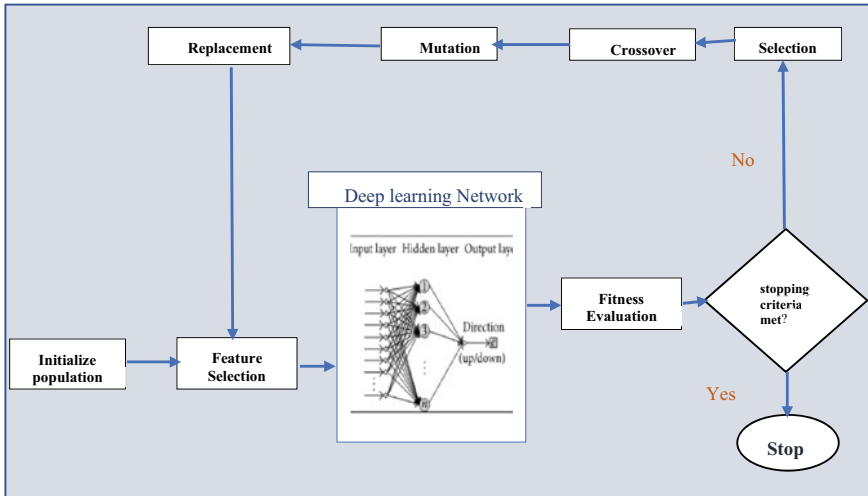


Fig. 1 Steps involving building process of GA and DNN

4 Experimental Study and Results

The experimental studies carried out for this proposed approach is described in this section. The hybrid DNN + GA model has been developed using Python packages (Tensor Flow, Scikit-Learn and Keras). TensorFlow is used to train and calculate accuracy of the prediction model. Scikit-Learn is used to read and split the dataset for training and testing purpose. Keras provides the possibility to speed up experimentation cycles on CPU and GPU. All experiments have been conducted using a laptop consisting of the following configuration: Corei7-6500U CPU, × 64 based processor and 16 GB RAM. The PROMISE datasets have been used for training and prediction.

4.1 Dataset Details

In PROMISE repository consists of five primary datasets, namely: AR1, AR3, AR4, AR5, AR6. Since they are all related having similar attributes (e.g. loc, comment loc, cyclomatic complexity), we have decided to combine the dataset altogether. The dataset consists of 29 features, and 1050 records, out of which 70% will be used for training and the remaining 30% used for testing. A random state of 65 is used to ensure that each experiment splits the dataset with the same record in every set to acquire appropriate calculation of prediction accuracy for the model.

The datasets are categorized as follows:

- (i) LOC counts (total_loc, blank_loc, comment_loc, code_and_comment_loc, executable_loc, unique_operands, unique_operators, total_operands, total_operators): Defines numbers of lines of code
- (ii) Halstead (vocabulary, length, volume, level, difficulty, effort, error time): Based on number of operators and operands
- (iii) McCabe (cyclomatic_complexity, cyclomatic_density, decision_density, design_complexity, design_density, normalized_cyclomatic_complexity, formal_parameters): This keeps a measure of the number of possible alternative paths through the code
- (iv) Others (branch_count, decision_count, call_pairs, condition_count, multiple_condition_count)

4.1.1 Fitness Evaluation Using Performance Evaluation Metrics

For the performance of the defect prediction model, the following metrics have been used using these annotations which are as follows:

See Table 1

5 Performance Analysis and Discussion of Results

5.1 Experimental Scenario 1

In first instant, experiments were conducted whereby only the DNN has been taken into consideration. These experiments are conducted for AR1, AR3, AR4, AR5, AR6 (combined dataset) with all 29 attributes in the dataset. The efficiency of the DNN

Table 1 Evaluation metrics

Metric	Description	Formula
Accuracy	Used for the determination of chromosomes selection and for performance measurement of the hybrid prediction model	$(TP + TN)/\text{Total number of samples used}$
Recall	The percentage result that have correctly been classified by our algorithm	$TP/(TP + FN)$
Precision	Defined as the proportion of occurrences predicted as defective which actually are defective	$TP/(TP + FP)$

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

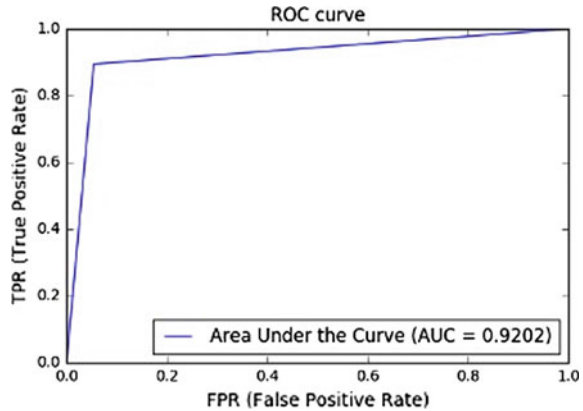
Table 2 Result for statistical performance analysis using deep neural network (DNN)

Precision	Recall	Accuracy
0.895	0.895	87.21%

Table 3 Confusion matrix using deep neural network (DNN)

Actual class	Predicted class	
Defective	5	17
Non-Defective	5	243

Fig. 2 ROC analysis curve



prediction model is evaluated and presented statistically in Table 2 and in a confusion matrix in Table 3. The time taken to run the algorithm was 31.34 s

Figure 2 depicts the ROC curve that shows the performance of the classification model by plotting the true positive and false positive rate; achieving an accuracy of 92.21% for DNN.

5.1.1 Experimental Scenario 2

Secondly, using the same settings and configurations, experiments has been conducted using the proposed hybrid model, and the results are presented in Table 4.

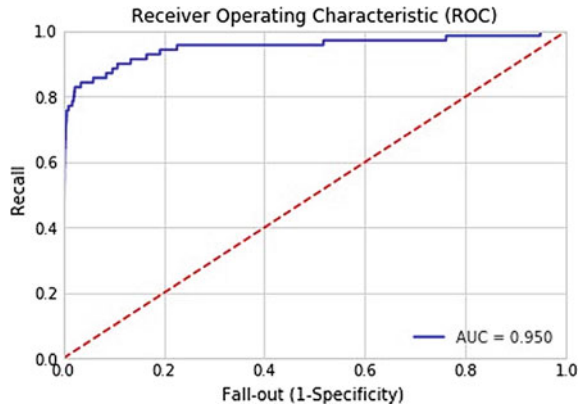
Table 4 Result for statistical performance analysis using proposed hybrid approach (DNN + GA)

Precision	Recall	Accuracy
0.896	0.896	92.21%
Actual class	Predicted class	
	Defective	Non-Defective
Defective	3	5
Non-Defective	6	258

Table 5 Confusion matrix using deep neural network (DNN) and GA

Actual class	Predicted Class
Defective	Non-Defective

Fig. 3 ROC analysis curve



The attributes that had been considered for GA to perform feature selection on DNN was LOC counts and Halstead only. The time taken to run the algorithm was around 12 h. (Table 5)

Figure 3 depicts the ROC analysis curve which takes into consideration both the true positive rate and false positive rate where proposed hybrid approach illustrates better performance when compared to DNN with an accuracy rate of 95%.

The result of experimental study shows that the proposed hybrid approach provides reliable performance that can be used for software defect prediction model. The result produced for the hybrid model has achieved accuracy of 92.21% while DNN an accuracy of 87.21%.

6 Conclusion and Future Work

In recent years, early prediction of software defects methods along with their problems and applications are emerging rapidly. This paper presents a hybrid approach for software defect prediction using Deep Neural Network (DNN) classification scheme combined with Genetic Algorithm (GA) using benchmark dataset from PROMISE repository. The performance of this hybrid approach when compared with a conventional DNN shows an increase of around 5% with regards to prediction accuracy.

Future research is highly applicable for this current study where this methodology implemented can be improved by using real-time application datasets. Furthermore, there are some requirements to consider such as overfitting phenomena and noise factors when designing the neural network. Thus, parameters of the learning functions

for the neural network should be selected properly for better optimization of hyper parameters in the networks. In addition, control parameters like crossover rate and mutation rate of genetic algorithm should be taken into consideration in order to derive suitable combinations to enhance performance of the model.

References

1. Hassan, A.E.: Predicting faults using the complexity of code changes. In: Proceedings of the 31st International Conference on Software Engineering, pp. 78–88. IEEE Computer Society (2009)
2. Azar, D., Vybihal, J.: An ant colony optimization algorithm to improve software quality prediction models: case of class stability. *Inf. Softw. Technol.* **53**(4), 388–393 (2011)
3. Ayon, S.: Neural network based software defect prediction using genetic algorithm and particle swarm optimization. pp. 1–4. (2019). <https://doi.org/10.1109/icasert.2019.8934642>, 2019
4. Turhan, B., Menzies, T., Bener, A.B., Di Stefano, J.: On the relative value of cross-company and within-company data for defect prediction. *Empirical Softw. Eng.* **14**(5), 540–578 (2009)
5. Bishnu, P.S., Bhattacharjee, V.: Software fault prediction using quad tree-based K-means clustering algorithm. *IEEE Trans. Knowled. Data Eng.* **24**(6), 1146–1150 (2012)
6. Catal, C., Diri, B.: A systematic review of software fault prediction studies. *Expert Syst. Appl.* **36**(4), 7346–7354 (2009)
7. Elish, K.O., Elish, M.O.: Predicting defect-prone software modules using support vector machines. *J. Syst. Softw.* **81**(5), 649–660 (2008)
8. Hassan, F., Farhan, S., Fahiem, M.A., Tauseef, H.: A review on machine learning techniques for software defect prediction. *Technical J. Univer. Eng. Technol. (UET) Taxila Pakistan* **23**(2), 2313–7770 (2018)
9. Gondra, I.: Applying machine learning to software fault-proneness prediction. *J. Syst. Softw.* **81**(2), 186–195 (2008)
10. Gray, D., Bowes, D., Davey, N., Sun, Y., Christianson, B.: Using the support vector machine as a classification method for software defect prediction with static code metrics. In: *Engineering Applications of Neural Networks*, pp. 223–234. Springer, Berlin (2009)
11. Hall, T., Beecham, S., Bowes, D.: A systematic literature review on fault prediction performance in software engineering. *IEEE Trans. Softw. Eng.* **38**(6), 1276–1304 (2012)
12. Henein, M.M.R., Shawky, D.M., Abd-El-Hafiz, S.K.: Clustering-based Under-sampling for software defect prediction. In: *13th International Conference on Software Technologies (ICSOFT)*, pp. 185–193. (2018)
13. Khoshgoftaar, T.M., Gao, K.: Count models for software quality estimation. *IEEE Trans. Reliab.* **56**(2), 212–222 (2007)
14. Perreault, L., Berardinelli, S., Izurieta, C., Sheppard, J.: Using classifiers for software defect detection. In: *26th International Conference on Software Engineering and Data Engineering (SEDE)* (2017)
15. Rasneet, K.C., Iqbal, S.: Latest research and development on software testing techniques and tools. *Int. J. Current Eng. Technol.* **4**(4) (2014)
16. Yusta, S.C.: Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognit. Lett.* **30**(5), 525–534 (2009)
17. Kim, S., Whitehead, E.J., Zhang, Y.: Classifying software changes: clean or buggy? *Softw. Eng. IEEE Trans.* **34**(2), 181–196 (2008)
18. Wang, S., Liu, T., Tan, L.: Automatically learning semantic features for defect prediction. In: *Proceedings of the International Conference on Software Engineering*, May 14–22, pp. 297–308. (2016)
19. Shivaji, S., James Whitehead, E., Akella, R., Kim, S.: Reducing features to improve code changebased bug prediction. *IEEE Trans. Softw. Eng.* **39**(4), 552–569 (2013)

20. Suzuki, M., Tsuruta, S., Knauf, R.: Structural diversity for genetic algorithms and its use for creating individuals. In: IEEE Congress on Evolutionary Computation, Cancun, pp. 783–788 (2013)
21. Puranika, S., Deshpande, P., Chandrasekaran, K.: A novel machine learning approach for bug prediction. In: 6th International Conference on Advances In Computing and Communications, ICACC, 6–8 September (2016)
22. Menzies, T., Greenwald, J., Frank, A.: Data mining static code attributes to learn defect predictors. *IEEE Trans. Softw. Eng.* **33**(1), 2–13 (2007)
23. Menzies T., Turhan B., Bener A., Gay G., Cukic B., Jiang, Y.: Implications of ceiling effects in defect predictors. In: Proceedings of the 4th international workshop on Predictor models in software engineering, pp. 47–54. ACM (2008)
24. Wahono, R.S., Herman, N.S., Ahmad, S.: Neural network parameter optimization based on genetic algorithm for software defect prediction. *Adv. Sci. Lett.* **20**, 1951–1955 (2014)
25. Yang, X., Lo, D., Xia, X., Zhang, Y., Sun, J.: Deep learning for just-in-time defect prediction. In: QRS'15: Proceedings of the International Conference on Software Quality Reliability and Security (2015)
26. Yang, X., Lo, D., Xia, X., Zhang, Y., Sun, J.: Deep learning for just-in-time defect prediction. In: IEEE International Conference on Software Quality, Reliability and Security (QRS15), pp. 17–26. (2015)

Application of Classifier for Breast Cancer Cell Detection



Ashutosh Mishra, Unnati Mantry, Dibyasha Garhnayak, Sourav Panda, Rasmita Rautray, Rasmita Dash, and Rajashree Dash

Abstract Due to increased number of breast cancer cases, there is rise in concern about the disease. So, in today's era, machine learning classification can play a vital role in classifying the degree of malignancy of the cancer. This paper uses four different classifiers such as KNN, logistic regression, naïve Bayes and decision tree to classify the cancer into two classes of benign and malignant. The proposed model is simulated over Wisconsin diagnostic breast cancer dataset and evaluated by accuracy metric. The model classifies the disease based on the most accurate classifier. The KNN model shows significantly better result than the other classifier.

Keywords Breast cancer · KNN classifier · Naïve Bayes · Decision tree · Logistic regression

1 Introduction

Breast cancer is the most appearing female disease and the most familiar cause of demise in women. It is very common in females over the age of 45 years. As per the National Cancer Institute (NCI), breast cancer is mostly found in females aged between 55 and 64 years. It is found increasing specifically in the less developed countries where most of the cases are determined in later phases due to lack of early detection programmes. But if the tumour is found and treated early, then breast cancer is most often curable. Nowadays, breast cancer deaths have reduced by 1/3rd or more as compared to past three decades, and this became possible due to increased screening along with improved early treatment of the disease. Early detection helps to remove the tumour which stops the growth of the cell and spreading it further. So, the main objective of our work is to help its users to easily classify the cells as malignant or benign. The user has to input the details of the tumour (i.e. various attributes), and then, the model will speculate whether the tumour is cancerous or not.

A. Mishra (✉) · U. Mantry · D. Garhnayak · S. Panda · R. Rautray · R. Dash · R. Dash
Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be)
University, Bhubaneswar, Odisha, India
e-mail: mishra.ashutosh04@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_19

221

The paper is structured as follows: Sect. 1 contains the introduction part and Sect. 2 deals with the literature survey. The proposed system, the workflow and the techniques used are discussed in Sect. 3. Section 4 comprises of the metric used for validation. The experiment and result analysis are discussed in Sect. 5. Section 6 comprises of the conclusions and future scope.

2 Literature Survey

Nowadays, breast cancer is one of the most prevalent forms of disease observed in females. It claims many lives almost in every age group of females. Machine learning proves to be promising in medical fields, particularly classifying and predicting outcomes of many diseases. Cancer happens to be one such disease which can successfully be classified into various outcomes by using machine learning approach. Breast cancer is also one of the most seen forms of cancer responsible for deaths in females all over the world. The whole process of appearance of early symptoms and cancer stage detection to successfully cure the disease is a daunting task for doctors. So, in this work, various classifiers like KNN, naïve Bayes, decision tree and logistic regression are used to classify the affected cells as malignant or benign.

The first technique used in this work is KNN classifier, which is a widely used machine learning approach that is used for classification as well as regression. It has been applied in different fields of research. In [1], KNN classifier was used to classify flower images based on different texture features. There were 25 classes of flowers to which it had been classified. The results showed that the texture features gave the best result using KNN classifier. This technique has also been used in the field of medical science. The authors of [2] used KNN to diagnose liver disease where researchers used artificial neural network as well. The dataset used in [2] was the liver disease dataset which gave good performance with KNN classifier. Similarly, this classifier has been used for melanoma detection in [3], which is based on the input image. It is classified as cancerous and non-cancerous. The image is pre-processed, enhanced, segmented and using KNN classifier; it is classified as malignant or benign and was observed to receive a better quality rate by using KNN classifier as discussed in [3].

Brain cancer classification is one of the toughest jobs today. The KNN classifier has been of great use in this field as discussed in [4]. It was used to categorize MRI brain images as usual and unusual in which 50 images were classified and gave an amazing accuracy rate of 98%, which was the best among the rest of the techniques.

The author of [5] have used a different kind of classifier called naïve Bayes classifier for text classification as it gives better results in problems having multiple classes. The author classified the type of news articles in Indonesian language by implementing ECS algorithm, and there were various groups used in the system such as economics, entertainment, lifestyle, technology, sports, etc. The accuracy obtained from the results was 63% and was found out to be the best.

Naive Bayes classifier was also used in sentiment classification of product reviews as discussed in [6] where the users were grouped into positive or negative based on

the textual information and the dataset “movie review”, is classified into positive and negative sentiments using the classifier. Similarly in [7], this classifier has also been used for sentiment analysis of twitter data, that is interpretation of emotions that whether it is positive, negative or neutral. At first, the dataset was pre-processed and then the adjective was obtained from the dataset, which had some specific meaning (known as feature vector), and then, the feature vector list was chosen and techniques like naive Bayes classifier, maximum entropy and SVM were applied which then extracted the synonyms and similarity for the content feature.

Naïve Bayes classifier also plays an important role in the predicting serious illness like lung cancer survivability using classification algorithms as discussed in [8] where data was obtained from the Oxford dataset and various classification techniques like random forest, naive Bayes classifier were used to predict the endurance rate of cancer.

One such technique used in classification is decision tree. It is considered as one of the most prominent classification methodologies applied in data mining and is classified in two phases (I) tree building and (ii) tree pruning phase. Researchers have applied it on “Yahoo! Japan” web pages to construct decision trees as in [9]. This approach used a rule-based inductive classification. The authors of [10] used decision tree for student performance prediction and evaluation where decision tree C4.5 was used and evaluated to classify the predicted value, which gave more explanation on whether the students are going to fail or pass at the end of the academic. Decision tree is also adopted in the field of image classification where it is used to group the image according to its quality as discussed in [11], which was executed on the MATLAB decision tree and gave an accuracy of 81%. It evidently explains that the decision tree achieves better accuracy.

Logistic regression is a widely used technique for classification. There are many applications of this technique used to classify different data pertaining to different fields of study. Logistic regression is really beneficial for classifying data with more than one outcome, preferably two. These include researches like profiling primary health care workers which was conducted for three developing countries chosen on the basis of their geographical and cultural divergence and also to use logistic regression to determine fluctuations in distribution of healthcare workers in Tanzania and Nicaragua as discussed in [12]. Logistic regression could successfully classify the healthcare workers in three countries and give an idea about the healthcare sector of the countries. There was also a study done in [13] to build a model which would be optimal for estimating credit risk for commercial banks using suitable tools. This method is also adopted in fields of geology such as predicting landslide patterns. Various classifiers along with logistic regression such as SVM, bagging and re-bagging were effectively used in [14] which resulted in proving the logistic regression to be very effective along with the other two classifiers in appropriate classification of the problem. The authors of [15] paired logistic regression with ANN that helped in determining land use factors influenced by afforestation in New Zealand. This contributed significantly to the urban planning in the country.

3 Proposed System

The proposed prototype provides a data model to classify the Wisconsin breast cancer dataset based on four classifiers and finds accuracy before accepting user data for final classification as discussed in Fig. 1.

Dataset Loading and Importing required functions

Different relevant functions were imported before working on the dataset to provide set the stage for implementing different functionalities provided by python-3. Then, the Wisconsin dataset was loaded to the programme for further work.

Data Processing

Here, the dataset is examined for any anomalies or missing values and processed accordingly to sanitize the data. The 'id' column is dropped as it is not necessary in the process of classification. Also, the worst values and standard errors of each feature are dropped. Only the significant mean or average values are chosen in order to work with only relevant data and avoid certain error and complex calculations. Thereafter, it is checked if the dataset contains any null value. As stated earlier, it is correctly found out that this dataset has no null values to be handled separately. It is

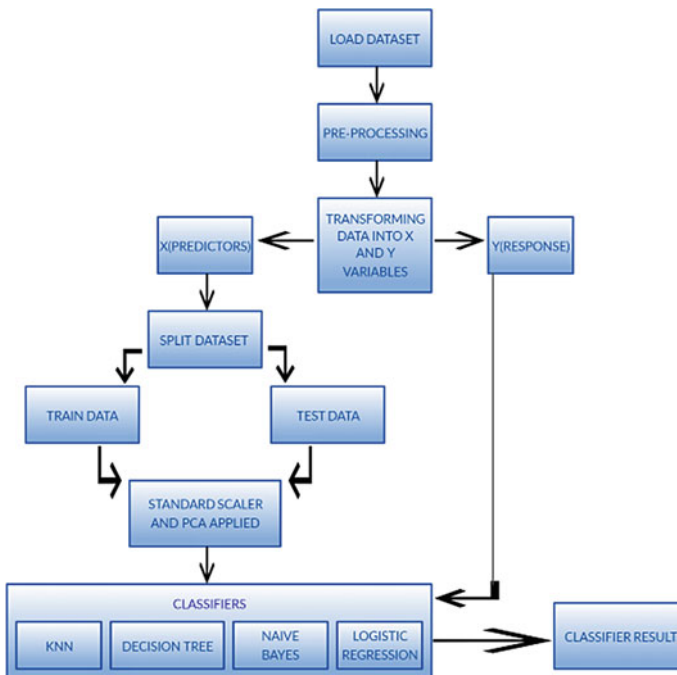


Fig. 1 Workflow of proposed system

also very important to check for missing values. As no missing values were found, the data types of each attribute was known by the above function.

Transforming Data

The dataset is divided into independent variable(y) and dependent variables(x) or the responses and predictors. As the response values are string values, they should be converted into proper numerical labels to deal with them. Hence, they were mapped as: M(malignant) = 1, B(benign) = 0.

Evaluating model

The dataset is splitted into training and validation sets in the ratio of 80:20. For this sake, k -fold cross-validation is adopted, where $k = 10$. Then, as the predictors contain highly variable values, they had to be scaled so that they have similar standard deviation to make them more easy to handle and give promising results. Principal component analysis (PCA) is applied to find out the best features to be used that actually affect the response and filter out unnecessary values to obtain accurate results and avoid errors.

Finally, the different models are tested for their accuracy, and the results were obtained as follows.

Hence, the best classifier was found out to be KNN with 93.85% accuracy, with logistic regression giving it a close competition with 93.84% accuracy.

User data classification

All ten attributes of the cancer cell mass is taken and classified based on KNN classifier into the two classes.

3.1 Techniques Used

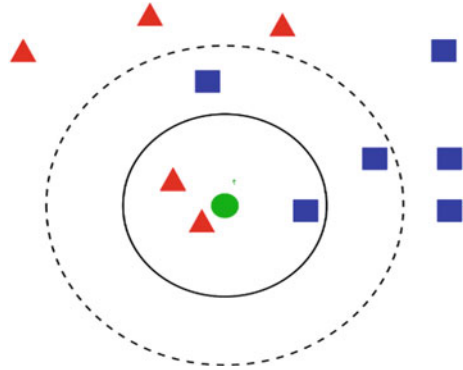
3.1.1 KNN Classifier

KNN classifier is a category of supervised learning. Supervised learning is a machine learning technique which learns on the basis on some input data that is labelled with some given output. KNN classifier is a machine learning method that is used for classification and regression. In this work, KNN classifier is used for classification.

Figure 2 illustrates classification using KNN. The data in the above figure is classified into two classes, namely red triangles (Class A) and blue squares (Class B). The class of the green circle needs to be found out by using KNN classifier. K neighbours of the green circle must be found by using the formula in Eq. (1) that is Euclidean distance.

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

Fig. 2 Classification using KNN classifier



$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \tag{1}$$

Using Eq. (1), K nearest neighbours are found out, and based on the classes of the neighbours, the class of the new data is found out. Larger value of K gives more accuracy. So, it is important to wisely choose the value of K . KNN is widely used and is very easy to implement. It can be used for classification and moreover is more accurate as compared to other machine learning techniques. So, we are using KNN classifier in this work for breast cancer cell classification.

3.1.2 Naïve Bayes Classifier

Naive Bayes is an amazingly powerful algorithm which helps in categorizing the data. It is a type of supervised learning which is applicable for classification and is based on Bayes Theorem. It helps to find the probability of an object based on certain features whether it belongs to particular group or not.

Bayes Equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

Using Bayesian probability terminology, Eq. (2) can be written as:-

$$\text{Posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \tag{3}$$

The class having the maximum posterior probability specified in Eq. (3) is the prediction outcome. It is user friendly and fast to predict the class of dataset.

3.1.3 Decision Tree Classifier

A decision tree is a tree like structure in which there are nodes out of which the internal nodes represent a check on an attribute (a condition), each branch of the tree gives the outcome of the test, and finally leaf nodes gives class level after computing everything of all the possible solutions to a decision based on certain conditions.

It is commonly used in operations research indicatively in decision analysis. It is based as powerful tool in machine learning. It is based on a flowchart like structure.

3.1.4 Logistic Regression

Logistic regression is a widely used classification technique and a supervised model. This technique is used where the traditional method of linear regression fails, when the response variable has more than one levels. This is mostly useful when the response is binary. To generalize, the independent variable can assume more than one values, which may be ordinal, nominal, ratio or interval values. Hence, it is very useful in classifying breast cancer data, which obviously has different responses, that is, malignancy or benign. The responses should be clearly of different types without any ambiguity to be segregated into different classes. Before applying the algorithm, it is checked that the dataset is sanitized, which means there should be no correlations among the predictors and no outliers in the data. The central idea of logistic regression is to calculate the log odds of the event, the unit of which is called logit or logistic unit. The mathematical representation of logistic regression can be formulated as:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (4)$$

So, Eq. (4) is the backbone logistic function that lies in the algorithm, which also ensures that the values come out to be between 0 and 1. This work exploits the advantage and employs a binary logistic regression model. As the name suggests, the response variable is binary; that is, it has two distinct values: malignant (0), benign (1) and the labels are mapped to respective values using suitable python functionality. If the curve is plotted, it would form a characteristic sigmoid(S-shaped) curve. This method is popularly used in most of the machine learning works and is very practical in medical applications, thereby predicting outcomes of many diseases effectively.

4 Metric Used for Validation

Accuracy is a metric for evaluating various classifiers. Different classifiers have different measures of definitude. To include various classification methods it is required to single out the best among them to provide the developers with the clarity

and quantitatively stating the accuracy of different classifiers before choosing the best among them. It forms a basic method for evaluating models to select the best while employing them in the end.

Accuracy is a parameter that can be calculated as number of correctly predicted responses out of the total number of predicted responses. Mathematically, it can be formulated as follows:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (5)$$

For a classifier having binary responses, the accuracy metric can be calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

The above parameters can be stated as:

1. True positive is an event where the model precisely predicts the positive class.
2. True negative is an event where the model accurately predicts the negative class.
3. False positive is similarly an event where the model predicts inaccurately the positive class.
4. False negative is an event where the model makes a conjecture wrongly about the negative class.

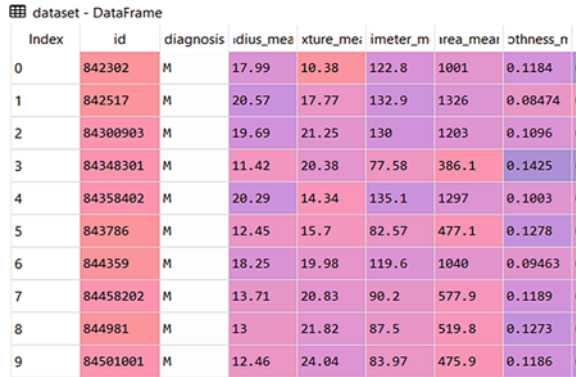
5 Experiment and Result Analysis

The dataset adopted in this work is the “Wisconsin Diagnostic Breast Cancer Dataset”. This extensive set of relevant data was developed by the University of Wisconsin, in the year 1995. It has 32 columns or features and 569 rows or samples. The originator of this dataset is Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian. The digitized picture of a fine needle aspirate (FNA) of a breast mass was studied, and the measurements were derived for the data of the rows. These are nothing but the information about assessment of the cell nuclei which is present in the sample is shown in Fig. 3. It has 569 rows and 33 columns. As the data has lot of unnecessary data, it needs to be pre-processed. After pre-processing, the data is reduced to 10 columns as shown in Fig. 4.

The output column needs to be separated. Figure 5 shows the output column. The output column needs to be labelled by binary values that are 0 and 1 in order to implement the classification.

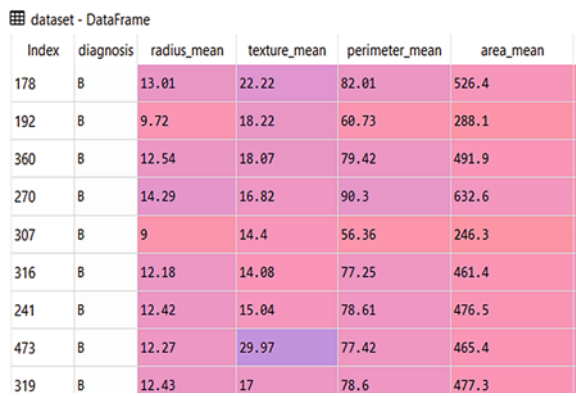
The data was then transformed using PCA. Then, the accuracies of the all the models were calculated.

Fig. 3 Breast cancer dataset



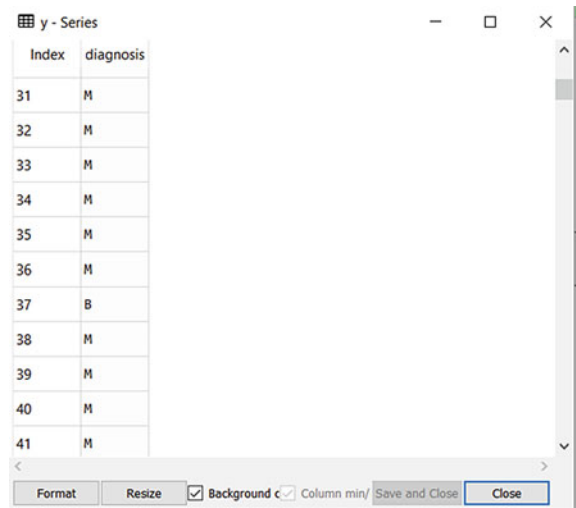
Index	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_norm
0	842302	M	17.99	10.38	122.8	1001	0.1184
1	842517	M	20.57	17.77	132.9	1326	0.08474
2	84300903	M	19.69	21.25	130	1203	0.1096
3	84348301	M	11.42	20.38	77.58	386.1	0.1425
4	84358402	M	20.29	14.34	135.1	1297	0.1003
5	843786	M	12.45	15.7	82.57	477.1	0.1278
6	844359	M	18.25	19.98	119.6	1040	0.09463
7	84458202	M	13.71	20.83	90.2	577.9	0.1189
8	844981	M	13	21.82	87.5	519.8	0.1273
9	84501001	M	12.46	24.04	83.97	475.9	0.1186

Fig. 4 Dataset after pre-processing



Index	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
178	B	13.01	22.22	82.01	526.4
192	B	9.72	18.22	60.73	288.1
360	B	12.54	18.07	79.42	491.9
270	B	14.29	16.82	90.3	632.6
307	B	9	14.4	56.36	246.3
316	B	12.18	14.08	77.25	461.4
241	B	12.42	15.04	78.61	476.5
473	B	12.27	29.97	77.42	465.4
319	B	12.43	17	78.6	477.3

Fig. 5 Output column (y)



Index	diagnosis
31	M
32	M
33	M
34	M
35	M
36	M
37	B
38	M
39	M
40	M
41	M

Based on these values, using the best classifier, the data was being classified. The user is prompted to enter the values of the necessary attributes, after which it classifies it as malignant and benign and gives it as a result.

6 Conclusion and Future Scope

This paper focuses on the breast cancer dataset to classify it into two groups: malignant (cancerous) and benign (non-cancerous). The data was pre-processed in order to remove the missing values. Subsequently, the data was split into training and validation data. Then, the data was transformed by the help of principal component analysis (PCA) to find out the best features to be used that actually affect the response and filter out unnecessary values to obtain accurate results and avoid errors. Using the given dataset, by using various machine learning techniques, like KNN classifier, naïve Bayes classifier, decision tree (CART) and logistic regression, the accuracy of classifying the data was found out. Then, by using the best classifier, the new data was classified. However, it was found out that KNN classifier gives the best accuracy as compared to the other classifiers. Also, logistic regression gave a close fight to KNN classifier. So using these machine learning algorithms, the breast cancer data was classified as malignant or benign. This work can be used not just by doctors to detect whether a cell is cancerous or not, but also it can be used by normal people to know if the cells are malignant or benign.

In this work, cancer cells are classified as malignant or benign based on certain attributes entered by the user. Here, the dataset that is used is constant, so it gives the same accuracy rates in almost every case. Hence, by adding a database connectivity, we can intend to store the data that are entered by the user and the class to which it is been classified. Then, this data can be further included in the dataset for training and testing to get better results.

References

1. Guru, D.S., Sharath, Y.H., Manjunath, S.: In: Features and KNN in Classification of Flower Images: Recent Trends in Image Processing and Pattern Recognition, Special Issue (2010)
2. Lin, R.H. ve Chuang,C.H.: A hybrid diagnosis model for determining the types of the liver disease. *Comput. Biol. Med.* **40**, 665–670 (2010)
3. Selvaraj, D. , Kumar, D.A., Dhinakaran, D.: Melanoma detection using hybrid classifier. *Int. J. Adv. Res. Sci. Eng.* **4** (Special Issue(01)), 1018-1033 (2015)
4. Machhale, K., Nandpuru, H.B., Kapur, V., Kosta, L.: MRI brain cancer classification using hybrid classifier(SVM-KNN). In: 2015 International Conference on Industrial Instrumentation and Control(ICIC), Pune, pp. 60–65. (2015)
5. Utari, 1., Muhammad, I., Medyawati, H.: Classification of news types by implementing enhanced confix stripping stemmer. *Int. J. Eng. Technol. Manage. Res.* **6** (Special Issue(05)), 135–141 (2019)

6. Baid, P. et al.: Sentiment analysis of movie reviews using machine learning techniques. *Int. J. Comput. Appl.* **179**, 45–49 (2017)
7. Matharasi I, P.B., Senthilrajan, A.: Sentiment analysis of twitter data using naïve bayes with unigram approach. *Int. J. Sci. Res. Public.* **7**(Special Issue(05)), 337–341 (2017)
8. Krishnaiah, V., Narsimha, G., Chandra, N.S.: Diagnosis of lung cancer prediction system using data mining classification techniques. *Int. J. Comput. Sci. Inf. Technol.* **4**(Special Issue(01)), 39–45 (2013)
9. Tsukada, M., Washio, T.: In: Conference on Artificial Intelligence Young Researcher Paper (Institute of Scientific and Industrial Research), vol. 5 (Issue 4) (1998)
10. Ojo G.F., Afolabi I.T.: Student's performance prediction using decision tree. *Computer and Information Sciences*, vol. 8 (Issue 7), Covenant University, Ota, Ogun State (2019)
11. Anushya, A.: Quality recognition of images using classifiers. *Int. J. Comput. Sci. Mobile Comput.* **9**(1) (2020)
12. Barden-O'Fallon, J., Angeles, G., Tsui, A.: Imbalances in the health labour force: an assessment using data from three national health facility surveys. *London School of Hyg. Trop. Med.* **10**, 80–90 (2006)
13. Chilingaryan, F.: Credit risk measurement: probability of default for Russian banks (2009–2016). *Manoogian Simone College of Bus. Econ.* **7**, 34–43 (2018)
14. Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth Syst. Sci.* **5**(Special Issue), 853–862 (2005)
15. West, T.A.P., Monge, J.J., Dowling, L.J.: Comparison of spatial modelling frameworks for the identification of future afforestation in New Zealand. *Landscape Urban Plan.* **198**, 1–7 (2020)

Apriori-Backed Fuzzy Unification and Statistical Inference in Feature Reduction: An Application in Prognosis of Autism in Toddlers



Shithi Maitra, Nasrin Akter, Afrina Zahan Mithila, Tonmoy Hossain, and Mohammad Shafiul Alam

Abstract Weak Artificial Intelligence (AI) allows the application of machine intelligence in modern health information technology to support medical professionals in bridging physical/psychological observations with clinical knowledge, thus generating diagnostic decisions. Autism, a highly variable neurodevelopmental condition marked by social impairments, reveals symptoms during infancy with no abatement with time due to comorbidities. There exist genetic, behavioral, neurological actors playing roles in the making of the disease and this constructs an ideal pattern recognition task. In this research, the *Autism Screening Data (ASD)* for toddlers was initially exploratorily analyzed to hypothesize impactful features which were further condensed and inferentially pruned. An interesting application of the business intelligence algorithm: *Apriori* has been made on transactions consisting of ten features and this has constituted a novel preprocessing step derived from *market basket analysis*. The huddling features were fuzzily modeled to a single feature, the membership function of which evaluated to the degree to which a toddler could be called autistic, thus paving the way to the first optimized Neural Network (NN). Features were further eliminated based on statistical *t*-tests and *Chi*-squared tests, administering features only with *p*-values < 0.05 —giving rise to the second and final optimized model. The research showed that the unremitted 16-feature and the optimized 5-feature models showed equivalence in terms of maximum test accu-

S. Maitra (✉)
Sheba.xyz, Dhaka, Bangladesh
e-mail: shithi30@gmail.com

N. Akter
East West University, Dhaka, Bangladesh
e-mail: nipa.ete@gmail.com

A. Zahan Mithila · T. Hossain · M. Shafiul Alam
Ahsanullah University of Science and Technology, Dhaka, Bangladesh
e-mail: azmithila2014@gmail.com

T. Hossain
e-mail: tonmoyhossain.cse@ieee.org

M. Shafiul Alam
e-mail: shafiul.cse@aust.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_20

racy: 99.68%, certainly with lower computation in the optimized scheme. The paper follows a ‘hard (EDA, inferential statistics) + soft (fuzzy logic) + hard (forward propagation) + soft (backpropagation)’ pipeline and similar systems can be used for similar prognostic problems.

Keywords Autism spectrum disorder · Exploratory data analysis · Apriori algorithm · Fuzzy modeling · Membership functions/values · Inferential statistics · *t*-test · *Chi*-squared (χ^2) independence test · ANOVA-test

1 Introduction

Autism is a neurodevelopmental condition marked by stereotypical, compulsive, repetitive, ritualistic behavior and sometimes by limited interests and tendency towards self-harm, in extreme cases—noticed in early childhood with no diminutive effect along with the increase of age. The said disorder shows a severity-gradient [1] and hence the term Autism Spectrum Disorder (ASD) is coined. The developed world is inflicted more with its scourge and almost 1.5% of children were diagnosed autistic in 2017 [2]. While early intervention is necessary to groom the affected individuals for better self-care, the UK National Autism Plan for Children presently endorses an assessment as lengthy as 30 weeks [3].

Artificially intelligent *Clinical Decision Support Systems (CDSS)* can draw demarcations between neurologically sound and unhealthy subjects given a proper knowledge-base. A CDSS can be developed and deployed using clinical data mining which is the process of extracting medical insights from diagnostic data. This paper implements the steps of building such a system—which can significantly spare the time for prognosis, the working hours and efforts of a physician and can make caregivers mentally, financially prepared—by:

- mining a soundly, consensually collected dataset that captures both historical, behavioral aspects
- performing an intensive preprocessing and producing three variants of the screening dataset
- modeling the data to an appropriate algorithm for generating the most practical prognosis in light of known cases.

It is expected that a bulky set of toddlers’ diagnostic data may be uncertain, imprecise, partially true and an exact solution might be infeasible—hence the appeal of soft computing methods. The beauty of this recent development in computing is that it has a humanoid, heuristic process of giving a useful, optimal solution. The research at hand complementarily applies two components of soft computing for the prognosis of a probable ASD trait, namely: fuzzy logic and neural networks (NNs). The layered, hierarchical structure of a neural network can recognize complex patterns spread through multiple dimensions by iteratively refining some initial set of parameters using back-propagation.

Personality traits in a particular cohort (in our case, the autistic toddlers) often show up in bundles, due to which association rule mining can serve the purpose of identifying the frequent traits occurring together in toddlers’ behavior. *Apriori* algorithm (Agrawal and Srikant 1994) can be used for defining such rules, which may club several features together by aggregation. This research attempts to structure the screening data in a transactional form for the application of *Apriori*. The clubbed outcome of such basket analysis is then fuzzily modeled, to capture the vague and imprecise effect.

This era of big data demands the usage of clean, dimensionally reduced data since an increased dimensionality demands greater numbers of examples for sound training of a soft, predictive model—otherwise known as the curse of dimensionality. Hence this research first produces a conglomeration of as many as ten features and further prunes the remaining features based on inferential statistics. Concretely, this paper poses a supervised binary classification problem upon extensive, novel preprocessing of toddlers’ data, to label them as neurologically typical or autistic instances following a hard-soft-hard-soft pipeline. A concise review of existing literature, rendering of the proposed methodology, followed by tabulation and explanation of results concludes and constructs the paper.

2 Literature Review

The recent academic literature invested in prognosticating autistic behavior encompasses multi-sourced data collection of random-control groups, its analysis (both statistical and predictive), elimination of redundant features and finally generating the output using both structured data, sequence models and images (Fig. 1). A gradual unraveling of such methods to our purpose is demonstrated using four paradigms as stated below.

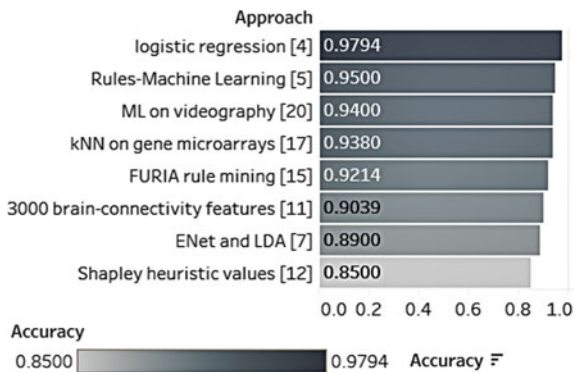


Fig. 1 A synoptic review of related literature

2.1 Data Garnering and Predictive Analyses Using Basic ML

Thabtah [4] firstly developed a mobile application through which he could construct the ASD dataset which contained information of toddlers, children, adolescents and adults. The effort created a motley dataset which contained data from different countries, languages, ethnicities and introduced the Q_{10} -chat. He deployed Naive Bayes and logistic regression which earned him an accuracy of 97.94% and a recall of 98%. Next, Thabtah and Peebles [5] developed a medical decision support system introducing Rules Machine Learning, which gave both diagnosis and insights. This is an advancement that utilized their legacy Q_{10} -chat questionnaire [4] which has also been used in this research. This earned them the accuracy of a highest 95% and recall of a highest 97%.

Sarkar et al. [6] drilled down to the toddlers' group and built a tablet-run application that assessed the severity of autistic behavior. The pilot evaluation showed potential in detecting the syndrome with an F1-score of 0.94. Duda et al. [7] classified between ASD, ADHD using 15 feature-like responses that were collected through a rigorous, holistic data collection procedure. They dealt with imbalanced data and obtained an accuracy of 0.89 ± 0.01 using ENet and LDA classifiers.

2.2 Approaches to Eliminate Redundant Features

Achenie et al. [8] applied a feed-forward neural network on the M-CHAT-R dataset for the screening of autism in toddlers. The reduced the scope of features from 20 to 18 to a final 14 and measured its effectiveness gender/ethnicity-wise. They introduced maternal educational period as a feature and found 99.72% accuracy. Thabtah et al. [9], as a sequel, proposed Variable Analysis (Va) that helped shed off features based on correlations.

Abbas et al. [10] accumulated both structured and graphical content and trained two different classifiers, to combine them finally to produce both conclusive and inconclusive outcomes, after much feature-scrambling. Kong et al. [11] extracted brain connectivity data from images and capped the number of features to 3000 in a descending sequence of F1-score, achieving a 90.39% accuracy. Tariq et al. [12] applied Shapley heuristic values to determine the importance of features which led them to 85% accurate outcomes.

2.3 Applications Using Fuzzy Logic

Farsi et al. [13] argued that Fuzzy Cognitive Maps (FCMs) have limitations in handling a great level of uncertainty due to the causal inferences they tend to make and use Interval Agreement Approach to assign weights to the links of FCMs in order to

model the uncertainties better in predicting ASD. Al-diabat [14] applied fuzzy rule mining and extracted 29 (11 ‘yes’ + 18 ‘no’) fuzzy rules, contending FURIA to be the most effective with an accuracy of 90%. He identified the most frequently occurring features to be the most influential. Khan and Alshara [15] similarly endorsed FURIA with 92.14% accurate predictions.

2.4 Usage of Data from Variegated Sources

Xiao et al. [16] intended to avoid reporting bias for ASD screening and found random forest classifier as the best-performing on regional cortical thickness data obtained by neuroimaging. This performed better than volume and area-based data, using 20 highest importance regions. Kim et al. [17] examined blood gene expression profiles and used microarray data to first verify distinguishability using clustering and finally found 93.8% accuracy using *k*NN. Karan et al. [18] examined electroencephalogram (EEG) signals’ data to classify among autistic and neuro-typical subjects with 71% precision.

Abbas et al. [19] applied machine learning to structured data obtained from children’s parents’ responses to questionnaires and also to short snippets of videos obtained from their homes. They proposed an extension of their work for diagnosing other neurodevelopmental conditions as well. Tariq et al. [20] observed 30 behaviors from 3-min home videos of children with and without ASD traits and found > 94% accuracy by modeling them to ML algorithms.

This paper is a different endeavor in that it attempts association rule mining to find a grouping tendency among features and hence justifiably clubs them up to a fuzzy feature to find comparable accuracy using fewer features, hence combining ML, feature reduction and fuzzy logic.

3 Proposed Model

The screening data for autism in toddlers [4] was researched in a modular fashion: with the first module performing an intense preprocessing on the raw data, the second module finding and fitting an appropriate predictive model and the final one evaluating the efficacy of the former two in generating the desired outcome (Fig. 2).

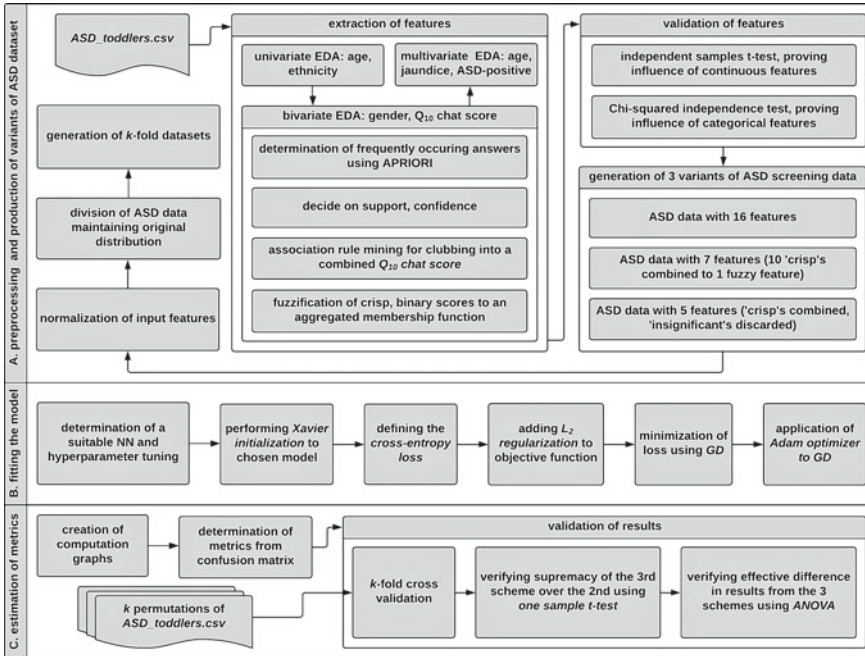


Fig. 2 Workflow for implementing the proposed autism-decision support system in toddlers

3.1 Preprocessing ASD Data and Producing Variants of the Dataset

Exploratory data analysis (EDA) is a preliminary visual summarization of a dataset for formulating hypotheses to be further tested and for developing insights into what mathematical model should fit the data best. In this piece, we employ EDA in three modalities: univariate, bivariate and multivariate—depending on how many variables are under study. We use EDA to initially hypothesize an attribute as a feature and further bolster this hypothesis by inferential statistics. The features thus found are transformed to fit a predictive model, introducing *Initial Data Analysis (IDA)* to the process, which is encompassed by EDA. We firstly explore singularly (hence, univariate) the impact of the continuous variable: age and secondly analyze the casual correlation of the discrete variable: ethnicity upon a toddler being autistic. The analysis (Fig. 3) reveals that among the culturally variegated data, toddlers from a European descent (34.34%) are the most affected with the condition followed by Asians (29.12%), whereas children from South Asian, Hispanic, Latin American ancestry are the least inflicted (less than 5%). The age-range of 12–36 months being definitive of toddlers, the exploratory analysis reveals 24–36 months to be appropriate for a positive diagnosis (Fig. 4).

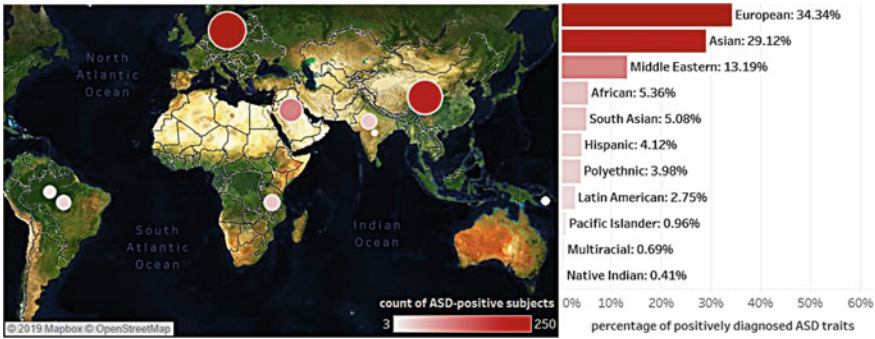


Fig. 3 Geographic distribution of ASD-positive toddlers among different ethnic groups

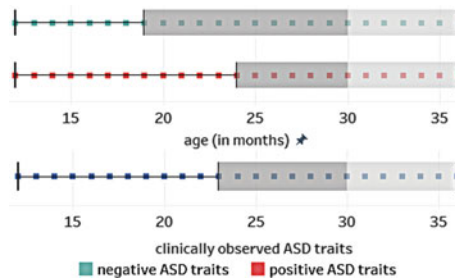


Fig. 4 Box-and-whisker plot demonstrating effective age-ranges for prognosis

Next, we delve into a bivariate analysis in that we examine the interaction between ASD traits and Q_{10} -chat score, gender. We discover that the 10 questions, answered by parents or surrogate caregivers, elicit an affirmative response for neurologically sound subjects (except for Q_7). We examine the inverse answers and find that there exists a considerably large gap when it comes to the unsound subjects (Fig. 5). We perform an *Apriori*-based aggregation to the data to generate the Q_{10} -chat score and find its positive correlation with ASD-positive subjects. We further find that despite a 69.73% share of male toddlers in the dataset, a greater 73.35% of the ASD-positive cases are among the males (Fig. 6).

It is understandable that a representative, combined feature instead of 10 different features would be convenient and more computationally efficient for running expensive soft computing algorithms. According to the jargon of database management, we treat the dataset as a relation and apply *Apriori* association rule mining and find frequently occurring positive answer-sets. There exists its application in *market basket analysis*, but the application here is relevant because a tendency of appearing together had been detected in EDA (Fig. 5). The hyperparameters (Fig. 7): support ≥ 0.02 and confidence ≥ 0.8 , have led to 6 association rules (Fig. 8) with full confidence that justify the amalgamation of the answers to a single feature.

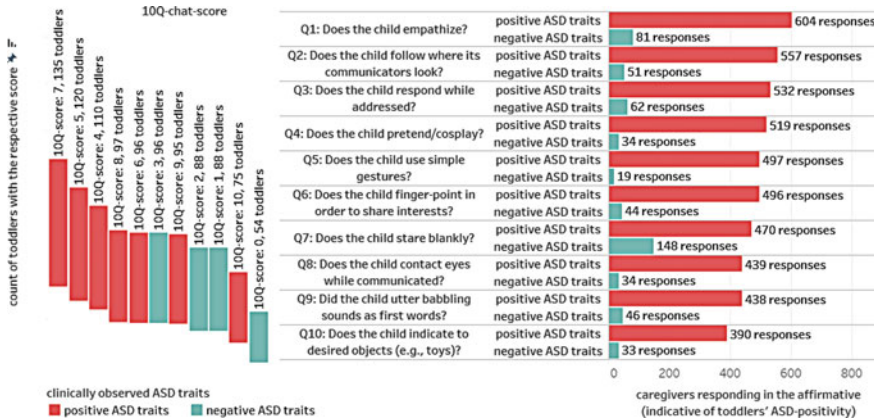


Fig. 5 Toddlers scoring more (more than 3) in the Q_{10} chat seem more likely to have autism

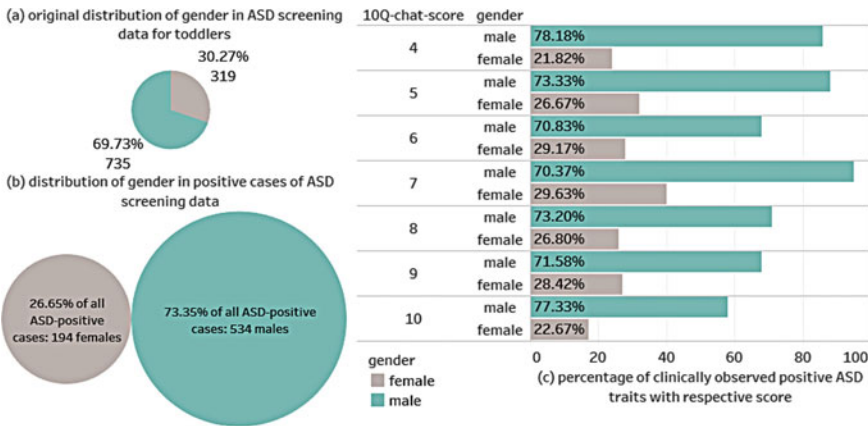


Fig. 6 Males showing a greater tendency towards a positive diagnosis

Fuzzy models are capable of representing and utilizing partly true, vague and imprecise information mathematically. It is particularly useful in medical decision-making frameworks because it can appropriately define the degree to which a symptom is being visible. After discovering that Q_{10} answers appear frequently together using *Apriori*, we define a fuzzy process in the following way (Fig. 9) which transformed 10 attributes into an enveloped feature:

- First, we take as input crisp binary answers and fuzzify them using a simple, discrete membership function $\mu_{\text{autistic}}(\text{toddler})$ of aggregation.
- Then we execute applicable rules from the rule-base we define for different severity-levels of ASD [21] and calculate the answers' average, which serves as a fuzzy output value (discrete membership value, support(autistic) = {toddler | $0 \leq \mu_{\text{autistic}}(\text{toddler}) \leq 1$ }).

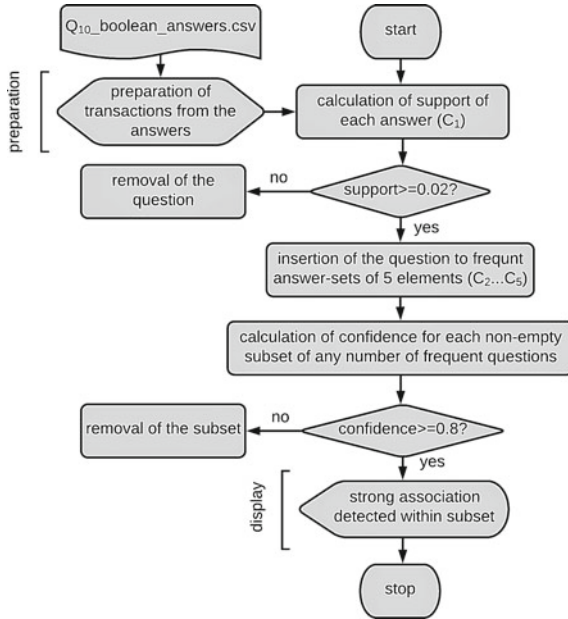


Fig. 7 ASD-specific *Apriori* algorithm devised and deployed for effective clubbing of Q_{10} answers

- Finally, the said average serves as a crisp output once we convert it into percentage and this defines the extent to which a toddler can be called autistic according to its Q_{10} -score.

Having concluded the bivariate analysis, we explore the relationships individually between age, family history, neonatal jaundice history, caregivers’ association and positive ASD traits (Fig. 10). The multivariate analysis (Fig. 10a) exposes that among the subjects with positive ASD traits, 29.53% have had a history of neonatal jaundice while among normal subjects jaundice was less pervasive, 22.39%. If family-history (Fig. 10b) or the test-administering caregiver (Fig. 10c) plays a role in ASD is unclear from EDA, and is being left for statistical inference to clarify. EDA thus helped to investigate and hypothesize features, which were further validated/invalidated by statistical inference.

Hypothesized features need validations that the tendencies shown by EDA will indeed hold if further data-points are added and that their impact on the classes (autistic/typical) is not due to chance. Inferential statistics does just that by applying the *Chi*-squared independence test on discrete and independent samples *t*-test on continuous variables. The *t*-test verified that the sample means showed statistical evidence of being considerably segregated while the *Chi*-squared test compared existing and expected frequencies for statistical independence. We recognize an attribute as a feature only upon getting statistically significant (p -value < 0.05) results (Tables 1 and 2).

```
> inspect(head(sort(itemsets, by="support"), 20))
  items      support  count
[1] {a1,a5,a6,a7,a9} 0.2125237 224
[2] {a4,a5,a6,a7,a9} 0.2115750 223
[3] {a1,a4,a6,a7,a9} 0.2115750 223
[4] {a1,a2,a4,a6,a7} 0.2096774 221
[5] {a1,a4,a5,a6,a7} 0.2077799 219
[6] {a1,a2,a5,a6,a7} 0.2068311 218
[7] {a1,a2,a6,a7,a9} 0.2011385 212
[8] {a3,a4,a6,a7,a9} 0.1992410 210
[9] {a1,a4,a5,a7,a9} 0.1982922 209
[10] {a3,a4,a5,a6,a7} 0.1963947 207
[11] {a1,a4,a5,a6,a9} 0.1963947 207
[12] {a3,a4,a5,a7,a9} 0.1925996 203
[13] {a1,a3,a4,a6,a7} 0.1925996 203
[14] {a3,a4,a5,a6,a9} 0.1916509 202
[15] {a1,a2,a6,a7,a10} 0.1888046 199
[16] {a1,a2,a4,a7,a9} 0.1888046 199
[17] {a1,a2,a4,a5,a7} 0.1878558 198
[18] {a1,a2,a4,a5,a6} 0.1850095 195
[19] {a4,a6,a7,a8,a9} 0.1840607 194
[20] {a5,a6,a7,a8,a9} 0.1840607 194

> inspect(strong_rules)
  lhs      rhs support  confidence lift
[1] {a2,a3,a5,a7,a8,a9,a10} => {a4} 0.07685009 1.0000000 1.951852
[2] {a2,a3,a5,a6,a7,a9,a10} => {a4} 0.08918406 1.0000000 1.951852
[3] {a1,a2,a3,a5,a7,a8,a9,a10} => {a4} 0.07495256 1.0000000 1.951852
[4] {a2,a3,a5,a6,a7,a8,a9,a10} => {a4} 0.07210626 1.0000000 1.951852
[5] {a1,a2,a3,a5,a6,a7,a9,a10} => {a4} 0.08633776 1.0000000 1.951852
[6] {a1,a2,a3,a5,a6,a7,a8,a9,a10} => {a4} 0.07115750 1.0000000 1.951852
[7] {a2,a3,a5,a7,a9,a10} => {a4} 0.09582543 0.9901961 1.932716
[8] {a1,a2,a3,a5,a7,a9,a10} => {a4} 0.09108159 0.9896907 1.931730
[9] {a1,a3,a5,a7,a8,a9,a10} => {a4} 0.08918406 0.9894737 1.931306
[10] {a3,a5,a6,a7,a8,a9,a10} => {a4} 0.08918406 0.9894737 1.931306
```

Fig. 8 Output generated from applying hyperparametrically tuned *Apriori*

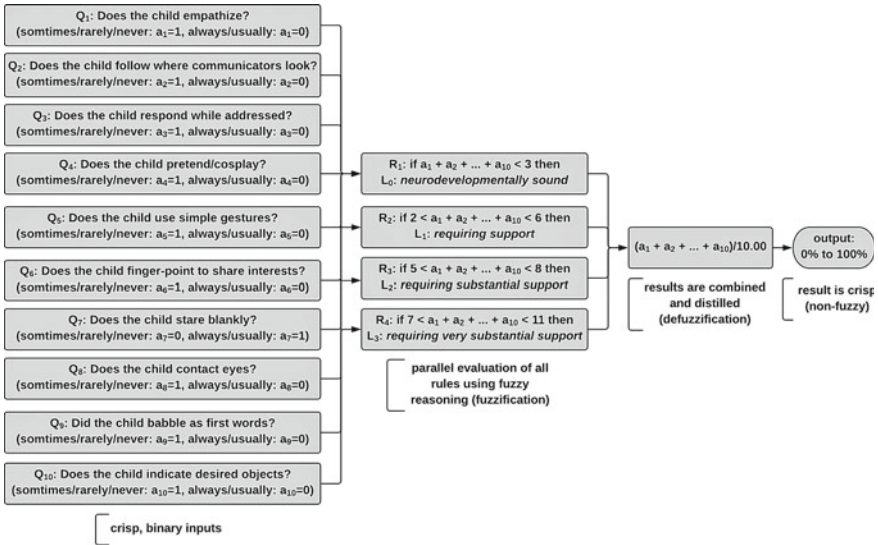


Fig. 9 Application of fuzzy logic for the conversion of crisp inputs into a fuzzy input

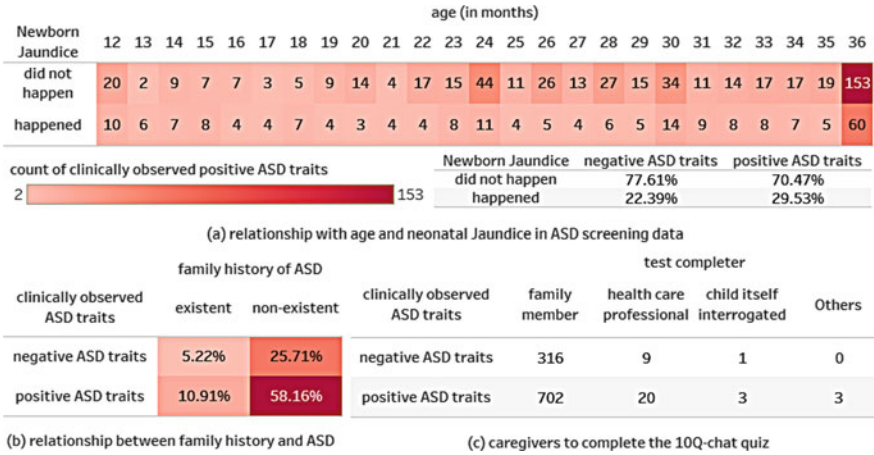


Fig. 10 Neonatal jaundice may play a role in ASD, while the roles of family, test-administrator remain unknown

Table 1 Welch Two Sample *t*-test results

Continuous features	<i>t</i> -score	Degrees of freedom	<i>p</i> -value	$H_0: \mu_1 = \mu_2$	$H_a: \mu_1 \neq \mu_2$
10 questions' quiz-score	-55.072	1006.4	<2.2E-16	Reject	Retain
Age (months)	-2.0323	537.95	4.26E-02	Reject	Retain

Table 2 Pearson's χ^2 -test results

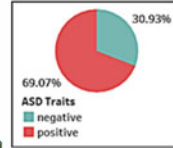
Discrete features	χ^2	Degrees of freedom	<i>p</i> -value	H_0	H_a
Ethnicity, ASD traits	43.571	10	3.93E-06	Reject	Retain
Gender, ASD traits	14.044	1	1.79E-04	Reject	retain
Jaundice history, ASD traits	5.427	1	0.01983	Reject	Retain
Test responder, ASD traits	1.4153	3	7.02E-01	Retain	reject
Family history, ASD traits	0.12094	1	7.28E-01	Retain	Reject

Fig. 11 R-script to partition data, maintaining the original class-distribution (shown in inset) within ASD screening data

```

1 # setting working directory
2 setwd("F:/Autism Research")
3
4 # reading data into a dummy dataframe
5 dummy <- read.csv("autism_toddler.csv")
6
7 # shuffling data within the dummy frame
8 dummy <- dummy[sample(1:nrow(dummy)), ]
9
10 # separating autistic/mentally sound subjects
11 dummy_ones <- dummy[dummy$group==1, ]
12 dummy_zeros <- dummy[dummy$group==0, ]
13
14 # preparing test set with 30% data
15 # 31% of test data
16 test_zeros <- dummy_zeros[1:98, ]
17 # 69% of test data
18 test_ones <- dummy_ones[1:218, ]
19 test <- rbind(test_zeros, test_ones)
20
21 # preparing training set with 70% data
22 # 31% of training data
23 train_zeros <- dummy_zeros[99:326, ]
24 # 69% of training data
25 train_ones <- dummy_ones[219:728, ]
26 train <- rbind(train_zeros, train_ones)
27
28 # assembling prepared dataset
29 dummy2 <- rbind(train, test)
30 write.csv(dummy2, "autism_toddler_1.csv", row.names=FALSE)

```



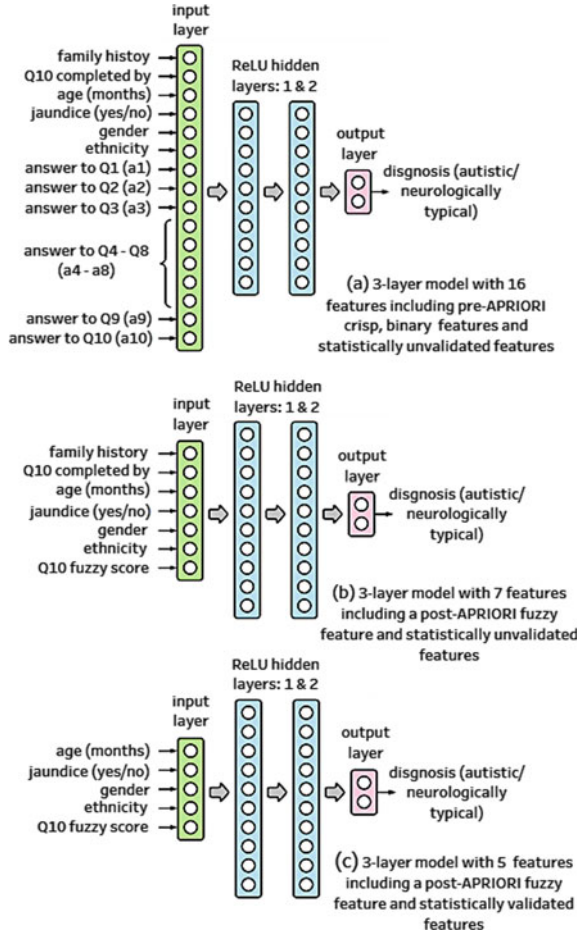
Preprocessing thus far leaves us with three variants of the ASD toddlers' data: one with all features intact, another with the 10 Q_{10} -chat responses converged to a single feature and the other finally excluding statistically insignificant (Tables 1 and 2) features (test administrator, family history). We generate k -fold ($k = 5$) permutations of each variant for getting an unbiased estimate of the metrics. For gradient descent to converge following an unflinching trajectory, we normalize the input numerals within $[0, 1]$ for a fair comparison. Finally, we split the data fairly into training (70%) and test (30%) sets with each set maintaining the representative distribution (69% autistic + 31% typical) of the two classes (Fig. 11).

3.2 Finding, Fitting, Tuning Predictive Model

We choose to fit a neural network (Fig. 12) that takes as input the numeric representation of both categorical and continuous inputs, propagates them through *ReLU*-activated hidden layers and finally maps them to a *SoftMax* classification layer. Hyperparameters, which impact performance greatly alongside the architecture, were tuned in the following way:

- **the number of layers, neurons:** The best-fitting NNs shrunk in architectural complexity as we gradually narrowed down on features as shown in (Fig. 12a–c).
- **the number of epochs:** For the highest refinement of parameters, the models were trained for 250 epochs.

Fig. 12 Evolutions in the neural networks employed for the research



- **learning rate, α** : A small learning rate of 0.001 was maintained not to overshoot minima.
- **regularization parameter, λ** : To prevent overfitting by penalizing the parameters, this was set to 0.08.
- **size of minibatch**: Minibatch gradient descent was run using 256 training examples at a time so as to not run out of primary memory.

The weights, mapping the neurons hierarchically from one layer to the other, were initialized using *Xavier* initialization assuming the inputs to hail from a *Gaussian* or uniform distribution. The cross-entropy loss, appended by an L_2 regularizer (to prevent overfitting) has been optimized for the classification problem. In the equation below, n , $y^{(n)}$, $\hat{y}^{(n)}$, i , λ , L and w are representative of count of training examples, gold labels for separate examples, model's predicted labels, sequence of a layer's activation, regularization parameter and weights being refined, respectively; with F

denoting *Frobenius* norm.

$$-\log L(\{y^{(n)}\}, \{\hat{y}^{(n)}\}) = \sum_n H(\{y^{(n)}\}, \{\hat{y}^{(n)}\}) + \frac{\lambda}{2n} \sum_L \|w^{[L]}\|_F^2 \quad (1)$$

An initial set of parameters θ is refined by optimizing loss $J(\theta)$ through running gradient descent [21] repeatedly for a specified number of epochs or until convergence—parallelly for all features i.e., for $j = 0, 1, \dots, n$ where α is the learning rate. For m training examples $(x^{(i)}, y^{(i)})$, where $h_\theta(x^{(i)})$ is the machine-prediction, gradient descent is run like the following:

Repeat until convergence {

$$\theta_j = \theta_j + \alpha \sum_{i=1}^m [y^{(i)} - h_\theta(x^{(i)})] x_j^{(i)} \quad (\text{for every } j) \quad (2)$$

}

We apply the gradient-based *Adam* optimizer to optimize gradient descent—harmonizing present parameters with lower-order moments. We select the exponential decays, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a very small number $\epsilon = 10\text{E}-08$ for preventing division by 0.

3.3 Estimating Performance Metrics

The ML framework of *TensorFlow* uses graph theory’s computation graphs (Fig. 13) for defining its sessions with the circular nodes rendering operations and the rectangular ones denoting operators on one-hot representations. The research places the predictions thus computed in a contingency table having both classes along both its dimensions, to evaluate medically important metrics: accuracy, precision (proportion of truly correct autistic identifications), recall (proportion of the actually autistic, classified correctly) and F1-score (harmonic mean of precision and recall). This cross-tabular layout is called a confusion matrix (Fig. 15).

One sample t -test and ANOVA-test have been used to strengthen that the performance of the three neural networks was systematically different. We use the one-sample t -test in order to compare the mean of the second scheme with $k = \text{fivefold}$ accuracies of the third scheme. We finally applied ANOVA which null hypothesized equality of the average F1-scores of the three schemes, $H_0: \mu_1 = \mu_2 = \mu_3$ (no association exists). To simplify, the ANOVA/ t -statistic calculated the ratio of the variance between and the variance within the random $k = 5$ -sample groups. The greater the ratio, the more the probability of the alternative hypothesis, H_a (association exists), being justified.

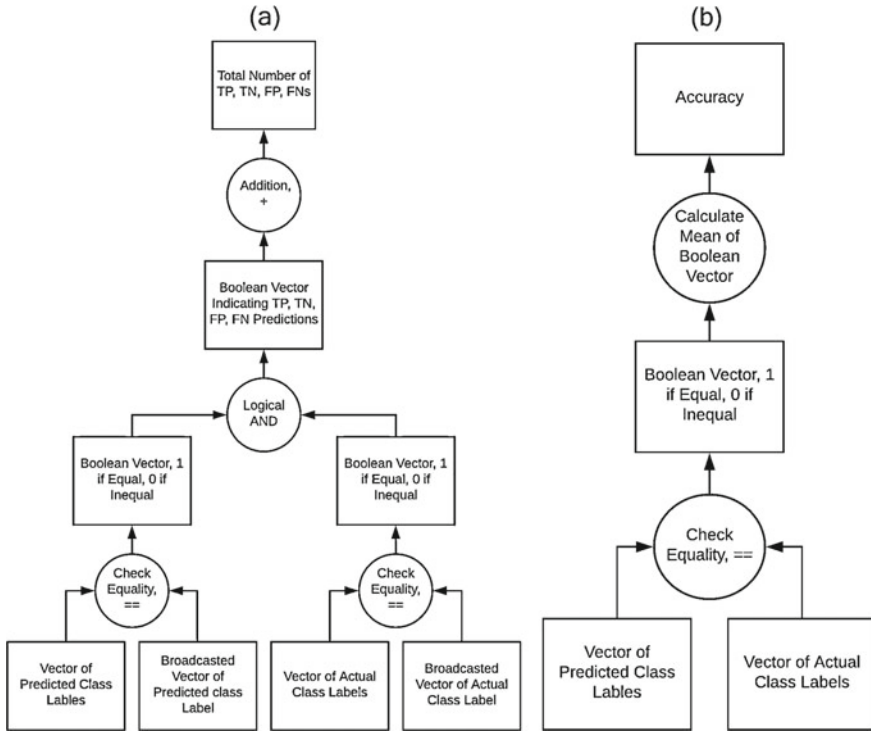


Fig. 13 a generalized computation graph for determining entries associated with confusion matrix, b computation graph for the computation of accuracy

4 Experimental Results and Discussion

The experimental results produced in this study draw much from the methods followed since the results influenced the methods to devolve to the final, most successful, highly optimized model (Fig. 12c). The results initiate a funnel-like pavement which gradually accumulates the fragmented modules to a single, meaningful piece. The methodology, intertwined with the results, show a gradual reduction in the complexity of the neural networks and the number of features—lowering them from 16 to 7 to a final 5, whilst not compromising the accuracy (Table 4), 99.77% (maximum).

Before delving into the practical results evaluated upon the models, we initially shed light on the results obtained from preprocessing the data. We first applied the *Apriori* algorithm on the transactional forms of the Q_{10} -chat answers and found a high tendency among them of being together (Fig. 8). To recapitulate the methods, observing the ten strong rules reveals this tendency and hence we perform a fuzzy conglomeration of these to a single feature. Next, we shed off features based on t/Chi -metrics (Tables 1 and 2) and find the prior features (Table 3).

Table 3 The finally extracted features with p -values in increasing order, defining their statistical priority

Statistical priority	Discrete/continuous feature	Inferential p -value
1	10 questions' quiz-score	$<2.2E-16$
2	Ethnicity	$3.93E-06$
3	Gender	$1.79E-04$
4	Jaundice history	0.01983
5	Age (months)	$4.26E-02$

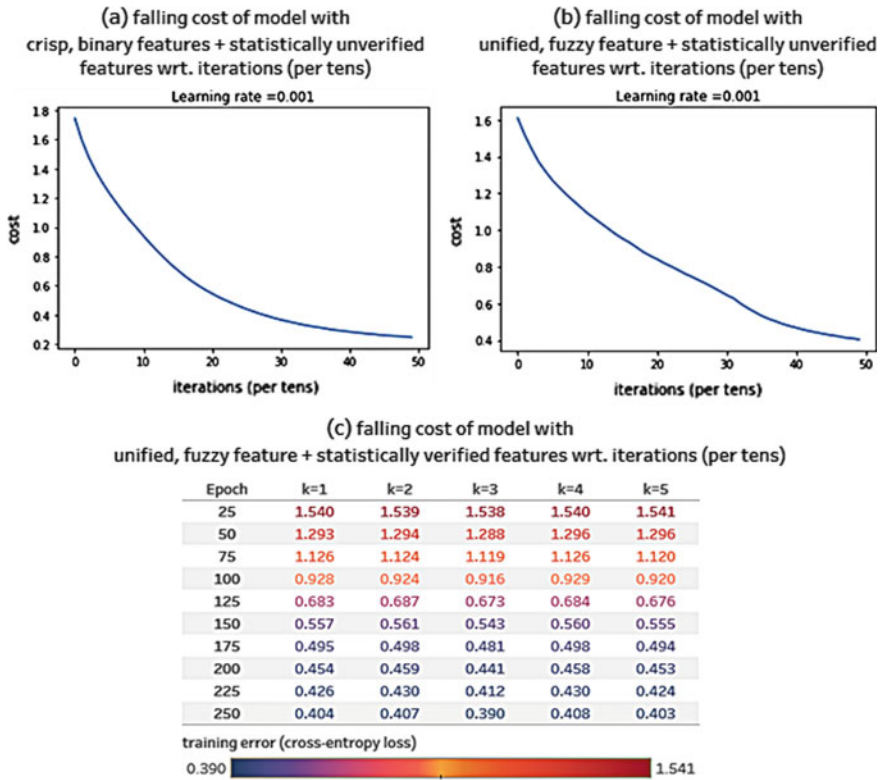


Fig. 14 Learning curves learned upon training the different schemes, with the loss plotted once per ten epochs

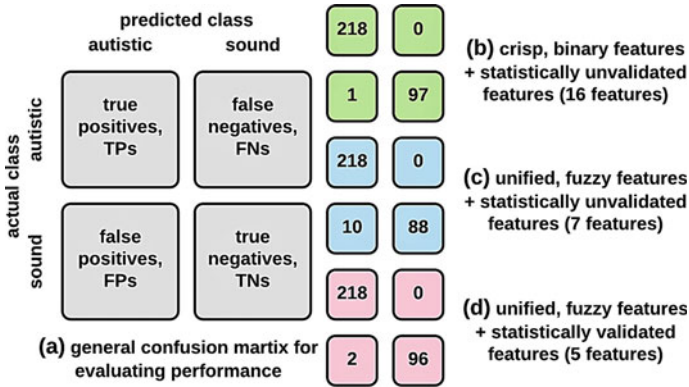


Fig. 15 Confusion matrices filled against $k = 1$ -st cross-validation set for each of the schemes

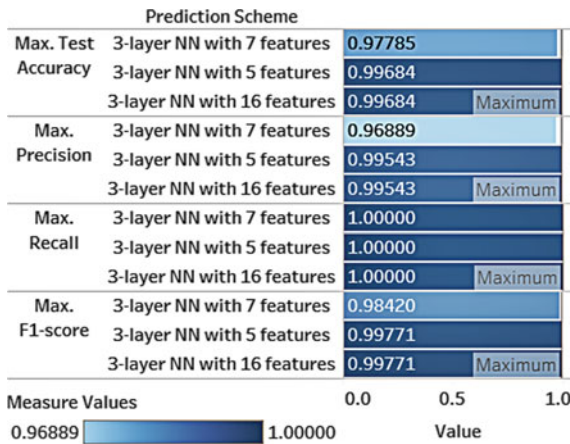


Fig. 16 Comparison among the proposed schemes' results

First, we attempt to show that we conducted a sound training of the ML models. With a minuscule learning rate of 0.001, the learning curves obtained showed a gradual reduction in the loss function with each of 250 epochs (Fig. 14a, b). For the final model, we choose to show each of $k = 5$ cross-validations converging to a minimal error with the warmer shades showing higher errors that gradually cooled down to errors as low as 0.390 (Fig. 14c). The training led us to machine-generated predictions that filled out the confusion matrices (shown for $k = 1$ in Fig. 15).

For each optimized model, we evaluate k -fold cross-validated results with $k = 5$ (Table 4) to make sure that the models are showing consistent performance. The results were obtained upon using an identical set of hyperparameters and showed brilliance in the test accuracies given that above 95% is generally held as an excellent performance by any ML algorithm. We perform a comparative analysis among the maximum of the metrics evaluated using each model and find that the results of the

Table 4 Raw results and metrics delivered by the models for $k = 5$ validation-sets each

Prediction scheme	k -fold	Optimized training loss	Test accuracy	Precision	Recall	F1-score
3-layer NN with 16 features	–	0.259586	0.9968355	0.9954338	1	0.9977117
3-layer NN with 7 features	1	0.426662	0.9683544	0.95614034	1	0.97757847
3-layer NN with 7 features	2	0.417105	0.9778481	0.9688889	1	0.984198651
3-layer NN with 7 features	3	0.432018	0.9746835	0.96460176	1	0.981981977
3-layer NN with 7 features	4	0.404772	0.9746835	0.96460176	1	0.981981977
3-layer NN with 7 features	5	0.529996	0.9778481	0.9688889	1	0.984198651
3-layer NN with 5 features	1	0.403833	0.99050635	0.98642534	1	0.993166287
3-layer NN with 5 features	2	0.40695	0.9936709	0.9909091	1	0.995433795
3-layer NN with 5 features	3	0.389566	0.9968355	0.9954338	1	0.997711676
3-layer NN with 5 features	4	0.40799	0.99050635	0.98642534	1	0.993166287
3-layer NN with 5 features	5	0.402822	0.9936709	0.9909091	1	0.995433795

final 5-feature model showed equal promise as the 16-feature model, albeit with less computation (Fig. 16).

After having tabulated the results we demonstrate that the third model, i.e. the second optimized scheme with 5 features took less time to be trained than the first optimized model, i.e. the model with 7 features (Fig. 17). We show the comparison employing all the $k = 5$ validations and find a decrement in the average training time in the model with 5 features. The intuition behind such results is that a model using fewer features will require less time for backpropagation since there will be fewer features to calculate the contribution in the total error.

We finally apply statistical inference-tests to make sure that the supremacy of one model over the other actually holds water. We apply one-sample t -test on all $k = 5$ cross-validated accuracies of the second optimized scheme (Fig. 12c) against the average accuracy of the first optimized scheme (Fig. 12b) and find t -statistic = 15.501 with a p -value = 0.0001011 (Table 5), repudiating the null hypothesis and proving the alternative hypothesis that the 5-feature model works significantly and consistently better. We apply a statistical ANOVA (Table 6) on the F1-scores of the three models and find: $F(2, 12) = 96.98$, p -value = $3.91\text{E}-08 \ll 0.05$, leading us to accept the safe conclusion that the models are indeed performing with measures separating them from each other. We finally visualize our efforts in comparison

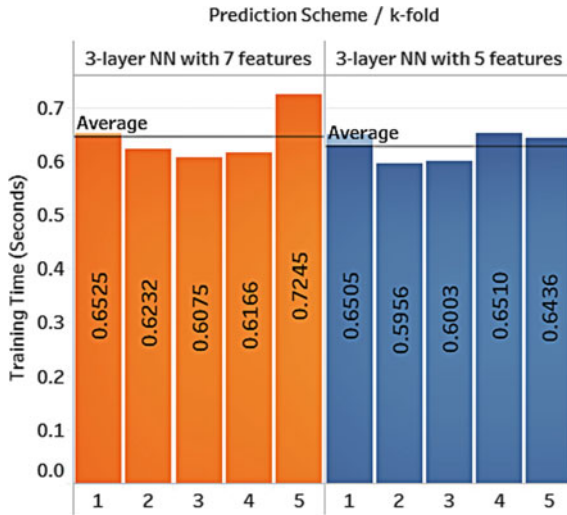


Fig. 17 Training time in the final model is reduced

Table 5 *t*-test results verifying better results yielded by the second optimized model over the first

One sample <i>t</i> -test paramters/Metrics	Evaluations
Comparison of avg. accuracy (model using 5 features) with	<i>k</i> -fold accuracies of model using 7 features
<i>t</i> -statistic	15.501
Degrees of freedom	4
<i>p</i> -value	0.0001011
95% confidence interval	0.9897505 to 0.9963255
Sample mean	0.993038
Alternative hypothesis, H_a : true mean is not equal to 0.9746835	True

Table 6 ANOVA-test verifying systematic difference in the models' performances

ANOVA (Analysis of Variance) test metrics	Values
Degrees of freedom for numerator (ind)	2
Degrees of freedom for denominator (residuals)	12
Sum of squares of numerators (ind)	0.0007059
Sum of squares of denominators (residuals)	0.0000437
Mean of squares of numerators (ind)	3.53E-04
Mean of squares of denominators (residuals)	3.60E-06
Analysed value	96.98
<i>p</i> -value, $Pr(> F)$	3.91E-08

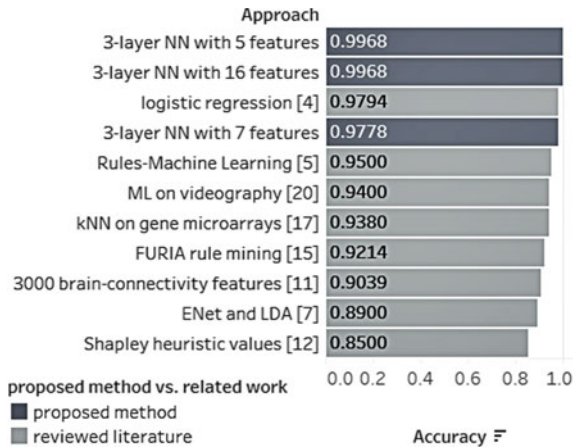


Fig. 18 Comparison among proposed and reviewed methods for the detection of autism in toddlers

with reviewed literature in terms of accuracy and find equivalent or in some cases, superior results (Fig. 18). The point of the discussion is that by initially combining some features into a fuzzy one, we get a tremendous accuracy and by further pruning of insignificant features, we manage to get even better performance, equivalent to what we had previously achieved by retaining all the features—obviously with less computation.

5 Conclusion

The problem of prognosticating autism has so far been a lengthy medical process that this study strives to overcome by taking mostly social and interactional elements into account. The novelty of this study has been the application of business intelligence procedures into a neuro-psychological problem, which was obtained through the following milestones:

- The study uses the *Apriori* algorithm not for producing the final output or merely for generating rules, the algorithm has rather obtained endorsement as a preprocessing step for creating better, fewer representative features.
- The paper has shown the effectiveness of the *Apriori*-based clubbing by converting the club into a fuzzy feature, the membership values of which constituted the foremost important feature.
- The research has reckoned the impact of features by using statistical inference and has assigned priorities to them using no machine learning method, hence delineating the effectiveness of statistical preprocessing.
- The endeavor has shown that hard computing (statistical preprocessing, association rule mining), if applied prior to soft computing (neural networks), can signifi-

cantly reduce the load on soft computing while maintaining the same scintillating performance.

The research-work presented adds value to not only computer science or data science, but also attempts to help the psychological or neurological community to investigate the aspects of autism at a young age. The work has future prospects of being extended over varieties of age-groups and the limitation of a class-imbalance can be overcome from a data-scientific front. The study lends a message to the research community to explore novel preprocessing techniques instead of making their first leap to machine learning.

References

1. <https://www.verywellhealth.com/what-are-the-three-levels-of-autism-260233>. Accessed on 6th March 2020
2. Lyall, K., Croen, L., Daniels, J., Fallin, M.D., Ladd-Acosta, C., Lee, B.K., Park, B.Y., Snyder, N.W., Schendel, D., Volk, H., Windham, G.C.: The changing epidemiology of autism spectrum disorders. *Ann. Rev. Public Health* **20**(38), 81–102 (2017). Mar
3. Dover, C.J., Le Couteur, A.: How to diagnose autism. *Arch. Disease Childhood* **92**(6), 540–5 (2007). Jun 1
4. Thabtah, F.: An accessible and efficient autism screening method for behavioural data and predictive analyses. *Health Inform. J.* **25**(4), 1739–55 (2019). Dec
5. Thabtah, F., Peebles, D.: A new machine learning model based on induction of rules for autism detection. *Health Inform. J.* **29**, 1460458218824711 (2019). Jan
6. Sarkar, A., Wade, J., Swanson, A., Weitlauf, A., Warren, Z., Sarkar, N.: A data-driven mobile application for efficient, engaging, and accurate screening of ASD in toddlers. In: *International Conference on Universal Access in Human-Computer Interaction*, vol. 15, pp. 560–570. Springer, Cham (2018)
7. Duda, M., Haber, N., Daniels, J., Wall, D.P.: Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Transl. Psychiatry* **7**(5), e1133
8. Achenie, L.E., Scarpa, A., Factor, R.S., Wang, T., Robins, D.L., McCrickard, D.S.: A machine learning strategy for autism screening in toddlers. *J. Develop. Behav. Pediatrics* **40**(5), 369–76 (2019). Jun 1
9. Thabtah, F., Kamalov, F., Rajab, K.: A new computational intelligence approach to detect autistic features for autism screening. *Int. J. Med. Inform.* **1**(117), 112–24 (2018). Sep
10. Abbas, H., Garberson, F., Glover, E., Wall, D.P.: Machine learning approach for early detection of autism by combining questionnaire and home video screening. *J. Am. Med. Inform. Assoc.* **25**(8), 1000–7 (2018). Aug
11. Kong, Y., Gao, J., Xu, Y., Pan, Y., Wang, J., Liu, J.: Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing* **9**(324), 63–8 (2019). Jan
12. Tariq, Q., Fleming, S.L., Schwartz, J.N., Dunlap, K., Corbin, C., Washington, P., Kalantarain, H., Khan, N.Z., Darmstadt, G.L., Wall, D.P.: Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: development and validation study. *J. Med. Internet Res.* **21**(4), e13822 (2019)
13. Al Farsi, A., Doctor F, Petrovic, D., Chandran, S., Karyotis, C.: Interval valued data enhanced fuzzy cognitive maps: towards an approach for autism deduction in toddlers. In: *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6. IEEE, 9 July 2017

14. Al-Diabat, M.: Fuzzy data mining for autism classification of children. *Int. J. Adv. Comput. Sci.Appl.* **9**(7), 11–7 (2018). Jul 1
15. Khan, S., Alshara, M.: Fuzzy data mining utilization to classify kids with autism. *IJCSNS* **19**(2), 147 (2019). Feb
16. Xiao, X., Fang, H., Wu, J., Xiao, C., Xiao, T., Qian, L., Liang, F., Xiao, Z., Chu, K.K., Ke, X.: Diagnostic model generated by MRI-derived brain features in toddlers with autism spectrum disorder. *Autism Res.* **10**(4), 620–30 (2017)
17. Kim, S.H., Kim, I.B., Oh, D.H., Ahn, D.H.: Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. *Eur. Neuropsychopharmacol.* **1**(27), S1090 (2017). Oct
18. Karan, P., Pirouz, M.: EEG Analysis for Predicting Early Autism Spectrum Disorder Traits
19. Abbas, H., Garberson, F., Glover, E., Wall, D.P.: Machine learning for early detection of autism (and other conditions) using a parental questionnaire and home video screening. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 3558–3561. IEEE, 11 Dec 2017
20. Tariq, Q., Daniels, J., Schwartz, J.N., Washington, P., Kalantarian, H, Wall, D.P.: Mobile detection of autism through machine learning on home video: a development and prospective validation study. *PLoS Med.* **15**(11) (2018)
21. Ng, A.: CS229 Lecture notes. *CS229 Lecture Notes* **1**(1), 1–3 (2000)

Developing a Framework for Generating Hotel Recommendation



Md. Shafiu Alam Forhad and Mohammad Shamsul Arefin

Abstract Recommendation Systems (RS) are utilized in a variety of areas. RS provide suggestions to items that a particular user is most likely interested in. There has been a huge amount of research in this field. An intelligent approach to generate recommendations by using heterogeneous data is proposed in this paper. For increasing the performance of the recommendation system, we consider the surrounding environments of the considered hotels. Feedback system is considered as a computer program in information technology (IT). The analysis of the feedbacks provided by the users is essential for the improvement of accuracy RS. By analyzing both textual and numeric feedbacks, a better suggestion will be provided to the users.

Keywords Recommendation Systems · Point of Interest · Collaborative Filtering · Category Tree · Score Generation

1 Introduction

Recommendation Systems are information processing systems which usually make use of either or both collaborative filtering and content-based filtering and other systems such as knowledge-based systems. In [1], a simple and original RS makes recommendations to the active user. This recommendation system is based on items that other users with similar tastes liked in the past. Extraction of features from textual reviews is necessary. Recommenders are mainly applied in a homogeneous way. Movie recommenders like Netflix, news recommenders like Google News, as well as music recommenders each focuses on a single specific application domain. Homogeneity is a major limitation given the growing multiplicity of web applications. To achieve true recommendations, processing and analyzing large heterogeneous web

Md. Shafiu Alam Forhad (✉) · M. S. Arefin
Department of Computer Science & Engineering, CUET, Chattogram, Bangladesh
e-mail: forhad0904063@cuet.ac.bd

M. S. Arefin
e-mail: sarefin@cuet.ac.bd

data are necessary. The data can be come from various sources including text, images, videos, etc.

Many well-known e-commerce companies such as TripAdvisor [2], Booking.com [3], Agoda [4], Expedia [5], and Foursquare [6] sell their products and services via Internet. In e-Tourism [7–9], methods based on ontology are used extensively.

We have developed a framework for hotel recommendation by considering the surrounding environments of the hotels, numeric and textual feedbacks, etc. This development task also involves some preprocessing steps based on natural language processing.

The rest of the paper is summarized as follows: A brief review of related work is provided in Sect. 2. In Sect. 3, a detail description of our proposed framework is presented. The experimental result is presented in Sect. 4. Finally, a conclusion section is provided in Sect. 5.

2 Related Work

Shamsul Arefin et al. [10] proposed a technique for suggesting hotels to the users. For choosing a hotel for recommendation, they considered the reviews of the user's about hotels, coexistence of some facilities within the surrounding areas, etc. They performed different experiments and their results are well applicable for the users.

In [11], they introduced an efficient method for recommending hotels for the users. They considered the coexistence of other facilities such as museum, art gallery, and other facilities in the nearby areas. Various recommender systems (RSs) based on collaborative filtering [12], content-based filtering [13] and the combination of both [14] (hybrid) have been developed in the last two decades to provide a personalized recommendation to the users. These recommender systems are being used in the various online platforms such as Netflix, Amazon, and eBay to increase their sales output. A hybrid recommendation systems have been developed while combined the features of collaborative and content-based filtering to solve some of their shortcomings. However, there are several limitations still exists in these approaches. In [15] they proposed an algorithm that combined the multi-criteria ratings of items and items semantic information. Their proposed algorithm is very efficient in dealing with the new item and sparsity problems with respect to the recommendation and coverage prediction accuracy. The performance of this algorithm against larger data sets is the limitation. In [16], they introduced a multi-criteria recommendation system and the proposed framework is able to unsupervisedly extract relevant aspects from the review. Their proposed methods are not encoding the characteristics extracted from users reviews. In the last decade, there are a variety of multi-criteria recommendation systems [12, 15–18] which have been developed to solve the problem of traditional recommendation systems. The insufficiency of rating data makes this method unsatisfactory [19]. These methods are still not adequate when target user has little historical data.

In [20], they worked on multi-dimensional rating information. They demonstrated that the integration of item and user-based methods leads to accurate recommendations. They also showed that all dimension ratings are not equally important for prediction. In [21], they have considered a technique for recommending hotels. Most of typical hotel recommendation systems use a vector space model to represent hotels. They proposed a technique that may automatically decide attributes of vector space. In most of the systems, users will submit natural language comments, during which users can express numerous impressions concerning hotels. They have utilized natural language comments that have submitted by users. They analyzed texts in comments by using latent Dirichlet allocation (LDA) and extracted representative topics concerning hotels from the texts automatically. They further analyzed the texts sentiment for every extracted topic for every hotel.

In [22], they proposed a system which is based on CF and Rankboost algorithm. By using this recommendation system, a user can find a hotel efficiently and quickly. In [23], they proposed an opinion mining system. Their proposed system collects user reviews and comments from the web. Their system is capable of classifying, detecting and retrieving reviews on the web. It is also capable of generating comprehensive overviews from those comments. In [24], they proposed a system which recommends by using the traveler preferences. In [25], they proposed a new technique which builds user profiles from users review texts. By using topic profile collaborative filtering, it profiles to filter other review texts with the eyes of the user. This filtering provides a far better mean average error (MAE) once predicting ratings. It also provides a far better approximation of user preference orders. In [26], they proposed a new method using SVM classifier with completely different user given rating reviews based on mostly feature selection. Their challenge is to detect of spam contents in users review. The challenge is also to use the study of this problem to solve. To get any restaurants positive and negative reviews, the system has turned the online available reviews interesting. They have researched for sentiment analysis of user reviews and have applied support vector machines to find optimal results. In [27], their proposed method has a tendency to specialize in the analysis values given by contributors whose preferences are just like the user's preference.

In [28], collaborative user and item filtering techniques was used in combination with sentiment classification. They used sentiment classification results as feedback. Their proposed technique helps in the case where an item has textual reviews but no ratings. The recommendation results are more accurate compared to recommendation.

systems based solely on filtering techniques (Fig. 1).

3 System Architecture and Design

Data are collected from different hotel booking websites. Then the items which do not satisfy some criteria over some features are eliminated. A user can view systems different facilities from systems facilities database. We analyzed numerical ratings

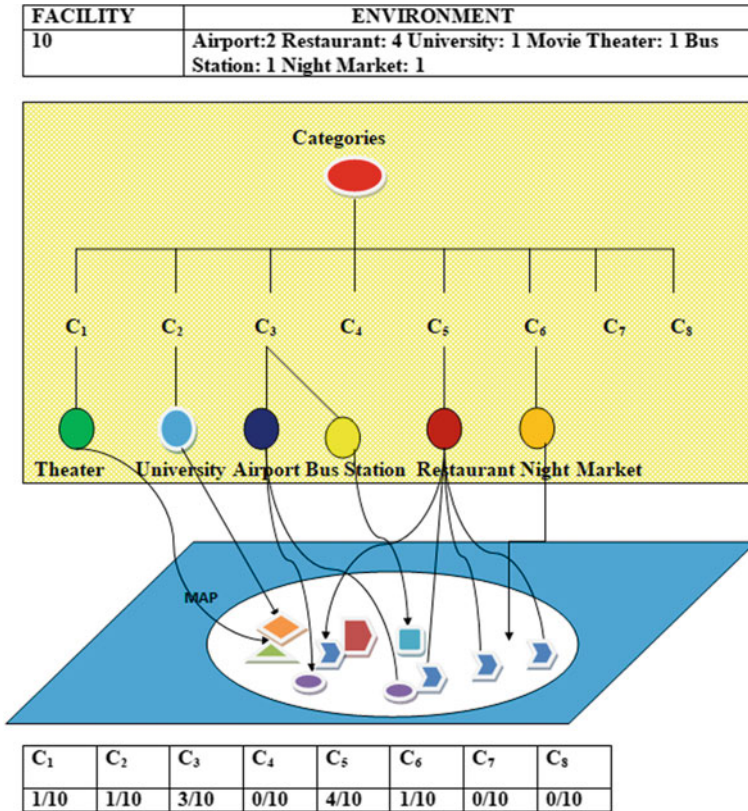


Fig. 1 An Example of Evaluating Hotel’s Surrounding Environment

and textual feedback. Corresponding numerical scores are calculated from the written one. Stop words are removed from the textual feedbacks. Some special characters and symbols are also removed from the texts. Finally, the score generation module generate scores for different reviews of the given datasets. Then the numerical scores and scores generated from textual feedback are aggregated. Finally, the average scores are calculated for each considered hotel of the dataset. The system architecture of our system is shown in Fig. 2. Data crawling module, data storing module, user activity module, feedback analysis module, keyword extraction module, recommendation module, etc., are the different modules used in our system.

3.1 Data Storing and Processing

Data storing and processing module consist of the sub-modules: collect feedback, feedbacks separator, feedbacks database creator, stop word removal, keywords

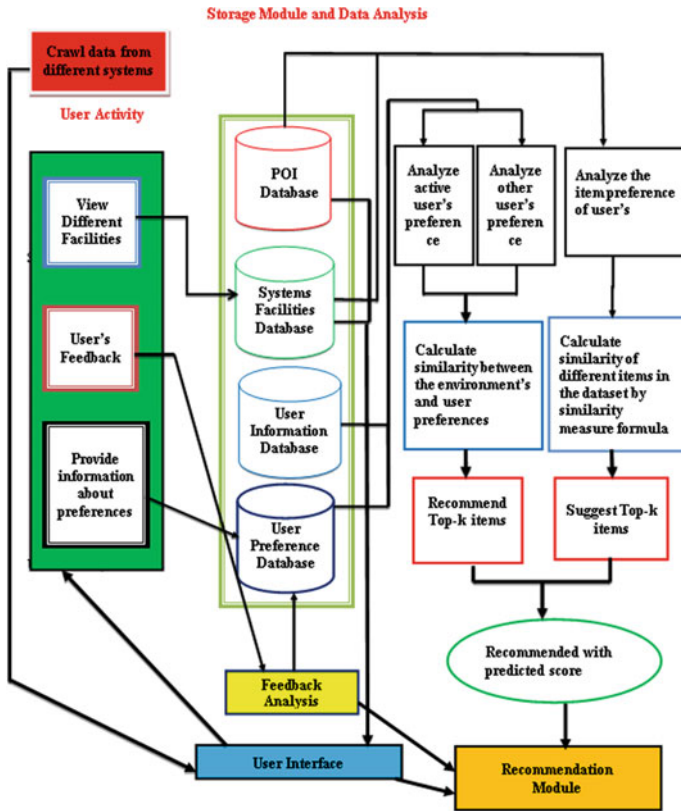


Fig. 2 Developing a Framework for Generating Hotel Recommendation

extractor and morphological analyzer sub-modules. Storage module stores information processed by database storing and processing module. Three storages are there, Feedbacks databases will store all written feedbacks Stop word database will keep the list of stop words and extracted keywords from written feedbacks are stored in keyword database. This total storage is required for the next score generation module to process the written feedbacks.

3.2 Point of Interest (POI)

POI is a specific point location that specifies the latitude and longitude by considering a certain map datum. POIs database is used in our system. This database is used for evaluating the surrounding environments of hotels. We classified facilities according to various categories. It is shown in Fig. 3a as a category tree. The objects of the

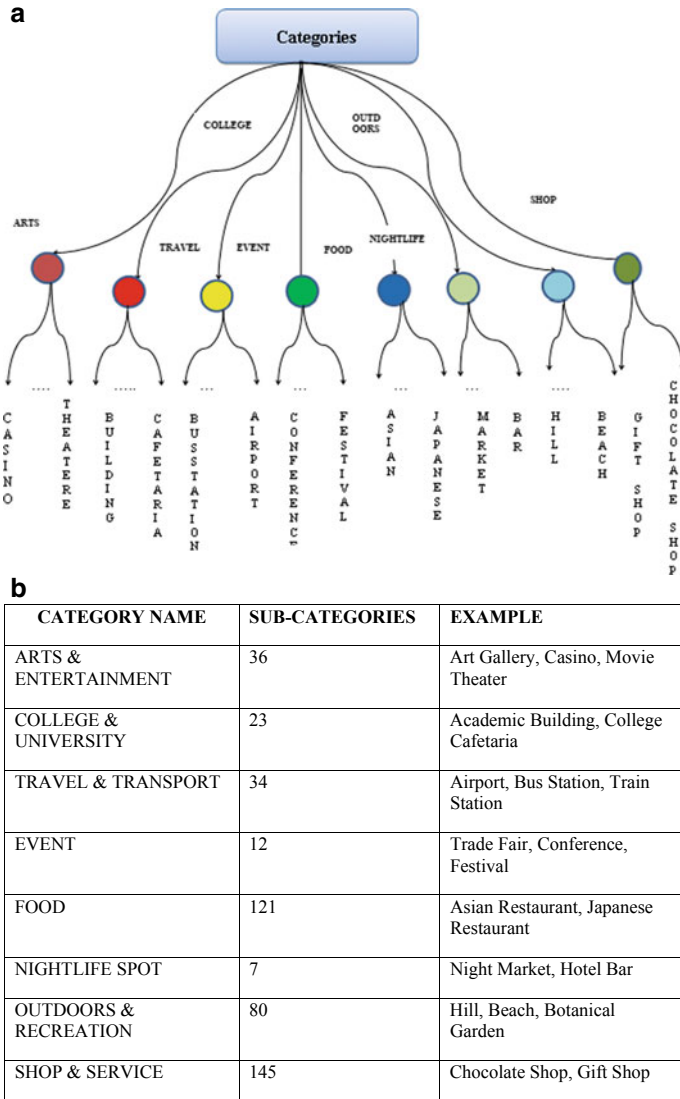


Fig. 3 a Category Tree, b Location Categories in [6]

facilities are indicated by the leaf nodes and the internal nodes contain the facility types.

We have used the information from ‘www.foursquare.com’ to draw this category tree. There are many sub-categories under the specific categories. By taking into consideration the information of the category tree, the scores of each considered hotels are calculated. Our system infers the user’s preferences by monitoring the different actions of users. Our systems provide better recommendations by analyzing

user's feedback. In Fig. 1, a motivating example of evaluating hotel's surrounding environments is shown. In Table 1, an example of item-item similarity calculation is shown for a specific user. At first, our system will find the preference list of a user. The preference scores of different categories of different facilities are calculated with the help of the user past stayed hotels. Then by using the similarity measure formula, our system provides recommendations for top-k similar hotels to the user.

3.3 Keyword Extraction and Refinement

Keyword extraction module extracts the root of every word using morphological analyzer. Here, the words from the review texts are taken as input and the root of the words are generated as an output. The process of finding the root of a word is called stemming. To stem the postfixes from the terms of the written comment, the terms were checked against a postfix list. After completing this step, specific keywords are stored as features. Features are kept in a database called 'Keywords' as features database.

3.4 Score Generation

A framework to generate scores from written feedbacks is shown in Fig. 4. We create a list of positive and negative words, and then we categorized the words as five categories. For the category 'Excellent,' we assigned five point for those words which fall in this category. The words which are in the category 'Terrible' will be given a score equal to one. For each review, the score is obtained by dividing the total scores for the review to the total number of categorized words. After calculating score for each review, an average score is generated for each hotel in the sample dataset. We only consider near about three hundred reviews of nine different hotels of tripadvisor for the purpose of score generation. In future, we will consider more reviews from different hotel booking websites to provide better recommendations to the users.

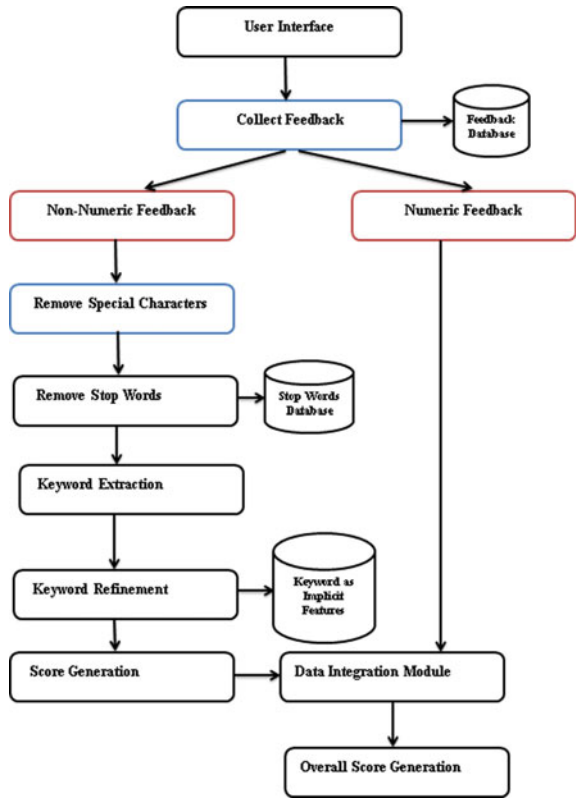
4 Experimental Results and Analysis

Data are collected from different hotel booking websites. In Fig. 5, a comparative analysis of the performance of the hotels from the user provided reviews collected from TripAdvisor and Expedia booking sites are shown. Then we analyze comments and scores are generated from the comments. We analyzed near about three hundreds written comments from TripAdvisor for nine hotels. In future we will try to experiment more data from two or more booking websites. In Fig. 6, the average scores generated from comments for each hotel and the average scores of the numerical

Table 1 Item-Item Similarity Calculation

HID	Arts and Entertainment	College and University	Traavel and Transport	Event	Food	Nightlife Spot	Outdoors and Recreation	Shop and Service	Similarity
1	0.230769	0	0.384615	0	0.076923	0.076923	0.076923	0.153846	0.723969
2	0.4	0	0.5	0	0	0	0.1	0	0.705115
3	0.363636	0	0.090909	0	0.090909	0.090909	0.090909	0.272727	0.865066
4	0.181818	0	0.363636	0	0	0	0.272727	0.181818	0.793941
5	0.153846	0	0.307692	0	0	0	0.307692	0.230769	0.802084
6	0.3	0	0.4	0.1	0	0	0.1	0.1	0.746174
7	0.357143	0	0.142857	0.071429	0.071429	0.071429	0.214286	0.071429	0.979432
8	0.166667	0	0.25	0	0.166667	0	0.25	0.166667	0.854501
9	0.4	0	0.3	0	0	0	0.1	0.2	0.837969
10	0.166667	0	0.277778	0.166667	0.055556	0	0.222222	0.111111	0.810622

Fig. 4 A Framework for Feedback to Score Generation



rating of each hotel is shown. In Tables 2 and 3 keywords for different categories & two examples of score generation are shown.

5 Conclusions and Future Work

In this paper, we introduced a hotel recommendation system. As the booking hotels via web-based services are becoming popular day by day, there is a need to consider heterogeneous data for providing better recommendations. Our proposed hotel recommender framework consists of data crawling module, data storage module, feedback analysis module, etc. The proposed framework not only processes user’s feedback, but also considers surrounding environments of the hotels. Firstly, we utilized POIs database to obtain the surrounding environment of considered the hotels that a particular user may be interested in. Then, the hotels that do not satisfy some criteria are removed from the list. Then the numerical ratings and non-numerical feedbacks, etc., provided by the different users of the system are considered. We used the NLTK library in our system to identify polarity of the textual reviews.

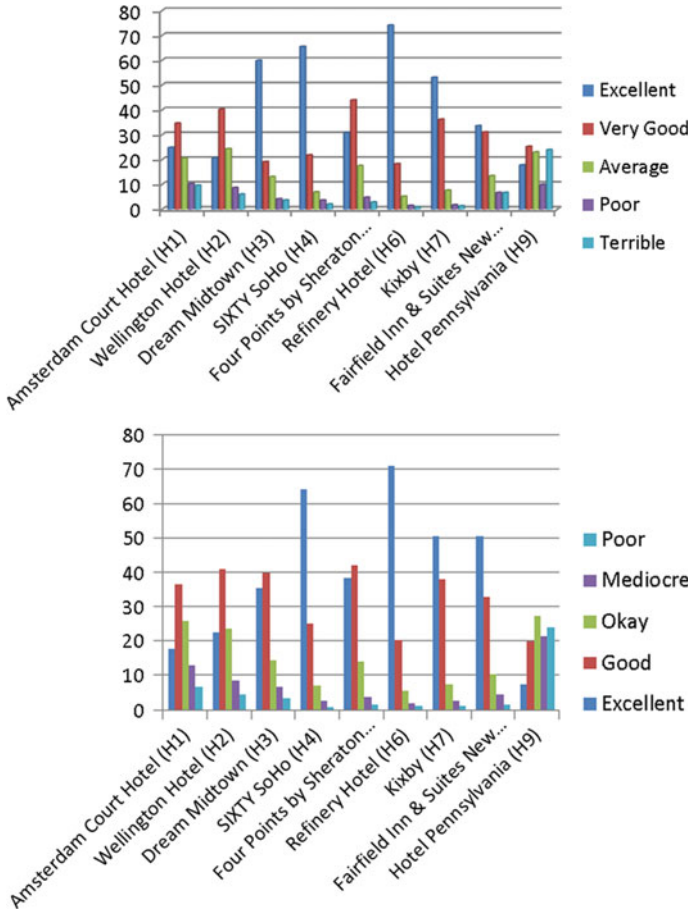


Fig. 5 Comparative analysis of the user provided reviews for different data sources of different hotel booking sites (TripAdvisor and Expedia)

Scores generated from the textual reviews and play an important role in the recommendation system. In future, we will try to consider data from two or more hotel booking web sites to generate better recommendations.

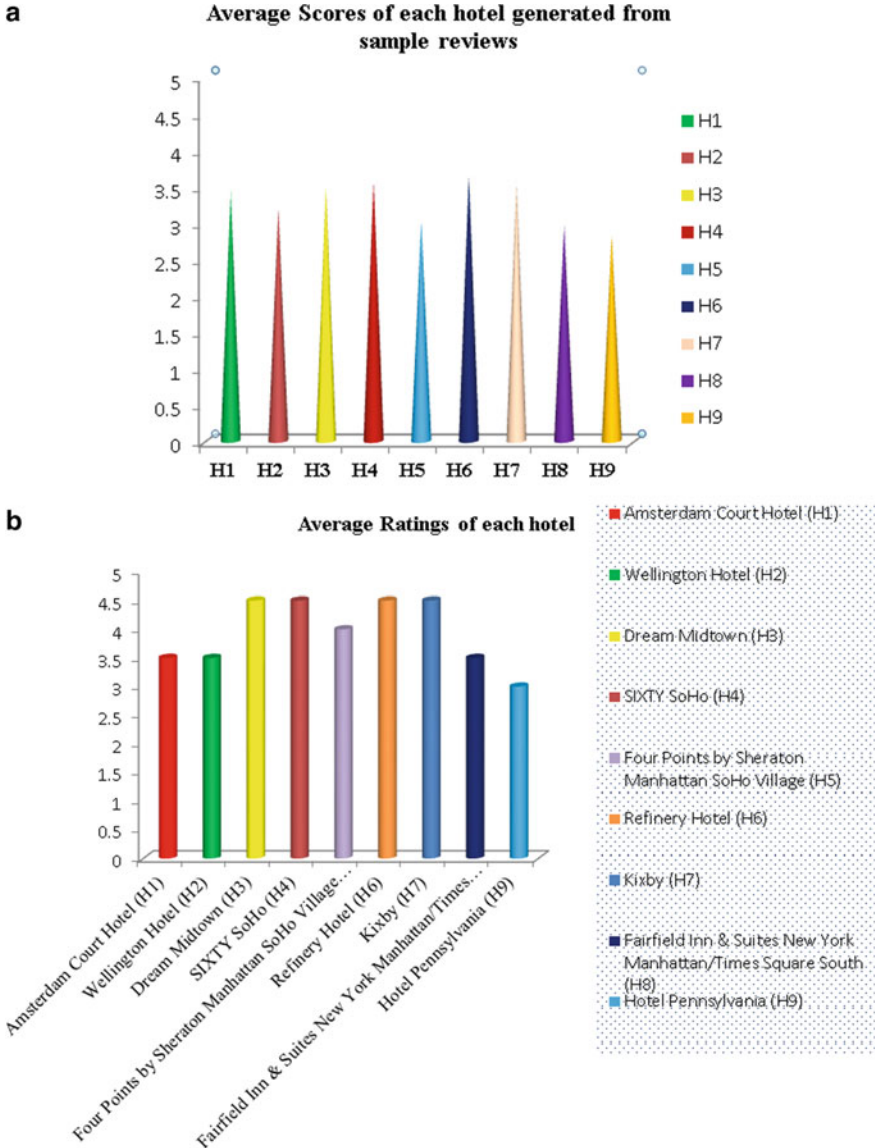


Fig. 6 a Average scores of each hotel generated from the sample reviews of the nine hotels. **b** Average numerical ratings of each hotel (data collected from www.tripadvisor.com)

Table 2 Keywords for different category of words

S. No.	Category	Keywords	Score
1	Excellent	Excellent, best, ...	5
2	Very good	Happy, nice, ...	4
3	Average	Average, ordinary, ...	3
4	Poor	Poor, afraid, difficult, ...	2
5	Terrible	Danger, horrible, terrible...	1

Table 3 Example of score generation

UID	Feedbacks	Generated Score
1	Very comfortable hotel. Excellent and friendly and	$(5 + 5 + 4)/3 = 5$
2	Squeaky floors difficult to go to the bathroom at night without waking my husband	2

References

- Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. *Commun. ACM* **35**(12), 61–70 (1992)
- TripAdvisor: www.tripadvisor.com
- Booking.com: www.booking.com
- Agoda: www.agoda.com
- Expedia: www.expedia.co.in
- Fourquare: www.foursquare.com
- Sebastia, L., Garcia, I., Onaindia, E., Guzman, C.: E-tourism: a tourist recommendation and planning application. *Int. J. Artif. Intell. Tools* **18**(5), 717–738 (2009)
- Sebastia, L., Giret, A., Garcia, I.: A multi agent architecture for tourism recommendation. In: Demazeau, Y., Dignum, F., Corchado, J.M., et al. (eds.) *Trends in Practical Applications of Agents and Multiagent Systems*, vol. 71 of *Advances in Intelligent and Soft Computing*, pp. 547–554. Springer, Berlin (2010)
- Garcia, I., Sebastia, L., Onaindia, E.: On the design of individual and group recommender systems for tourism. *Expert Syst. Appl.* **38**(6), 7683–7692 (2011)
- Shamsul Arefin, M., et al.: Recommending hotels by social conditions of locations. In: *Intelligent Systems Reference Library*, pp. 91–106 (2015)
- Chang, Z., et al.: Hotel Recommendation Based On Surrounding Environments. In: 2013 IIAI International Conference on Advanced Applied Informatics (IIAIAI) (2013)
- Shambour, Q., et al.: An item-based multi-criteria collaborative filtering algorithm for personalized recommender systems. *Int. J. Adv. Comput. Sci. Appl.* **7**(8) (2016)
- Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: *The Adaptive Web*, pp. 325–341
- Bogers, T., Den Bosch, A.V.: Collaborative and content-based filtering for item recommendation on social bookmarking websites, vol. 532
- Adomavicius, G., Kwon, YoungOk: New Recommendation techniques for multicriteria rating systems. *IEEE Intell. Syst.* **22**(3), 48–55 (2007)
- Adomavicius, G., et al.: Multi-criteria recommender systems. In: *Recommender Systems Handbook*, pp. 769–803 (2010)
- Shambour, Q., Lu, J.: Integrating multi-criteria collaborative filtering and trust filtering for personalized recommender systems. In: 2011 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MDCM) (2011)

18. Manouselis, N., et al.: Revisiting the multi-criteria recommender system of a learning portal. *CEUR Workshop Proceedings*, vol. 896, pp. 35–48 (2012)
19. Jakob, N., et al.: Beyond the stars. In: *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion (TSA'09)* (2009)
20. Jannach, D., et al.: Recommending hotels based on multi-dimensional customer ratings. In: *Information and Communication Technologies in Tourism*, pp. 320–331 (2012)
21. Zhang, Z., Morimoto, Y.: Collaborative hotel recommendation based on topic and sentiment of review comments (2017)
22. Huming, G., Weili, L.: A hotel recommendation system based on collaborative filtering and rankboost algorithm. In: *2010 Second International Conference on Multimedia and Information Technology* (2010)
23. Kasper, W., Vela, M.: Sentiment analysis for hotel reviews, pp. 45–52 (2011)
24. Fasahte, U., et al.: Hotel Recommendation system. *Imperial J. Interdisc. Res.* **03**(11), 317–324 (2017)
25. Cristian Musat, C., et al.: Recommendation Using Textual Opinions, pp. 2684–2690 (2013)
26. Kharadi, B., Patel, K.: Opinion mining of restaurant review by sentiment analysis using SVM. *Int. J. Innov. Res. Comput. Commun. Eng.* **5**(3) (2017)
27. Takuma, K., et al.: A hotel recommendation system based on reviews: what do you attach importance to? In: *2016 Fourth International Symposium on Computing and Networking (CANDAR)* (2016)
28. Jayashree, R., et al.: Recommendation system with sentiment analysis as feedback component. In: *Proceedings of Sixth International Conference on Soft Computing for Problem Solving, Advances in Intelligent Systems and Computing* 547, pp. 359–367. Springer, Singapore (2017)

Rumor Source Identification on Social Networks: A Combined Network Centrality Approach



Abhijit Das and Anupam Biswas

Abstract The source identification of any particular rumor in the social network is a critical task due to its intricate network connectivity. In this paper, we have analyzed network centrality measures to detect the rumor source on real social network datasets. We used susceptible-infected (SI) model to construct a graph of infected nodes initiated from a single source and proposed a combined network centrality approach (CNCA) to identify that source. Our approach combines well-known rumor centrality value with betweenness centrality value to maximize the likelihood of source detection probability. Simulations were performed on varied network structures with varied complexities to analyze the performance of the proposed approach. We compared our approach with rumor centrality, betweenness centrality, and Jordan centrality approaches. We observed that a combination of rumor centrality and betweenness centrality approach outperforms the individual centrality measures on source identification.

Keywords SI model · Rumor source detection · Betweenness centrality · Jordan centrality · Rumor centrality

1 Introduction

Web 2.0 is a participatory web where end users participate in generating content on a reliable and straightforward environment [1]. This participatory culture on the Internet has enabled the evolution of virtual social networks known as social media, where each user gets connected to others, regardless of geographical location. This social media has widely integrated into our daily routine, and its usage may vary from one individual to another [2]. Malicious users mostly use this platform to circulate

A. Das (✉) · A. Biswas
Department of Computer Science and Engineering, National Institute of Technology Silchar,
Silchar, Assam, India
e-mail: abhijitd87@gmail.com

A. Biswas
e-mail: abanumail@gmail.com

rumors for economically extensive coverage [3]. Susceptibility of social media users to rumors forces our society toward severe threats. Simultaneously, it reduces the belief of people in the content shared in social media.

Serious consequences may occur from the spread of rumors in a social network [3, 4]. In 2019, the share price of the UK's Metro Bank dropped by 11% due to the circulation of false rumors on WhatsApp and Twitter. Rumor stated that "Metro bank in UK was close to collapse and encouraging customers to empty their accounts and safety deposit boxes" [5]. In 2019, a rumor circulated on the social network, which mentioned that "final Naga settlement has been arrived at and will be announced soon Naga accord is being signed unilaterally." This rumor created anxiety and concern in some parts of India. Immediately, the Ministry of Home Affairs (MHA) released notifications to clarify that all stakeholders will be duly consulted and their concerns will be taken into consideration for any settlement on Naga issues [6]. The rumor source needs to be detected immediately to counter the spread and identify the purpose behind this malicious activity. Rumor source identification also helps law enforcement agencies to identify the culprit and act accordingly.

The process of information diffusion in the social network can be modeled using mathematical epidemiological models for analysis. An approach in detecting the rumor source is to spread the rumor on the network of nodes using the existing epidemiological model. The infected network of nodes can be analyzed to identify the probable source based on the diffusion process. In this study, susceptible-infected (SI) model is used as an underlying information propagation model to study rumor source identification on social network represented by a graph $G(V, E)$. SI model is used to infect a subset of nodes $V_N \subseteq V$, and the snapshot of an infected graph is analyzed to identify source using network centrality measures like betweenness centrality, Jordan centrality, and rumor centrality. Betweenness centrality and rumor centrality score of the infected nodes are combined to maximize the probability of source detection and termed our approach as a combined network centrality approach (CNCA). The combination of two centrality measures to identify source adds novelty to this study as per the literature reviewed. Given an undirected graph of infected social network $G_N = (V, E)$, where V is the set of infected nodes, E is the set of edges, and N is the number of nodes. Our goal is to minimize the probability of error in single rumor source identification, with prior knowledge of the underlying information propagation model.

The rest of the paper is organized as follows: Sect. 2 describes the related work, and Sect. 3 presents preliminaries on the information diffusion model and network centrality measures. Our proposed approach is described in Sect. 4, and Sect. 5 presents the obtained results on implementation. Finally, we conclude our paper in Sect. 6.

2 Related Work

Prior works related to rumor can broadly be classified into two categories: (I) rumor source identification and (II) counter rumor diffusion. Diffusion of rumors in the network may be initiated from either a single source or multiple sources. An approach to identify a single rumor source is based on the state of each node in the network. At an instant, a node can be either of the three states: susceptible, infected, or recovered [7]. According to the literature, network observation is of three types: complete observation, snapshot observation, and sensor observation [8]. Complete observation considers the exact state of each node in the network at time t while snapshot observation considers only the nodes infected at time t . Sensor observation is a little different from the other two categories. A designated sensor node is deployed in the network to periodically collect states of other nodes network with their infection time. In [9], the author tried to identify the rumor source using snapshot observation with the assumption that rumor is spread in the network using the SI model. They defined a new centrality measure called rumor centrality and formulated the problem as the maximum likelihood (ML) estimator. They performed simulation on real and synthetic networks and observed that rumor centrality performs better than distance centrality in general graphs, but for trees performance of both rumor centrality and distance centrality are equivalent. In [10], the authors placed sensors in the network to identify the location of the rumor source. Their approach has a computational complexity of $O(n)$ on trees, where n is the total no. of infected nodes.

In [11], the authors used the SI model as rumor diffusion initiated from a single source under snapshot observation. They modeled a maximum a posteriori (MAP) estimator and proposed a local rumor center for providing high probability in source identification. In [12], the authors studied rumor source identification without prior information of the number of sources under snapshot observation using the SI model. They proposed an estimator whose probability touches one with the increase in the number of nodes. When the number of infection sources is known to be two, the estimator calculates in $O(n^2)$, where n is the number of infected nodes. In [13], the authors proposed the NETSLEUTH algorithm and employed the minimum description length (MDL) principle to detect the best set of source nodes while considering the graph of infected nodes. In [14], diffusion of rumor in the network is based on the susceptible-infected-recovered (SIR) model initiated from a single source. The authors proposed a reverse infection algorithm to find the best path leading to the given snapshot and node with minimum infection eccentricity called Jordan infection center is the source node that leads to that best path.

In [15], the authors proposed a two-stage process to identify the infection source, which first detects a likely set of rumor sources and then uses the Markov random field method to identify the source node from a set of suspected nodes. SI model was considered for information diffusion in the network and formulated the problem as the maximum likelihood (ML) estimator. In [16], the authors explored the identification of multiple rumor sources in general networks that are highly loopy. They obtained the Hashimoto or non-backtracking matrix from the infected network and proposed

a heuristic approach based on the dominant eigenvalue of that matrix. The proposed approach neither converts the graph into tree nor partitions it into non-overlapping parts. In [17], the authors proposed graph convolutional network-based source identification (GCNSI), a deep learning-based model to identify multiple rumor sources close to the actual source. Their approach does not require prior knowledge of the underlying information diffusion model. Also, they proposed an input generation algorithm to convert labels of nodes from integer to vector.

3 Preliminaries

In this section, relevant information propagation model and network centrality measures are discussed in detail that is necessary for the understanding of CNCA.

3.1 Susceptible – Infected (SI) Model

The susceptible-infected (SI) model is chosen as the information propagation model in an undirected graph $G(V, E)$ where V is the set of vertices, and E is the set of edges of the form (u, v) where $u, v \in V$.

At any instant t , each node can be in one of the two possible states: susceptible (S) and infected (I) [18, 19]. The state transition diagram of nodes in the network is represented by Fig. 1. We assume that initially, there is only one node at an infected state, which is considered as the rumor source v^* , and all other nodes are in the susceptible state. Once a node becomes infected, it remains in the infected state forever, and it keeps on infecting its neighbors' nodes with probability β , which defines the rate constant of the spread of infection. The logistic growth equation of the SI model can be written as in Eq. (1).

$$\frac{dv'}{dt} = \beta(1 - v')v' \tag{1}$$

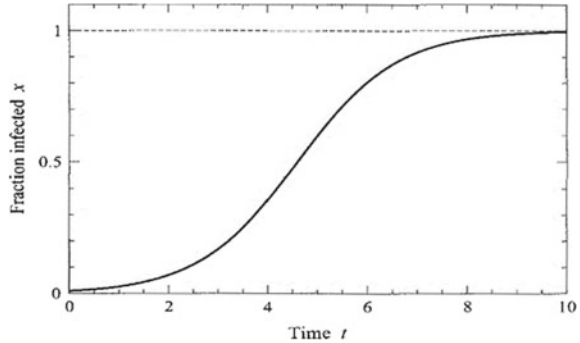
where v' is the fraction of infected nodes, and $(1 - v')$ is the fraction of susceptible nodes. Then Eq. (1) can be solved using the standard method to give Eq. (2).

$$v'(t) = \frac{v'_0 e^{\beta t}}{1 - v'_0 + v'_0 e^{\beta t}} \tag{2}$$

Fig. 1 Representation of SI model



Fig. 2 Logistic growth curve of the SI model



where v'_0 is the value of v' at $t = 0$. Equation (2) gives an S-shaped “logistic growth curve” for a fraction of infected nodes, as shown in Fig. 2. As most of the nodes are initially in a susceptible state, so the curve grows exponentially and then saturates when most of the nodes get infected with rumor, and it becomes difficult to find any susceptible node.

3.2 Betweenness Centrality

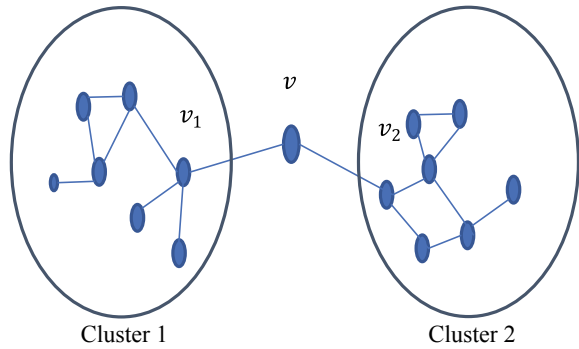
The betweenness centrality (BC) is a measure to count the number of the distinct shortest path passes through a particular node [20, 21]. Equation (3) gives the betweenness centrality of a node v .

$$B(v, G) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

where σ_{st} denotes the number of shortest paths between nodes s and t in the graph G , and $\sigma_{st}(v)$ denotes the number of shortest paths between nodes s , and t that goes through v . The node v in Fig. 3 connected two nodes v_1 and v_2 from cluster 1 and cluster 2, respectively. As all shortest paths from one cluster to the other pass through v , so node v has the highest betweenness centrality than all other nodes in the network.

The importance of node v here is that its removal from the network will partition the entire network into two disconnected clusters. Therefore, node v plays an active role in the entire network in information propagation. Suppose, an attack in the network, initiated from *cluster 1*, putting a firewall at v prevents the *cluster 2* from being infected. However, if the attack initiates from node v , attacks will be propagated in both the clusters. Thus, nodes with high BC can have high importance in the spread of viruses, rumor, and fake news in the network.

Fig. 3 Representation of betweenness centrality



3.3 Rumor Centrality

Let us consider an undirected network $G(V, E)$ where V is the set of nodes, and E is the set of relationships among them. Suppose, the rumor started in the network G at time 0 from the node, say v^* and propagates in the network using the SI model as discussed above. Let us take a snapshot of the graph G at some point of time and consider the connected infected subgraph of G denoted by G_N , and N is the number of nodes in G_N . Rumor centrality of a node v denoted by $R(v, G_N)$ measures the number of distinct ways, and the infected graph G_N can be formed starting from the source node v and is given by Eq. (4) [9].

$$R(v, G_N) = N! \prod_{u \in G_N} \frac{1}{T_u^v} \tag{4}$$

where T_u^v is the number of nodes in the subtree with node u as the root and node v as the source. Notation T_u^v is illustrated with an example shown in Fig. 4. The node with the highest rumor centrality value is known as the rumor center of the graph. Moreover, it is the maximum likelihood (ML) estimator of the original rumor source v^* for regular trees [9].

Fig. 4 Illustration of notation T_u^v

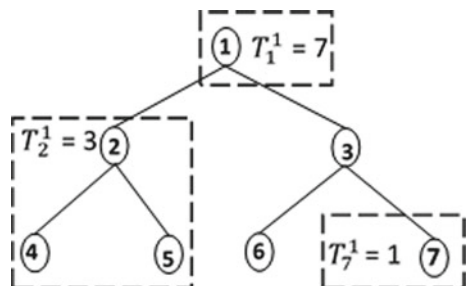
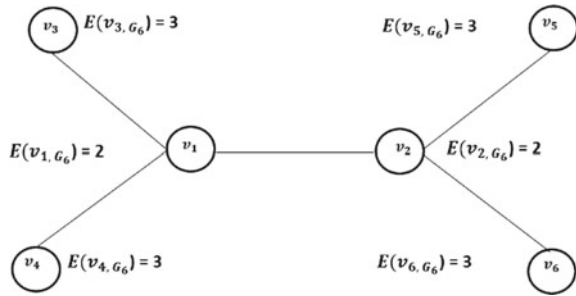


Fig. 5 Eccentricity of the vertices of the graph G_6



3.4 Jordan Centrality

The eccentricity of a vertex v in an infected graph G_N is the distance from v to the farthest vertex u in G_N and is denoted by $E(v, G_N)$, as in Eq. (5) [22]. The distance between two vertices v_i, v_j denoted by $d(v_i, v_j)$ is the number of edges in the shortest path between them. The vertices with minimum eccentricity in G_N are known as Jordan center of G_N denoted by \tilde{j} as in Eq. (6) [23].

$$E(v, G_N) = \max_{v_i \in G_N} d(v, v_i) \tag{5}$$

$$\tilde{j} \in \arg \min_{v \in G_N} E(v, G_N) \tag{6}$$

Let us consider an undirected graph G_6 with six nodes to illustrate Jordan center, as shown in Fig. 5. Nodes v_1 and v_2 in G_6 have minimum eccentricity value than all other nodes in the graph. Thus, v_1 and v_2 are known as Jordan center or Jordan infection center [24] of the graph G_6 .

4 Proposed Approach

More than one node may have a similar centrality score in the graph of a vast complex social network. Thus, an infected graph can have more than one rumor centers or Jordan centers. So, it becomes difficult to select a single source out of many nodes having similar centrality values. We propose to break the tie to a certain level by combining existing centrality measures. We choose to combine betweenness centrality and rumor centrality to minimize the probability of error in rumor source identification. Thus, we named our algorithm as a combined network centrality approach (CNCA).

The proposed CNCA algorithm is given in algorithm 1. We consider an infected graph $G_N(V, E)$ as an input to our algorithm, and it returns an estimated source

denoted as \hat{s} . For all the nodes in graph $G_N(V, E)$, rumor centrality value, betweenness centrality value, and CNCA value are calculated in steps (1–4). Rumor centrality measure, $R(v_i, G_N)$, is calculated using the dynamic message-passing algorithm described in [9] and storing it in a list $\hat{r}[i]$ in (step 2). In step 3, we then calculated betweenness centrality, $B(v_i, G_N)$, of all the nodes in the infected graph using the default functions of the NetworkX library [25] and stored it in a list $\hat{b}[i]$. Later, we combined both the centrality values and stored in a list $\hat{c}[i]$ in (step 4). In (step 6), the highest combined centrality value is calculated from all CNCA values in list \hat{c} . Then the index of the node having the highest CNCA value is returned as the estimated source.

Algorithm 1. Combined Network Centrality Approach (CNCA)

Input: An infected subgraph $G_N(V, E)$

Output: \hat{s} , an estimated source

```

1: for all  $v_i \in G_N$  do
2:    $\hat{r}[i] \leftarrow R(v_i, G_N)$ 
3:    $\hat{b}[i] \leftarrow B(v_i, G_N)$ 
4:    $\hat{c}[i] \leftarrow \hat{r}[i] + \hat{b}[i]$ 
5: end for
6:  $\hat{s} \leftarrow \arg \max(\hat{c})$ 

```

5 Results and Discussions

The identification of the rumor source in the social network is implemented using PyCharm IDE of Python. Initially, a real-world dataset is chosen to create network model $G(V, E)$ based on the adjacency matrix. SI model is used to spread a rumor on the network from a source node selected at random to create an infected graph $G_N(V, E)$. Minimum ten nodes are made infected in every experiment. On the adjacency matrix of the infected graph, all approaches are applied to estimate the rumor source. This rumor source is then compared with the actual source selected at random initially during rumor propagation. For every set of infected nodes, the above process is run for fifty times. Now, the error percentage is calculated by dividing the number of times error occurred by fifty for every set of infected nodes. The number of infected nodes is then incremented to either five or ten depending on the size of the graph, and the entire process is repeated for the new set of nodes. Results obtained specific to each network are discussed in the below sections.

Properties of the selected real-world datasets to create network $G(V, E)$ and simulation parameters used to create infected graph $G_N(V, E)$ are shown in Table 1. In the karate network, source identification using the CNCA algorithm has less errors than all other implemented algorithms. When the number of infected nodes is

Table 1 Properties of graph $G(V, E)$ and simulation parameters in $G_N(V, E)$

Dataset	Properties of $G(V, E)$			Simulation parameters of $G_N(V, E)$		
	No. of nodes	No. of edges	Average degree	Min infected nodes	Max infected nodes	No. of trials
Karate	34	78	4.5882	10	30	50
Dolphin	62	159	5.1290	10	60	50
Power Grid	2152	2824	2.6245	10	120	50
Facebook	4635	55,583	23.9840	10	180	50

15, Jordan centrality performs better, but errors observed to be higher for all other nodes, and the results of the analysis are shown in Fig. 6. In the dolphin network, CNCA has less errors than all other approaches. Error percentage of Jordan centrality touches 100 in almost all cases. Betweenness centrality performed better when the number of infected nodes is fifteen; the results of the analysis are shown in Fig. 7. In the power grid network, source identification using the CNCA algorithm has less errors than all other implemented algorithms except for the number of nodes 40 and 110, which is similar to other approaches. Error percentage of Jordan centrality was observed to be the highest for almost all sets of infected nodes except for the set 40 where it performed better than all others. The results of the analysis are shown in Fig. 8. In the Facebook network, the CNCA algorithm has less errors in almost all the sets of infected nodes. When the number of infected nodes is 15, the rumor centrality performed better than all other implemented algorithms, and the results of the analysis are shown in Fig. 9.

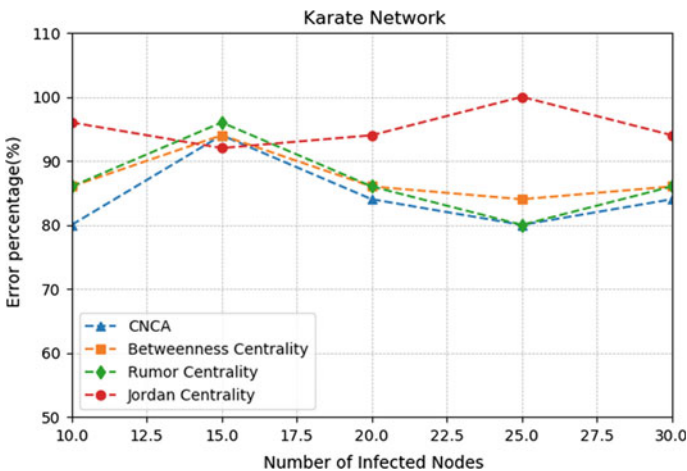


Fig. 6 Results of simulation on karate network

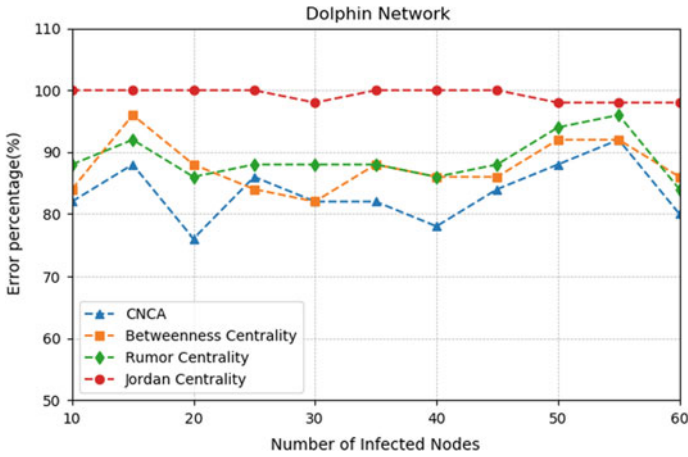


Fig. 7 Results of simulation on dolphin network

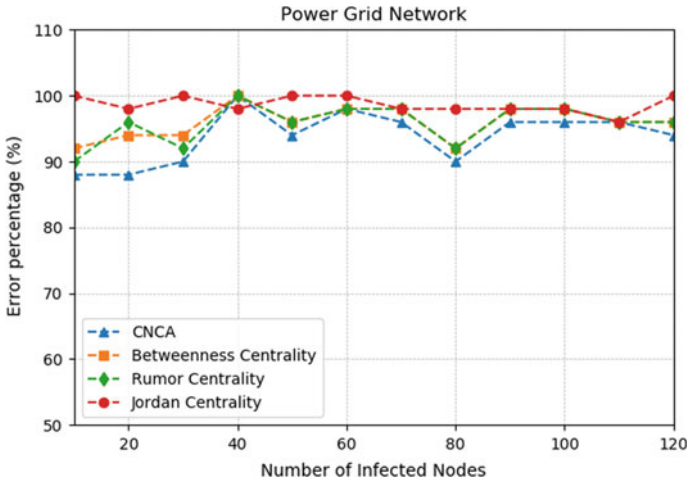


Fig. 8 Results of simulation on the power grid network

6 Conclusion

CNCA combines the rumor centrality and betweenness centrality to give new centrality value. The estimated source chosen is the node having a maximum of CNCA value. The results obtained by the CNCA algorithm are compared with three other existing algorithms, like rumor centrality, betweenness centrality, and Jordan centrality. It is observed that the CNCA algorithm has less errors in almost every experiment performed.

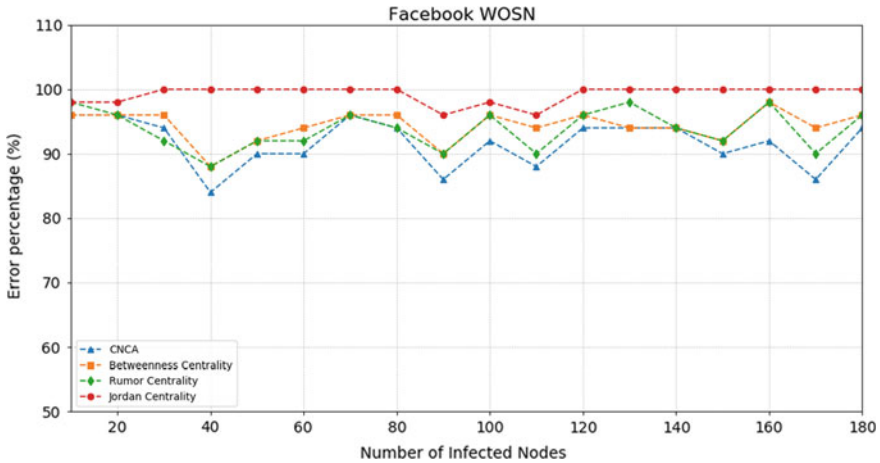


Fig. 9 Results of simulation on Facebook network

CNCA algorithm needs to be evaluated using other information diffusion models and on a large number of infected nodes to check its performance. In this study, CNCA combined only two centrality measures betweenness centrality and rumor centrality so, future work on combinations of other centrality measures can be done for rumor source identification on the social network.

References

- Blank, G., Reisdorf, B. C.: The participatory web: a user perspective on Web 2.0. *Inform. Commun. Soc.* **15**(4), 537–554 (2012)
- Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: a data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint [arXiv:1809.01286](https://arxiv.org/abs/1809.01286) (2018)
- Doerr, B., Fouz, M., Friedrich, T.: Why rumors spread fast in social networks. *Commun. ACM* **55**(6), 70–75 (2012)
- Zhou, X., Zafarani, R.: Fake news: a survey of research, detection methods, and opportunities. arXiv preprint [arXiv:1812.00315](https://arxiv.org/abs/1812.00315) (2018)
- Binham, C.: Companies fear rise of fake news and social media rumour. <https://www.ft.com/content/4241a2f6-e080-11e9-9743-db5a370481bc> (2018). Accessed on 30 Dec 2019
- Govt rejects rumors circulating in social media regarding final Naga settlement issue. <https://ddnews.gov.in/national/govt-rejects-rumours-circulating-social-media-regarding-final-naga-settlement-issue> (2019). Accessed on 30 Nov 2019
- Shelke, S., Attar, V.: Source detection of rumor in social network—a review. *Online Social Networks Media* **9**, 30–42 (2019)
- Jiang, J., Wen, S., Yu, S., Xiang, Y., Zhou, W.: Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Commun. Surv. Tutor.* **19**(1), 465–481 (2016)
- Shah, D., Zaman, T.: Rumors in a network: who’s the culprit? *IEEE Trans. Inf. Theory* **57**(8), 5163–5181 (2011)

10. Pinto, P.C., Thiran, P., Vetterli, M.: Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.* **109**(6), 068702 (2012)
11. Dong, W., Zhang, W., Tan, C. W.: Rooting out the rumor culprit from suspects. In: Proceedings of the IEEE International Symposium on Information Theory, pp. 2671–2675 (2013)
12. Luo, W., Tay, W.P., Leng, M.: Identifying infection sources and regions in large networks. *IEEE Trans. Signal Process.* **61**(11), 2850–2865 (2013)
13. Prakash, B. A., Vreeken, J., Faloutsos, C.: Spotting culprits in epidemics: how many and which ones? In: Proceedings of the IEEE 12th International Conference on Data Mining, pp. 11–20 (2012)
14. Zhu, K., Ying, L.: Information source detection in the SIR model: a sample-path-based approach. *IEEE/ACM Trans. Network.* **24**(1), 408–421 (2014)
15. Shi, C., Zhang, Q., Chu, T.: Source identification of network diffusion processes with partial observations. In: Proceedings of the IEEE 36th Chinese Control Conference, pp. 11296–11300 (2017)
16. Pan, J., Zhang, W.: Identifying rumor sources using dominant eigenvalue of Nonbacktracking matrix. In: Proceedings of the IEEE Global Conference on Signal and Information Processing, pp. 748–752 (2018)
17. Dong, M., Zheng, B., Quoc Viet Hung, N., Su, H., Li, G.: Multiple rumor source detection with graph convolutional networks. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 569–578 (2019)
18. Keeling, M.J., Eames, K.T.: Networks and epidemic models. *J. R. Soc. Interface* **2**(4), 295–307 (2005)
19. Newman, M. E. J.: *Networks An Introduction*. Oxford University Press, Oxford (2010)
20. Zaman, T. R.: Information extraction with network centralities: finding rumor sources, measuring influence, and learning community structure. Doctoral dissertation, Massachusetts Institute of Technology (2011)
21. Barthelemy, M.: Betweenness centrality in large complex networks. *Eur. Phys. J. B* **38**(2), 163–168 (2004)
22. Deo, N.: *Graph theory with applications to engineering and computer science*. Courier Dover Publications (2017)
23. Wasserman, S., Faust, K.: *Social network analysis: Methods and applications*, vol. 8. Cambridge University Press, Cambridge (1994)
24. Zhu, K., Ying, L.: A robust information source estimator with sparse observations. *Comput. Soc. Netw* **1**(3), 1–21 (2014)
25. Hagberg, A., Swart, P., Schult, D.: Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference, pp. 11–15. (2008)

Sentiment Polarity Detection on Bengali Book Reviews Using Multinomial Naïve Bayes



Eftekhar Hossain, Omar Sharif, and Mohammed Moshuiul Hoque

Abstract Recently, sentiment polarity detection has increased attention to NLP researchers due to the massive availability of customer's opinions or reviews in the online platform. Due to the continued expansion of e-commerce sites, the rate of purchase of various products, including books, is growing enormously among the people. Reader's opinions/reviews affect the buying decision of a customer in most cases. This work introduces a machine learning-based technique to determine sentiment polarities (either positive or negative category) from Bengali book reviews. To assess the effectiveness of the proposed technique, a corpus with 2000 reviews on Bengali books is developed. A comparative analysis with various approaches (such as logistic regression, naive Bayes, SVM, and SGD) also performed by taking into consideration of the unigram, bigram, and trigram features, respectively. Experimental result reveals that the multinomial naive Bayes with unigram feature outperforms the other techniques with 84% accuracy on the test set.

Keywords Bangla language processing · Sentiment polarity detection · Feature extraction · Book reviews · Machine learning

E. Hossain

Department of Electronics and Telecommunication Engineering, Chittagong University of Engineering and Technology, Chittagong 4349, Bangladesh
e-mail: eftekhar.hossain@cuet.ac.bd

O. Sharif · M. Moshuiul Hoque (✉)

Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong 4349, Bangladesh
e-mail: moshiul_240@cuet.ac.bd

O. Sharif

e-mail: omar.sharif@cuet.ac.bd

1 Introduction

Automatic detection of sentiment polarity is one of the notable research issues in opinion mining and natural language processing. The number of opinions or reviews on social media, blogs, and online platforms are growing enormously due to the substantial growth of Internet uses and its uncomplicated access via e-devices. Online marketers facilitated their purchaser to convey feeling about the procured items or services to boost up purchaser contentment. The new purchasers prefer to read earlier posted product reviews, and they can make their decision to buy a particular item based on these reviews. Sentiment detection is a technique to estimate the point of view of a user on a specific matter. It categorizes the polarity of the text content (in the form of tweets, reviews, comments, posts, or bulletins), as the positive, neutral, or negative [12].

Recently, the purchasing habits of people are increased exponentially via online/e-commerce sites, and the book is one of the most selling products online. The amenities of the online book platform is that both the customers and authors can get a sight of how readers and reviewers responded to a particular book. Thus, book reviews may play an essential role in creating attraction or aversion on a particular book and help them in deciding to purchase this book. It may become a common scenario among the customers to read the reviews before buying a book. A book with an abundant amount of positive reviews can succeed in gaining the customer's attention and faith towards the publishers as well as authors. Nevertheless, it is a very complicated and time-consuming task to scrutinize every review manually from a considerable amount of reviews. Thus, this time-consuming task puts concealment for an automated method that preserve comprehend the contextual polarity of reviewer's annotations scripted at various social media groups as well as online book shops. An automatic sentiment polarity detection technique can use for better decision making, efficient manipulation, and understating customer feelings/opinion on a particular book. Sentiment analysis is associated with labeling a specific sentence with positive and negative annotation. The task of sentiment analysis is momentous to the business community, where user reviews need to take account to know the sustainability of their product and to identify the general feel of the consumers towards that product.

The sentiment polarity detection broadly classifies as sentiment knowledge-based and statistical-based [5]. In the sentiment knowledge-based approach, a sentiment dictionary is used to execute specific computation of words that bear sentiment in the content and figure out the exploration of sentiments from text[1]. On the other hand, the statistically-based approach picks a text analysis technique to extract appropriate linguistic features of the text and utilizes a machine learning technique to consider the sentiment analysis as a classification problem [20]. In our work, we use the statistical-based approach. As far as we concern, there is no study has been conducted to date on sentiment analysis from book reviews in the Bengali language. To address this issue, our contributions to this work can remark in the following:

- Develop a corpus consisting of 2000 Bengali book reviews, which are labeled as positive and negative sentiments.

- Develop a supervised machine learning model using naive Bayes technique to categorize the sentiment from book reviews into positive or negative polarities.
- Analyze the effect of several n -gram features on naive Bayes, logistic regression, SVM, and SGD algorithms by using develop dataset.
- Analyze the effectiveness of the proposed technique on different distributions of the developed dataset.
- Assess the outcome of the proposed technique by comparing with the other machine learning techniques on the developed dataset.

The remaining of the paper is arranged following: Sect. 2 represents related work. A brief description of the developed corpus is outlined in Sect. 3. Section 4 provides the suggested framework with a detailed explanation of its constituent parts. Section 5 states the experimentation with the analysis of findings. Finally, the paper is concluded with a summary in Sect. 6.

2 Related Work

Sentiment analysis or detection from text is a well-studied research issue for highly resourced languages like English, Arabic, and other European languages [10]. Srujan et al. [15] presented a random forest technique to analyze sentiment of amazon book reviews. They used TF-IDF to extract n -gram features and obtained the accuracy of about 90.15%. Akshay et al. [9] explored machine learning classifiers for analyzing sentiment of the restaurant reviews. They have obtained the highest accuracy of 94.5% for their dataset. A positive and negative sentiments detection model is developed on cell phone reviews using SVM, which achieved an accuracy of 81.77% [14]. Fang et al. [4] presented a model for analyzing sentiment on online product reviews where a sentiment score formula along with three classifiers was used for the categorization of text polarity. Chinese text-based sentiment analysis is developed using naive Bayes, where a sentiment dictionary is used for classification [18]. A classifier model is proposed in the Arabic language for grouping reviews of social networks [7]. A deep learning-based emotion detection technique is developed by Xu et al. [19] on the medical/health dataset, which achieved 85% accuracy on emotional fatigue.

Since Bengali is an under-resourced language, the amount of e-text contents or reviews in the Bengali language (primarily related to books) is quite limited. In addition to that, no benchmark dataset is available on sentiment classification of book reviews in the Bengali language. Due to these barriers, very few research activities have been conducted in this area of Bangla language processing (BLP), which are mainly related to sentiment detection from news, restaurant reviews, product reviews, social media reviews, micro-blogging comments, and so on. Rumman et al. [2] proposed a sentiment classifier for movie reviews written in Bengali text. They have explored different machine learning techniques over a small review dataset. Their dataset shows excellent performance for SVM and LSTM model with an accuracy of 88.90% and 82.42%, respectively. An SVM-based sentiment analysis on the Ban-

glaseh Cricket dataset is developed, which achieved 64.59% accuracy [11]. Sarkar et al. [13] presented a sentiment classification system on the Bengali tweet dataset, where SVM and multinomial naive Bayes classifiers used for the classification task. Their system achieved the highest 45% accuracy for SVM classifier over n -gram and SentiWordNet features. A recent technique to analyze sentiment from the Bengali text was proposed by Taher et al. [16], where various n -gram features incorporated with the SVM classifier. They have obtained maximum accuracy of 91.68% for linear SVM. A supervised model proposed to determine the positive and negative sentiments from Facebook status written in Bengali [8]. This system achieved 0.72 f -score using a multinomial naive Bayes classifier with bigram features. Another model of sentiment analysis on the Bengali horoscope corpus is proposed based on SVM, which achieved 98.7% accuracy using unigram features [6]. It observed that most of the research activities have been conducted so far, considered a small dataset of Bengali sentiment analysis. Furthermore, these considered movie reviews and social blog comments for sentiment analysis. As far as we are aware that none of the work has done yet on sentiment analysis on Bengali book reviews. This work proposes a technique for detecting positive and negative sentiments from online book reviews written in Bengali.

3 Dataset Preparation

Bengali is one of the resource-poor languages due to its insufficient e-contents and unavailability of the standard dataset. To serve our purpose, we developed our dataset on sentiment polarity by collecting data from the available web resources such as blogs, Facebook, and e-commerce sites. We endorsed a technique of developing a dataset, as explained by Das et al. [3].

Table 1 shows a negative and positive review as examples. The developed dataset contained 2000 textbook reviews, and these reviews randomly divided in 03 distributions: training set (T_R), validation set (V_D), and test set (T_S), respectively. Table 2 exhibits the summary of the developed data set.

Table 1 Sample dataset

Sample reviews	Sentiment
অতি অসাধারণ একটা ডার্ক হরর। এক নিঃশ্বাসে পড়ে শেষ করার মতো বই। খুবই ভালো। (An extraordinary dark horror. This is a breath taking book. That's good)	Positive
এটা আসলে বই ছিল না এক ধরনের বাজে রসিকতা তা বোধহয় সমরেশ মজুমদার বলতে পারবেন। অখাদ্য। পুরোই মেজাজটাই খারাপ হয়ে গেল। (Only shomresh majumder can tell, was it really a book or a kind of crap joke. Disgusting. The whole mood went bad.)	Negative

Table 2 Summary of the dataset

Dataset attributes	T_R	V_D	T_S
Number of documents	1600	200	200
Number of words	29,079	4458	5234
Total unique words	8336	1193	1518
Size (in bytes)	974,848	79,872	70,656
Number of sentences	6728	1137	876

4 Sentiment Polarity Detection Framework

Figure 1 depicts the suggested technique for sentiment polarity detection/classification. This framework composes of three major phases: pre-processing, feature extraction, and classification, respectively.

4.1 Preprocessing

Text reviews $r^{(i)}$ of the corpus $\mathfrak{R}[]$ can processes in several steps. To illustrate the pre-processing steps, a sample review $r =$ “ ইহা একটি অসাধারণ বই !! ...!! ” have selected from the corpus.

1. **Removal of redundant characters:** punctuation symbols, special character,s and numbers are removed from the collected reviews. These are considered redundant as they do not bear any sentiment information. After removing the redundant characters, the sample review becomes $r =$ [ইহা একটি অসাধারণ বই].

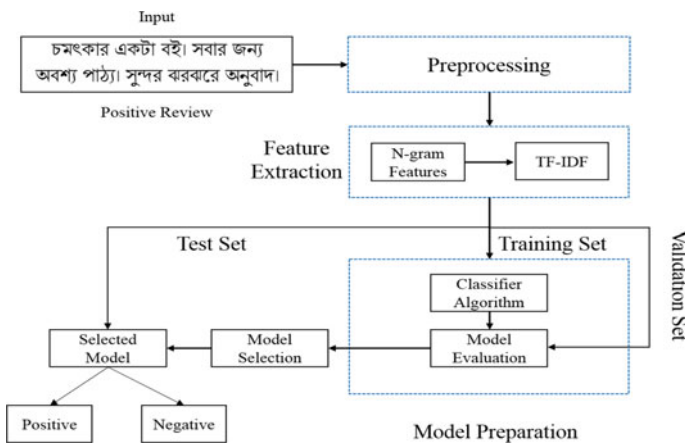


Fig. 1 Proposed framework for sentiment polarity detection

2. **Tokenization:** is the process of partitioning a review into its constituent words. We get a vector of words $r = \{w^{<1>}, w^{<2>}, \dots, w^{<l>}\}$ by tokenizing a review r of l words. $r = [\text{'ইহা'}, \text{'একটি'}, \text{'অসাধারণ'}, \text{'বই'}]$.
3. **Stop words removal:** A word $w^{(i)}$ has been removed from a review r by matching it with a stop word list $S[] = \{s_1, s_2, \dots, s_t\}$ having t stop words. A word $w^{(i)}$ which did not contribute to fix whether a review is positive or negative has been considered as a stop word such as conjunction, pronoun, and preposition. At this step the sample review becomes $r = [\text{'অসাধারণ বই'}]$.

A clean corpus has been developed by performing all preprocessing steps on reviews.

4.2 Feature Extraction Techniques

The feature extraction process determines the degree of success of any machine learning-based system. N-gram features of the text have experimented with the proposed approach. Table 3 shows unigram, bigram, and trigram features for a sample review (লেখকের উপস্থপনা বেশ চমৎকার) with total system features.

A common way of extracting “*tfidf*” feature-value from a text is computed by Eq. 1 [17].

$$tfidf(w^{(i)}, r^{(i)}) = tf(w^{(i)}, r^{(i)}) \log \frac{N}{|\{r \in \mathfrak{R} : w \in r\}|} \quad (1)$$

where $tfidf(w, r)$ represents the value of word $w^{(i)}$ in review $r^{(i)}$, $tf(w, r)$ means the frequency of $w^{(i)}$ in review $r^{(i)}$, N represents the total number of reviews, $|\{r \in \mathfrak{R} : w \in r\}|$ indicates the number of reviews containing $w^{(i)}$.

The combination of n -gram features and *tfidf* values was used as the features of the proposed system. Table 4 represents a small fragment of *tf – idf* values for n -gram features for some arbitrary reviews.

Features are represented in a two-dimensional space where reviews $\mathfrak{R} = r^{(1)}, r^{(2)}, \dots, r^{(n)}$ in rows and n -gram features are represented in columns. Columns are divided into three parts, and these are (i) Uni-gram feature set $U = \{w_1, w_2, \dots, w_u\}$ (ii) Bi-gram feature set $B = \{w_1, w_2, \dots, w_b\}$ (iii) Tri-gram feature set $T = \{w_1, w_2, \dots, w_t\}$. The number in each cell holds the *tfidf* value of word w_i in the feature set $\{U, B, T\}$ for a review r_i . Each row corresponds to the feature vector $F = F_V[1], F_V[2], F_V[3], \dots, F_V[n]$ for reviews $\mathfrak{R} = r^{(1)}, r^{(2)}, r^{(3)}, \dots, r^{(n)}$ of the corpus.

Table 3 Illustration of various N-grams features

Sample Text	লেখকের উপস্থপনা বেশ চমৎকার	Total Features
Uni-gram feature	'লেখকের', 'উপস্থপনা', 'বেশ', 'চমৎকার'	2364
Bi-gram feature	'লেখকের উপস্থপনা', 'উপস্থপনা বেশ', 'বেশ চমৎকার'	14,612
Tri-gram feature	'লেখকের উপস্থপনা বেশ', 'উপস্থপনা বেশ চমৎকার'	19,588

Table 4 A small fragment of feature space

\mathfrak{R}, C	Uni-gram			Bi-gram			Tri-gram		
	w_1	...	w_u	w_1	...	w_b	w_1	...	w_t
$r^{(1)}$	0.45	...	0.231	0.31	..	0.431	0.57	...	0.35
$r^{(2)}$	0.25	...	0.51	0.24	...	0.253	0.53	...	0.57
$r^{(3)}$	0.63	...	0.09	0.13	...	0.356	0.20	...	0.46
$r^{(4)}$	0.17	...	0.12	0.65	...	0.134	0.036	...	0.35
$r^{(5)}$	0.29	...	0.32	0.37	...	0.541	0.076	...	0.35

Classifier Model Preparation The proposed work aims to develop a sentiment classifier to categorize a book review either in positive or negative category. N -gram features with $tf - idf$ (described in Sect. 4.2) mainly used for the model preparation. The extracted features segmented into three random distributions, and each distribution applied to a particular stage of the model preparation, such as classification and model evaluation. **Classification** In this stage, a set of classifier models $M[] = \{m_1, m_2, \dots, m_7\}$ has been developed by applying different learning algorithms on the training set $T_R = \{r^{(1)}, r^{(2)}, \dots, r^{(x)}\}$. These algorithms are logistic regression (LR), decision tree (DT), random forest (RF), multinomial naive Bayes (MNB), KNN, SVM, and stochastic gradient descent (SGD), respectively.

4.2.1 Model Evaluation

For the purpose of evaluation, tenfold cross-validation has been performed on each of the model $M[i]$. Among these models, the best four (LR, SVM, MNB, and SGD) will be chosen based on their cross-validation accuracy, which is done by algorithm 1. These models will be evaluated on different parameters by using the validation set $V_D = \{r^{(1)}, r^{(2)}, \dots, r^{(y)}\}$ to find out the desired one. Finally, test set $T_S = \{r^{(1)}, r^{(2)}, \dots, r^{(z)}\}$ will be used to assess the selected model.

Algorithm 1 Finding best 4 models based on cross-validation scores.

```

1: Initialize  $CV_m \leftarrow [], i \leftarrow 0$ ;    " $CV_m =$  Set of cross-validates model"
2: Initialize  $CV_s \leftarrow [s_1, s_2, \dots, s_7]$ ;    " $CV_s =$  List of cross-validation score"
3: Sort( $CV_s, CV_s + 7$ )
4:  $CV_m.append(CV_s[i])$ 
5: if  $i == 4$  then
6:   exit
7: else
8:    $i \leftarrow i + 1$ 
9:   go to step 4
10: end if

```

5 Results Analysis

The performance assessment of the proposed technique is performed by using several graphical and statistical measures such as confusion matrix, f_1 score, recall, precision, ROC, and precision versus recall curve. For the development of the sentiment classification model initially, seven classifiers have been selected for the training. Tenfold cross-validation has been done over all the classifiers using n -gram features. The trained classifiers are LR, MNB, RF, DT, KNN, SVM, and SGD. Among these seven trained classifier models, only four classifiers provide acceptable cross-validation accuracy and thus selected them for further evaluation over validation data. Table 5 shows the tenfold cross-validation accuracy over the n -gram features for all the classifiers. The result indicates that that KNN, DT, and RF algorithms achieved the lower accuracy for all cases of n -gram features.

Selected four classifier models assessed by using the validation dataset. Table 6 illustrates the performance of four classifiers in terms of accuracy, precision, recall, and f_1 measures. It observed that in the case of the uni-gram feature, MNB achieved the highest accuracy of about 87%. For bi-gram features, both LR and SGD achieved the highest accuracy of about 80%, whereas MNB provided the lowest accuracy (74%). On the contrary, for tri-gram features, all four classifiers provided the lowest accuracy (only 70%) than the bi-gram and tri-gram features. Although the number of features increased in tri-gram, the limited number of reviews in each class suppressed the chances of occurring features in several reviews. As a result, the overall accuracy decreased. The result of the analysis shows that multinomial naive Bayes for unigram features outperformed the other classifiers and features extraction methods.

For all classifiers, the experiment performed again for graphical analysis. Figures 2, 3, and 4 show ROC and PR curve for the selected four classifier models with unigram, bigram, and trigram features, respectively. In the case of unigram features, MNB provides the highest AUC at 89.6% and average precision at 91.2%. On the other hand, for bigram features, all the classifiers jointly give good AUC and AP score for ROC and PR curves, respectively.

Table 5 Tenfold cross-validation results

	Classifier	Uni-gram	Bi-gram	Tri-gram
Accuracy(%)	LR	83	77	66
	KNN	58	54	60
	DT	69	77	59
	RF	73	77	68
	MNB	88	78	69
	SVM	81	77	63
	SGD	77	72	68

Table 6 Performance measures of classifiers

Features	Classifier	Accuracy (%)	Precision (%)	Recall (%)	f_1 score (%)
Uni-gram	LR	81	81	81	81
	MNB	87	89	86	86
	SVM	81	81	81	81
	SGD	78	82	76	76
Bi-gram	LR	80	81	81	80
	MNB	74	84	71	70
	SVM	76	81	78	76
	SGD	80	84	82	80
Tri-gram	LR	61	60	95	73
	MNB	64	61	95	75
	SVM	68	89	49	63
	SGD	60	58	95	73

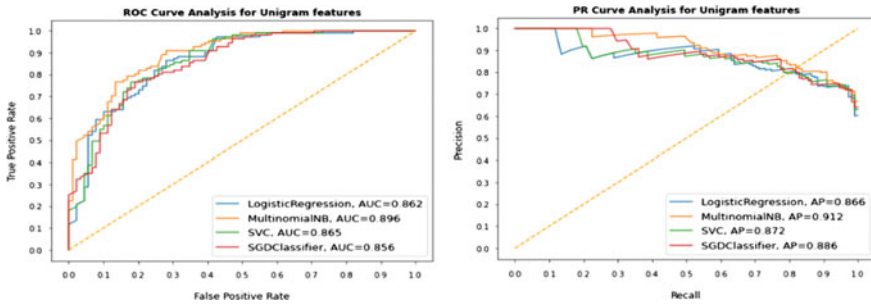


Fig. 2 Classifiers performance on Uni-gram features

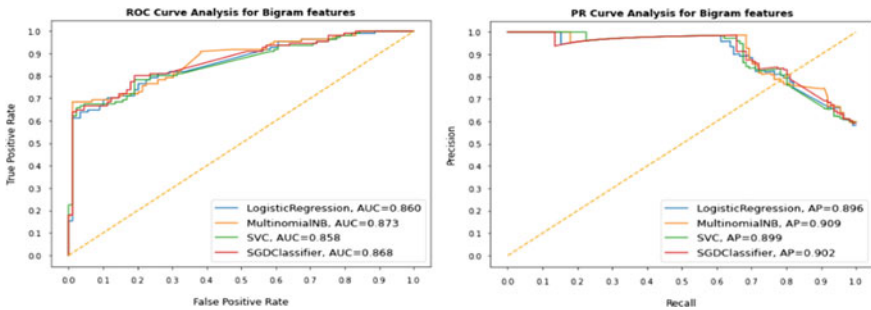


Fig. 3 Classifiers performance on Bi-gram features

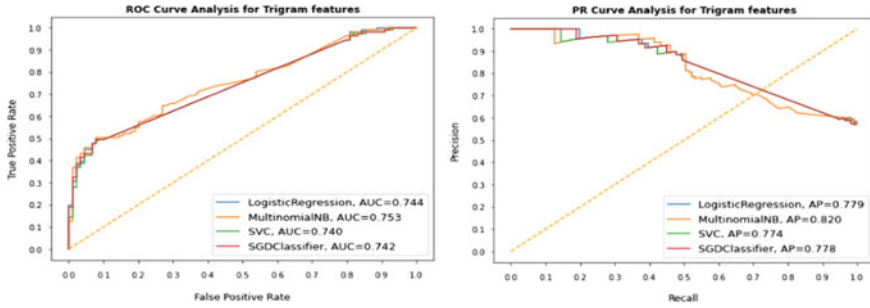


Fig. 4 Classifiers performance on Tri-gram features

Table 7 Summary of evaluation of the proposed technique

Sentiment polarity (C)	Precision	Recall	f_1 -score	Support
Negative (C_n)	0.91	0.72	0.81	89
Positive (C_p)	0.81	0.95	0.87	111
avg./total	0.86	0.83	0.84	200

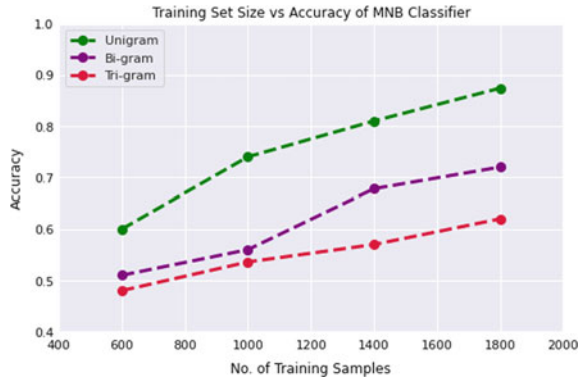
In the case of trigram features, the AUC and AP scores show the lowest score for every classifier model. The low AUC and AP scores reported by the classifiers imposed the fact that in trigram features. Many of the strong sentiment carrying words and phrases misrepeated in most of the reviews.

Detailed analysis of the results brings the notation that the MNB with unigrams features gives the highest accuracy, AUC, and AP scores in comparison with all the other classifier models and gram features. Thus, a multinomial naive Bayes classifier selected as the final model. Table 7 depicts the detailed classification reports of the MNB on the test set.

From the classification report, it observed that the value of f_1 -score is 0.81 and 0.91, respectively, for positive and negative sentiment class. From the analysis of the results, it concluded that the proposed model successfully classifies 91% of negative, and 81% of positive book reviews and achieved 84% overall accuracy on the test set.

The overall classification performance of the proposed system is oscillating between 60 and 85%. It is due to the lack amount of training samples, and it can be improved further by including more data on the training phase. Figure 5 shows the impact of the number of training samples on the model performance. This figure reveals that the accuracy increases with the increase in training dataset for unigram features.

Fig. 5 Effect of classifier performance on training sets



6 Conclusion

This paper presents a machine learning-based sentiment classification framework that can determine the sentiment into positive and negative categories from the Bengali book reviews by exploiting the various feature extraction techniques. The combination of $tf - idf$ values with n -gram features considered as the best feature of the proposed framework. These extracted features are applied to classify the inherent sentiment of the Bangla text reviews using various ML techniques. Among these classifiers, multinomial naive Bayes with unigram features provided the highest accuracy of 87 and 84% for validation and test datasets, respectively. Analyzing the sentiment of a book is very helpful for the authors and publishers as it can aid them by providing an abstract that what readers think about a book. In the future, deep learning techniques and sophisticated feature extraction methods such as word2Vec or Glove may apply for improved accuracy.

References

1. Cambria, E., Grassi, M., Hussain, A., Havasi, C.: Sentic computing for social media mark theory. *Multimedia Tools Appl.* **59**(2), 557–577 (2019)
2. Chowdhury, R.R., Hossain, M.S., Hossain, S., Andersson, K.: Analyzing sentiment of movie reviews in bangla by applying machine learning techniques. In: *International Conference on Bangla Speech and Language Processing* (2019)
3. Dash, N.S., Ramamoorthy, Naicker, L.: *Utility & Application of Language Corpora*, pp. 17–34 (2019)
4. Fang, X., Zhan, J.: Sentiment analysis using product review data. *J. Big Data* **2**(1), 5 (2015)
5. Feng, Z.: Hot news mining and public opinion guidance analysis based on sentiment computing in network social media. *Personal Ubiquitous Comput.* **23**, 373–381 (2019)
6. Ghosal, T., Das, S.K., Bhattacharjee, S.: Sentiment analysis on (Bengali horoscope) corpus. In: *2015 Annual IEEE India Conference (INDICON)*. pp. 1–6. IEEE (2015)

7. Hammad, M., Al-awadi, M.: Sentiment analysis for Arabic reviews in social networks using machine learning. In: *Information Technology: New Generations*, pp. 131–139. Springer, Berlin (2016)
8. Islam, M.S., Islam, M.A., Hossain, M.A., Dey, J.J.: Supervised approach of sentimentality extraction from Bengali Facebook status. In: *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pp. 383–387. IEEE (2016)
9. Krishna, A., Akhilesh, V., Aich, A., Hegde, C.: Sentiment analysis of restaurant reviews using machine learning techniques. In: *Emerging Research in Electronics, Computer Science and Technology*, pp. 687–696. Springer, Berlin (2019)
10. Le, H.S., Van Le, T., Pham, T.V.: Aspect analysis for opinion mining of Vietnamese text. In: *2015 International Conference on Advanced Computing and Applications (ACOMP)*. pp. 118–123. IEEE (2015)
11. Mahtab, S.A., Islam, N., Rahaman, M.M.: Sentiment analysis on bangladesh cricket with support vector machine. In: *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–4. IEEE (2018)
12. Rahman, A., Dey, E.K.: Datasets for aspect-based sentiment analysis in Bangla and its baseline evaluation. *Data* **3**(2), 1–11 (2018)
13. Sarkar, K., Bhowmick, M.: Sentiment polarity detection in Bengali tweets using multinomial naïve bayes and support vector machines. In: *2017 IEEE Calcutta Conference (CALCON)*. pp. 31–36. IEEE (2017)
14. Singla, Z., Randhawa, S., Jain, S.: Sentiment analysis of customer product reviews using machine learning. In: *2017 International Conference on Intelligent Computing and Control (I2C2)*, pp. 1–5. IEEE (2017)
15. Srujan, K., Nikhil, S., Rao, H.R., Karthik, K., Harish, B., Kumar, H.K.: Classification of Amazon book reviews based on sentiment analysis. In: *Information Systems Design and Intelligent Applications*, pp. 401–411. Springer, Berlin (2018)
16. Taher, S.A., Akhter, K.A., Hasan, K.A.: N-gram based sentiment mining for Bangla text using support vector machine. In: *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–5. IEEE (2018)
17. Wu, S., Jia, J.K.: Feature extraction based on improved feature weighting algorithm. In: *Recent Trends in Decision Science and Management*, pp. 703–708. Springer, Berlin (2020)
18. Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y., Wu, X.: Chinese text sentiment analysis based on extended sentiment dictionary. *IEEE Access* **7**, 43749–43762 (2019)
19. Xu, J., Hu, Z., Zou, J., Bi, A.: Intelligent emotion detection method based on deep learning in medical and health data. *IEEE Access* **8**, 3802–3811 (2020)
20. Zhou, G., Zhu, Z., Hu, X.T.: Cross-lingual sentiment classification with stacked autoencoders. *Knowl. Info. Syst.* **47**(1), 27–44 (2016)

Real-Time Facial Emotions Analysis in Videos



Babajee Phavish, Suddul Geerish, Armoogum Sandhya, and Foogooa Ravi

Abstract Deciphering human facial expressions is an integral part of achieving a seamless human to machine communication. This may assist in various cognitive tasks and convey important information such as emotions, intentions and opinions in a medium where machines can understand beyond the traditional binary form of communication. In this paper, we present a significant progress in this direction. We adapted a machine learning approach based on a neural network to help recognize human facial expressions in real time. With a core revolving around a multi-layered convolutional network responsible to train a model, we were able to successfully detect frontal faces from a video stream and encode, in real time, the seven emotions of joy, disgust, neutral, anger, fear, sadness and surprise.

Keywords Classification · Real time · Emotions · Facial expressions recognition (FER) · Videos

1 Introduction

Face to face communication is a process which involves real-time image processing occurring at a very high rate. Our human vision system can detect those emotions almost impeccably as the average human can accurately guess up to around 150 frames per second (FPS). The uncertainty level at this time frame is considerable. The concept of machines capable of handling face to face communication on this basis demands robust real-time perceptive capabilities. We present significant progress in the development of this type of system, one that can autonomously identify faces from a video stream and code the selected facial expression dynamics simultaneously. We present a purely contactless method to analyze and decode facial features with the help of a webcam or video device which allows emotion recognition of the user by the machine. This further serves to accomplish the goal of joint action which is going on between the artificial system and the human. “Joint action should be improved on

B. Phavish (✉) · S. Geerish · A. Sandhya · F. Ravi
University of Technology, Mauritius, La Tour Koenig, Pointe-aux-Sables, Republic of Mauritius
e-mail: p.babajee@umail.utm.ac.mu

the side of the artificial system by adapting emotionally to the human for better use of data being shared in both ways” [1].

In this research, we present the encouraging outcomes of a completely user-independent system for real-time recognition of basic human expressions irrespective of age, gender or genetic heritage. The system automatically detects frontal faces in the video stream and codes each frame as one of the seven basic human emotions, namely neutrality, anger, disgust, fear, joy, sadness and surprise [2]. Our work differs from previous research in that it is fully automated with simultaneous lossless processing and result visualization. This further optimizes processing time, a key element for real-time applications. Explicit detection and alignment of internal facial features to capture faces for processing have been made redundant in our proposed system which represents a significant saving in processing time and enhances the performance of real-time applications.

2 Related Work

A traditional facial expressions recognition (FER) system constitutes of three key steps: facial detection, feature extraction, and classification of the facial expression [3]. Real-time recognition solutions employ techniques like Viola–Jones algorithms [12], motion history image [4], skin color model [4], support vector machine, SVM [5], neural network or active appearance model (AAM) [6]. Most of the algorithms for facial detections depend on the features on the face to locate the global position of the face and features on the face [7]. Face detection refers to the preprocessing stage where face regions are located [8]. Facial feature extraction aims to find the most fitting representation of facial images. Performance of real-time applications is critical. Thus, we adopted an experimental approach with a phase by phase improvement on the prototype of our system stemming from a previously published system [2] which mainly consisted of emotional analysis on static local images. Our approach is based on a geometric feature-based method which extracts both the shapes and the locations of facial components information using an edge detection framework [9]. For the facial expression classification, a probabilistic classifier was used to identify different expressions relative to the extracted facial features. A Convolutional neural network is used, with the neurons in the output later defined by the number of facial expressions targeted by our scope. The results are, subsequently, evaluated using the accuracy rates of similar past work in the area of facial emotion recognition. We integrated the video processing library of Keras [10] to individually splice and process each frame. The classification is done without any delay, with the result displayed simultaneously.

3 Proposed Approach

Figure 1 represents the experimental approach of our system. It consists firstly of training a model based on a dataset of images containing a mixed set of facial expressions [11]. The images require preprocessing, including face detection, followed by feature extraction. Once the model is trained, it can be used for making predictions based on a new input facial image. The preprocessing portion mainly comprises a cascade of classifiers, each containing a subset of filters similar to Haar basis functions. For the second part of our approach, live video output is to be processed to use the trained model. We opted to go forward with our image-based training to leap to video processing. The preprocessing stage, where we are employing the Viola–Jones algorithm [12], will split the frames of the video being streamed in and treat the frames the same way as it would to a static image. We are well aware that the testing of the system would be problematic with blurry or out-of-focus image frames, however, the Viola–Jones algorithm only allowed frames with recognized and localized faces to go through the system.

3.1 Face Detection and Feature Extraction

The first step processing of an image is to extract the useful facial areas of a human being that will facilitate the feature extraction process and the classification process. The preprocessing is a twofold approach using the Viola–Jones algorithm. An image frame is first cropped using a Haar feature selection [13] to choose only the face area and to eliminate unimportant surrounding data. The cropped image is then converted to gray scale. The next step after detecting a face from a given image is the feature extraction process that helps isolate the important facial areas. Two methods can be used for feature extraction: analytic approach and holistic approach [14]. The holistic approach uses a raw facial image as input, while the analytic approach focuses on some of the important facial features detected and extracted from the face. As illustrated in Fig. 2, the analytic approach is being used in our research, where

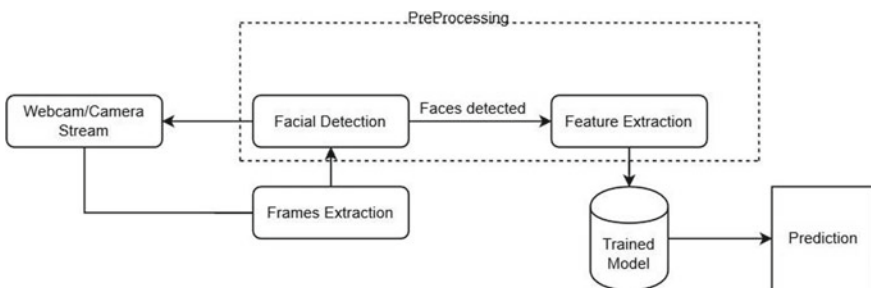


Fig. 1 System architecture

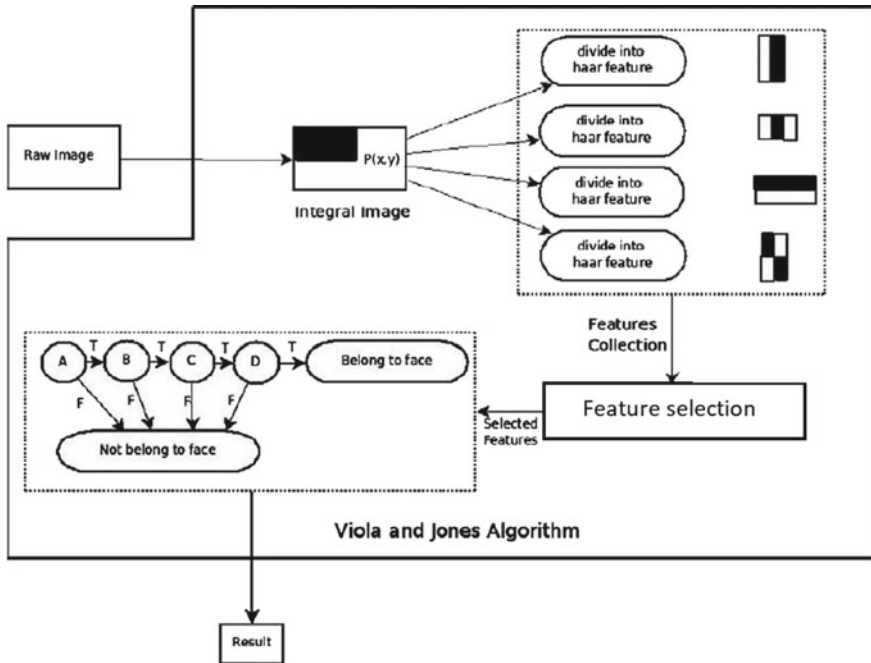


Fig. 2 Viola-Jones algorithm [15]

we extract selected features that are obtained using edge detection from the image [15]. The result is then fed to our model.

3.2 Video Processing

Our approach to video classification is more than just a simple image classification. It consists of classifying one video frame at a time to simply reduce the video classification problem to an image classification problem. The videos are processed in such a way that each instance of the video frames is split into separate frames. This processing is however strictly required to select frames containing data useful in our context. An issue is solved by the first part of our preprocessor, omitting frames without faces detected. All frames have very similar conditions.

3.3 Dataset

We trained and tested our system on the FER2013 dataset [11] as it is among the biggest open-sourced datasets for emotion recognition with a total of 32,398 individually trained images. The FER 2013 dataset is preprocessed, consisting of cropped grayscale images in the size of 48 by 48 pixels. The dataset takes the form of a CSV file with three columns. The first two represent “emotion” and “pixels” and the last its usage. The “emotion” column contains a numeric code from 0 to 6 representing the emotion emitted in the image which is represented below.

Each of the seven emotions is categorized with a unique key [0 to 6].

- 0 Anger: 4593 images
- 1 Fear: 5121 images
- 2 Sadness: 6077 images
- 3 Disgust: 547 images
- 4 Joy: 8989 images
- 5 Neutrality: 6198 images
- 6 Surprise: 4002 images.

3.4 Training Process

For our study, we used 70% of the FER2013 dataset as input for the training phase, to which feature extraction was applied. Key attributes of the image are computed and stored as feature vectors which represent the essential properties detected in the faces. This preprocessing step allows for reduced data sizes as just a handful of important features from an entire image are selected and more concise information is obtained from feature selection. The classification rate is traditionally determined in the testing phase. The testing and training phases both follow the same steps of feature extraction and classification. Classification is however different for the testing phase as the features are tested against the model constructed in the training phase. The outcome of this step yields a prediction score which indicates the emotion predicted by the model.

3.5 Experiment and Result

The model was trained and tested (2) with a 79.8% accuracy. Using a Core i7 7700HQ system with an Nvidia GTX 1050TI driver and 16 GB of memory, the test was solely conducted on static image files on the previous iteration. For this paper, we extended the testing to video using the laptop’s inbuilt 5mps camera, as well as a Logitech C920 PRO HD Webcam which provided 1080p video stream at 30fps. Both results,

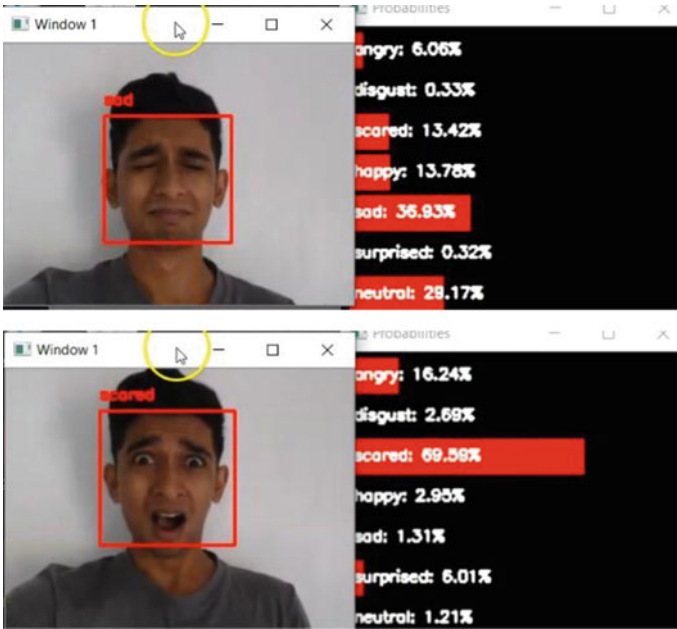


Fig. 3 Results

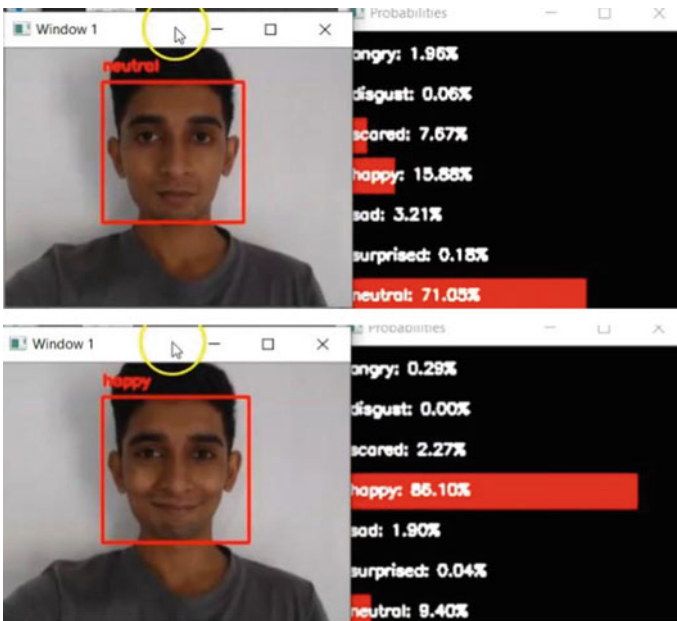


Fig. 4 Results

displayed in Figs. 3 and 4, were extremely encouraging with a seamless smooth analysis of all emotion classes and illustration of the prediction result in real time.

4 Conclusion

The aim of this paper was to assess the real-time application and efficiency of facial emotion recognition on a video stream. We can concur that the promising results show that this form of application is ready to be used and adapted for commercialization. FER systems have a wide spectrum of applications as nonverbal cues are an essential form of communication, and the addition of a video processing layer will further widen the spectrum of application to real-time emotion analysis in job interviews or even criminal investigation as a potential to complement existing systems like polygraph testing.

References

1. Sebanz, N., Bekkering, H., Knoblich, G.: Joint action: bodies and minds moving together (2006)
2. Babajee, P., Suddul, G., Armoogum, S., Foogooa, R.A.: Machine learning approach for detecting facial expressions. *Rev. Bus. Technol. Res.* **15** (2019)
3. Turabzadeh, S., Meng, H., Swash, R.M., Pleva, M., Juhar, J.: Real-time emotional state detection from facial expression on embedded devices. In: *Proceedings of the 2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, Luton, UK, pp. 46–51, 16–18 Aug 2017
4. Huang, X., Lin, Y.: A vision-based hybrid method for facial expression recognition. In: *The proceedings of the 1st International Conference on Ambient Media and Systems* (2008)
5. Michel, P., Kaliouby, R.E.: Real time facial expression recognition in video using support vector machines. In: *Proceedings of ICMI 2003* (2003)
6. Datcu, D., Rothkrantz, L.: Facial expression recognition in still pictures and videos using active appearance model. In: *Proceedings of International Conference on Computer Systems and Technologies (CompSysTech'07)* (2007)
7. Shan, K., Guo, J., You, W., Lu, D., Bie, R.: Automatic facial expression recognition based on deep convolutional-neural-network structure. *IEEE 15th International Conference on Software Engineering Research, Management and Application*, pp. 123–128 (2017)
8. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(6), 681–685 (2001)
9. Revina, M., Emmanuel, W.: A survey on human face expression recognition techniques. Department of Computer Science, Christian College, Sunadaranar University, Tirunelveli, Tamil Nadu, India (2018)
10. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Real-time face detection and expression recognition: development and application to human–computer interaction. In: *Proceedings of the CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, Madison, WI (2003)
11. Goodfellow, I., Erhan, D., Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Lee, D., Zhou, Y., Rameiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shave-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J.,

- Romaszko, L. Xu, B., Chuang, Z., Bengio. Y.: Challenges in representation learning: a report on three machine learning contests (2013)
12. Viola, P., Jones, M.: Robust real-time object detection. Technical Report CRL 20001/01, Cambridge Research-Laboratory (2001)
 13. Jayalakshmi, J., Mathew, T.: Facial Expression recognition and emotion classification system for sentiment analysis. Mar Baselios College of Engineering and Technology. Trivandrum, India (2017)
 14. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell* **22**(12), 1424–1445 (2000)
 15. Theng, B.: Portable Real Time Emotion Detection System for the disabled. *Expert Systems with Applications*, Malaysia (2010)

An Extended Genetic Algorithm-Based Prevention System Against DoS/DDoS Flood Attacks in VoIP Systems



Sheeba Armoogum and Nawaz Mohamudally

Abstract VoIP frameworks need protections against unexpected organized dangers. The plausibility of such dangers begins from the progressing convergence of media transmission and IP arrange foundations. The flood attack is now considered as the most unsafe threat. In this paper, a new model for mitigation is proposed using a genetic algorithm that is capable to detect distributed attacks with convincing performance after evaluation using known parameters. The model can detect two unexpected behaviors of attackers. The results after performance analysis show that the detection rate is almost 100% for an attack rate of 40 messages per second and above. The false alarm rate is also good (around zero) while the sensitivity is near to 100%.

Keywords Denial of service (DoS) · Distributed denial of service (DDoS) · Flood attacks · Session initial protocol (SIP) · Genetic algorithm (GA)

1 Introduction

Voice over IP (VoIP) innovation has picked up its improvement universally. The quick alter is presently exceptionally for developing nations. The Global Market Insights Inc. [13] advertise a forecasted entrance of 12% for the period 2019 and 2025. The estimated global share will rise from USD 20 billion (2018) to USD 55 billion [7]

Since the introduction of the Session Initiation Protocol (SIP) [23], the voice conversation system using IP networks received positive praise by enterprises. This protocol which works at the application layer can be implemented easily by organizations. Recently, there has been a boost by open-source providers, like Asterisk [3], MiniSipServer [17], and OpenSIPs [20] for providing their SIP-based products to

S. Armoogum (✉)
University of Mauritius, Reduit, Mauritius
e-mail: s.armoogum@uom.ac.mu

N. Mohamudally
University of Technology, La Tour Koenig, Mauritius, Mauritius
e-mail: alimohamudally@umail.utm.ac.mu

businesses. For example, for the latest SIP product (MiniSipServer), someone does need technical knowledge to install the SIP server. The server can accommodate up to 500 users and can work on Linux, Windows, and Android platforms. Many product providers produce their SIP-based softphones as well as open-source applications.

However, it is also problematic when something is simple. Attackers have noticed the weakness of this protocol to conduct malicious activities via networks and the Internet. Since VoIP technology was still immature, the expert community via the taxonomy [27] report has warned enterprises of the huge number of voice threats. At many summits and meetings, researchers, industry experts, and engineers have demonstrated how far this system is vulnerable.

Recently, Azeez et al. [4] have given a comprehensive report on threats and on the detection and prevention techniques to mitigate the attacks. The most complex threat is the flood attack and this probably the reason why many researchers focus on this type of threat. Nazih et al. [18] conduct a recent work to combat against flood attacks. The authors claim that “further for being the most hazardous threat to VoIP system, it is very easy to launch.” The flood attack is when an illegitimate user sends a huge volume of messages to a system to paralyze the server as the latter will be unable to process bulks of requests in limited time. Once this is achieved, the entire network is down which can cause severe losses to a business enterprise. Nowadays, tools are available to launch these attacks from a remote location.

Like mentioned by the research community [12, 15, 26], “these messages are neither tampered nor malformed and hence are considered to be legitimate and can easily penetrate the system by tricking the server.” Among the different versions of the flood attacks, the INVITE flood attack is the most destructive one [14]. It exhausts the network by sending a huge number of INVITE messages to weaken either the server or its users. Since all users are constantly in communication with the SIP server, it is very easy to keep it busy for several seconds or few minutes until its memory is overflowed or the CPU is over-utilized or the network bandwidth is fully saturated. Now, it could be that the attacker is a user agent client (UAC), who may deliberately or unintentionally carry out malicious activities in a network. If the activity is not deliberate, then the UAC has become a zombie by a remote master attacker or a real attacker and a handler [1]. Indeed, the handler is an attacker that has gained control of multiple user agents to inject malicious codes for launching flood attacks. Such kind of attack is called a distributed flood attack which is very challenging since it is difficult to predict a behavioral pattern of an attacker.

For nearly two decades, VoIP providers and the academic community are developing tools (hardware or software) to mitigate the attacks [10]. There have been efforts on designing full-fledged defense systems, consisting of dedicated firewalls with sophisticated filtering and iptables utility, and intrusion detection systems (IDS) like the Snort IDS together with monitoring systems for an attempt to track down the illegitimate messages. Others have developed additional detection and mitigation applications (statistical, signature-based, anomaly-based, rule-based) as a second line of defense to support the firewall and protect the SIP server [4, 18, 25]. Statistical methods are used to develop either anomaly-based or signature-based IDS and intrusion detection and prevention systems (IDPS). Recently, many researchers are

developing hybrid IDS to overcome misuse and anomaly-based suspicious events. For instance, an anomalous self-learning system is developed by Rieck et al. [22] to detect deviations from genuine events in SIP-based systems.

The central subject matter of the paper is to develop an IDPS by extending the functions of a genetic algorithm to defend against the flood attacks. The new model, which we have named as the extended GA prevention (e-GAP) system, has two main objectives: Firstly, to prevent malicious activities by analyzing pattern behaviors by using a previously defined knowledge, and secondly, to pay attention to issues related to processing time, detection rate and other quality issues which will be addressed during the performance analysis phase.

We organize the outstanding sections as follows: Sect. 2 covers a list of reviews and related work conducted. We present our model in Sect. 3 and the testbed and methodology will be elaborated in Sect. 4. We present our results and performance analysis in Sect. 5 before providing a conclusion in the last section.

2 Literature Survey

Intrusion detection strategies are undergoing advanced changes basically after the introduction of artificial intelligence and machine learning using a statistical approach to train systems.

Chen [9] has developed a tool to address the DoS attacks by modifying original-state machines to detect the anomalies using four parameters of the messages. Based on the findings, it is a simple algorithm that works perfectly to detect fraudulent activities. However, the work is constrained to stateful servers only.

A hybrid method was developed to detect the distributed attacks to SIP-based network [16]. The IDS addresses the issue of high flow rate and high processing that SIP servers have difficulties to handle. The statistical method uses the sliding window to achieve low false alarm values for high attack rates.

Wan et al. [28] presented a defense algorithm to mitigate flooding attacks using a queue analysis model. The method can also alleviate the server's influence on these attacks by relaxing its service time. The advantage of this method is that the response time cannot be easily discarded. However, the authors did not specify the types of flood attacks.

An evolutionary approach using a GA technique was used to develop an IDS which can address the problem of a high false alarm rate triggered by illegitimate attackers and handlers [11]. This system can detect existing and new intrusions to send to a firewall for further action using a set of defined rules. During the same year, there was another effort by Netesan et al. [19] to improve the performance of an IDS using an algorithm known as AdaBoost, a machine learning algorithm. This method uses other machine learning algorithms to improve the attack detection rate, thereby decreasing the number of false alarms.

Another queueing method based on adaptive rules is developed by Basem et al. [6] to address flood attacks against SIP systems. The authors have proposed a complete

defense structure where the IDS is been integrated. Results show that the system works with improved performance compared to their previous work.

Ahmed and Ali [2] have developed a signature and anomaly-based method, capable to reduce the effects of flood attacks. According to the authors, reports indicate that IDSs developed are inefficient as most of them have a detection rate of less than 95.5% with a false positive are of 1.8%. Their model makes use of the data mining technique to train the system to reduce these two figures.

Saini et al. [24] claim that the number of distributed denial of service attacks to networks to devoid business of services is rising every day. This has motivated the researchers to develop a system that can address different types of flood attacks. The model which uses a machine learning algorithm is compared with other algorithms (random forest and Naïve Bayes algorithms). Results show that the machine learning model is feasible to use to give support to the firewall.

In this study, the same aims are met using our genetic algorithm-based model.

3 The Mitigation Model

In this section, we propose the e-GAP method to address the problem of flood attacks.

The e-GAP system is been based on the fundamental concept of a genetic algorithm to detect possible attacks and prevent the attacks by eliminating the attackers and notifying the monitoring system. The process begins with an initial population, which is a two-dimensional array with each individual characterized by a set of parameters (Source IP, Destination IP, and Info). The fitness functionality determines the attackers. An attacker is been identified with the highest fitness score. In the e-GAP algorithm, an individual with the least or fitness value one is considered as a legitimate user to the system.

The algorithm initiates with an initial population and the said population performs a crossover by the ‘divide and swap’ concept at a crossover point. The resultant two subarrays (lists) go through a mutation process with the selected target individual. In the mutation process, the target individual is been processed to find the fitness value by the method of identifying a repetitive number of times the individual is present in the selected list. In case the fitness score is greater than 3, then the said target individual is added to a list (SIPQuarantine), otherwise, the target individual is added to another list (SIPLegitimate). The mutation process returns the merged processed subarrays (lists), which is the new population for the next generation. The process continues by selecting a new target individual from the new population with crossover and mutation method until no further individuals are been left in the population.

The e-GAP algorithm is implemented in Python platform as depicted below.

The algorithm shows how it can address the problem of a single flood, distributed flood with pattern behaviors.

e-Gap Algorithm:

1. Input the SIP call dataset **SourceSIPcsv** from the Firewall in csv format
2. Cleanse the **SourceSIPcsv** and create a two-dimensional list **SourceSIPList** where each element of the list has the values of Source, Destination, and Info.
3. Cleanse the time field to the millisecond and assign the **WindowSize** variable as the difference between the last time record and the first-time record of the dataset.
4. Declare **SIPQuarantine** and **SIPLegitimate** as two empty global list variables.
5. Initialise the **initialPopulation** list to **SourceSIPList** and **numGeneration = 0**.
6. Select **Target** variable as the first element of the **initialPopulation** list.
7. Perform **CrossOver** method with **initialPopulation** as the parameter value:
 - a. Divide the **initialPopulation** list into two halves (parentF, parentM) at the mid-index of the list.
 - b. Assign **crossPoint** as the mid-index of the largest list, which is the list with the highest list size.
 - c. Swap elements of lists parentF and parentM at the (crossPoint - 1) and (crossPoint + 1)
 - d. Return lists parentF, parentM.

8. Perform **Mutation** method with **parentF**, **parentM**, **Target** and **WindowSize** as the parameter values:
 - a. Assign variable fitness to 0.
 - b. Loop list parentF and parentM from 0 to the size of each list:
 - i. Compare Target with loop elements.
 - ii. Increment fitness by 1 for each successful comparison of source, destination, and Info values of the Target variable.
 - iii. Increment loop index by 1.
 - c. If (WindowSize < 500ms and fitness = 1), (WindowSize < 1000ms and fitness = 2), (WindowSize < 2000ms and fitness = 3), (WindowSize < 4000ms and fitness = 4), (WindowSize < 8000ms and fitness = 5), (WindowSize < 16000ms and fitness = 6), then add Target to the list SIPLegitimate.
 - d. If else (WindowSize < 500ms and fitness >1), (WindowSize < 1000ms and fitness >2), (WindowSize < 2000ms and fitness > 3), (WindowSize < 4000ms and fitness > 4), (WindowSize < 8000ms and fitness > 5), (WindowSize < 16000ms and fitness > 6), then add Target to the list SIPQuarantine.
 - e. Remove Target from both lists parentF and parentM and create a newPopulation by merging parentF and parentM.
 - f. Return newPopulation
9. Loop from 0 to the size of the **newPopulation** list until newPopulation size =0:
 - a. Assign Target as the first element of the **newPopulation** list.
 - b. Perform Steps 6, 7, and 8.
 - c. Increment numGeneration by 1.
10. Send unique sourceIPs from the SIPQuarantine list to the Firewall.
11. Deep analyse the list SIPQuarantine for the target users of attack.
12. Deep analyse the SIPLegitimate List for false positive and false negative messages.
13. Inform the system administrator with the updated data of legitimate sourceIPs and quarantine SourceIPs.

4 The VoIP Defense System Architecture

In this section, a full set of groundwork to validate the IDS will be explained.

4.1 Methodology and Testbed

To validate our model, a proposed architecture is set as illustrated in Fig. 1. This test is conducted in a separate network to avoid problems of security to other adjacent networks.

As mentioned earlier, the SIP server is mounted using MiniSIPServer products. The server is installed on a Linux Ubuntu (version 18.04) machine (Intel Quad-Core i7 processor), working at a processor frequency of 3.3 GHz and supporting Gigabit Ethernet connection with an installed RAM of 8 GB. The firewall and the e-GAP IDPS are installed on an Intel Core i5 processor computer of 4 GHz of speed with a memory of 16 GB. We use a computer (Intel Core i7 processor 4 GHz and 8 MB memory) to install our penetration tool to generate flood attacks. In this same computer, two softphones (MiniSIPphone 7.3 and Zoiper) are installed. We use two mobile devices and two laptops to install additional softphones (MiniSIPphone 7.3, Zoiper, and LinPhone).

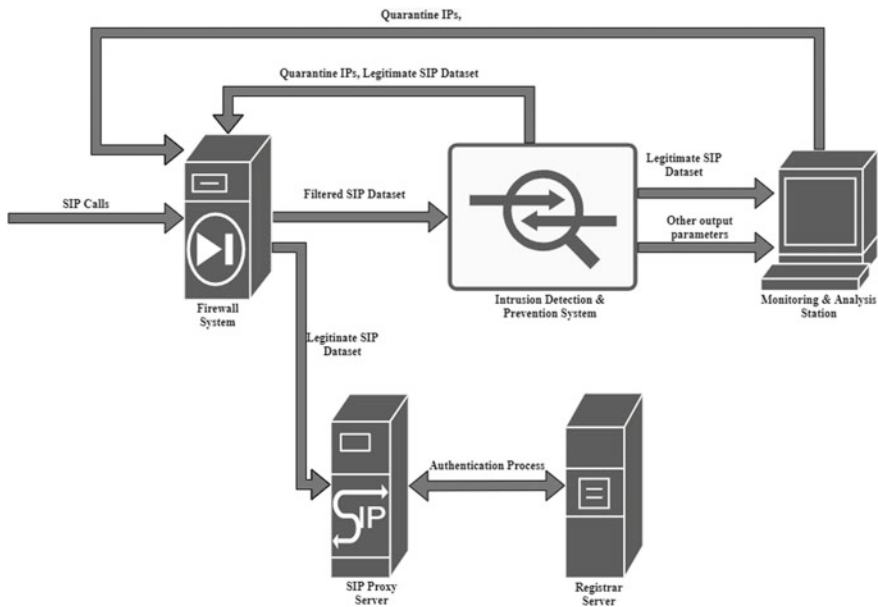


Fig. 1 Defense system

Data that are collected in the firewall in a CSV file containing the five fields (Time, Source, Destination, Protocol, and Info) are sent to the e-GAP model. The IDPS will analyze the messages to detect the illegitimate attacks, following which will be sent to the firewall for blocking new intruders having the same signatures. The bad packets are also sent to the monitoring and analysis station to conduct further analysis so that a system administrator can take necessary action if need be.

4.2 Dataset and Attack Scenarios

As mentioned earlier, there is no public dataset available for testing threats, except the list from Bad Packets LLC [5]. To simulate our work, we prepare our INVITE messages using the flood message generator and by initiating genuine calls using the three types of softphones. The list is obtained from the firewall (CSV format). To make the dataset more representative, messages are created using the list from Bad Packets LLC [5].

Based on the SIP specifications [23], INVITE messages will be verified every 500 ms as per response time till timeout (32 s) in case there is no successful call. For this study, the experiment will be carried out from 20 messages per second (mps) to 200 mps for 16 s. Furthermore, the window size will vary and will be deviated from the response time set at the firewall. The window size will be calculated during the running of the e-GAP system.

There are different attack scenarios catered by the e-GAP model. It first caters to single flood attacks, that is, an attacker flooding a destination node. We simulate as well when an attacker sends sequential messages (one per slot) during the 16 s. Another unusual pattern, which we have assumed, is when an illegitimate message is sent every alternate slot, thereby making the detection difficult. Similarly, for distributed attacks, multiple attackers are executing the two above mentioned behaviors.

5 Experimental Analysis and Discussion

The analysis is conducted based on metric parameters, which include the detection rate, false positive alarms, and system recall. Besides, we compute the amount of processing time of the e-GAP model for each attack rate.

5.1 Experimental Metrics Used

As mentioned in the previous section, the measurement will be conducted for 32 slots of 500 ms. Hence, the system is trained every 1/32th = 3.125% of the total packets entering the IDPS.

We evaluate the performance of the new model by recording the values of the true positive (TP), false positive (FP), true negative (TN), and the false negative (FN).

The performance parameters are calculated as follows:

- (i) Attack detection rate is given by

$$\text{Detection Rate (DR)} = \frac{\text{Total attacks detected}}{\text{Total attacks injected}} \tag{1}$$

- (ii) False positive rate or false alarm rate is the probability that the model detects a genuine message as an attacker.

$$\text{False Alarm Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2}$$

- (iii) System recall or system sensitivity is defined as the true positive rate (TPR), that is, it measures the percentage of actual positives that are correctly been identified by our IDPS. A high sensitivity means high accuracy of the IDPS.

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

5.2 Performance Evaluation

We use the cross-sectional and the before-and-after during the experiment. We also conduct three experiments using different datasets of the same size with the almost same number of genuine calls. The experiments are repeated for four attack rates (Table 1). Table 1 shows the total messages injected for different attack rates.

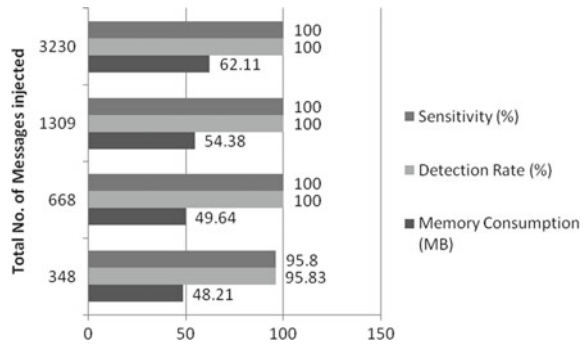
Table 1 Attack scenarios

Attack rates	Number of genuine messages	Number of attacks injected	Total number of messages
20	28	320	348
40	28	640	668
80	29	1280	1309
200	30	3200	3230

Table 2 Performance results

Total messages	Total response time (ms)	Memory consumption (MB)	Detection rate (%)	False alarm rate (%)	Sensitivity (%)
348	5.76	48.21	95.83	0.00	95.80
668	5.81	49.64	100.00	0.00	100.00
1309	6.10	54.38	100.00	0.00	100.00
3230	6.50	62.11	100.00	0.00	100.00

Fig. 2 Performance results



The findings are tabulated after computing the response time for different attack rates (the execution time of the IDPS), the overall memory usage to process each dataset, the detection rate, the false positive alarms, and the system sensitivity (Table 2).

Table 2 and the consolidated chart below (Fig. 2) show several variations (refer to the three above equations) when the number of messages increases. The total processing time increases with an increase in the number of messages injected. The memory consumption follows a linear regression curve instead of regular linear change, which is perfect for an IDPS. It is observed that the probability that our model falsely detect legitimate messages as attackers is near to zero percent when the attack rate is greater than 40 mps. There is rise in the detection rate to 100% with an increase in attack rate above 40 mps. Finally, the sensitivity of our model is determined by computing the true positive rate. Again, the outcome shows a near 100% above 40 mps, hence, proving that our model is feasible for integrating in a defense system.

6 Conclusion

In this paper, we present a new mitigation technique using a modified genetic algorithm to reduce the flood attacks. The results show that the e-GAP model can mitigate

both types of attacks. Two behaviors are successfully been implemented. The results indicate that the IDPS can classify a false alarm of approx zero. Therefore, when benchmarking with the work cited by Ahmed and Ali [2], our model is giving an excellent performance as the detection rate is nearly 100% with a false alarm rate of near 0% and a mean sensitivity value of around 98.8% due to the deep analysis method for 16 s which justifies that it can be used in a defense system.

In our future work, we will further enhance the IDPS by conducting stress tests using higher attack rates. Also, other types of attacks will be considered (malformed and tampering) to strengthen the firewall and hence protect the SIP server.

References

1. Ahmad, W., Singh, D.: VoIP security: a model proposed to mitigate DDoS attacks on SIP based VoIP network. In: A Multi-Disciplinary Research Book, pp. 37–48. Researchgate (2018)
2. Ahmed, M.R.A., Ali, F.M.A.: Enhancing hybrid intrusion detection and prevention system for flooding attacks using decision tree. In: International Conference on Computer, Control, Electrical, and Electronics Engineering. IEEE (2019)
3. Asterisk: Ready to get started (2020). Available at <https://www.asterisk.org/>. Accessed 10 May 2020
4. Azeez, N.A. et al.: Intrusion Detection and Prevention Systems: An Updated Review. Springer, Berlin (2020)
5. Bad Packets, L.: Meaningful Intelligence for an Evolving Cybersecurity Landscape (2020). Available at <https://badpackets.net>. Accessed 4 June 2017
6. Basem, B., Ghalwash, A.Z., Sadek, R.A.: Multilayer secured SIP based VoIP architecture. *Int. J. Comput. Theory Eng.*, 453–462 (2015)
7. Bhutani, A., Wadhvani, P.: Voice over Internet Protocol (VoIP) Market Size by Type (Integrated Access/Session Initiation Protocol). Global Market Insights (2019)
8. Bouzida, Y., Mangin, C.: A framework for detecting anomalies in VoIP networks. In: IEEE, Third International Conference on Availability, Reliability, and Security (2008)
9. Chen, E.Y.: Detecting DoS attacks on SIP systems. In: IEEE Workshop on VoIP Management and Security. IEEE (2006)
10. Coulibaly, E., Liu, L.H.: Security of VoIP networks. In: 2nd International Conference on Computer Engineering and Technology (ICCET) (2010)
11. Dhak, B.S., Lade, S.: An evolutionary approach to intrusion detection system using genetic algorithm. *Int. J. Emerg. Technol. Adv. Eng.* **2**(12), 632–636 (2012)
12. Ehlert, S., Wang, C., Magedanz, T. Sisalem, D.: Specification-based denial-of-service detection for sip voice-over-ip networks. In: The Third International Conference on Internet Monitoring and Protection. IEEE (2008)
13. Global Market Insights: Insights to Innovation (2019). Available at <https://www.gminsights.com/industry-analysis>. Accessed 20 Oct 2019
14. Hussain, I., Djahel, S., Zhang, Z., Naït-Abdesselam, F.: A comprehensive study of flooding attack consequences and countermeasures in Session Initiation Protocol (SIP). *J. Secur. Commun. Netw.* 4436–4451
15. Lahmadi, A., Fester, O.: A framework for automated exploit prevention from known vulnerabilities in voice over IP services. *IEEE Trans. Netw. Serv. Manag.*, 114–127 (2012)
16. Li, W., Guo, W., Luo, X., Li, X.: On Sliding Window Based Change Point Detection for Hybrid SIP DoS Attack. In: IEEE Asia-Pacific Services Computing Conference. IEEE (2010)
17. MyVoIPApp: MiniSIP Server—Setup IP PBX step by step (2018) Available at: https://www.myvoipapp.com/docs/faq/setup_ippbx_for_small_business_step_by_step/index.html. Accessed 23 Jan 2018

18. Nazih, W., et al.: Efficient detection of attacks in SIP based VoIP networks using linear L1-SVM classifier. *Int. J. Comput. Commun. Control* **14**(4), 518–529 (2019)
19. Netesan, P., Balasubramanie, P., Gowrison, G.: Improving the attack detection rate in network intrusion detection using Adaboost Algorithm. *J. Comput. Sci.* 1041–1048(2012)
20. OpenSIPS (2017): Available at <https://www.opensips.org/>. Accessed 5 June 2017
21. Ormazabal, G., Nagpal, S., Yardeni, E., Schulzrinne, H.: Secure sip: a scalable prevention mechanism for dos attacks on sip based VoIP systems. In: *Principles, Systems, and Applications of IP Telecommunications* (2008)
22. Rieck, K., Wahl, S., Laskov, P., Domschitz, P., Müller, K.R.: A Self-learning System for Detection of Anomalous SIP Messages, pp. 90–106. Springer (2008)
23. Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A.: SIP: session initiation protocol, RFC 3261. IETF Network Working Group, IETF (2002)
24. Saini, P.S., Behal, S., Bhatia, S.: Detection of DDoS attacks using machine learning algorithms. In: *7th International Conference on Computing for Sustainable Global Development*. IEEE (2020)
25. Semerci, M., Cemgil, A.T., Sankur, B.: An intelligent cybersecurity system against DDoS attacks in SIP networks. *J. Comput. Networks* **136**, 137–154 (2018)
26. Tang, J., Cheng, Y., Yong, H.: Detection and prevention of SIP flooding attacks in voice over IP networks. In: *INFOCOM Conference*, pp. 1161–1169. IEEE (2012)
27. VOIPSA: VoIP Security and Privacy Threat Taxonomy. Security and Threats (Official Doc.) (2005)
28. Wan, X.-Y., Li, Z., Fan, Z.-F.: A SIP DoS flooding attack defense mechanism based on priority class queue. In: *IEEE International Conference on Wireless Communications, Networking and Information Security* (2010)

Entropy Based Cluster Selection



Arko Banerjee, Arun K. Pujari, Chhabi Rani Panigrahi, and Bibudhendu Pati

Abstract Clustering has emerged as a method of unsupervised partitioning of a given set of data instances into a number of groups (called clusters) so that instances in the same group are more similar among each other with respect to instances in other groups. But there does not exist a universal clustering algorithm that can yield satisfactory result for any dataset. In this work we consider an ensemble (collection) of clusterings (partitions) of a dataset obtained in different ways and devise two methods that judiciously select clusters from different clusterings in the ensemble to construct a robust clustering. The superior performances of the proposed methods over well-known existing clustering algorithms on several benchmark datasets are empirically reported.

Keywords Clustering · Clustering ensemble · Consensus clustering · Entropy

1 Introduction

Cluster analysis or clustering [1] is a machine learning technique that unsupervisedly recognizes groups (clusters) of *similar* data instances collected from fields such as marketing, social network, bio-medical etc. [2]. In the literature numerous clustering

A. Banerjee (✉)

College of Engineering and Management, Kolaghat, West Bengal, India

e-mail: arko.banerjee@gmail.com

Biju Patnaik University of Technology, Rourkela, Odisha, India

A. K. Pujari

Central University of Rajasthan, Ajmer, Rajasthan, India

e-mail: arun.k.pujari@gmail.com

C. R. Panigrahi · B. Pati

Rama Devi Women's University, Bhubaneswar, Odisha, India

e-mail: panigrahichhabi@gmail.com

B. Pati

e-mail: patibibudhendu@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_26

algorithms [3] have been proposed that conceptually differ among themselves for clustering datasets having a different distribution of instances. But since there does not exist a universal clustering algorithm that can provide an acceptable result for any dataset [4], clustering ensemble (also referred to as consensus clustering) [5] techniques have been proposed in the literature that optimally combine the results of different clusterings to yield a qualitatively better and robust clustering solution. But since finding a consensus clustering in the space of clusterings is an NP-hard problem, several heuristics are being proposed to find an acceptable consensus, which can be broadly categorized into hyper-graph based, information theory based, mixture model (EM) based, voting based and co-association based methods.

In this work, we propose a new way of arriving at a consensus clustering from an ensemble of clusterings with a true number of clusters. We present a concept of consensus clustering that can be formed by selecting clusters from the existing clusterings in the ensemble. Every cluster in the ensemble can be assigned with an entropy measure representing its reliability in the ensemble. A cluster having lower entropy can be considered being of higher priority in forming a consensus. Hence selection of clusters and overlapping problem (if any) of selected clusters can be resolved by prioritizing clusters having lower entropy values to form the final consensus clustering. In a second approach we propose to select a cluster having the lowest entropy from an ensemble of clusterings and iteratively the process is repeated after removing the data instances of the selected clusters to construct the final clustering. In the empirical section, we explain the significance of our proposed methods and show their superiority over well-known existing clustering methods on several benchmark datasets.

The outline of the rest of this paper is as follows. The related work is written in Sect. 2. Section 3 presents the problem statement and the proposed methods are described in Sect. 4. Section 5 empirically demonstrates the performance of the proposed algorithms and Sect. 6 concludes the paper.

2 Related Work

In recent years a significant number of research papers have been published on new algorithms of consensus clustering [7]. We give a brief overview of the well-known clustering ensemble algorithms in this section.

Fred et al. [5] combined ensemble of clusterings produced in different ways into a co-association matrix (also called similarity matrix) and the final consensus clustering was obtained by applying hierarchical single-link algorithm on the co-association matrix. The method was referred to as Evidence Accumulation (EAC) method. Huang et al. [8] applied average-linked agglomerative clustering technique on a cluster-level weighted co-association matrix to derive a consensus clustering. The proposed method was referred by them as Locally Weighted Ensemble Accumulation (LWEA) method, in which the said cluster-level weight depended on a user defined parameter, and the best result was reported in the experiment. Strehl et al. [6] represented the

problem of clustering ensemble as a combinatorial optimization problem in terms of shared mutual information and tried to solve the problem by representing the clustering ensemble as a hyper-graph, where each clustering is an hyper-edge. They proposed three different algorithms, namely Cluster-based Similarity Partitioning (CSPA), Hyper Graph Partitioning (HGPA), and Meta-Clustering (MCLA). Huang et al. [10] implemented Crowd Agreement Estimation and Multigranularity Link Analysis to solve the clustering ensemble problem. They proposed two algorithms namely, weighted evidence accumulation clustering (WEAC) and graph partitioning with multi-granularity link analysis (GP-MGLA). Huang et al. [10] computed Probability Trajectory based Similarity between clusters in the clusterings and proposed two algorithms namely, Probability Trajectory Accumulation(PTA) and Probability Trajectory Based Graph Partitioning(PTGP). PTA was an agglomerative clustering algorithm that depended on two user-defined parameters. In our experimental comparison the best result of the method is reported. Nguyen et al. [12] presented iterative voting techniques to find a clustering ensemble in which the principal operation was to learn the closest cluster from each data instance in every iteration. On this basis they proposed their algorithms, namely Iterative Voting Consensus (IVC), Iterative Pairwise Consensus (IPC) and Iterative Probabilistic Voting Consensus (IPVC). Fern et al. [9] proposed Cluster and select(CAS) method that used NMI [6] measure to group similar clusterings from the initial ensemble. The final ensemble was created by selecting a representative from each group of clusterings. Finally, CSPA was used as the clustering ensemble method on the final ensemble. The size of the final ensemble was a user-defined parameter, and the best result was reported in the experiment.

3 Problem Statement

Let D be the set of N data instances $D = \{d_1, d_2, \dots, d_N\}$. We are given with an ensemble of M clusterings, $E = \{P_1, P_2, \dots, P_M\}$ of D , where a clustering P_t ($t = 1, 2, \dots, M$) on D is defined as a set of clusters $P_t = \{C_1^t, C_2^t, \dots, C_K^t\}$ such that $C_i^t \subseteq D$, $C_i^t \cap C_j^t = \phi, (i \neq j)$ and $\bigcup_{i=1}^K C_i^t = D$, ($i, j = 1, 2, \dots, K$). The goal of clustering ensemble is to find a clustering $P^* = \{C_1^*, C_2^*, \dots, C_k^*\}$ that is a sort of median of the given ensemble of clusterings P_1, P_2, \dots and P_M . P^* is called the consensus clustering.

In this work we approach the problem of consensus clustering by assigning an entropy value to individual cluster in the ensemble. To do that let us define the joint probability $p(C_i^t, C_j^s)$ that signifies the level of agreement cluster C_i^t (in clustering P_t) has with cluster C_j^s (in clustering P_s). By level of agreement we mean the number of elements both clusters have in common. Then,

$$p(C_i^t, C_j^s) = \frac{|C_i^t \cap C_j^s|}{|C_i^t|} \quad (1)$$

Let the entropy measure $\text{Ent}(C_i^t, C_j^s)$ represents the entropy of a cluster C_i^t w.r.t. C_j^s (and vice versa) and is computed as in Huang et al. [8] (it is to note that, our proposed methods do not depend on a specific surprisal measure),

$$\text{Ent}(C_i^t, C_j^s) = -p(C_i^t, C_j^s) \cdot \log_2 p(C_i^t, C_j^s) \quad (2)$$

Then the total entropy of the cluster C_i^t w.r.t all clusters in the ensemble E is

$$\text{Ent}(C_i^t) = - \sum_{\substack{s=1 \\ s \neq t}}^M \sum_{j=1}^K p(C_i^t, C_j^s) \cdot \log_2 p(C_i^t, C_j^s) \quad (3)$$

Less the total entropy of a cluster more it agrees with other intersecting clusters in the ensemble. Hence a clustering, containing clusters having minimum possible entropy values, is a sort of median of the entire ensemble which approximates the consensus clustering P^* . In our work we propose two heuristics to achieve the said goal. It is also to note that, Eq. 3 tends to allot lower entropy to finer clusters than that of coarser clusters. In a trivial case, a cluster having a single data instance has an entropy value of zero.

4 Proposed Methods

In our first approach we put all the clusters of each clustering in the ensemble into a set say C . Since each clustering contains K clusters and there are total M number of clusterings in the ensemble E , there will be $K \cdot M$ number of clusters in C . We consider a collection, say S , of all possible subsets of C , where each set consists of K clusters and covers all the data instances in D . Cluster overlapping may happen in a set if all the K clusters are not mutually exclusive. Out of all the sets in S we consider the set, say \overline{C}^* , in which the sum of entropy values of all the constituent clusters is minimum. If all the K clusters in \overline{C}^* are mutually exclusive, then \overline{C}^* is the desired consensus clustering P^* . Else the cluster overlapping problem is solved by prioritizing clusters having lower entropy value. More precisely, if two clusters are not disjoint, the cluster having less Entropy value will be retained, whereas the intersection part of the said two clusters is removed from the other cluster. It is to note that, reduction of a cluster reduces its degree of disagreement with other clusters in the ensemble; as a result the reduced cluster tends to have a lower entropy value than that of its previous version. The final set, without overlapping clusters, is the desired consensus clustering P^* . The proposed solution of the overlapping problem can actually decrease the entropy of the clustering (which is the sum of the total entropy values of its constituent clusters), hence resulting in a better consensus. We write down the said heuristics in form of an algorithm and we call it Entropy based Cluster Selection (ECS) Consensus Clustering. The output of the algorithm cannot be empty, as the existing clustering with the lowest entropy in the ensemble

is also a possible output. Such output can occur only when the selected clusters are non-overlapping. The time complexity of the proposed algorithm can be reduced by solving it using dynamic programming.

Algorithm 1 Entropy Based Cluster Selection (ECS)

Input Ensemble of Clusterings $E = \{P_1, P_2, \dots, P_M\}$
 Output Consensus Clustering P^* with K clusters

- 1: $C = \{C_i^t : C_i^t \in P_t, t = 1 \dots M, i = 1, \dots, K\}$
- 2: $S = \{\bar{C} \subseteq C : |\bar{C}| = K \wedge \bigcup_{j=1}^K C_{i_j}^{t_j} \subseteq D\}$
 where, $\bar{C} = \{C_{i_j}^{t_j} : C_{i_j}^{t_j} \in C, j = 1, \dots, K, i_j \in \{1, \dots, K\}, t_j \in \{1, \dots, M\}\}$
- 3: $\bar{C}^* \leftarrow \arg \min_{\bar{C}} \left(Ent(\bar{C}) = \sum_{j=1}^K Ent(C_{i_j}^{t_j}) : \bar{C} \in S \right)$
 $\{Ent(.) \text{ defined in Eq. 3}\}$
- 4: *WOLOG* Let $\bar{C}^* = \{C_1^*, C_2^*, \dots, C_K^*\} \ni Ent(C_u^*) \leq Ent(C_{u+1}^*),$
 $u = 1, 2, \dots, K - 1$
- 5: $T \leftarrow C_1^*$
- 6: **for** $2 \leq l \leq K$ **do**
- 7: $C_l^* \leftarrow C_l^* \setminus T$
- 8: $T \leftarrow T \cup C_l^*$
- 9: **end for**
- 10: $P^* \leftarrow \bar{C}^*$

We propose an alternative approach in generating a robust and good quality clustering from an ensemble of clusterings. In our second approach we follow the following steps to come up with the final clustering.

1. Consider the set of all clusters C in an ensemble of clusterings each having K number of clusters.
2. Select the cluster in C having the lowest entropy value.
3. Remove the data instances contained in the selected cluster from the dataset D .
4. Generate a new ensemble of clusterings each having $K = K - 1$ number of clusters.
5. Repeat steps 1–4 until the remaining data instances in D is grouped into a single cluster.

The above method cannot be categorized as a traditional consensus clustering method. It iteratively generates the final clustering from a sequence of an ensemble of clusterings with a reducing number of clusters. In this method the quality of the final clustering depends more on the ensemble generation process than ECS. We write down this method in form of an algorithm and we call it Maximum Entropy Cluster Selection (MECS) Clustering.

5 Experimental Result

In this section, we empirically show the efficiency of our proposed methods over other existing well-known clustering ensemble algorithms. First, we discuss the data sets and the basic settings of our experiments used in the evaluation process.

5.1 Datasets

In our experiments, we use nine well known real-world labeled datasets, namely, *Iris*, *Wine*, *Glass*, *Image Segmentation (IS)*, *Ecoli* and *Steel Plates Faults (SPF)*. Table 1 displays the details of the datasets. All the datasets are taken from the UCI machine learning repository [11].

Algorithm 2 Maximum Entropy Cluster Selection (MECS)

Input Dataset D and number of desired clusters K

Output Consensus Clustering P^* with K clusters

- 1: Initialization: $P^* \leftarrow \emptyset$
 - 2: **for** each $K \geq l \geq 2$ **do**
 - 3: Generate Ensemble of Clusterings $E^l = \{P_1^l, P_2^l, \dots, P_M^l\}$ on D
 where, $P_i^l = \{C_1^i, C_2^i, \dots, C_l^i\} \ni C_i^i \subseteq D, C_i^i \cap C_j^i = \phi, (i \neq j)$
 $\wedge \bigcup_{i=1}^l C_i^i = D, (i, j = 1, 2, \dots, l)$
 - 4: $C_l \leftarrow \arg \min_{C_i^i} (Ent(C_i^i) : C_i^i \in P_i^l, i = 1, \dots, l, t = 1, \dots, M)$
 $\{Ent(.) \text{ defined in Eq. 3}\}$
 - 5: $D \leftarrow D \setminus C_l$
 - 6: $P^* \leftarrow \{P^*, C_l\}$
 - 7: **end for**
 - 8: $C_1 \leftarrow D$
 - 9: $P^* \leftarrow \{P^*, C_1\}$
-

Table 1 Description of datasets

Dataset	#Instances	#Attribute	#Class
<i>Iris</i>	150	4	3
<i>Wine</i>	178	13	3
<i>Glass</i>	214	10	7
<i>IS</i>	2310	19	7
<i>Ecoli</i>	336	8	8
<i>SPF</i>	1941	27	7

Table 2 Comparison of performance with NMI

Method	Iris			Wine			Glass		
	$E/20$	$E_d/10$	$E_d/5$	$E/20$	$E_d/10$	$E_d/5$	$E/20$	$E_d/10$	$E_d/5$
ECS	0.7515	0.59412	0.59201	0.4287	0.40556	0.40125	0.36227	0.34828	0.34356
MECS	0.7462	0.57227	0.56718	0.3988	0.38917	0.3912	0.35986	0.34231	0.34029
LWEA	0.74944	0.53154	0.54595	0.41973	0.39167	0.39726	0.36122	0.34839	0.34341
EAC	0.74944	0.53154	0.542	0.41973	0.39167	0.39714	0.36182	0.34821	0.34261
WEAC-AL	0.74998	0.53514	0.55392	0.4287	0.39167	0.39445	0.36182	0.33849	0.34143
GP-MGLA	0.74998	0.53514	0.55392	0.4287	0.39167	0.39273	0.36114	0.35165	0.33545
PTA-AL	0.49068	0.22804	0.34885	0.26771	0.22225	0.23623	0.35885	0.34679	0.33576
PTA-CL	0.49068	0.2024	0.32519	0.21513	0.15742	0.23529	0.35684	0.34824	0.34253
PTA-SL	0.413	0.30307	0.3361	0.29405	0.21686	0.24043	0.34524	0.31615	0.32587
PTGP	0.68158	0.53225	0.54479	0.40422	0.39555	0.39867	0.36501	0.35776	0.34848
CSPA	0.64280	0.50197	0.51379	0.38122	0.34395	0.30911	0.34046	0.32114	0.32071
HGPA	0.59815	0.46718	0.47819	0.35480	0.32012	0.28769	0.31687	0.29889	0.29848
MCLA	0.67935	0.53052	0.54302	0.40290	0.36351	0.32670	0.35983	0.33941	0.33895
IVC	0.62025	0.48437	0.49578	0.36785	0.33189	0.29828	0.32853	0.30988	0.30946
IPVC	0.66308	0.53205	0.52568	0.39324	0.36454	0.31626	0.35120	0.34037	0.32812
IPC	0.6678	0.52151	0.53378	0.39606	0.35734	0.32115	0.35372	0.33364	0.33319
CAS	0.64778	0.50407	0.51502	0.38418	0.34540	0.30985	0.34311	0.32249	0.32146
Method	IS			Ecoli			SPF		
	$E/20$	$E_d/10$	$E_d/5$	$E/20$	$E_d/10$	$E_d/5$	$E/20$	$E_d/10$	$E_d/5$
ECS	0.54861	0.51435	0.45085	0.55961	0.55415	0.55427	0.10418	0.09202	0.08059
MECS	0.54275	0.51249	0.44988	0.55323	0.55172	0.55283	0.10386	0.09179	0.08021
LWEA	0.55191	0.51723	0.45754	0.55608	0.55225	0.55393	0.10377	0.08482	0.080289
EAC	0.53113	0.51381	0.43187	0.55855	0.55287	0.55426	0.10421	0.092756	0.08176
WEAC-AL	0.49775	0.47101	0.42107	0.54755	0.54024	0.531045	0.10336	0.093972	0.08349
GP-MGLA	0.54541	0.50775	0.44019	0.54085	0.54016	0.53679	0.10513	0.091142	0.072813
PTA-AL	0.54527	0.50486	0.4484	0.51711	0.51647	0.51454	0.10434	0.093604	0.073398
PTA-CL	0.55638	0.50981	0.44589	0.51195	0.50937	0.50806	0.1058	0.09519	0.073283
PTA-SL	0.36147	0.36147	0.34005	0.54009	0.54663	0.54719	0.088353	0.076516	0.064218
PTGP	0.54559	0.50305	0.44856	0.52732	0.52449	0.52489	0.10635	0.090923	0.079137
CSPA	0.51454	0.47442	0.42303	0.49731	0.53647	0.60210	0.10030	0.08575	0.07463
HGPA	0.47888	0.44154	0.39372	0.46285	0.49929	0.56037	0.09335	0.07981	0.06946
MCLA	0.54380	0.50140	0.44709	0.52559	0.56698	0.63634	0.10600	0.09063	0.07888
IVC	0.49650	0.45779	0.40820	0.47987	0.51766	0.58099	0.09678	0.08274	0.07202
IPVC	0.53077	0.50282	0.43281	0.51299	0.55339	0.62108	0.10630	0.09088	0.07910
IPC	0.53457	0.49289	0.43950	0.51667	0.55735	0.62554	0.10420	0.08909	0.07754
CAS	0.51854	0.47642	0.42403	0.50118	0.54064	0.60678	0.10072	0.08611	0.07495

5.2 Ensemble Generation

The clusterings in the ensemble are generated by applying the K-means algorithm [1] with different random initializations. The K-means method is considered here as it is widely used in the clustering ensemble studies in the literature. We generate an ensemble of 100 clusterings for all datasets in which clusterings may repeat. We denote E as the entire ensemble, E/x as the ensemble formed by randomly selected x percentage of clusterings from E , E_d as the distinct partitions in E , and E_d/x would mean a randomly selected x percentage clusterings from E_d . We perform tests on three types of ensembles $E/20$, $E_d/10$ and $E_d/5$. We run all the clustering ensemble algorithms, discussed in Sect. 2, ten times for each dataset and report the average performance. The performances (validity) of the methods are measured with the help of Normalized Mutual Information (NMI) [6], by comparing their result with the ground truth information available with each dataset. The range of NMI is between 0 and 1 and a larger value indicates better quality clustering w.r.t ground-truth information. Let P^* be the consensus clustering and G is the ground-truth clustering. The NMI score of P^* given G is defined as follows:

$$\text{NMI}(P^*, G) = \frac{\sum_{i=1}^K \sum_{j=1}^K N_{ij} \log \frac{N_{ij} N}{N_i^{P^*} N_j^G}}{\sqrt{\sum_{i=1}^K N_i^{P^*} \log \frac{N_i^{P^*}}{N} \sum_{j=1}^K N_j^G \log \frac{N_j^G}{N}}}, \quad (4)$$

where P^* and G have K (true) number of clusters and N is the total number of instances in the data set. $N_i^{P^*}$, N_j^G and N_{ij} are the numbers of data instances in the i th cluster of P^* , j th cluster of G and in both i th cluster in P^* and j th cluster in G , respectively.

5.3 Evaluation

We report the experimental outcomes in Table 2. The proposed *ECS* achieves highest NMI scores on the *Iris*, *Wine* and *Ecoli* datasets. It performs second best and third best in the case of *Glass* and *IS* datasets, respectively. The performance of the proposed *MECS* is inferior to that of *ECS*, but superior to most of the well-known methods.

6 Conclusion

In this work, we introduce the concept of cluster selection based consensus clustering technique. We demonstrate that judiciously selecting a set of clusters from the ensemble can result in a better consensus. Evaluations on different datasets show the proposed approaches are efficient and effective in improving the quality of consensus

as compared to the existing approaches when the true number of clusters is considered. In the future, we would like to explore other possibilities towards consensus using our proposed concept of cluster selection.

References

1. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Engle wood Cliffs (1988)
2. Aggarwal, C. C. , Reddy, C.K.: Data Clustering: Algorithms and Applications, 1st ed. Chapman and Hall, London (2013). ISBN 9781466558212
3. Xu, D., Tian, Y.A.: Comprehensive survey of clustering algorithms. *Ann. Data. Sci.* **2**, 165–193 (2015)
4. Kleinberg, J.: An impossibility theorem for clustering. In: Proceedings of Advanced in Neural Information Processing Systems (2002)
5. Fred, A.L.N. , Roli, F., Kittler, J.: Finding consistent clusters in data partitions. In: Proceedings of 3rd International Workshop on Multiple Classifier Systems, vol. 2364, pp. 309–318 (2002)
6. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2002). <https://doi.org/10.1162/153244303321897735>
7. Zhou, Z.: Ensemble Methods: Foundations and Algorithms. CRC Press, Boca Raton, FL, USA (2012)
8. Huang, D., Wang, C., Lai, J.: Locally weighted ensemble clustering. *IEEE Trans. Cybern.* **48**(5), 1460–1473 (2018)
9. Fern, X.Z., Lin, W.: Cluster ensemble selection. *Statistical Anal. Data Mining* **1**(3), 128–141 (2008)
10. Huang, D., Lai, J., Wang, C.: Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis. *Neurocomputing* **170**, 240–250 (2015)
11. Dua, D., Graff, C.: UCI Machine learning repository. University of California, School of Information and Computer Science, Irvine, CA (2019). <https://archive.ics.uci.edu/ml>
12. Nguyen, N., Caruana, R.: Consensus clusterings. In: Proceedings of IEEE International Conference on Data Mining (ICDM), pp. 607–612 (2007). <https://doi.org/10.1109/ICDM.2007.73>

Human Activity Recognition Using Machine Learning: A Review



Ankita Biswal, Sarmistha Nanda, Chhabi Rani Panigrahi,
Sanjeev K. Cowlessur, and Bibudhendu Pati

Abstract With the enrichment of technologies, humans want to maximize automation by reducing the manpower and time, Human Activity Recognition (HAR) has a heterogeneous broad range of significant applications such as health care, theft detection, work monitoring in an organization and detecting emergencies. Various machine learning (ML) classification algorithms are applied on publicly available HAR datasets to recognize human activities in the literature. In this work, we have identified different HAR datasets involving different levels of activities and the methods for acquisition of data. We have also done a detailed review of HAR approaches with the implementation of various ML classifiers along with specific future directions in this area.

Keywords HAR · Sensor-based · Vision-based · Machine learning classifiers

1 Introduction

Humans are addicted to technology, and it has become a compulsive behaviour to them in the current era. With this vast enrichment of technical things, everyone wants to make the work easy with little effort and time. HAR is one of the classification

A. Biswal · S. Nanda (✉) · C. R. Panigrahi · B. Pati
Department of Computer Science, Rama Devi Women's University, Bhubaneswar, India
e-mail: sarmisthananda@gmail.com

A. Biswal
e-mail: ankitaangel94371@gmail.com

C. R. Panigrahi
e-mail: panigrahichhabi@gmail.com

B. Pati
e-mail: patibibudhendu@gmail.com

S. K. Cowlessur
Department of Software Engineering, Sanjeev K Cowlessur, Université Des Mascareignes (UdM),
Beau Bassin-Rose Hill, Mauritius
e-mail: scowlessur@udm.ac.mu

problems in the research area that can be implemented in many application areas such as assisted living, health care, emergency assistance, work monitoring in an organization, theft detection [1–4].

HAR is the problem of recognizing or classifying human activity. The activities of humans are classified from the data that is collected over a certain period and can be of *sensor-based* or *vision-based* [5]. In sensor-based data, the data is collected by using sensors such as accelerometers, gyroscopes and magnetometers. Video streaming or image data are examples of vision-based HAR data [6]. In either case, the data are collected for a while and then various ML classification algorithms such as random forest, logistic regression and support vector machine (SVM) are used to classify the activities [7, 8]. Once the activity is recognized with higher accuracy, the system can take suitable action as per requirement. The field of HAR is one of the contemporary topics due to the availability of low power consuming and low-cost sensors. These sensors can be embedded in objects, the human body, or in the environment to receive the raw data [9].

The rest part of the paper is organized as follows: Sect. 2 describes various HAR datasets that have been used in various applications and the methods of data acquisition, Sect. 3 presents a literature survey on HAR and Sect. 4 discusses some specific future directions and Sect. 5 concludes the paper.

2 HAR Datasets

The collection and sampling of data are some of the important aspects of accurate classification of activities. Human activities can be classified into hierarchical structure that indicates the levels of activity [10].

Primitive-level activity: Primitive-level activities are atomic actions such as stretching one-arm and raising the left leg. *Actions/Activities*: It is more complex than a primitive level. A collection of one or more primitive-level activities creates action such as walking, jogging and running. *Interaction-Level activity*: This level includes one or more person and an object, for example brushing, eating, cooking, playing games, etc.

The aim of HAR is to notice the daily behaviours of the people through the analysis of observations obtained and their neighbouring environments of living [11]. Data are collected from different types of sensors such as object sensors, body-worn sensors [12, 13] and ambient sensors. Other than sensors, activity recognition can also be done through vision-based systems in which data are collected from videos or images with the help of surveillance cameras. The detection of these activities can solve thousands of problems with automation systems to make life easier. The HAR dataset can be broadly classified into two types: *sensor-based* and *vision-based* and are described in the following subsections.

2.1 Sensor-Based Dataset

Kwon et al. used temporal segmentation [14] in their data collection system in which they segmented data of accelerometer before extraction. Chernbumroong et al. [15] collected the data from accelerometers, gyroscope with the wrist-worn sensors and sampled at a rate of 33 Hz. Zigel et al. [16] presented a system for fall detection where data is collected from the fall of a doll “Rescue Randy” and the system is based on sound sensing and floor vibration, uses signal processing and pattern recognition algorithm to distinguish between fall events and other events.

Table 1 presents different sensors that have been used in HAR datasets with their respective codes, and Table 2 presents some commonly used abbreviations for different sensor-based datasets.

For HAR, the sensor-based data is used in two ways, i.e. the data either may be collected by the researchers themselves or publicly available dataset. In HAPT and UCI datasets [21, 26–28], low-pass Butterworth filters with a 50 Hz cut-off frequency have been used. DHOP contains the motion data of 14 healthy older (66–86) people [19]. The activities were scripted using batteryless wearable sensors on top of their clothing. The collected data were noisy due to the use of the passive sensor. In most

Table 1 Different sensors with their code used in sensor-based datasets

Device	Code
Accelerometer	A
Body-worn sensor	B1
Batteryless body-worn sensor	B2
Gyroscope	C
Ambient sensor	D
Smartphone with built-in accelerometer and gyroscope	E
Smart watch	F
Object sensor	G

Table 2 Sensor-based dataset names with their abbreviations

Dataset name	Abbreviations used
Activity recognition from single chest-mounted accelerometer [17]	ARFCM
UCI dataset [18]	UCI
Dataset_Healthy_Older_People [19]	DHOP
Human activities and postural transitions [20]	HAPT
UCI-Human activity recognition [21]	UCI-HAR
Dataset for ADL recognition with wrist-worn accelerometer[22]	ADL-dataset
Mobile health [23]	MHEALTH
Activity recognition system based on Multisensor data fusion[24]	AReM
Wireless sensor data mining [25]	WISDM

Table 3 Sensor-based publicly available datasets

Dataset	Year	Device	Sampling	Attributes	Instances	Activities	Activity level
WISDM [25]	2019	E, F	20 Hz	6	15,630,426	18	Interaction
AReM [24]	2017	B1	20 Hz	6	42,240	7	Action/activities
UCI [18]	2016	E	50 Hz	561	5744	6	Action/activities
DHOP [19]	2016	B2	N/A	9	75,128	4	Interaction
HAPT [20]	2014	E	50 Hz	561	10,929	12	Action/activities
ADL_Dataset [22]	2014	A	32 Hz	3		14	Interaction
MHEALTH [23]	2014	B1	50 Hz	23	120	12	Action/activities
UCI-HAR [21]	2013	E	50 Hz	561	10,299	6	Action/activities
ARFCM [17]	2012	A	52 Hz	4	1,926,626	7	Interaction
OPPORTUNITY [29]	2010	D, G, B1	32 Hz	242	2551 701,366	7	Interaction

applications, only primitive-level datasets are collected by the authors themselves or the datasets [29] which have actions/activities or interaction levels they might include a category of gestures (naming varies according to the authors) which have the atomic actions.

The publicly available datasets for HAR are listed in Table 3. The data in AReM were collected from wearable sensors through a gateway which was using instances of communication to transmit data to other modules of the distributed system [24]. MHEALTH [23] data collected using used three wearable Shimmer2 sensors were used for data collection. OPPORTUNITY dataset has seven categories of activities containing 39 sub-categories. As OPPORTUNITY dataset captured real-world issues, there was no post-processing involved except alignment of sensor channels into common matrix format meaning the data of multiple sensors were synchronized then unit conversion of sensor data was performed [29].

2.2 Vision-Based Datasets

Vision-based datasets are more complex than sensor-based datasets as it contains both spatial and temporal information. Within a video, each frame holds important information, i.e. called spatial and also the context of that frame relative to the frames before spatial in time is temporal [30]. This type of extra information requires different network architecture as well as larger memories and more computation [31]. Authors in [32–34] presented the dataset of human activity recognition with the clips having resolution of 320×240 , 160×120 , 180×144 , respectively. The Human Motion DataBase (HMDB51) and Willow-Actions resolutions are 240 and 500 pixels, respectively. The vision-based dataset can be categorized into two types such as *static vision-based dataset* and *dynamic vision-based dataset*.

Table 4 Vision-based datasets

Dataset	Year	Sequences	Resolution	Sampling	Type	No. of activities	Level
UCF101 [32]	2012	13,320	320 × 240	25 fps	Dynamic	101	Interaction
HMDB_51 [36]	2011	6474	240 pixels	30 fps	Dynamic	51	Interaction
Willow-actions [35]	2010	911-images	At most 500 pixels	NA	Static	7	Interaction
KTH [33]	2004	2391	160 × 120	25 fps	Dynamic	6	Interaction
Weizmann [34]	2005	10	180 × 144	50 fps	Dynamic	10	Action/activities

Static Vision-based Dataset: Static as the name suggests these datasets are still in nature. These datasets include images of subjects. Willow-Action mentioned in Table 4 is an example of a static dataset. It includes Flickr images which represent variations of natural human actions in terms of camera viewpoint, clothing, human pose, occlusions and scene background [35].

Dynamic Vision-based Dataset: In this type of dataset, the subject changes its position constantly or have some movement in regular interval. Basically, this type of dataset contains videos where the participant or the temporal information changes at certain intervals. For this, the datasets KTH [37], HMDB51, Weizmann and UCF are sampled at a rate of 25, 30, 50 and 25 fps, respectively. Table 4 summarizes the publicly available vision-based datasets available in the literature.

3 Literature Study

In this section, a detailed review of HAR applications is done by considering different parameters such as dataset used, different ML classifiers applied, environmental setups, and the classification accuracy percentage obtained from different studies and is summarized in Table 5.

In [38], authors used classifiers such as FNN, PEF, PDF and PTN on different HAR datasets (MHEALTH, WISDM and SPAR) and from the obtained results by the authors it was found that FNN gives better results as compared to the other considered classifiers. For training the embedding, the categorical cross-entropy loss and triplet loss were used by the authors. They also found that subject triplet selection is a better option to improve the training. Yang et al. [39] presented a system of dual path CNNRNN cascade network (DPCRCN) which achieves end-to-end learning for classification. The authors used Adam as an optimizer with the ReLU activation function.

Wang et al. [40] used kinetic, HMDB and UCF101 datasets. The weights of the branch were learned in a supervised manner with standard backpropagation and the relation schema was integrated with an appearance branch which formed

Table 5 Summary of HAR applications

Author/year	Dataset used	Tools/framework	Classifiers used	Best classifier	Accuracy in %
Burns et al. [38]	MHEALTH, WISDM and SPAR	Keras, Seglearn, scikit-learn	FCN, PEF, PDF, PTN	PTN	MHEALTH: 99.9% WISDM: 91.3% SPAR: 99.0%
Yang et al. [39]	AReM	Dual path convolutional neural network, LSTM, Fully connected Layer, softmax	DPCRCN, LR, SVM, RF, XgBoost, DPCN, LSTM (6 layers), LI-SEN, IDNNs	DPCRCN	99.97
Wang et al. [40]	Kinetics, UCF101, HMDB51	Two stream CNN, 3D CNNs, ARTNets, SMART Blocks	ARTNet, C3D	ARTNet	Kinetics: 78.7 UCF101: 94.3 HMDB51: 70.9
Jain et al. [41]	UCI HAR, Physical activity Sensor data	Feature level fusion, Score level fusion	Multiclass SVM, k-NN	SVM	UCI HAR: 97.12, physical activity sensor data: 96.83
Denton et al. [42]	KTH	LUA	AE-LSTM, DRNET	DRNET	93.9
Walse et al. [43]	WISDM	WEKA, Adaboost.M1	Random Tree, Decision Stump, Hoeffding Tree, J48, RF, REP Tree	J48	97.83
Jiang et al. [44]	UCF101, CCV	VGG 19 network, CNN_M Model	CNN, LSTM, CNN + LSTM (Spatial and Motion) + Audio	CNN + LSTM (Spatial & Motion) + Audio	UCF101: 90.3, CCV: 82.4
Kutlay et al. [45]	MHEALTH	WEKA	MLP, SVM	MLP	91.70

SMART BLOCK and captured spatiotemporal information. ARTNet was constructed by stacking up multiple SMART BLOCKs which learned hierarchical spatiotemporal features. Jain et al. [41] used two publicly available sensor-based dataset that is UCI-HAR and physical activity sensor data for activity recognition. Authors used SVM and kNN as classifiers where SVM performed best for both datasets. The proposed approach does feature extraction by using a histogram of gradient and

centroid signature-based Fourier descriptor. Then feature and score-level fusion were combined for information fusion.

Denton et al. [42] presented a model based on a pair of encoders that factored video into content and pose. DRNET learned disentangled image representation means influences the temporal coherence of video and a novel adversarial loss to learn a representation that factorized each frame into a stationary part and temporally varying component. Walse et al. [43] used the MetaAdaboost.M1 classifier for classification with selected decision trees (random tree, Hoeffding tree, J48, Decision Stump, random forest, REP tree) classifier with default parameters. Each classification model was examined with tenfold cross-validation. In this, MetaAdaboost.M1 with J48 outperformed all other decision trees combined with MetaAdaboost.M1.

Jiang et al. [44] in their work first extracted motion, spatial and audio features using CNN. Then presented modelling of temporal dynamics in videos with long short-term memory (LSTM) as LSTM's two-stream approach focuses only on appearance and short-time motion information, which ignores long-term temporal dynamics in videos. In the next step, regularized feature fusion network was designed to model feature correlations and finally the detailed design of contextual refinement using two LSTMs and the regularized feature fusion network. Kutlay et al. [45] used the MHEALTH dataset for classification using multilayer perceptron (MLP) and SVM classifiers. The results obtained by the authors indicate that MLP gave an accuracy of 91.70% with 10 cross-validations. The MLP has 13 output units and 18 hidden nodes labelled Sigmoid node 0–30 [45]. The weights were given for each variable that fed into each sigmoid node.

Authors in [14] collected the data using a smartphone with an accelerometer sensor. They considered 11 activities from 3 locations such as 5 office activities, 3 kitchen activities and 3 outdoor activities. The temporal segmentation and feature extraction are done on the collected data. The ANN as classifier and ReLU as activation function are used in this model. Without location information, the model found an accuracy of 90%, whereas with location information the model found 95% accuracy. Authors in [15] tried to assist elderly people in their daily livings by recognizing their activities. MLP, RBF and SVM classifiers are used in the proposed model, which gives more than 90% classification accuracy. Here it is also provided that temperature sensor and altimeter can be combined with an accelerometer for data collection to give a better result. Authors in [46] proposed a model based on semantic reasoning and ontological modelling. The real-time data was collected by authors using multiple sensors. In this knowledge-driven approach, domain knowledge was used in the first phase. The second phase used ontologies for explicit context, and in the third phase, semantic modelling is used. Both fine-grained and coarse-grained activities are taken into consideration. In this approach, 94.44% average accuracy is achieved while recognizing the activities.

Zhu et al. [47] considered a smart assisted living system for assisting the elderly and disabled persons. They addressed the daily activity recognition and hand gesture recognition problem, neural network classification is implemented for gesture spotting, and for context-based based recognition, they used hierarchical hidden Markov model. In this model, they are claiming accuracy of 98.3%. In [48], a 3D video

tracking system is created to recognize human behaviour such as walking, sitting and speech. Cameras and ambient sensors are used to record the data. This framework uses the HMM classifier to classify the data. Here the overall recognition rate is 88.79%. Elderly people are passing through a high risk of falls. Considering this, authors in [16] presented an automated fall detection system. “Rescue Randy”, a human mimicking doll is used for human fall data simulation. In this case, they used a Bayes decision rule classifier and the experimental results obtained by the authors indicate a fall detection specificity of 98.6% and sensitivity of 97.5%.

4 Conclusion and Future Work

Classification is a well-known problem in ML. In this review, the authors tried to identify the publicly available benchmark datasets both vision-based and sensor-based data for HAR. They have also studied different HAR applications with implementation of different ML classification algorithms with consideration of various parameters like the type of dataset, classifiers used, best classifier, tools, types of sensors used and classification accuracy.

From this study, it is found that activities of HAR datasets are classified using many ML classifiers. But, the problem is to find the best fit classifier for a considered dataset to accurately identify human activity. The neural network is the current trend and is used for classification. Google’s assistant is one of its implementations and becoming robust day by day by learning its own. It has mostly five phases such as perceptron, activation function, cost function, optimization and backpropagation. The activation function can be step function, Sigmoid function, hyperbolic tangent, ReLU, softmax, etc. In place of cost function or loss function, cross-entropy, minimum likelihood, maximum likelihood, etc., may be used for HAR. The ML classifier can also be combined with soft computing approaches such as genetic algorithm, ant colony optimization and particle swarm optimization to improve the accuracy of the classifiers.

Acknowledgement This work is funded by CURIE Grant, Department of Science and Technology, Govt. of India under Grant no. DST/CURIE/01/2019(G) dt. 20.05.2019.

References

1. Lloret, J., Canovas, A., Sendra, S., Parra, L.: A smart communication architecture for ambient assisted living. *IEEE Commun. Mag.* **53**(1), 26–33 (2015)
2. Gannapathy, V.R., Ibrahim, T., Fayeez, A., Zakaria, Z., Othman, A.R., Latiff, A.: Zigbee-based smart fall detection and notification system with wearable sensor (e-SAFE). *Int. J. Res. Eng. Technol. (IJRET)* **2**(08):337–344
3. Lieser, P., Alhamoud, A., Nima, H., Richerzhagen, B., Huhle, S., Böhnstedt, D., Steinmetz, R.: Situation detection based on activity recognition in disaster scenarios. In: ISCRAM (2018)

4. Tripathi, R.K., Jalal, A.S., Agrawal, S.C.: Suspicious human activity recognition: a review. *Artif. Intell. Rev.* **50**(2), 283–339 (2018)
5. Kangas, M., Konttila, A., Lindgren, P., Winblad, I., Jämsä, T.: Comparison of low-complexity fall detection algorithms for body attached accelerometers. *Gait Posture* **28**(2), 285–291 (2008)
6. Poppe, R.: A survey on vision-based human action recognition. *Image Vis. Comput.* **28**(6), 976–990 (2010)
7. Fleury, A., Vacher, M., Noury, N.: SVM-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. *IEEE Trans. Inf Technol. Biomed.* **14**(2), 274–283 (2009)
8. Yang, J.-Y., Wang, J.-S., Chen, Y.-P.: Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. *Pattern Recogn. Lett.* **29**(16), 2213–2220 (2008)
9. Chahuara, P., Fleury, A., Portet, F., Vacher, M.: Using Markov logic network for on-line activity recognition from non-visual home automation sensors. In: *International Joint Conference on Ambient Intelligence*, pp. 177–192. Springer, Berlin (2012)
10. Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., Li, Z.: A review on human activity recognition using the vision-based method. *J. Healthcare Eng.* (2017)
11. Antar, A.D., Ahmed, M., Ahad, M.A.R.: Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: a review. In: *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pp. 134–139. IEEE (2019)
12. Casale, P., Pujol, O., Radeva, P.: Human activity recognition from accelerometer data using a wearable device. In: *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 289–296. Springer, Berlin (2011)
13. Casale, P., Pujol, O., Radeva, P.: Personalization and user verification in wearable systems using biometric walking patterns. *Pers. Ubiquit. Comput.* **16**(5), 563–580 (2012)
14. Kwon, M.-C., Choi, S.: Recognition of daily human activity using an artificial neural network and smartwatch. *Wirel. Commun. Mob. Comput.* (2018)
15. Chernbumroong, S., Cang, S., Atkins, A., Hongnian, Yu.: Elderly activities recognition and classification for applications in assisted living. *Expert Syst. Appl.* **40**(5), 1662–1674 (2013)
16. Zigel, Y., Litvak, D., Gannot, I.: A method for automatic fall detection of elderly people using floor vibrations and sound—Proof of concept on human mimicking doll falls. *IEEE Trans. Biomed. Eng.* **56**(12), 2858–2867 (2009)
17. Casale, P., Pujol, O., Radeva, P.: BeaStreamer-v0. 1: a new platform for multi-sensors data acquisition in wearable computing applications, pp 532–562 (2012)
18. Davis, K., Owusu, E., Bastani, V., Marcenaro, I., Hu, J., Regazzoni, C., Feijs, L.: Activity recognition based on inertial sensors for ambient assisted living. In: *2016 19th International Conference on Information Fusion (Fusion)*, pp. 371–378. IEEE (2016).
19. Torres, S., Luis, R., Visvanathan, R., Hoskins, S., Van den Hengel, A., Ranasinghe, D.C.: Effectiveness of a batteryless and wireless wearable sensor system for identifying bed and chair exits in healthy older people. *Sensors* **16**(4), 546 (2016)
20. Reyes-Ortiz, Jorge-Luis, Luca Oneto, Alessandro Ghio, Albert Samá, Davide Anguita, and Xavier Parra. “Human activity recognition on smartphones with awareness of basic activities and postural transitions.” In: *International Conference on Artificial Neural Networks*, pp. 177–184. Springer, Cham (2014)
21. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: A public domain dataset for human activity recognition using smartphones. In: *ESANN* (2013).
22. Bruno, B., Mastrogiovanni, F., Sgorbissa, A.: A public domain dataset for ADL recognition using wrist-placed accelerometers. In: *the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 738–743. IEEE (2014)
23. Banos, O., Garcia, R., Holgado-Terriza, J.A., Damas, M., Pomares, H., Rojas, I., Saez, A., Villalonga, C.: mHealthDroid: a novel framework for agile development of mobile health applications. In: *International Workshop on Ambient Assisted Living*, pp. 91–98. Springer, Cham (2014)

24. Palumbo, F., Gallicchio, C., Pucci, R., Micheli, A.: Activity Recognition system based on Multisensor data fusion (AReM) dataset (2017)
25. Weiss, G.M.: WISDM smartphone and smartwatch activity and biometrics dataset. UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set (2019)
26. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Energy efficient smartphone-based activity recognition using fixed-point arithmetic. *J. UCS* **19**(9), 1295–1314 (2013)
27. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: International Workshop on Ambient Assisted Living, pp. 216–223. Springer, Berlin (2012)
28. Reyes-Ortiz, J.L., Ghio, A., Parra, X., Anguita, D., Cabestany, J., Catala, A.: Human activity and motion disorder recognition: towards smarter interactive cognitive environments. In: ESANN (2013)
29. Roggen, D., Calatroni, A., Rossi, M., Holleczeck, T., Förster, K., Tröster, G., Lukowicz, P., et al.: Walk-through the OPPORTUNITY dataset for activity recognition in sensor rich environments. Helsinki, Finland, May 2010
30. <https://blog.coast.ai/continuous-online-video-classification-with-tensorflow-inception-and-a-raspberry-pi-785c8b1e13e1>. Last accessed on 10 June 2020
31. <https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5>. Last accessed on 10 June 2020
32. Soomro, K., Zamir, A.R., Shah, M.: A dataset of 101 human action classes from videos in the wild. Center for Research in Computer Vision 2 (2012)
33. Schudt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 3, pp. 32–36. IEEE (2004)
34. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE international conference on computer vision (ICCV'05) vol. 1, vol. 2, pp. 1395–1402. IEEE (2005)
35. Delaire, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations (2010)
36. Jhuang, H., Garrote, H., Poggio, H., Serre, T., Hmdb, T.: A large video database for human motion recognition. In Proceedings of IEEE International Conference on Computer Vision, vol. 4(5), p. 6. (2011).
37. <https://www.nada.kth.se/cvap/actions/>. Last accessed on 10 June 2020
38. Burns, D.M., Whyne, C.M.: Personalized Activity Recognition with Deep Triplet Embeddings. arXiv preprint [arXiv:2001.05517](https://arxiv.org/abs/2001.05517) (2020)
39. Yang, C., Jiang, W., Guo, Z.: Time series data classification based on dual path CNN-RNN cascade network. *IEEE Access* **7**, 155304–155312 (2019)
40. Wang, L., Li, W., Li, W., Van Gool, L.: Appearance-and-relation networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1430–1439 (2018)
41. Jain, A., Kanhangad, V.: Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sens. J.* **18**(3), 1169–1177 (2017)
42. Denton, E.L.: Unsupervised learning of disentangled representations from video. In: Advances in Neural Information Processing Systems, pp. 4414–4423 (2017)
43. Walse, K.H., Dharaskar, R.V., Thakare, V.M.: A study of human activity recognition using AdaBoost classifiers on WISDM dataset. *Inst. Integr. Omics Appl. Biotechnol. J.* **7**(2), 68–76 (2016)
44. Jiang, Y.-G., Zuxuan, Wu., Tang, J., Li, Z., Xue, X., Chang, S.-F.: Modeling multimodal clues in a hybrid deep learning framework for video classification. *IEEE Trans. Multimedia* **20**(11), 3137–3147 (2018)
45. Kutlay, M.A., Gagula-Palalic, S.: Application of machine learning in healthcare: analysis on mhealth dataset. *Southeast Europe J. Soft Comput.* **4**(2) (2016)

46. Chen, L., Nugent, C.D., Wang, H.: A knowledge-driven approach to activity recognition in smart homes. *IEEE Trans. Knowl. Data Eng.* **24**(6), 961–974 (2011)
47. Zhu, C., Sheng, W.: Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. *IEEE Trans. Syst. Man Cybern. Part A: Syst. Humans* **41**(3), 569–573 (2011)
48. Brdiczka, O., Langet, M., Maisonnasse, J., Crowley, J.L.: Detecting human behavior models from multimodal observation in a smart home. *IEEE Trans. Autom. Sci. Eng.* **6**(4), 588–597 (2008)

Advanced Computer Networks and Algorithms

Ensuring Secure Communication from an IoT Edge Device to a Server Through IoT Communication Protocols



Roshni Vidya S. Boolakee, Sandhya Armoogum, Ravi Foogooa, and Geerish Suddul

Abstract The Internet of things (IoT) is seen as one of the next Internet revolutions. More and more IoT devices are connecting to the Internet. These devices sense the physical world and perform some tasks in response to events occurring. Such devices are identified through unique IP addresses and often have to send data which might be sensitive and confidential over the network. Thus, the security of the data packets sent by IoT devices over the Internet has to be considered. The Constrained Application Protocol (CoAP) supports the RESTful HTTP functionalities and has been proposed specifically for IoT devices which are characterized by low processing power and energy. However, CoAP uses UDP protocol and must rely on the Datagram Transport Layer Security (DTLS) and sometimes on IPSec for security. In this paper, the focus has been on the practical evaluation of the different mechanisms that can be used to secure data packets being sent by IoT devices, to ensure that communication from the IoT node to the cloud/server is not compromised; and data is encrypted ensuring full end-to-end security. The IoT technologies communication protocols considered are transport layer and application layer protocols HTTP, CoAP, and FTP. A testbed was implemented which allow connectivity between a resource-constrained device, namely a Raspberry Pi, and a cloud server. Using this test bed, the different communication protocols for IoT secure connectivity have been assessed.

Keywords Internet of things · IoT · Constrained Application Protocol (CoAP) · Security · Transport layer · Energy consumption

R. V. S. Boolakee (✉) · S. Armoogum (✉) · R. Foogooa (✉) · G. Suddul (✉)
University of Technology, Mauritius, La Tour Koenig, Pointes-Aux-Sables, Mauritius
e-mail: rboolakee@umail.utm.ac.mu

S. Armoogum
e-mail: sandhya.armoogum@umail.utm.ac.mu

R. Foogooa
e-mail: rfoogooa@umail.utm.ac.mu

G. Suddul
e-mail: g.suddul@umail.utm.ac.mu

1 Introduction

Since the Internet of things (IoT) has been unleashed in the 1999, it is constantly growing, and today, it is being adopted in the fields of healthcare, agriculture, smart homes, smart cities, and military. With the advent of IPv6, connected objects can be uniquely identified and they ubiquitously connect to each other, and in our lifetime, we will experience life with a trillion-node network. These devices have sensing capabilities which allow collection of data on a real-time basis. Such data is often transmitted over the Internet through IoT communication protocols to servers. Since these devices will be intrinsically linked to human lives in the near future, security considerations are a high priority and there is a need for security mechanisms to enhance the security in IoT. IoT devices are resource-constrained equipment as compared to other devices connected in terms of power availability, storage, and processing capability. Thus, they may be more vulnerable to certain threats as they do not have the capability to sustain security mechanisms which have been developed for computer systems. In 2014, in a security report by HP, it was found that there was an average of twenty five vulnerabilities per IoT device. 80% of devices did not use strong passwords, typical password being “admin” or “password,” 70% did not encrypt data communications and 60% of devices had vulnerable firmware or user interfaces [14]. These can result in privilege escalation to hackers, who seek to gain control of these devices which lack security mechanisms and consequently exploiting them to cause attacks such as denial of service (DoS) attacks, distributed DoS (DDoS) attacks, buffer overflow, and battery exhaustion. A DDoS attack launched against the control systems, which regulated the heating of two apartment buildings in eastern Finland, resulted in the residents being left in the cold (with temperatures well below freezing) in October 2016 [1]. Such electricity disruptions due to cyber-attacks have also been reported in Israel in January 2016 [2] and in Turkey in December 2016 [3].

Smart homes are becoming more popular. Smart homes consist of a range of sensors (e.g., temperature, motion, light, etc.), systems (e.g., heating, lightning, security, etc.), and devices (e.g., smart home appliances such as smart meters, smart lights, smart thermostat, washing machine, refrigerator, microwave, etc.), which can be monitored, controlled, and automated through a smart phone or computer, locally or remotely, via the Internet. Such digital devices can communicate and regulate themselves based on the sensed physical state of the home. In [4], the authors identified the different attacks possible in an IoT-based smart home such as physical intrusion leading to burglary or ransom, death (i.e., block signal to sound alarm in case of fire), leak of behavioral patterns, leak of private data as IoT devices collecting and processing private information, extortion, and direct financial loss.

The key goal of IoT communication technologies is to provide seamless communication anywhere and anytime through any type of network [11–13]. However, it is essential to choose an appropriate method of communication to ensure accuracy and data integrity, efficient performance, energy consumption, and bandwidth usage. When data is transferred from one node to another, it is also important to secure these data with the use of cryptography. However, due to the resource constraint of IoT

devices, the amount of energy consumed when implementing encryption algorithms is an important consideration.

In this paper, we aim to investigate the communication protocols that can be used for transporting data from an IoT device to a server on the cloud and their security. The focus of this work is mainly on the CoAP protocol. A test bed is implemented which involves sensors connected to a Raspberry Pi (RP). The RP is itself connected to the cloud via the Internet. In Sect. 2, some related works are described; Sect. 3 presents our design and implementation of the test bed. Results of our evaluations are discussed in Sect. 4, and finally, Sect. 5 concludes the paper.

2 Background Study

2.1 CoAP

Communication protocols that can be used to transport data in IoT networks include HTTP, CoAP, and MQTT [8]. HTTP is the protocol used to transfer data between client and servers on the web and it can be used with SSL/TLS for security (https). Similarly, Constrained Application Protocol (CoAP), designed by the Internet Engineering Task Force (IETF), is meant for constrained-resource devices as a synchronous request/response web transfer protocol [5]. It was designed mainly to be lightweight machine-to-machine (M2M) asynchronous data exchange protocol and as a replacement of HTTP protocol for IoT network [9]. The CoAP protocol is based on the REST model and can be represented as consisting of two layers (*Messages* and *Request/Response*) running on the UDP transport layer and providing service to the upper application layer. Unlike HTTP, CoAP runs over UDP instead of TCP for unicast and multicast communication. The messages layer of CoAP interacts with UDP to transfer messages while the request/response layer manages application-related interactions. The UDP protocol being less reliable [6], CoAP supports four different message types namely, (1) confirmable, (2) non-confirmable, (3) acknowledgment, and (4) reset, to address for the unreliability of UDP protocol. Each CoAP message has a unique ID, which allows detecting message duplicates. When using CoAP, there is not predefined client and server roles. A device involved in machine-to-machine interactions can adopt both client and server roles at a particular point in time. CoAP uses the same commands as HTTP (GET, PUT, POST, and DELETE) to exchange data between the client and server sides on a one-to-one basis unlike MQTT which follows a publisher–subscriber approach.

In CoAP, a reliable message is sent as a confirmable (CON) message whereby the sender can be assured that the message will be received to the destination as such a CON message is sent multiple times until the receiver sends the acknowledgment message (ACK) which contains the same ID as the CON message. If the receiver cannot process the incoming request of the message, it responds with a reset message (RST) instead of the ACK. To support asynchronous communication, if the server

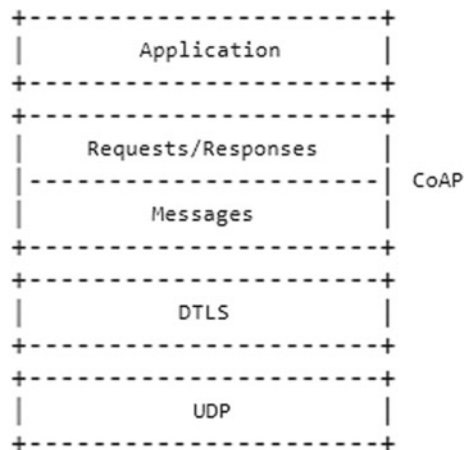
(receiver) cannot comply immediately with the request carried in a CON message, it can send an empty ACK message so that the client does not repeatedly send new CON requests. When the server is ready to send a response to this request, it will send a new CON message to the client. The client replies to the CON message with an ACK message. ACK message is not required for non-confirmable (NON) messages. Typically, real-time sensor values or application related data is exchanged as NON messages while commands are sent via CON messages..

2.2 Security in CoAP

Security is a required feature in IoT protocols to ensure the protection of commands and data which is transferred between IoT devices and/or the edge node or control server. HTTP often uses SSL/TLS for securing communication between client and server on the Internet. SSL/TLS supports authentication of communicating nodes, integrity of data, replay protection as well as confidentiality of data transferred. Given that TLS is implemented along with the TCP protocol and that CoAP uses the UDP protocol, TLS cannot be used to secure CoAP messages. CoAP uses Datagram TLS (DTLS) over UDP to provide security. DTLS supports the same security features as TLS as well as handling reordering and packet loss. Figure 1 depicts the layering when using DTLS with CoAP.

In this work, the *aiocoap* Python CoAP library was used to implement the CoAP protocol. Unfortunately, no suitable DTLS implementation library in Python was available at the time of this research work. There is the *tinyDTLS* Java library available for DTLS implementation but it implements limited functionality of DTLS.

Fig. 1 DTLS secured CoAP protocol stack



3 Testbed

The implemented test bed constitutes of an edge node, namely a Raspberry Pi 3 Model B (RP), which is connected to a Sense HAT (SH) device for collecting sensor data such as temperature, humidity, and pressure. Data collected from the SH by the RP is temporarily stored. Such data could be aggregated with other sensor data from various other sensors in an IoT network. Data is then to be transferred to the server on the Internet or cloud for storage, display, and analytics purposes. In this testbed, a Google Cloud server was implemented. Data received by the server is stored on a database (mySQL server). The database was named IoT project and consisted of four tables as follows: users, location, device, readings. Two roles are assumed: user and admin. Both the user and the administrator can access data that has been stored in the database. However, the system administrator can also interact directly with the RP via a VPN connection. The VPN connection is implemented using IPSec and IKEv2. The diagram in Fig. 2 depicts the system architecture.

RP is the CoAP client and the Google Cloud server the CoAP server. Figure 3 shows the state diagram of the CoAP client and Fig. 4 depicts the state diagram of the CoAP server.

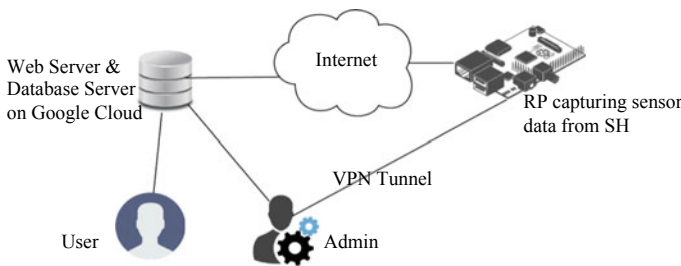


Fig. 2 System architecture

Fig. 3 State diagram of CoAP client

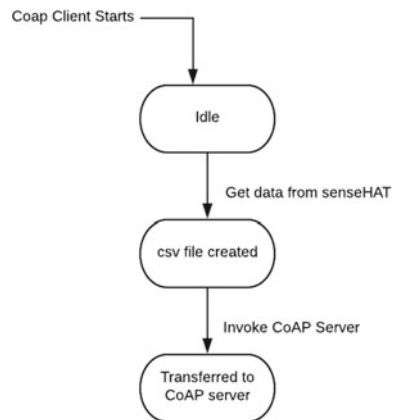
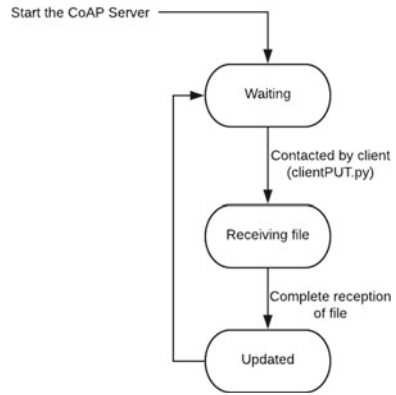


Fig. 4 State diagram of CoAP server



```

sense = SenseHat()
sense.clear()
deviceId="881"
date=time.strftime("%x")
time=time.strftime("%X")
t = sense.get_temperature()
p = sense.get_pressure()
h = sense.get_humidity()
t = round(t,1)
p = round(p,1)
h = round(h,1)
  
```

Date and time recorded when SenseHAT data is captured.

SenseHAT capturing data in the atmosphere.

Data captured is rounded to one decimal place.

Fig. 5 RP Code snippet for capturing sensor data

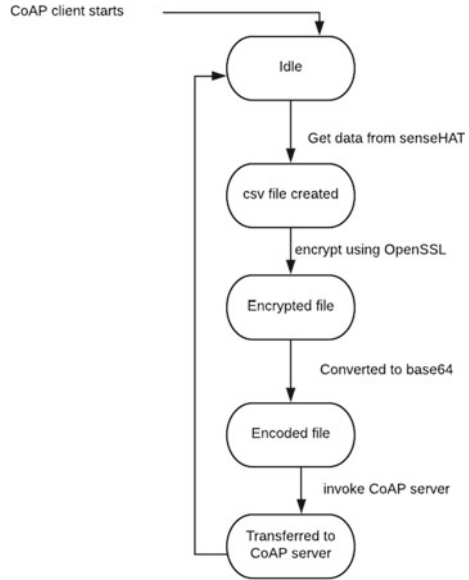
On the RP, SH data is captured as depicted in Fig. 5. The data captured (DeviceID, date, time, temperature, pressure, humidity) is concatenated and stored as a csv file (IoTdata.csv).

To ensure confidentiality of the sensor data, the OpenSSL is used. OpenSSL is the most popularly used toolkit for the SSL/TLS protocols. But it is also a general purpose cryptography library. Using the openssl library, the csv file was encrypted (encryptcsv.dat file). Further, to ensure that the encrypted file is readily received, it is encoded using base64 (encryptcsv.dec file). Figure 6 depicts the client side processing and Fig. 7 depicts the server side processing. Figure 8 shows the implementation of the encryption and encoding on the client side. The private key is saved in a file (private_key.dat) and is manually copied by the admin over the VPN connection. The private key is later configured on the server for decryption purposes.

The clientPUT.py script sends the encoded data to the server. A snippet of code is as shown in Fig. 9. The aiocoap package is used to implement CoAP on both the client and server [7, 10].

The server was created on a virtual machine (VM) with the Ubuntu operating system on the Google Cloud Platform (GCP) and it was allocated an IP address of 35.198.246.125. GCP supports firewall rules to allow or deny traffic to and from the VM instances set up, based on the port numbers and IP addresses configured. To allow ingress and egress traffic between the Raspberry Pi and the Google Cloud

Fig. 6 CoAP client side processing



server, the firewall rules were configured to allow duplex communication between the RP and the server using the HTTP, HTTPS, FTP, and CoAP protocol. FTP service was installed on the server and configured to allow file exchange (vsftpd.conf file).

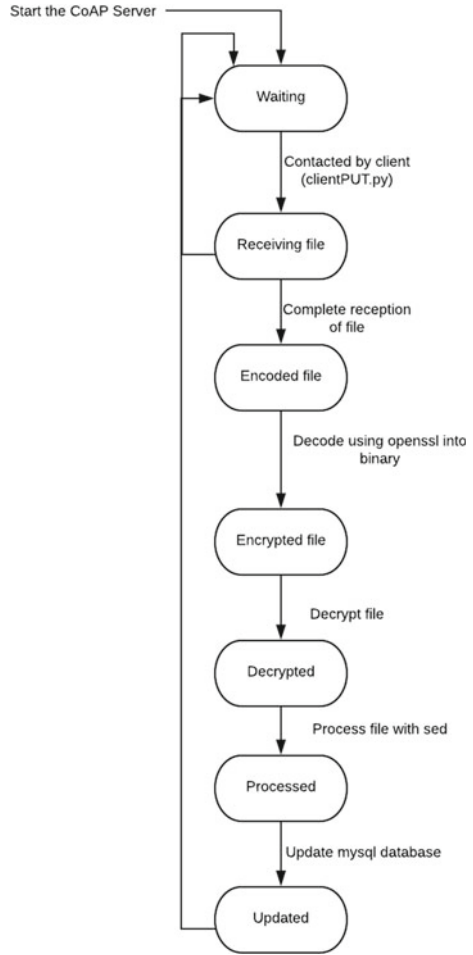
Similarly, the Apache web server (HTTP server) was installed on the VM (Ubuntu). The Apache web server registers itself with the uncomplicated firewall (UFW) and supports four application profiles namely Apache, Apache Full, Apache Secure, and OpenSSH. The Apache profile opens the port 80 (http traffic), while Apache secure opens only port 443 (SSL/TLS traffic). Because in this work, both HTTP and HTTPS is used, Apache Full is enabled (i.e., *sudo ufw allow "Apache Full"*) as it opens both port 80 and port 443. Furthermore, self-signed X.509 SSL certificates were generated using the OpenSSL library (2048 bits RSA keys). Aiocoap package is also installed to support CoAP communications. A shell script was written on the server to receive data from the RP. The file received is then decoded, decrypted and the data saved in the database.

Finally, the VPN connection is established between the admin browser and the RP using IPsec and IKEv2.

4 Experiments and Results

The aim of this research work is to compare the following protocols for transmitting sensor data from RP to a server: FTP, HTTP, HTTP/SSL, CoAP, and CoAP with data previously encrypted and encoded with OpenSSL and Base64. The experiments

Fig. 7 CoAP server side processing



```
openssl rsautl -encrypt -inkey public_key.pem -pubin -in iotdata.csv -out encryptcsv.dat
openssl base64 -in encryptcsv.dat -out encryptcsv.dec
for i in `seq 1 100`;
do
python3 clientPUT.py
echo "Running $i..."
done
```

Fig. 8 CoAP client side encryption and encoding

conducted and observations are shown in Table 1. In all experiments, Wireshark was used to sniff and capture data being transmitted.


```

with open('/home/pi/aiocoap/encryptcsv.dec') as csvfile:
    readCSV=csv.reader(csvfile,delimiter=',')
    for row in readCSV:
        str1=str1+row[0]+'***'

print("SenseHAT Data Sent to server")

str2=str.encode(str1)

payload = str2
request = Message(code=PUT, payload=payload, uri="coap://35.198.246.125/other/block")

response = await context.request(request).response
    
```

Fig. 9 CoAP client side transmission of data

Table 1 Experiments conducted using different protocols to transfer data from RP to server

Protocol	Experiment	Observation
FTP	Send file size of 10, 30, 50 MB	Successful transfer to server
		Plaintext data is seen
HTTP	Send file size of 10, 30, 50 MB	Successful transfer to server
		Plaintext data is seen
HTTP/SSL	Send file size of 10, 30, 50 MB	Successful transfer to server
		Payload is encrypted
CoAP	Send real-time sensor data	Successful transfer to server
		Plaintext data is seen
CoAP + Encryption + Encoding	Send real-time sensor data	Successful transfer to server payload is encrypted

5 Results and Analysis

Figure 10 depicts the time taken for transferring the files between the RP and the cloud server using FTP, HTTP, and HTTPS.

As can be observed, the time taken for transferring file using FTP is the fastest and the time taken when using HTTP with TLS is higher than when using HTTP. This is due to the burden of handshaking and encryption involved when using TLS. However, as seen in Table 1, data confidentiality is provided when using HTTPS. Figures 11 and 12 depict the percentage memory used for transferring data using HTTP and HTTPS, respectively. The percentage memory used by HTTPS is higher than with HTTP as expected. The average time to transfer files from the RP to the cloud server through FTP is approximately 17.5 s and the average memory used is approximately 84.13%. Using HTTP, the average time to transfer files from the RP to the cloud server is approximately 117 s and the average memory used is approximately 80.13%. Using HTTPS, the average time taken for communication is 141.33 s/MB and the average percentage memory used is 82.1 which is slightly

Fig. 10 Time taken to transfer files

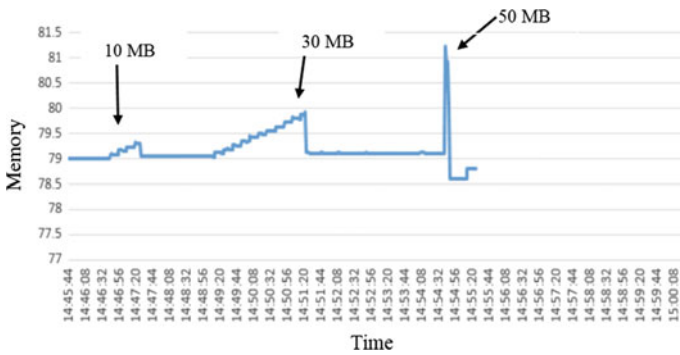
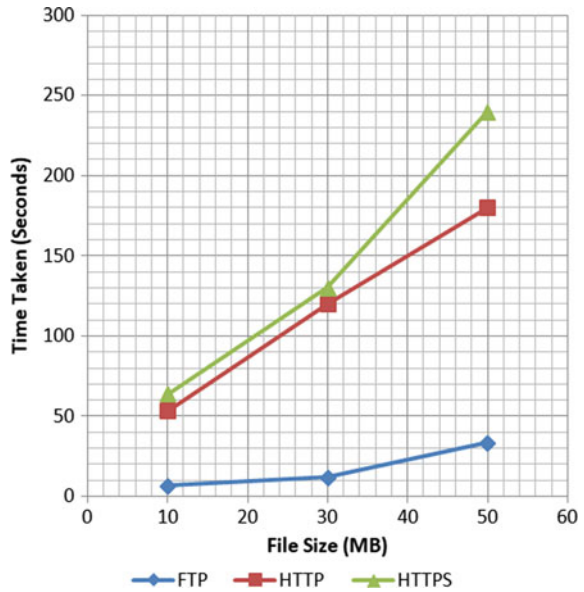


Fig. 11 Percentage memory used for communication using HTTP

higher than HTTP. It is also observed that the time taken increases almost linearly with increasing file size.

When experimenting with CoAP, as per the test bed described in Sect. 3, real-time data from the SH was transferred from RP to the cloud server. The size of the real-time sensor data transmitted on average was 0.0075 MB and the time taken for the data to be sent was around 57 s. The estimated average time for transferring data using CoAP is 12.6 s and the average percentage memory used is 82.5. Figure 13 depicts the comparative latency for the different protocols in milliseconds. As can be observed, the CoAP protocol can deliver data (encrypted data) at a faster rate than HTTP. In both cases, data confidentiality is maintained. Regarding memory usage,

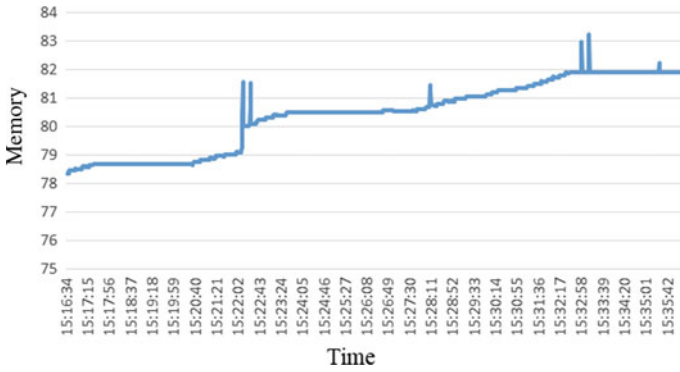


Fig. 12 Percentage memory used for communication using HTTPS

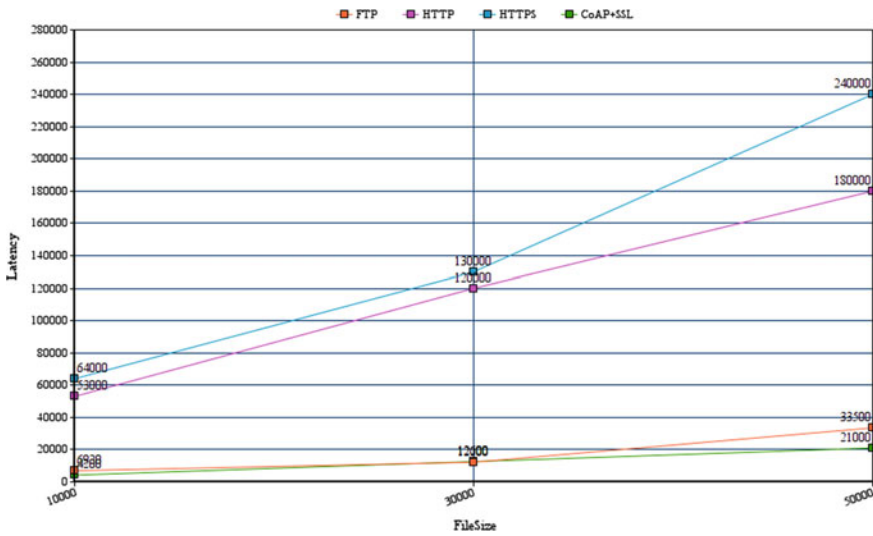


Fig. 13 Comparative latency

Figure 14 shows the comparative percentage of memory used for varying file size. The percentage memory when using CoAP is observed to be higher for small file size than when using the HTTPS protocol but lower for large file size.

6 Conclusions

This research work was based on the security of data being communicated between constrained IoT devices through the internet to a cloud server. An IoT device may not

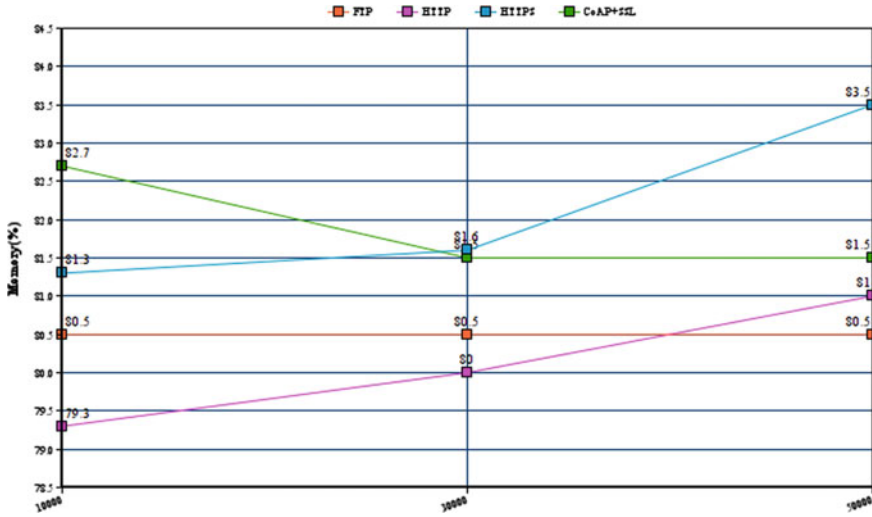


Fig. 14 Comparative percentage memory

have adequate resources to sustain the transport layer security mechanism (SSL/TLS) needed for the secure transmission of data across the internet. Therefore, this research attempts to study the adoption of the CoAP protocol of prior encrypted data. Using the CoAP protocol, data encrypted using the OpenSSL library was successfully transmitted from the Raspberry Pi (edge device which captures sensor data) to a server on the cloud. Data was encrypted prior to data transmission as no DTLS library implementation was available at the time of the project. To enable the proper transmission of encrypted data, the data also had to be encoded prior to transmission, base64 encoding was used. It was observed that CoAP has the lowest memory usage and latency compared to the HTTP and HTTPS protocols as the payload size increases. The performance of CoAP was observed to be closer to the performance of the FTP protocol which is optimized for file transfer. Overall, the CoAP protocol can be effectively used to transfer encrypted and encoded data at relatively low latency with acceptable system resources usage. This enables good performance and security of data transmitted between edge node and cloud server. Future work involves implementing CoAP with DTLS and comparing the performance in terms of time taken to transfer file and percentage memory used. File encryption and encoding will then be performed at the transport layer instead of the application layer.

References

1. Kumar, M.: DDoS Attack takes down central heating system amidst Winter in Finland (2016). [Online] Available at: <https://thehackernews.com/2016/11/heating-system-hacked.html>

2. Israel, T.: Israel's electric authority hit by 'severe' cyber-attack (2016). [Online] Available at: <https://www.timesofisrael.com/steinitz-israels-electric-authority-hit-by-severe-cyber-attack/>
3. Ankara and Kocaeli: Major cyber-attack on Turkish Energy Ministry claimed (2016). [Online] Available at: <https://www.hurriyetdailynews.com/major-cyber-attack-on-turkish-energy-ministry-claimed.aspx?PageID=238&NID=107981&NewsCatID=348>
4. Atlam, H.F., Alenezi, A., Alharthi, A., Walters, R.J., Wills, G.B.: Integration of cloud computing with internet of things: challenges and open issues. In: 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, 2017, pp. 670–675
5. Shelby, Z., Sensinode, Hartke, K.: Constrained Application Protocol (CoAP), draft-ietf-core-coap-18, 28 June 2013. <https://tools.ietf.org/html/draft-ietf-core-coap-18>
6. Albalas, F., Alsoud, M., Almomani, A., Almomani, O.: Security-aware CoAP application layer protocol for the internet of things using elliptic-curve cryptography. *Int. Arab J. Inform. Technol.* **15** (2018)
7. Canuto, L., Santos, L., Vieira, L., Gonçalves, R., Rabadão, C.: CoAP Flow signatures for the internet of things. In: 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), Coimbra, Portugal, pp. 1–6 (2019)
8. Martí, M., Garcia-Rubio, C., Campo, C.: Performance Evaluation of CoAP and MQTT_SN in an IoT Environment. In: Proceedings 13th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2019), vol 31, p. 49
9. Caturano, F., Jiménez, J., Romano, S.P.: Automated discovery of CoAP-enabled IoT devices. In: 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), Zagreb, Croatia, pp. 396–401 (2019). <https://doi.org/10.1109/ICUFN.2019.8806084>
10. aiocoap: The Python CoAP library, available at <https://aiocoap.readthedocs.io/en/latest/>, <https://git.azurewebsites.net/chrysn/aiocoap>
11. Singh, D., Pati, B., Panigrahi, C.R., Swagatika, S.: Security issues in IoT and their countermeasures in smart city applications. In: *Advanced Computing and Intelligent Engineering*, vol. 1089, pp. 301–313. Springer, Berlin (2020)
12. Rath, M., Pati, B.: Security assertion of IoT Devices using cloud of things perception. *Int. J. Interdiscip. Telecommun. Network.* **11**(4), 17–31 (2019)
13. Mishra, M., Choudhury, P., Pati, B.: Modified ride-NN optimizer for the IoT based plant disease detection. *J. Ambient Intell. Human Comput.* (2020). <https://doi.org/10.1007/s12652-020-02051-6>
14. Kovacs, E.: 70 Percent of IoT Devices Vulnerable to Cyberattacks: HP, July 29 (2014). available at <https://www.securityweek.com/70-iot-devices-vulnerable-cyberattacks-hp>

A QoI Assessment Framework for Participatory Crowdsourcing Systems



Ashley Rajoo, Kavi Kumar Khedo, and Utam Avinash Einstein Mungur

Abstract Participatory Crowdsourcing Systems have the potential to improve services in our daily life, such as health care, transportation and to monitor even the urban landscape using participatory sensing strategies. Data are the core mechanism that enables Participatory Crowdsourcing Systems to operate. It is very important to understand the evolution and relevance of data in Participatory Crowdsourcing Systems. Thus, this paper proposes a Quality of Information assessment framework which all Participatory Crowdsourcing Systems should strive to achieve to ensure data quality. The framework operates in a matrix schema that consists of four independent classes (horizontally) and has various dimensions within each class (vertically). The proposed framework will be flexible as it can incorporate new quality classes in the case of emerging technologies or domain areas. On the other hand, the vertical layer will have two subsections namely mandatory and desired features contained within a class.

Keywords Wireless systems · Participatory crowdsourcing systems · Quality of information

1 Introduction

A Participatory Crowdsourcing System (PCS) is characterized by a platform that links participant's mobile phones to a cloud service where data are collected and analyzed. In a PCS, data should be captured in real-time and the system should be able to define the relevance of the data collected for processing. This is so because the data collected within a PCS attempt to improve a domain area such as environmental, people or even task-oriented. Thus, there is a need to understand the Quality of Information (QoI) in a PCS which will define the relevance of data captured. Users often capture data while they are not connected to the crowdsourcing platform and, once connected, the data are sent to the cloud system. Therefore, timeliness and validity of the data

A. Rajoo (✉) · K. K. Khedo · U. A. E. Mungur
Faculty of Information, Communication and Digital Technologies, University of Mauritius,
Reduit, Mauritius
e-mail: ashleyrajoo10@yahoo.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_29

351

will need to be checked before they can be used. Furthermore, there are security issues that need to be catered for in a PCS. Hackers could connect to the PCS and gain vital information or even share malicious data. Participants should be able to connect to a PCS and post their views or suggestions. Thus it is very important to understand the evolution and relevance of data in a PCS. Most PCS tend to focus on the final results that are giving appropriate and adequate services after capturing data. Therefore, services provided tend to be partially fulfilled. It is critical to understand how data is treated from start to finish in such a system. PCS need to adhere to a framework which assesses quality factors before processing data. As a result, this paper proposes a flexible Quality of Information assessment framework for PCS. The framework operates in a matrix schema. The proposed framework will be flexible as it can incorporate new classes in case of emerging technologies or domain area and can also integrate and accept new dimensions of QoI.

This paper is organized as follows: Sect. 2 presents the research methodology and Sect. 3 outlines challenges for the participatory crowdsourcing system. Section 4 presents a taxonomy of quality of information related to PCS. Section 5 proposes the QoI assessment framework for PCS and Sect. 6 describes the metric for the different dimensions in the QoI assessment framework. Section 7 evaluate the Framework and Sect. 8 concludes this paper.

2 Research Methodology

The methodology used in this paper consists of four steps. Firstly, grouping of different PCSs in different domain areas based on the inclusion and exclusion criteria. From this selection process, two main research questions were identified, namely grouping of PCSs and QoI factors which are discussed in Sect. 3. Three major categories of PCSs were identified, namely (1) environmental participatory sensing systems, (2) people-oriented participatory sensing systems and (3) task-oriented participatory sensing systems. The data requirements and core features of each category were reviewed. Secondly, the different dimensions of QoI in these PCSs were analyzed. Thirdly, the common dimensions of QoI in the different domain areas of the PCSs were identified and grouped. From the selection process, while grouping the dimension of QoI, it has been noted that data had four main cycles namely data capture, data transmission, data storage, and data fusion. Fourthly a flexible QoI assessment framework that worked in a matrix arrangement was derived. The framework operated both horizontally and vertically. It was evaluated using an environmental PCS case study. Thus, this paper will help researchers in PCSs to have a complete view of the different QoI dimensions that influences PCSs. In this way, the proposed framework will provide a solid foundation for the assessment of QoI in PCSs and the framework can be integrated with different application domain areas.

3 Challenges for Participatory Crowdsourcing Systems

There are different dimensions of Quality of information which are core factors in a PCS. Over the years, researchers have been able to identify and define some quality factors in a PCS [4, 6, 8, 17]. These quality factors are highly dependent on the domain area where one dimension can have a different interpretation in a PCS. Thus, there is a need to define a quality assessment framework that can be integrated into a PCS. Moreover, it is important to understand the evolution and journey of data in a PCS [13, 14]. As mentioned in step three of the research methodology, it has been identified that data manipulation in any domain area of a PCS has four main cycles, namely: data capture, data transmission, data storage and data fusion. The challenge is to define each cycle's dimension of QoI and to ensure that the framework can incorporate them together. Additionally, when data are sent for processing, it is important to check if the data that have been captured are still relevant. For example, if a participant has captured data and has disconnected to the network, therefore it is important to check the reliability of the data. A research gap analysis in QoI for PCS has been performed where some core dimensions have been identified namely data reliability, accessibility, security, trustworthiness, accuracy and consistency.

4 Taxonomy of Quality of Information

In this section, a taxonomy of QoI in a PCS is discussed. Firstly, the QoI for PCS has been defined and secondly, the dimensions of QoI have been proposed.

4.1 *Quality of Information in PCS*

To understand QoI in PCS, the evolution and journey of data [13, 14] has to be analyzed. After having grouped the PCS in different domain areas, it has been identified that data has two major integrations in a PCS. These integrations have been categorized as horizontal and vertical functionalities. Horizontal functionality is a core mechanism defined as four classes namely, data capture, data transmission, data storage and data fusion. The four classes represent the usage of data in a PCS, that is, what is the importance and the role of data in the mechanism of a PCS.

The four classes are the followings:

1. **Data Capture:** refers to all requirements/standards/policies that must be fulfilled. This class assesses whether or not data need to be captured, irrelevant to a domain area. It will capture data in real-time and check if the expected requirements of a PCS meet the actual requirements of the PCS.

2. **Data Transmission:** prerequisite that QoI must be considered before any data or service transfer. This class determines the relevance of data to be sent to users or server and if there are any changes in the environment or time in a PCS.
3. **Data Storage:** applicability of data. This class assesses the importance of storing data as in a PCS, services are provided in real-time after data have been captured live. Therefore, storing any data will be a critical factor.
4. **Data Fusion:** data as a service/ service-oriented using primary data. This class will be responsible for bridging data from different sources as data is captured from different intervals and places with different technologies.

4.2 *Dimensions of QoI in Participatory Crowdsourcing Systems*

The vertical functionality is the baseline of each class. These baselines have been identified after grouping different PCSs and the dimensions are not dependent on any domain area. Table 1 shows the different dimensions [3, 10, 11] for participatory crowdsourcing systems.

5 **Proposed QoI Assessment Framework for PCS**

As mentioned in Sect. 1, the framework operates as a matrix schema. It consists of four independent classes in the horizontal layer. The proposed framework will be flexible to incorporate new classes in the case of emerging technologies or domain areas [7, 12, 15, 16]. Each class consists of a vertical layer with various dimensions that are either mandatory or desirable for a PCS to fulfill. The mandatory features are the critical dimensions that a PCS has to meet independently of any domain area. On the other hand, the desired features will be dependent on a domain area and will vary depending on the scenarios. For example, the desired features for an environmental PCS will differ from a people PCS. Nevertheless, both PCS will be required to fulfill the mandatory features. These amendments to the horizontal and vertical functionalities make the framework flexible to any domain. For the mandatory features, if one of the dimensions is missing, then the data journey that is relevant for the class will not be processed. Besides, there is an acceptable threshold which could be tolerated in case of a missing dimension which is discussed in Sect. 7. On the other hand, for desired features, if one of the dimensions is missing or partially fulfilled, depending on the domain area, the data will still need to be crosschecked for the mandatory features as only the core features are considered critical. Figure 1 shows a detailed view of all the dimensions mentioned for the framework.

From Fig. 1, all rectangles which are filled border are the mandatory features whereas the dotted border represents the desired features. A brief explanation is provided in Sect. 6 about the relevance of all the dimensions mentioned.

Table 1 Dimension of QoI

Dimensions	Definitions
Accessibility	The nature of having the option to be originated to or entered
Accuracy	It is a requirement that is desired before any processing to provide adequate services in a participatory application
Passable	Especially suitable or compatible level of information related to a task
Completeness	The state or state of having all the fundamental or suitable parts
Confidentiality	Degree to which data comply with policies that ensure they are accessed and used by authorized users
Concise representation	Marked by brevity of expression or statement related to information
Unambiguous data	Understanding of ideal components has been gained to handle information
Unaffected data	Information being correct and reliable
Interpretability	Information being clear using appropriate language
Objectivity	Information being unbiased
Real-Time Latency	It is the delay between input and output such that it is nonexistent. The latency must be low that is the system must react in real-time
Reliability	It is defined as the ratio of data being collected that produces correct output to the total number of participants
Relevancy	The quality or condition of being firmly associated or proper
Status	Information highly related to source and content
Security	Information being restricted from unauthorized persons
Rightness	Information related to time, up-to-date
Trustworthiness	It checks whether a malicious person will send irrelevant data or data affected by noise that is sent to the server
Understandability	Information that can be interpreted

6 Proposed Metric for Dimension in the QoI Assessment Framework

In this section, only the mandatory features of the framework are described and formulated. Dimensions are grouped from the four classes mentioned namely: data capture, data transmission, data storage, and data fusion. Two common constants are derived from past research which are AP (Active Participant) and TP (Total Participant) [1, 5, 9] in the PCS. These two constants are important because most of the dimensions are directly linked to the percentage of active users. As mentioned, only mandatory features have been explained and formulated whereas desired features will differ from domain area (Tables 2, 3, 4 and 5).

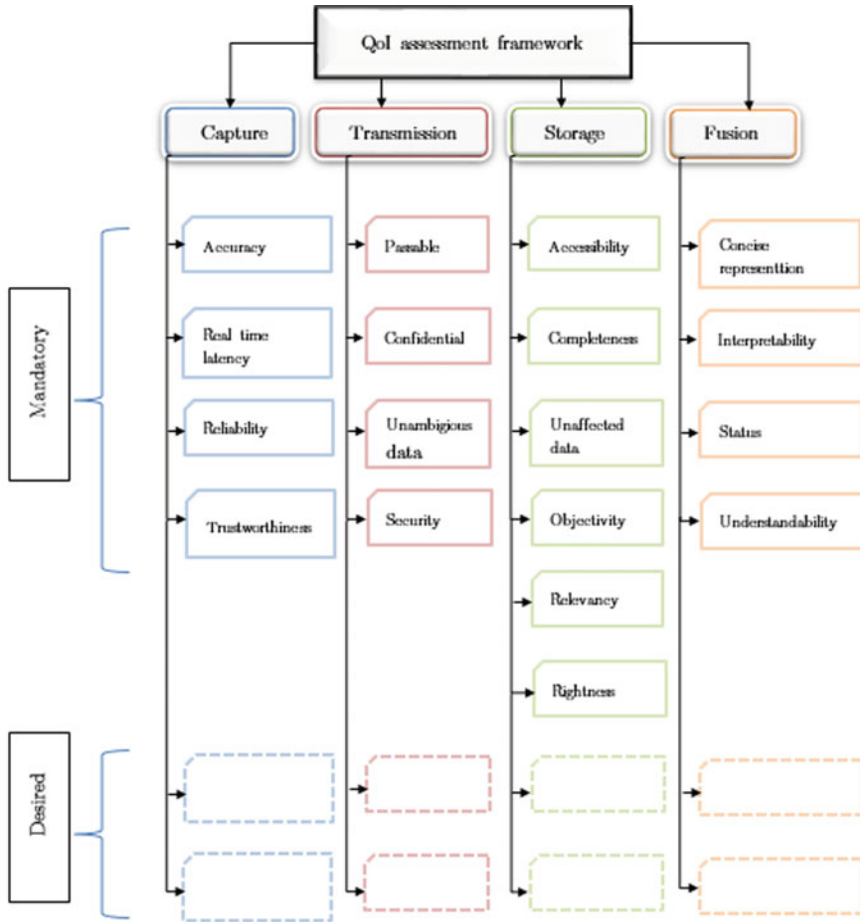


Fig. 1 Detailed view of the framework

7 Evaluation of the Proposed QoI Assessment Framework

This section describes the evaluation of the proposed QoI assessment framework related to an environmental PCS. The evaluation is conducted through a case study such that the framework is assessed horizontally and vertically. The case study is based on CrowdPic [2]: An interactive and Selective Picture Collection Framework for Participatory Sensing Systems. The PCS's contribution is to reduce the client-server communication cost by elimination of redundant image and strength is checking and removal of redundant image.

The QoI assessment framework is evaluated by answering five questions.

1. How data are assessed by the framework?

Table 2 Data capture

$\text{Accuracy} = \frac{\text{TEC}}{\text{TER}} * \frac{\text{AP}}{\text{TP}} * 100$	<p>TEC: Total events correctly detected by the participatory sensing system TER: Total events occurred in reality</p>	<p>Accuracy is linked with all events. The framework will take all event correctly occurred upon all events detected</p>
$\text{Real -Time Latency} = \frac{(\text{TRE}-\text{TAD})}{\text{MT}} * \frac{\text{AP}}{\text{TP}} * 100$	<p>TRE: Time of occurrence of a real event TAD: Time of arrival to the destination of the data MT: Maximum time in which the data becomes outdated</p>	<p>Real-time will subtract the time of transmission from the received to check whether data is outdated. The framework will take into consideration the latency</p>
$\text{Reliability} = \frac{(\text{TEC})}{(\text{TER})} - \text{TTAD} * \frac{\text{AP}}{\text{TP}} * 100$	<p>TEC: Total events correctly detected by the participatory sensing system TTAD: Total Time of arrival to the destination of the data TER: Total events occurred in reality</p>	<p>All events correctly occurred upon all events detected is subtracted from total time arrival of all corrected events</p>
<p>Trustworthiness = $\frac{\text{TADF}}{\text{ADC}} > \frac{\text{AP}}{\text{TP}} * 100$ AND Rightness</p>	<p>TADF: Total amount of data having correct format ADC: Amount of data collected</p>	<p>The summation of data having the correct format upon the amount of data collected is calculated. And it should have all the required standards and polices</p>

The assessment framework has two scenarios for the four classes individually. The first one is considered as a certainty, where a class is having more than 75% of the mandatory features fulfilled depending on the number of active participants. Therefore, data are processed in this scenario. The second one is defined as hold where data are not processed as from active participant (only 60% to 75% of mandatory features are fulfilled). For these two considerations, the framework is assessing data for one class at a time. Additionally, the framework will reassess the data and if certainty is achieved, data are processed but if the percentage is below hold, the class will discard the data.

2. What is the accepted threshold for mandatory features?
 The acceptable threshold for the mandatory features is above 80% as a user can share the same data a number of times. This percentage is critical as it is dependent on active participants. Thus, a user can share the data where all the variables have been fulfilled as per the metric mentioned in Sect. 6. However, the data is not considered if the constant, AP, active participant is too low. Additionally, once the data is accepted as per the policies, the data will be considered as per spatiotemporal requirements in case a user shares the same data.
3. What is the accepted threshold for the desired features?
 Compared to the mandatory features, the acceptable threshold for the desired features is 100% if one dimension is identified. Otherwise 50% if more than one

Table 3 Data transmission

$\text{Passable} = \frac{\text{ADU}}{\text{ADN}} * \frac{\text{AP}}{\text{TP}} * 100$	<p>ADU: Amount of data used by the participatory sensing system to meet the requirement in real-time ADN: Amount of data needed to meet present reality</p>	<p>The expected amount of data required by the system upon the actual amount of data required by the system is calculated</p>
$\text{Confidentiality} = \frac{\text{TPPF}}{\text{TCP}} * \frac{\text{AP}}{\text{TP}} * 100$	<p>Checklist of policies established in participatory sensing system Q1: Are data encrypted? Q2: Are data protected during storage? Q3: Are data manipulated by authorized users through access control? Q4: Is there any record for data being manipulated? Q5: Are users' session expired during inactivity? TPPF: Total Privacy policies fulfilled TCP: Total confidentiality policies</p>	<p>First a set of policies need to be checked. From the list, the total policies fulfilled is calculated upon the total policies set</p>
$\text{Unambiguous} = \frac{\text{ADT}}{\text{ADC}} * \frac{\text{AP}}{\text{TP}} * 100$	<p>ADT: Amount of data collected for a specific task ADC: Amount of data collected</p>	<p>All data related to one task are calculated upon the total amount of data collected</p>
<p>Security = Rightness AND Trustworthiness AND Reliability AND Accuracy</p>		<p>Security has four critical dimensions namely, Rightness, Trustworthiness, and Accuracy. These have been defined in Table 1</p>

Table 4 Data storage

$\text{Accessibility} = \frac{\text{ADAU}}{\text{ADA}} * \frac{\text{AP}}{\text{TP}} * 100$	ADAU: Amount of data accessed by users ADA: Amount of data to be accessed in a participatory sensing system	The amount of data accessed by users is calculated upon the amount of data the PCS required for a task
$\text{Completeness} = \frac{\text{TADF}}{\text{ADN}} * \frac{\text{AP}}{\text{TP}} * 100$	TADF: Total amount of data having correct format ADN: Amount of data needed to meet present reality	The total amount of data having the correct format is calculated upon the amount of data required by the PCS
$\text{Unaffected data} = \frac{\text{TADF}}{\text{ADC}} > \frac{\text{AP}}{\text{TP}} * 100$	TADF: Total amount of data having correct format ADC: Amount of data collected	The total amount of data having the correct format is calculated upon the amount of data collected
$\text{Objectivity} = \text{ADC} \geq \text{TP}$	ADC: Amount of data collected	The amount of data collected should be greater than the total participant in the PCS
$\text{Relevancy} = \frac{\text{ADU} - \text{TADF}}{\text{ADN}} * \frac{\text{AP}}{\text{TP}} * 100$	TADF: Total amount of data having correct format ADU: Amount of data used by the participatory sensing system to meet the requirement in real-time ADN: Amount of data needed to meet present reality	The difference between the expected amount of data required by the system and total amount of data having the correct format is calculated upon the amount of data needed by the PCS
Rightness = 100% if data collected meets all standards Rightness = 0%, if not met		All data captured meeting full standards by the PCS is considered as rightness. Else, data is not considered as rightness

dimension is identified. One consideration is that the desired features can vary. It is not fixed as the mandatory features. This is applicable to every single class. Based from this case study, the desired features for each class would be.

- Capture → Precision
- Transmission → Throughput, Affordability
- Storage → Precision, Affordability
- Fusion → Tunability

From the above, some of the dimensions can be interrelated which is not the case for mandatory features. Additionally, if the capture and data fusion are analyzed, the threshold would be 100% whereas, for data storage and transmission, the threshold would be 50%. For another domain, the desired features could change.

4. How flexible is the framework?

The framework can be reorganized if there is a need. For the mandatory field, only four classes have been identified due to existing research trends. But if a new technology or new domain areas have emerged, then additional classes can be added depending on the relevance of the data. This process will not affect

Table 5 Data fusion

<p>Concise representation = $\frac{ADN}{ADC} * \frac{AP}{TP} * 100$</p>	<p>ADN: Amount of data needed to meet present reality ADC: Amount of data collected</p>	<p>The amount of data required by the PCS is calculated upon the amount of data collected by the PCS</p>
<p>Interpretability = $\frac{ADF}{ADC} * \frac{AP}{TP} * 100$</p>	<p>ADF: Amount of data having correct the format ADC: Amount of data collected</p>	<p>The amount of data having correct format related to one task is calculated upon the amount of data collected</p>
<p>Status = $\frac{ADF}{ADN} * \frac{AP}{TP} * 100$</p>	<p>ADF: Amount of data having correct the format ADN: Amount of data needed to meet present reality</p>	<p>The amount of data having correct format related to one task is calculated upon the amount of data required by the PCS</p>
<p>Understandability = $\frac{TCAF}{TCA} * \frac{AP}{TP} * 100$</p>	<p>Checklist understandability attributes for participatory sensing system Q1: Are data collected having attributes, labels, and titles? Q2: Are data with labels and titles being intuitive and descriptive? Q3: Are data collected in the correct format? Q4: Data/information being distributed to users is in an appropriate format? Q5: Data collected are in a consistent order? TCAF: Total compressibility attributes fulfilled TCA: Total comprehensibility attributes</p>	<p>First a set of policies needs to be checked. From the list, the total comprehensive attributes fulfilled is calculated upon comprehensive attributes set</p>

any PCSs as each class works independently. For example, if a class has been identified and data are captured, nothing in terms of assessment is affected as the new class is not related to this existing one. If a new dimension is added to the mandatory feature of data capture, the framework will have the same principles as illustrated by Question number 2. For the desired features, it does not affect the framework as it will have the same principles as mentioned in Question number 3.

5. How scalable is the framework?

The framework can be considered as scalable because each class is processing data individually. As mentioned in the last part of Question number 2, data are continuously being assessed. Therefore, the more users are connected, the framework is not overloaded or is not resource intensive. Instead, the framework will continuously sense for data. Nevertheless, one important aspect to consider

is, at what interval should the data be sensed? This challenge will be considered for future upgrade of the framework.

8 Conclusion

A flexible and adaptive QoI Assessment framework is presented in this paper for PCSs. The framework has been documented and evaluated from concept to design. The contribution of this framework is to assess data quality in any PCS before data processing. This has been achieved after having grouped, analyzed and evaluated the different PCS. Moreover, a set of metrics for assessing the different QoI dimensions have been proposed. The framework has been reviewed through a case study where five prominent problems (review questions) are defined and answered. Scenarios with acceptable threshold for mandatory and desired features have been discussed. Furthermore, the framework is flexible enough to incorporate new QoI dimensions in the future. Finally, the proposed Quality of Information assessment framework provides a solid foundation for the assessment of data quality in Participatory Crowdsourcing Systems which will help researchers in PCSs to have a complete view of the different QoI dimensions that influences PCSs.

References

1. Auger, A., Exposito, E. and Lochin, E.: Generic Framework for qualitybased autonomic adaptation within sensorbased systems. In: International Conference on ServiceOriented Computing, p. 2132. Springer, Cham (2016)
2. Chen, H., Guo, B., Yu, Z., Huangfu, S., Nan, W., Wu, W.: Crowdpic: an interactive and selective picture collection framework for participatory sensing systems. In: 2014 IEEE International Conference on Computer and Information Technology, p. 512519. IEEE (2014)
3. Günther, L.C., Colangelo, E., Wiendahl, H.H., Bauer, C.: Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises. *Procedia Manuf.* **29**, 583–591 (2019)
4. Kosba, A.E., Saeed, A., Youssef, M.: Rasid: A robust wlan devicefree passive motion detection system. In: 2012 IEEE International Conference on Pervasive Computing and Communications, p. 180189. IEEE (2012)
5. Kumar, B., Rani, S., Singh, P.: A critical study of existing approaches based on quality of information attributes and metrics in wireless sensor network. *Int. J. Data Network Secur.* **4**(1), 152161 (2013)
6. Laaboudi, Y., Olivereau, A., Oualha, N.: June. An Intrusion Detection and response scheme for CP-ABE-encrypted IoT networks. In: 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS), pp. 1–5. IEEE (2019)
7. Lavric, A., Petrariu, A.I., Popa, V.: Long range sigfox communication protocol scalability analysis under large-scale, high-density conditions. *IEEE Access* **7**, 35816–35825 (2019)
8. Mashhadi, A.J., Capra, L.: September. Quality control for realtime ubiquitous crowdsourcing. In: Proceedings of the 2nd International Workshop on Ubiquitous Crowdsourcing, p. 58 (2011)
9. Panayiotou, T., Manousakis, K., Chatzis, S.P., Ellinas, G.: A data-driven bandwidth allocation framework with QoS considerations for EONs. *J. Lightwave Technol.* **37**(9), 1853–1864 (2019)

10. Parvathi, C., Talanki, S.: Bio-Inspired Scheme of Killer Whale hunting-based behaviour for enhancing performance of wireless sensor network. In: *Data Engineering and Communication Technology*, pp. 341–357. Springer, Singapore (2020)
11. Prathiba, B., Sankar, K.J., Sumalatha, V.: Novel framework of retaining maximum data quality and energy efficiency in reconfigurable wireless sensor network. *Int. J. Electr. Comput. Eng.* **2088–8708**, 9 (2019)
12. Rathore, P., Dhaka, K., Bose, S.K.: Network coding assisted multicasting in multi-hop wireless networks. *Comput. Commun.* **138**, 45–53 (2019)
13. Sachidananda, V., Khelil, A., Suri, N.: Quality of information in wireless sensor networks: a survey. In: *ICIQ (to appear)* (2010)
14. Setiyawati, H., Doktoralina, C.: The importance of quality accounting information management in regional governments in Indonesia. *Manage. Sci. Lett.* **9**(12), 2083–2092 (2019)
15. Vignesh, V., Premalatha, K.: Multi path interference aware reliable route path establishment in mobile adhoc network environment. *Cluster Comput.* **22**(5), 11029–11037 (2019)
16. Wu, Q., Mei, W., Zhang, R.: Safeguarding wireless network with UAVs: a physical layer security perspective. *IEEE Wirel. Commun.* **26**(5), 12–18 (2019)
17. Yang, M.: Software-defined networking-based adaptive resource allocation in optical networks (2019)

An Approach to Personalize VMware vSphere Hypervisor (ESXi) Using HPE Image Streamer



Richa and Jyoti Singh

Abstract Server virtualization technology is an automation method to control and monitor task running in multiple virtual machines. It uses software to divide a physical server into multiple virtual machines. Each virtual machine runs its own operating system. It increases the effective use of server. The VMware ESXi operating system can be personalized with required designs, updates, and patches. This study focuses on personalization of the ESXi operating system with the customized application stack. The bare metal compute hardware boots up directly into this application stack and is ready to host virtual machines on it. The HPE Image Streamer is used to host, configure, and serve the operating systems to HPE Synergy compute modules.

Keywords Virtual machine · Elastic sky X integrated (ESXi) · Virtual machine kernel · Virtual switch · Port group · Image streamer

1 Introduction

Server virtualization is a new technology in which a physical server is divided into multiple virtual servers using software application. A physical server acts as a host, and its memory storage is effectively used by more than one virtual client servers [1].

As per Chaubal [2], “The Architecture of VMware ESXi,” 2008, VMware ESXi, unlike the existing Linux-based hypervisors is not open source, it uses VMware’s management platform. This in turn calls for a detailed knowledge about its hardware requirements, VMkernel architecture and its networking layer, the deployment and management networks, multipathing with software or hardware iSCSI, etc., to be able to virtualize the existing HPE Synergy compute modules with the help of HPE

Richa (✉)

Hewlett Packard Enterprise, STSD, Bangalore, Karnataka 560016, India

e-mail: richa12.nitr@gmail.com

J. Singh

Chhattisgarh Professional Examination Board, Raipur, Chhattisgarh 492002, India

e-mail: jsraipur13@gmail.com

Image Streamer. This will bring down the time to service considerably, as, when the hypervisor boots up, it will have the customized application stack post personalization is done and is ready to host virtual machines [3]. The personalization here includes the setting up of hostname, domain name, and assign IP address to the management networks, multipathing the iSCSI by defining separate vSwitches and portgroups for high availability, etc., so that the same image can be used on multiple servers, with configurations tailored to the specific requirement, by the HPE Image Streamer.

1.1 Objective

A wide range of customers from various organizations worldwide and several partner teams from HPE deploy ESXi on scale setups. The aim is to make the process of getting the setup service ready for the host provision, efficient, and hassle-free, without the user having the need to know about the ESXi operating system in-depth or worry about the high availability of the host.

The manual effort put in to design and customize, applying update patches to each server in a scale system, makes it a cumbersome process, which may also result in human errors. Also, the vCenter is not configured to work with HPE Synergy out of box, which makes it difficult to run existing applications directly on it.

The objective of the proposed work is to give the facility of personalization of VMware vSphere hypervisor ESXi using HPE Image Streamer on server hardware. The aim is to decrease the time to service, by using developed scripts deployed using a single “Deployment Plan” and a single “Golden Image” on multiple hardware.

For the organizations working on large-scale server deployments, this solution is to enable them with the rapid deployment of servers, such that, they boot up directly with the personalized OS running on them, ready for further provisioning.

This paper elaborates the system architecture, describing various software used to implement this work, followed by the methodology stating the steps taken in order to achieve the required objective.

2 System Architecture

Implementation of the proposed work, requires HPE Synergy Image Streamer [4] driven by HPE Synergy Composer, HPE Synergy compute modules, converged network adapters, interconnect modules, and HPE 3PAR StoreServ.

Software requirement to accomplish this work include HPE OneView, HPE Image Streamer, VMware ESXi ISO image, and Guestfish.

HPE Image Streamer, the Synergy frame-embedded management appliance creates and configures remote OS boot volumes [5]. It is a secure personalization environment for configuring boot images.

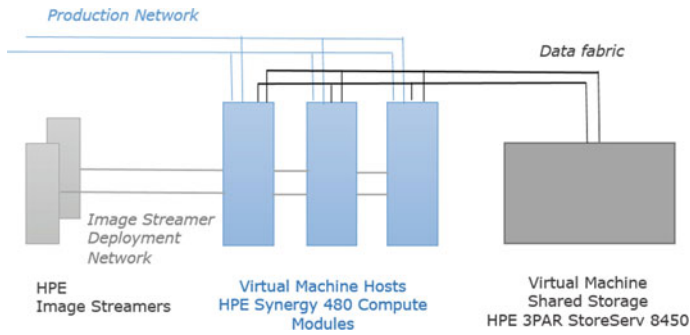


Fig. 1 Highly available virtualization deployment architecture

As seen in Fig. 1, the Image Streamer is running with the guestfish—the minimalistic OS that provides access to the libguestfs API. This will be used to discover the disk system partition (/dev/sda) on the ESXi OS where the customization can be done. It will run the scripts in order to set the hostname, domain name, and deployment network enabled with multipathing to access the data storage, management network, assigning static IP addresses or DHCP, etc., as per the user input, and configure it based on the physical ports available on the compute modules. Thus, the servers will boot up with all the configurations applied and ESXi ready to host virtual machines.

Figure 2 illustrates which physical network ports, on a single server will be assigned to the ESXi host virtual NICs, to be then added to the virtual switches created, and eventually to virtual machine kernel NICs, providing network to the guest machines hosted on the compute module.

3 Methodology

Let us consider an example to understand the workflow: HPE Synergy three frame enclosure, consisting of twelve compute modules. It has a pair of HPE Image Streamers for high availability of OS boot volumes. Out of the twelve servers, the user wants to install ESXi, on say eight servers. Each server is to be configured with the data and management network connections (Fig. 1). It should be enabled with multipathing so that multiple NICs can be used to provide failover and load balancing capabilities for the iSCSI connection between storage and host (Fig. 2). The ESXi deployed servers are to be then added to the vCenter cluster.

To enable this, the user need not to boot each server individually and configure all the settings manually on them. The intent of this project is to streamline this process. The procedure to be followed for the same is mentioned below step by step.

- Step 1: Boot a server with an empty OS volume of specified volume size say 20 GB.
- Step 2: Login to the HPE ILO and select the downloaded ESXi ISO of required version. Let the installation complete.

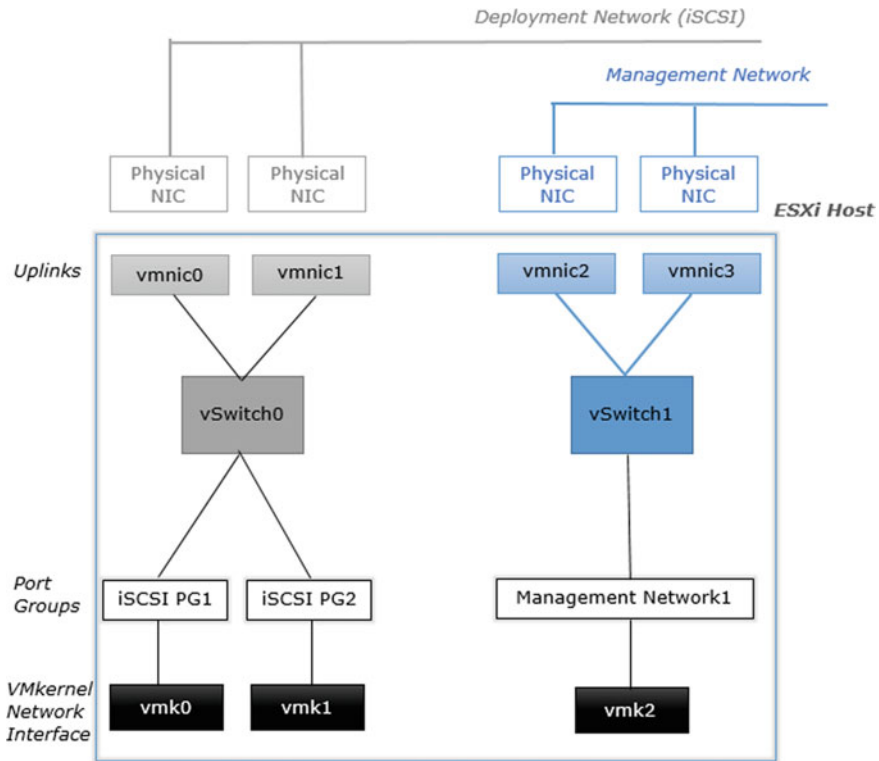


Fig. 2 Personalized network interfaces for a single HPE compute module with ESXi host deployed on it

Step 3: Power off the server gracefully. Now capture this OS volume from Image Streamer using scripts.

These scripts modify the image being captured by removing the previous configurations existing on the OS. This results in what is called an ESXi Golden Image which can be used for deployments manifold. It is a downloadable zip format image of the OS.

The Golden Image created on the appliance expands to Golden Volumes to prepare for rapid deployment. This single image can be used to deploy all the eight servers with customizations as per requirement.

Step 4: Create a Deployment Plan, which specifies a Build Plan and a Golden Image. A Build Plan is a group of PowerShell scripts called the Plan Scripts which perform OS Volume configuration. The user can choose among them what customizations are required.

The same Deployment Plan can be used to create several Server Profiles.

Step 5: Create eight server profiles, by selecting the same Deployment Plan specifying the server-specific configuration values and server hardware.

Step 6: Power on the blade servers. The OS Volumes get created very quickly via SmartClone thin replication of the Golden Volumes, for individual compute modules. Plan Scripts personalize OS Volumes to be server-specific based on the configuration values given for each hardware. The blade servers boot up with the customized ESXi OS, ready to be added to the cluster for further host provisioning.

(Note: Once personalized, the OS Volume becomes part of Server Volume Storage. HPE Composer automatically configures the server's iSCSI boot connection to its remote OS Volume.

HPE Image Streamer does not support OS deployments on compute modules that have UEFI secure boot enabled.)

Thus, the single Deployment Plan created in Step 4 with the single Golden Image created in Step 3 and the developed Plan Scripts provided can be used to deploy several blade servers, without the need of manual installation on each of them individually.

This project develops the PowerShell scripts (Plan Scripts) that facilitate the personalization of the ESXi OS.

The flowchart illustrating the implementation of personalization of ESXi OS is given in Fig. 3. The script when run, while deploying the compute module, mounts the ESXi boot partition on root. It checks for the updates on boot.cfg for both sda5 (bootbank) and sda6 (altbootbank) and determines which partition is the boot partition accordingly. All the configuration files (onetime.tgz, state.tgz, etc.) are stored in the bootbank. Once mounted, these can be accessed for personalization.

Post unpacking of onetime.tgz, the tar bundle is found to contain the local.sh script which is responsible for executing the personalization commands on the OS.

Multipathing is enabled (as described in Fig. 2) for high availability of the data network between the external storage device and the OS. It also provides load balancing, by distributing the I/O load across the various physical paths available, thus avoiding any bottlenecks. Configuring the iSCSI adapter to virtual machine kernel NICs (vmk NICs) achieve this by providing the failover and load balancing policy for each server.

Password of the root user, domain name, and host name are set as provided by the customer. Also, the option to enable or disable the SSH to the host is given to the user, in case one wants to access or restrict the console login.

Apart from these parameters, the management network connection needs to be provided, in order to access the internet (as shown in Fig. 2). There can be one or more physical network ports assigned for this. The option to either configure the management NIC with a static IP address, or to use Dynamic Host Configuration Protocol (DHCP) is given to the user. In case of configuring the NIC with a static value, certain attributes need to be specified—IP address, MAC address, DNS server IP, gateway, VLAN id, and netmask. Based on the requirement, more than one management NIC can be configured. It can be assigned to an active or standby mode based on the failover and load balancing policy set.

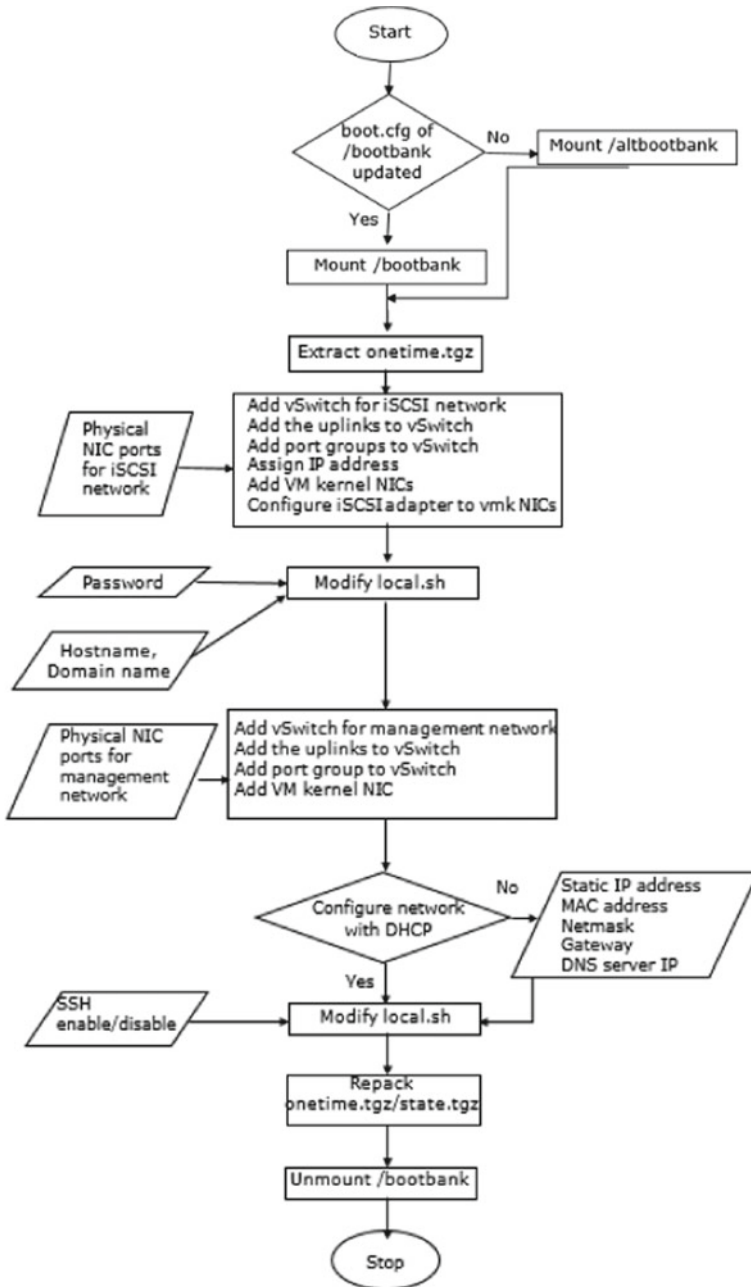


Fig. 3 Flowchart of implementation of ESXi personalization

Once the `local.sh` is modified with all the personalization commands, it is repacked as a part of the `onetime.tgz` or `state.tgz` configuration file. The bootbank is then unmounted.

4 Conclusion

VMware ESXi hypervisor is secure, reliable, and simplified application software. This is a small compact application program for optimal use of physical server hardware. It facilitates speedy installation of OS and applications on VMs. It helps in configuration, deployment, and implementation techniques.

ESXi OS is tiny program, which can be easily uploaded from a saved configuration file. ESXi system has better security and hardware reliability over other virtualization platforms.

This paper, by developing the scripts, facilitates the user, working on scale setups, deploying and configuring multiple blade servers simultaneously, to achieve the goal seamlessly, without getting into the manual work of customizing each server individually.

The admin or the client working on the server to make it ready for VM provision, need not to be having the ESXi internals knowledge. Using the scripts, the task of personalizing the OS has been made simple and user-friendly.

The proposed work successfully personalizes the VMware hypervisor ESXi using HPE Image Streamer. It is implemented and tested to the extent of correctness.

Acknowledgements My sincere thanks goes to Mr. Raghu Narasimha Murthy, Technical Expert, Hewlett Packard Enterprise, STSD, Bangalore, Mr. Tejaswi B. Rajeevalochanam, Senior Software Engineer V, Hewlett Packard Enterprise, Bangalore, Mr. Rajeev Hiremath, Project Manager, Hewlett Packard Enterprise, STSD, Bangalore, for their support and motivation throughout my work.

References

1. Nomnga, P., Scott, M.S., Nyambi, P.B.: A technical cost effective network-domain hosting through virtualization: a *VMware ESXi and vSphere client approach*. Int. J. Comput. Appl. (0975–8887), vol. 91(10) (2014)
2. Chaubal, C.: The architecture of VMware ESXi, VMware Whitepaper, Revision: 2008-10-24 WP-030-PRD-02-02
3. Budhprakash B.A., Bhattal, G.S.: A comparative study of various hypervisors performance. Int. J. Sci. Eng. Res. 7(12) (2016). ISSN 2229-5518
4. HPE Synergy Image Streamer 5.0 Support Matrix. Part Number: P01616-006. Published: August 2019. Edition: 1
5. HPE Synergy Image Streamer 4.2 User Guide. Part Number: P01617-005 Published: July 2019. Edition: 2

A Proposed IoT Architecture for Corals Research Using AI and Robotics



Wafiik Aumeer, Nabeelah Pooloo, and Rajeev Khoodeeram

Abstract Coral reefs are one of the most diverse ecosystems on the planet and are vital in providing nursery, spawning, refuge, and nurturing areas for a multitude of different organisms. However, coral reefs are degrading owing to climate change, overfishing, industrial pollution and invasive species. Thereafter, mass coral bleaching events and infectious disease outbreaks take place causing corals to die. This paper aims at contributing in the monitoring of corals in the Mauritian marine ecosystem by utilising a Hybrid Underwater Vehicle (HUV) autonomously for collection of oceanographic data as well as images and videos of surveyed coral reefs sites. Data collected will be transmitted to a base station via Cellular IoT communication for analytics using two methods of Artificial Intelligence (AI) namely Machine Learning and Deep Learning. The conceptual design is thoroughly described together with software applications.

Keywords Bleaching · IoT · HUV · Artificial intelligence · Machine learning · Deep learning

1 Introduction

Currently in Mauritius, data about corals and their surroundings are collected by qualified divers and marine scientists by the Mauritius Oceanography Institute (MOI). Marine scientists carry out ongoing research on corals involving detailed experi-

W. Aumeer (✉) · N. Pooloo · R. Khoodeeram
Université des Mascareignes, Beau Bassin-Rose Hill, Mauritius
e-mail: wkaumeer@student.udm.ac.mu
URL: <http://www.udm.ac.mu>

N. Pooloo
e-mail: rzpooloo@student.udm.ac.mu

R. Khoodeeram
e-mail: rkhoodeeram@udm.ac.mu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_31

371

ments, observations, taxonomic catalogues, sample collection and statistical analysis. However, numerous safety issues and problems arise with human intervention underwater, such as:

- Drowning due to diver panic or becoming unconscious.
- Decompression sickness due to presence of nitrogen bubbles inside body tissues when compressed air is breathed from the tanks while diving.
- High chance of repetitions of work.
- Shorter time for underwater exploration.
- Omnipresence of hazards: Collision with ships or site access.

This protocol also requires the identification and enumeration of hundreds of individuals belonging to hundreds of species and the accuracy of such visual-based assessments is highly dependent on conditions such as depth, dive duration and turbidity [6]. In this view, Underwater Vehicles (UVs) can aid in extending human capabilities to explore the sea environment. These specially designed robotic equipment are able to stay underwater much longer than human divers thus expanding time available for exploration. Divers who go underwater need to be qualified and equipped with robust and costly breathing support gears and other portable devices whereas UVs can be relatively simple and one-time investments.

One of the primary goals of contemporary coral reef ecology is to understand the dynamics of reefs in regard to global climate change. Research has been held back by limited access to underwater data, the duration of data acquisition and subsequent processing of obtained images [4]. To overcome these existing problems, a fully integrated IoT system for the UV is proposed, which collects different oceanographic data and media data in real-time. AI techniques which can reinforce and scale up processing of volumes of data, will then be used for data analytics.

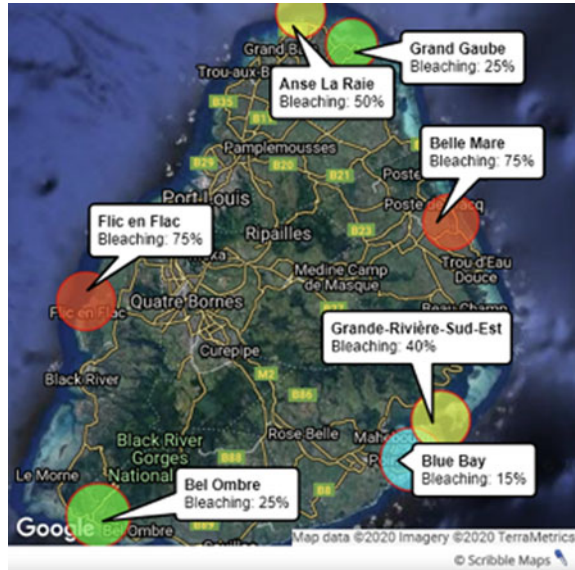
1.1 Coral Reefs in Mauritius

The coastline of Mauritius is 322 km long and the length of reef is 150 km covering an area of 300 km² [3]. Coral reefs play a major role in providing habitats for marine organisms while protecting vulnerable coastal communities. They particularly contribute by:

- Helping small fishes and other tiny creatures find their food and shelter.
- Controlling the levels of carbon dioxide in the atmosphere.
- Protecting coastal areas and communities from natural threats like tsunamis, cyclones and even sharks.
- Corals embellish lagoons which attract tourists in Mauritius.

However, coral reefs are particularly sensitive to the effects of climate change and industrial pollution [1]. As temperature increases due to global warming, mass coral bleaching events and infectious disease outbreaks take place causing the corals to

Fig. 1 Percentage of coral bleaching around the coast of Mauritius



die. From a survey done by the MOI in 2019, it was observed that more than 75% of corals around the lagoons of Mauritius were either partially or completely bleached due to an increase in temperature [3] as shown in Fig. 1.

The proposed system which consists of two parts namely Robotics and AI is discussed in the next section.

2 Architecture of the Proposed Framework

2.1 Concept of Proposed System

This paper proposes a solution which encompasses the use of a Hybrid Underwater Vehicle (HUV), based on Remotely Autonomous Vehicle (ROV) technology but with some Autonomous Underwater Vehicle (AUV) autonomy. The HUV collects vital data about coral reefs and transmits them to a Ground Station using IoT connectivity. Upon reception, those data are stored on the cloud and made accessible through a website and a mobile application thus enabling online data view. Additionally, the phenomenon of coral bleaching can be automatically monitored using embedded AI algorithms for image processing. An illustration of the proposed system is shown in Fig. 2.

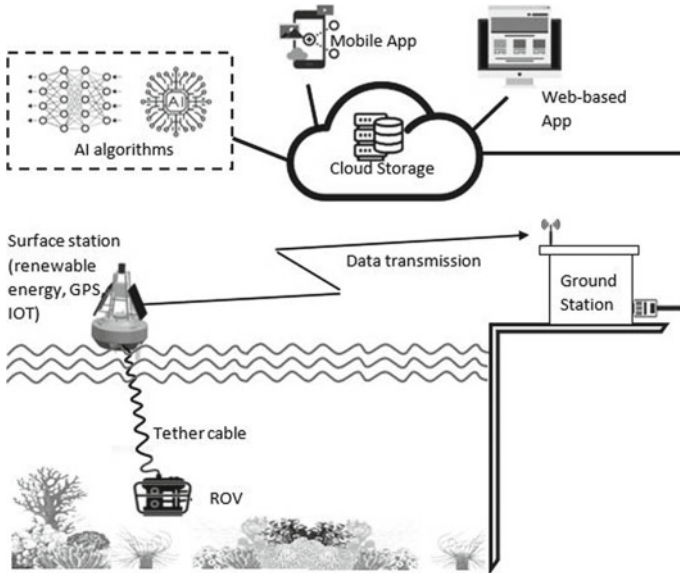


Fig. 2 Design of the proposed system consisting of the HUV and the data manipulation

2.2 Design of the Underwater Vehicle

A typical 3D representation of the ROV is first designed. The aim is to utilize low cost materials like PVC pipes and fittings and readily available system like Arduino microcontrollers and GoPro camera systems. This strategy allows for building a Proof of Concept (POC) before finalizing the UV. The main features of the ROV will be:

1. **Buoyancy:**

The type of buoyancy chosen is slightly positive to ensure that the ROV limits touching marine life in case of failure of the thrusters. The slightly positive buoyancy will be achieved by using two floats on top of each side of the frame of the ROV, while on the other hand, two hollow pipes of the bottom front and back of the ROV will evenly accommodate counter weights either as washers or lead bobs.

2. **Backup battery compartments:**

Sealed battery compartments are installed on the ROV to allow it to surface in case of emergencies.

3. **Camera housing and electronics compartment:**

This forms part of the ROV with a transparent dome in front. The latter houses the camera system while the rest of the section of the central tube is used to house the onboard controller.

4. **On board controller system:**

The onboard controller consists of a microcontroller and a System on a Chip (SoC) to control the thrusters and carry out the mission objectives. The onboard controller is also responsible to run simple navigation algorithm to permit smooth displacement of the ROV underwater.




5. **Tether cable:**

This cable acts both as a power transmission cable and also enables communication between the Surface Station and the ROV. The determining factors while designing the tether cable are its suitability for sea water, its strength, bandwidth and voltage drop. A maximum length 20m will be chosen, since it will be the depth that an entry level certified diver can dive to in case of any problems with the ROV.

6. **Operations:**

These range from simple GPS point ‘target and drop’ for close area survey to more elaborated waypoint missions. A few examples are illustrated in Table 1.

Table 1 Examples of possible waypoint missions

<p>Target and drop: HUV goes to a location and drops the ROV for a precise location survey</p>	
<p>Waypoint mission (line): ROV is dropped at fixed depth and moves together with surface station around a fixed trajectory defined by different GPS points (waypoints)</p>	
<p>Waypoint mission (area): HUV moves to GPS point. There ROV is dropped and it surveys the area by doing preset circles patters. After the survey mission, it surfaces and attaches back to the surface station. The HUV then moves to the next GPS point</p>	

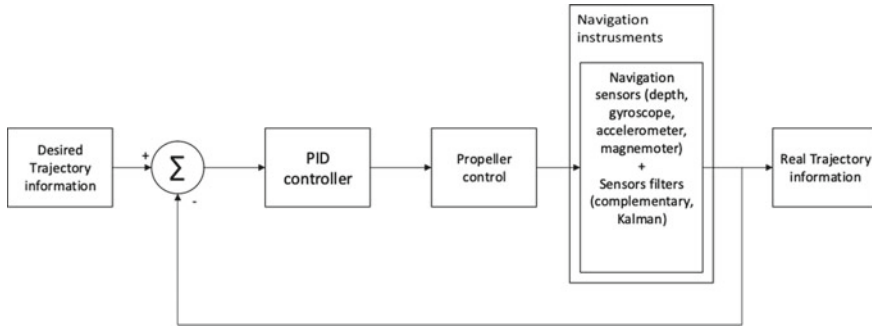


Fig. 3 Block diagram of enhanced navigation controller with PID control and complementary filter

7. Navigation and Control:

An enhanced navigation control system is designed to help the HUV system operate in a turbulent marine environment. It is therefore important that the HUV is able to position and orient itself to follow mission objectives. In addition, variations in data obtained from navigation sensor can make the navigation jittery (see Fig. 3).

2.3 Design of Surface Station

The Surface Station is the second part of the HUV system and communicates with the Ground Station. It houses devices like GPS transmitter, data logging and GSM modules. It also sends command signals to the ROV in case of emergencies (return home command or surface command). The features of the Surface Station are as follows:

1. **Receiving unit for tether cable:** The communication and power ends of the tether cable will be connected to an ethernet hub and to battery terminals inside the Surface Station.
2. **Renewable power source:** In order to power the ROV autonomously, solar panels will be installed on the Surface Station allowing charging of batteries. The system will consist of solar panels, charge controller and the battery system.
3. **IoT technologies:** This is discussed in detail in Sect. 3.

2.4 Design of Ground Station

The Ground Station will be on land and will house the computer system which will process the collected data. It will consist, as mentioned previously, of a GSM module acting primarily as a receiver which will obtain SMS/MMS data. The same module will be able to send emergency SMS commands to the HUV like Return to Home commands.

3 Integration of IoT and Artificial Intelligence

3.1 Design of IoT Structure

In this section, the use of IoT to connect the HUV to the cloud is described. The general configuration of the IoT structure is shown in Fig. 4.

3.1.1 Communication and Cloud Storage

Cellular IoT is chosen as it enables communication by utilizing existing mobile networks. GSM modules embedded in the system allow the use of SMS/MMS for communication in duplex to collect data and to send commands as well. The IoT device then allows connection to a suitably chosen cloud platform for the storage of images and videos captured by the HUV.

3.1.2 Mobile App

An Android mobile app allows researchers to mainly view collected information from the cloud. The features and functionalities are explained in Figs. 5 and 6.

3.1.3 Web-Based App

A Web application is developed for the public as well as for the researchers. Figures 7 and 8 shows the main pages and functionalities of the Web application.

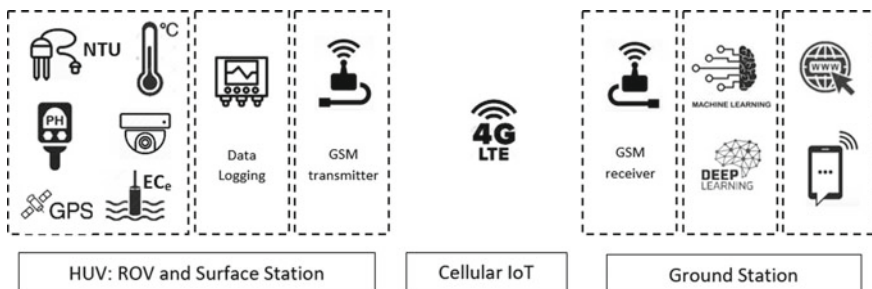


Fig. 4 IoT structure linking the HUV with the ground station

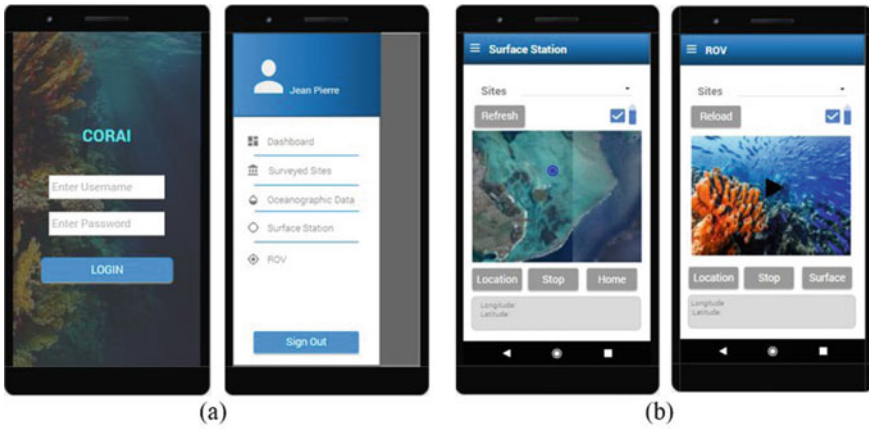


Fig. 5 **a** Dashboard options to get data from the cloud. **b** Localised site information are showcased, with possible interactions with the HUV

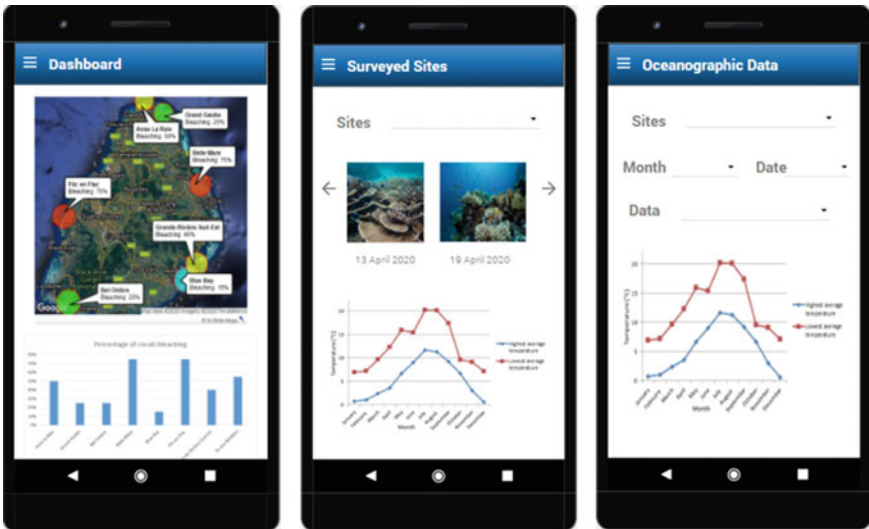


Fig. 6 On the dashboard, the user will have a global view of the percentage of coral bleaching in each region. Images captured by the ROV will be displayed on the app via the REST APIs

3.2 Use of AI Algorithms

The Identification of corals and detection of coral bleaching in underwater images can indeed be very challenging because (i) the colour and brightness of the image taken due to turbidity and depth of the water, (ii) the complex 3-dimensional structure of corals, (iii) presence of fishes and other organisms, (iv) the acquisition of a proper camera and imaging system.

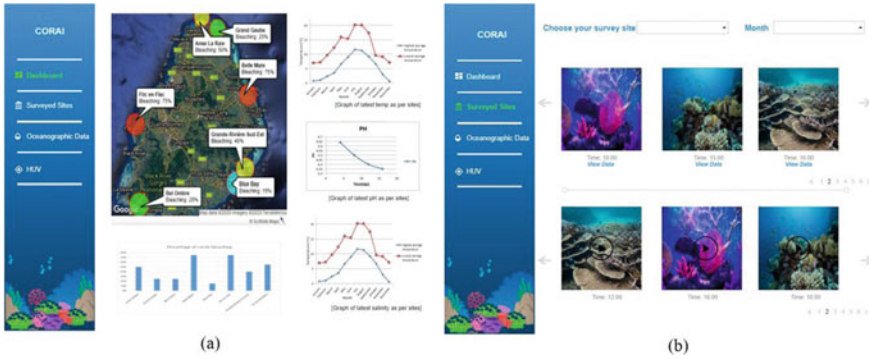


Fig. 7 **a** The dashboard will provide an overview of the data collected by the ROV in terms of graphs (such as latest temperature, pH, etc). A graph showing the percentage and levels of bleaching at different sites is also presented. **b** On this page, users will be allowed to go through the images and videos captured by the ROV. These are retrieved from the cloud storage



Fig. 8 **a** On the oceanographic data page, sensor data collected by the HUV will be presented in a table with its respective graph. **b** Finally, there is a page to display the location of the surface station on google map and the last video captured by the ROV

To overcome these setbacks, two algorithmic models of AI, namely Machine Learning and Deep Learning are investigated to enable the HUV system to autonomously analyse the image data and detect coral bleaching.

3.2.1 Datasets and Methods

Data collected by the HUV and pre-collected data from the MOI (as shown in Fig. 9) will be used to train the selected AI models. These data include physical, biological and chemical measurements for parameters such as temperature, conductivity, salinity, pH, turbidity, current/sea surface current pattern and genetic information on selected marine organisms among others.



Fig. 9 Data available around the coast of Mauritius by the MOI

Cleaning of data may be required for data preparation in case there are any missing values, errors or outliers. These data together with annotations obtained from the MOI are then used during the application of the AI techniques.

Training datasets created on the basis of underwater classes are different in terms of shape, color, texture, size, rotation, illumination, view angle, camera distance and light conditions [2]. Therefore, large dataset is needed to achieve a good performance. Video captured by the HUV will be analysed as frame of images and incorporated in the dataset as well.

3.2.2 Machine Learning

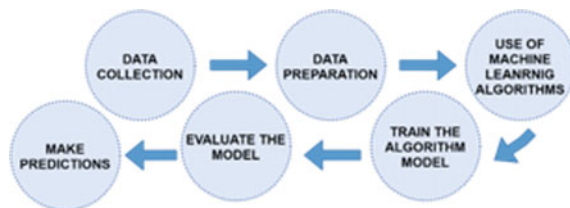
Machine learning algorithms, also called models, are mathematical expressions that represents data in the context of a problem. There are three types of machine learning: supervised learning, unsupervised learning and reinforcement learning. The general steps for machine learning algorithm are shown in Fig. 10. In this paper, we focus on supervised learning techniques.

Supervised Machine learning can be used as a technique to predict regions around the coast of Mauritius that are likely to be affected by coral bleaching. This is done by referring to past bleaching regions and other relevant data. The technique of supervised Machine Learning used is Classification. It consists of a training dataset which includes input variables and response variables. The response variable is also known as *class* or *category*. From these variables, a model is built to predict the response variable for new dataset captured by the HUV. In this case, the input variables are temperature, pH, salinity and turbidity; and the class is either bleaching region or non-bleaching region. This is a binary classification since there are only two options available for the class. Two classification algorithms will be explored namely Naive Bayes and Decision Trees.

Naive Bayes classifiers are built on Bayesian classification methods which assumes that an input variable is unrelated to other input variables even if they may depend on each other. For example, temperature is not related to pH, salinity nor turbidity. These properties independently contribute to the probability that a region is likely to have coral bleached or healthy corals. Using this theorem, the probability that a region will be bleached can be found, given that the input variables/features for a defined range of values have occurred. For instance, it is known that at a high temperature, corals start to bleach and thus the likelihood for the coral to bleach at that temperature may be calculated. Same applies for all the other features. The training data which is data from the MOI is separated by class during the application of the Naive Bayes. Using the statistics derived from these training data, the probability for a region to have coral bleaching can be calculated from new data collected by the HUV.

Another predictive model for classification is the Decision Tree Algorithm. A decision tree is similar to a flowchart tree structure where the internal node represents features, the branch denotes a decision rule and each leaf node depicts the result. The model uses a Top-Down approach where the topmost node, known as the root node, learns to partition based on feature value. It divides the tree in a recursive way

Fig. 10 Steps in the machine learning process



called recursive partitioning. Decision trees works from the root to some leaf node, with the leaf node providing the classification. Again, here the features will be the oceanographic data collected by the HUV and the classification will either be coral bleaching or non-bleaching. For example, one decision node can be the temperature at a specific region and its leaf node will be the decision made if the coral is considered bleached or not. From that temperature node, there can be another decision node for salinity at that specific region to deduce if the coral has the possibility of bleaching or not as the leaf node. Other features follow same procedure. The goal is to learn simple decision rules deduced from the data features.

3.2.3 Deep Learning

Deep learning method is a simple but nonlinear module to transform the expression at the natural input level to a higher and more abstract representation [5]. Deep Learning is used for the automatic processing and classification of the images of different ecological health of corals in order to detect coral bleaching. Two Deep learning techniques will be explored in this project namely CNN and RNN. These techniques are used for the image analysis of the various coral images captured by the HUV.

1. Convolutional Neural Networks

CNNs are made of layers of interconnected neurons and each neuron includes a ‘convolutional kernel’ that computes a set of mathematical operations (defined by ‘weights’) on the matrices of values describing the image (i.e. values for each color channel for each pixel) [6].

There are four steps for CNN:

a. Convolution

This part seeks at reducing the size of the image into a feature map for faster computations of the weights. This is done by using a 3×3 square pixel and multiply it with a filter. Convolution keeps the essential features of the coral image and excludes irrelevant noise.

b. Max Pooling

The purpose is to reduce the dimensionality of the input image to prevent overfitting and improve the computation speed. It divides the feature map into sub regions.

c. Flattening

The pooled feature map is flattened into a column since it will be introduced into an artificial neural network.

d. Full connection

The column data is inserted into the artificial neural network until it develops its output layer which represents the class into bleaching or non-bleaching. An example is illustrated in Fig. 11. The CNN will perform classification of the images obtained from the HUV after training with a large dataset of imagery with annotation of the condition of corals. The output result will either be bleaching or non-bleaching.

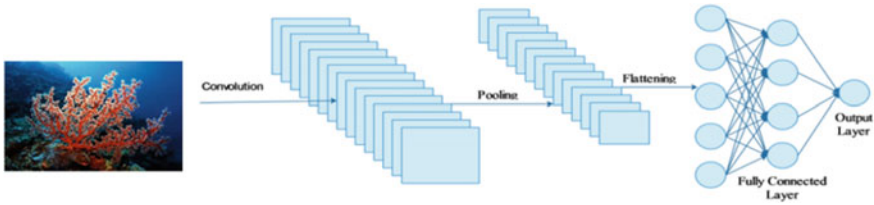


Fig. 11 CNN process for the coral image captured by the ROV

2. Recurrent Neural Networks

RNNs allow outputs from previous step to be used as inputs of the current step while having hidden states. These hidden states are the key features in RNN since they remember information about a sequence; compared to the traditional Neural Networks where all the inputs and outputs are independent to each other. The ability to remember information through time is useful in time series predictions. This is also termed as Long Short Term Memory (LSTM). The LSTM is composed of (i) Memory cell: it is where the input dwells, (ii) Input gate: it is where the input enters the cell and includes a ‘tanh’ activation function, (iii) Output gate: it filters and regulates the output of the function and (iv) Forget gate: this gets rid of previously stored information if it is not needed.

RNNs work by converting the independent activations into dependent ones by providing same weights and biases across all layers. This, in turn, memorises each preceding output by providing each output as input to the next hidden layer. Consequently, all the hidden layers can be joined into a single recurrent layer. Like CNNs, RNNs have been used as part of a model to generate descriptions for unlabelled images. The RNN process takes the coral images as input from the cloud storage and feeds them through the hidden layer to finally produce the output result with the bleaching or non-bleaching classification. Figure 12 depicts a RNN process.

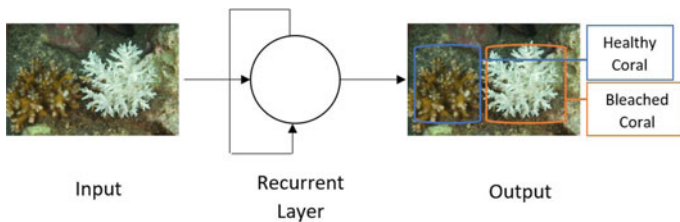


Fig. 12 Representation of RNN process for the coral image analysis

4 Expected Results

The main outcome of the project is the successful implementation of the automated system as a whole, yielding into a congruent method that is convenient for scientists, biologists and other interested parties to utilize in their field of work. It is thus expected that the HUV is correctly realised and deployed in real environmental conditions, with navigation autonomy and power sustainability. It must also be able to capture the defined data of the surveyed reefs for transmission to the Ground Station. The hardware harbouring the Cellular-IoT scheme needs to be particularly robust to sustain the transmission sensor measurements and imagery for further processing

Consequently, the proper algorithms must be designed, with proven capabilities of analysing the collected data autonomously using the methods of machine learning and deep learning. Based on the fruitful algorithm development, a meaningful and accurate classification of the images will thus be possible, aiding in identifying coral bleaching in surveyed sites and also to predict potential new episodes. Finally, the project will include the design and deployment of functional web-based and mobile application to illustrate succinct information of the conditions of the surveyed coral reefs based on the up-to-date information stream generated by the automated system.

5 Conclusion

The oceans contain roughly 1.34 billion cubic kilometres of water and according to the NOAA's National Ocean Service "more than 80% of our ocean is unmapped, unobserved and unexplored". In this context and with alarming cases of unprecedented mass degradation of corals, the monitoring of marine environment has more than ever been capital to ensure balance and sustainability for all. Aiming at tackling the current challenges and helping marine scientists in their risky work, this project has focused on proposing a fully- feasible system comprising of a HUV, appropriate AI algorithms together with web and mobile applications for automating the monitoring of the coral reefs around Mauritius in view of identifying coral bleaching areas. The proposed HUV design consists of a ROV, capable of capturing sensor data, images and videos, and also of a surface buoy which acts as a surface station powering the ROV and enabling transmission of collected data to a ground station using cellular IoT communication.

References

1. Hoegh-Guldberg, O., Mumby, P.J., Hooten, A.J., Steneck, R.S., Greenfield, P., Gomez, E., Harvell, C.D., Sale, P.F., Edwards, A.J., Caldeira, K., Knowlton, N., Eakin, C.M., Iglesias-Prieto, R., Muthiga, N., Bradbury, R.H., Dubi, A., Hatzioolos, M.E.: Coral reefs under rapid climate change and ocean acidification. *Science* **318**(5857), 1737–1742 (2007)

2. Mahmood, A., Bennamoun, M., Sohel, F., An, S.: Action recognition and prediction. In: Deep Learning for Coral Classification (2017)
3. Mauritius Oceanography Institute. Moi.govmu.org: Available at <http://moi.govmu.org/assets/pdf/OceanQuest-10> (2020)
4. Mehta, A., Ribeiro, E., Gilner, J., Van Woesik, R.: Coral reef texture classification using support vector machines. Available at <https://pdfs.semanticscholar.org/596e/f6f56e49b9e57a6cace30b8c662ad988ffbe.pdf> (2007)
5. Nguyen, H.-T., Lee, E.-H., Lee, S.: Study on the classification performance of underwater sonar image classification based on convolutional neural networks for detecting a submerged human body. *Sensors* **20**(1), 94. Available at <https://www.mdpi.com/1424-8220/20/1/94/html> (2020)
6. Villon, S., Mouillot, D., Chaumont, M., Darling, E.S., Subsol, G., Claverie, T., Villéger, S.: A deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecol. Inform.* **48**, 238–244 (2018)

Voice Password-Based Secured Communication Using RSA and ElGamal Algorithm



Prashnatita Pal, Bikash Chandra Sahana, S. Ghosh, Jayanta Poray,
and Amiya Kumar Mallick

Abstract Secured voice authentication-based communication is the main aim of this study. Here, eight speech keywords have been recorded and stored in computer memory. One speaker recognition model was used for voice password authentication. Then, the speech keywords were encrypted using private key. The message was encrypted using RSA or ElGamal algorithm. The message was modulated using FSK digital modulation technique and sent through the communication channel. The speech samples were demodulated and decrypted at the receiver. The received speech samples matched with the original transmitted voice samples. The equality ratio for this study is 0.6 and above. In this study, secured authentication technique has been adopted. After voice authentication, secured communication has been done successfully.

Keywords Speech recognition · Cryptography · RSA algorithm · ElGamal algorithm · Voice password

1 Introduction

In communication security, it consists of two main cryptographic processes. Firstly, a public key is used to convert an input speech into an unrecognizable encrypted output called cipher speech (encryption process), which makes it practically infeasible to recover the original speech without the encryption key. Speech communication is

P. Pal (✉) · B. C. Sahana

Electronics & Communication Engineering, National Institute of Technology, Patna, India
e-mail: prashnatitp@gmail.com

J. Poray

Computer Science & Engineering, Techno India University, Kolkata, India

A. K. Mallick

Retired, Electronics & Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India

S. Ghosh

University of Kalyani, Kalyani, India

used in several areas in our day to day life. To ensure the security in speeches and maintain its confidentiality with appropriate access control, integrity and availability are not only challenging, but also supported by diverse entities. In this context, protection of speeches using passwords or similar techniques from any misuse is an important consideration. In next step, a private key is used to convert the encrypted speech data back to its original form (through the decryption).

Today in the generation of electronic gadgets, the necessity to prevent data from miscreants is increasing day by day. Cryptography is the process of utilization of codes to prevent anyone from violating speech security. Speech protection can be accomplished by changing the original speech by any means to some other speech codes, so that if someone gets that speech by means of hacking, then also it must remain in useless bits of speech for that person. This process can be achieved by encrypting that speech by some means of algorithms which are known to the sender and on the other side the similar decryption algorithms must be known to only the desired receiver such that it can convert that encrypted speech back to the user understandable data or signal.

2 Literature Survey

RSA (Rivest, Shamir and Adleman, i.e., the name of the inventors) is among the most popular public key encryption algorithm [1] used to secure the speech communication process improve the protection mechanism the art is used to ensure speech communication security. Another popular algorithm is the ElGamal encryption system which is an asymmetric key encryption algorithm described by ElGamal [2]. Its security depends upon the ability of a hacker to compute discrete logarithms. Encrypted data can be transmitted using frequency shift keying or FSK, [3] which is a digital modulation technique. Here, using reflex klystron, the FSK is generated. In FSK, the frequency of the carrier signal varies as per the variation of the digital signal. The output frequency of FSK modulated wave remains high for a binary high input, and it remains low for its binary low counterpart. It plays an important role in long distance communication. Here, FSK is generated using reflex klystron [4]. Reflex klystron is basically a microwave generator where velocity modulation technique has been utilized to form a high energy density bunch of electrons which suitably reflect to generate high frequency RF oscillation in a re-entered cavity; in past (during 1950s and 1960s), in case of radar, it was used as local oscillator, and in case of microwave transmitters, it was used as a modulator. Moreover, the demodulation and decryption are done at the receiver end.

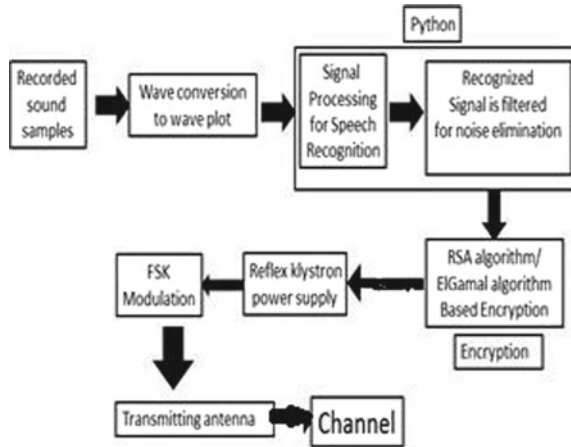
The conventional MFCC used in [5–8] has the limitation for eliminating a particular band noise that has by default poor recognition rates. We have used noise reduction using spectral gating filtering. This algorithm is based on the one out lined by audacity for the noise reduction effect. The disadvantage of [9, 10] is that it is able to recognize voice signals of very short times pan containing at the most one word. Our study deals not only with voice clips containing one word, but is also able to deal

with voice clips containing several words. This is an enormous advantage over the above-mentioned literatures. Since most of the voice recognition systems required in the present world needs to handle voice clips containing multiple words spanned over a long interval. In [11, 12], and [13], the accuracy of voice recognition was not sufficient; we have improved the recognition accuracy by employing a digital filter. In [15–17], and [18], it has been shown that scaling about the effect of voice is depends on the recognition rate. Moreover, this is measured by various disguising types. So, it is not easy to understand if a voice is disguised. Our system can easily recognize a disguised voice because it compares the input speech signal and therefore the reference speech signal and allows the communication to happen when equality ratio is over 0.6. Experimental results in our system have shown that a disguised voice could not achieve an equality ratio of a minimum of 0.6. [19] Used FPGA hardware platform which is costly. In [20] has used LabView programming model, which has several disadvantages like lot of memory, is needed and also time consuming. Developer edition is very costly. It also has debugging issues. We have used Python which is open source language and resource efficient. We used Jupyter notebooks embedded in Anaconda which is well-known software for executing Python programs. Several other IDEs are available which open source.

3 Methodology

Speech samples are recorded using mobile recorder in.mp3 format. But.wav format is desirable for working on the speech sample. Therefore, speech samples in.mp3 format are converted to.wav format using one converter application. Plotted the amplitude vs. time graph for each of the speech samples in MATLAB, which are uploaded into Python-based voice identification system. Jupyter notebooks embedded in Anaconda which is very well-known software for executing Python programs. Now, when a person speaks his speech is compared with the recorder speech samples. If it is matching with one of the speech samples, then he is an authenticated speaker and his speech is processed for transmission to the receiver. Otherwise, he is an unauthenticated speaker and transmission to the receiver will not take place. After a match is found, the spectrogram of the corresponding speech sample is plotted after eliminating the noise. RSA algorithm or ElGamal algorithm is applied on noise eliminated signal for encryption of speech signal. Digital signal was then passed through reflex klystron to convert this digital signal into modulated signal using frequency shift keying (FSK) and transmitted to the receiver as shown in Fig. 1. Reflex klystron will assign two types of frequencies, where the high frequency (f_1) is assigned for a high binary input value and the low frequency (f_2) range is associated with binary low input. From the characteristic of Reflex Klystron it is shown that Amplitude Modulation (AM) and Frequency Modulation (FM) are inherently inseparable. So here we are using pulse waveform. At the receiver, in order to identify f_1 and f_2 , FSK signal is passed through a coupler which divides the corresponding signal into two parts. These two parts contain both f_1 and f_2 frequencies in same phase. The resulting

Fig. 1 Voice authentication, encryption, and modulation block at transmitter



two signals are passed through two different resonating cavities of frequencies of f_1 and f_2 to identify them. The resulting two signals are then summed up using adder circuit to get the original speech signal. This signal is amplified and applied to digital to analog converter (DAC) to get back analog signal. RSA decryption algorithm or ElGamal decryption algorithm (whichever is applied) is applied on analog signal for decryption and get back original speech signal which is spoken as shown in Fig. 2.

The frequency shift keying (FSK) is a digital modulation technique, where the carrier signal frequency proportionately changes with the corresponding digital signal. This is a scheme, used for frequency modulation. Here, the range of output signals varies from high to low frequency associated with high and low signals, respectively. Figure 5 is the diagrammatic representation of FSK modulated waveform along with its input.

Fig. 2 Demodulation and decryption at receiver

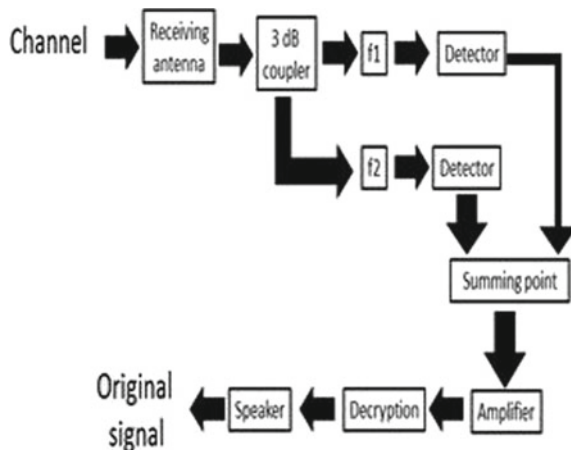
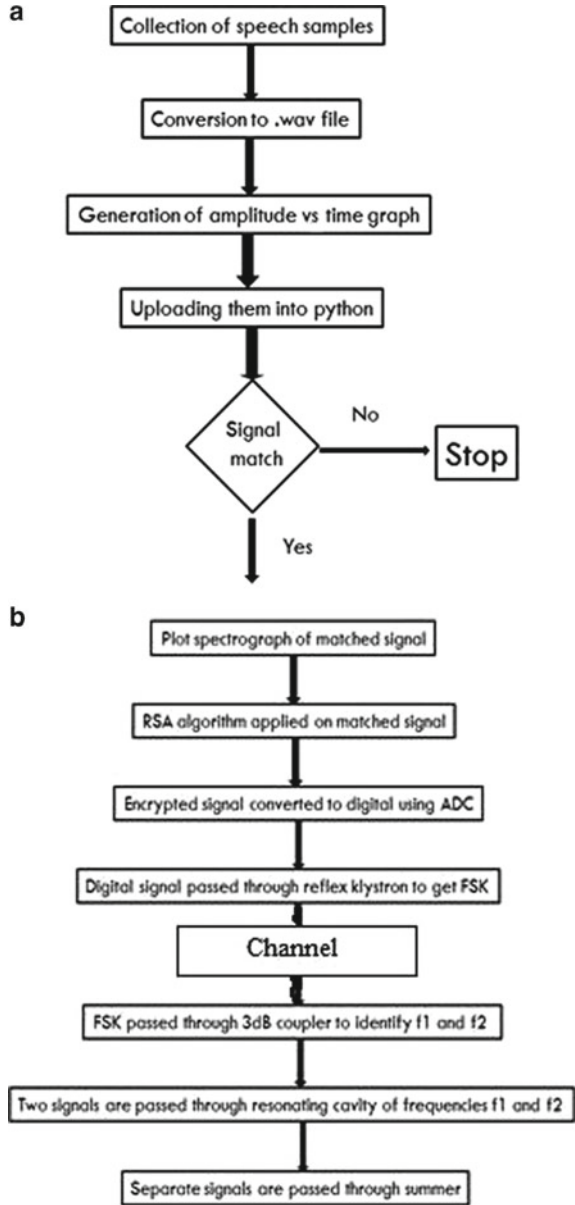


Fig. 3 a Flow chart of voice password authentication and modulation of speech signal. **b** Flow chart of voice password authentication and demodulation of speech signal



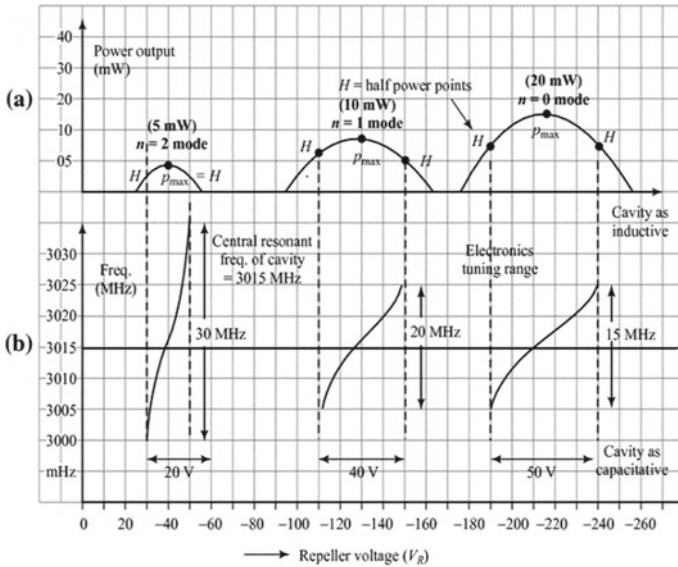
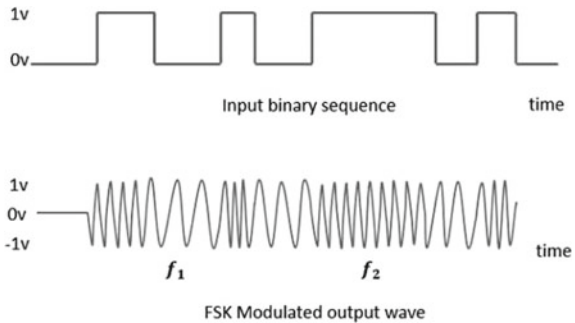


Fig. 4 Characteristic curve of reflex klystron

Fig. 5 Frequency shift keying



3.1 Encryption and Decryption Technique Using RSA and ElGamal Algorithm

In the year of 1977, the RSA [1] algorithm was designed and published by Ronald Rivest, Adi Shamir, and Leonard Adelman. Since then, this algorithm was used as one of the potential and widely popular public key encryption algorithm. This encryption standard has been used by Microsoft Explorer as well as Netscape Navigator web browsers to design the secure socket layer (SSL). Also, it has been used for encryption by Visa and MasterCard for secure electronic transactions (SET). This public key encryption system use the concept of prime number division arithmetic to transform a message (i.e., a sequence of numbers which is a product of two prime numbers)

into the corresponding cipher text. There are three major steps for RSA algorithm, namely, (1) Key generation, (2) Encryption, and (3) Decryption. These are depicted below:

3.2 Key Generation, Encryption, Decryption

Key Generation

Select p, q where p and q both prime, and $p \neq q$.

Calculate $n = p \times q$

Calculate $\phi(n) = (p - 1) \times (q - 1)$

Select integer 'e' such that $\gcd(\phi(n), e) = 1; 1 < e < \phi(n)$

Calculate $d, d \equiv e^{-1} \pmod{\phi(n)}$ or $d.e \equiv 1 \pmod{\phi(n)}$

Public Key: $PU = \{e, n\}$ and

Private Key: $PR = \{d, n\}$

Encryption

Plaintext: $M < n$

Cipher text: $C = M^e \pmod{n}$

Decryption

Cipher text: C

Plaintext: $M = C^d \pmod{n}$

The ElGamal algorithm [2] is another asymmetric key encryption cryptosystem. This uses the Diffie-Hellman key exchange mechanism, which was proposed by ElGamal [1985]. The ElGamal is used as a type of digital signatures scheme. Also, this is used as free GNU privacy guard software. The implementation of ElGamal scheme is based on any cyclic group G , which is a multiplicative group of integers using modulo arithmetic. The security of this scheme depends upon the difficulty level of a given problem. By nature, this problem is a type of discrete logarithm problem. Similar to RSA as per Fig. 6.

3.3 Proposed Methodology

1. Start
2. Speech samples are recorded using mobile recorder.
3. Conversion into .wav format for ease.

Fig. 6 ElGamal algorithm

<ul style="list-style-type: none"> Key Generation Select a large prime as a q Select x to be a member of the group $G = \langle Zq^*, X \rangle$, x must be "$1 \leq x \leq q - 1$" Select g to be a primitive root (generator) in the group $G = \langle Zq^*, X \rangle$ $y = g^x \text{ mod } q$ Public key $\leftarrow (g, y, q)$ Private key $\leftarrow x$
<ul style="list-style-type: none"> Encryption Select a random integer r in the group $G = \langle Zq^*, X \rangle$, r must be "$1 \leq r \leq q - 1$" $C_1 = g^r \text{ mod } q$ $C_2 = (p \cdot y^r) \text{ mod } q$ // p is the plaintext
<ul style="list-style-type: none"> Decryption $P = [C_2(C_1^{-x})^{-1}] \text{ mod } q$

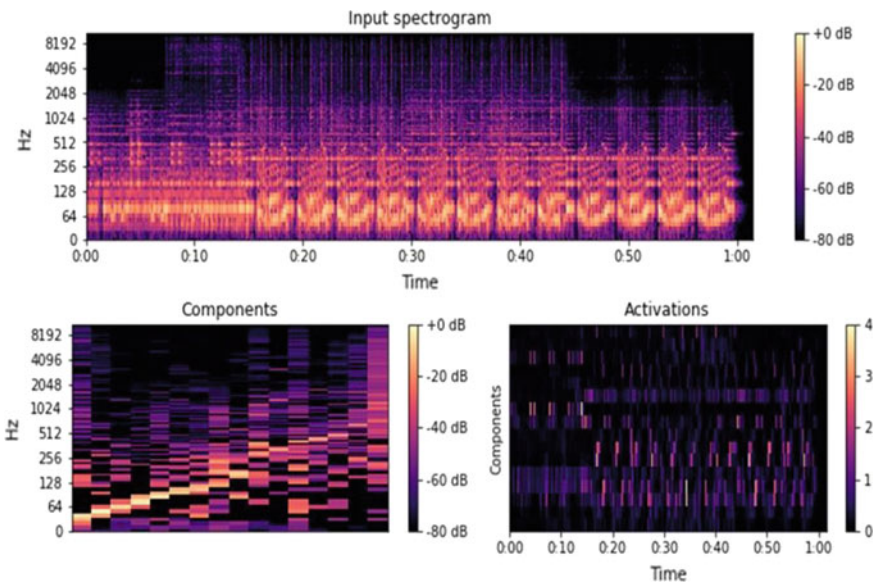


Fig. 7 Spectrograms of the recognized voice signal and its components

- Plotting the amplitude vs. time graph for each of the speech samples in MATLAB/Python
- Uploading it into Python for further analysis.
- Voice recognition is done by matching with a reference stored speech keyword database.
- After a match is found, the spectrograph of the corresponding speech sample is plotted in Python after eliminating the noise.
- RSA algorithm or ElGamal algorithm is applied on noise eliminated signal for encryption of speech signal.
- Digital signal was then passed through reflex klystron to convert this digital signal into frequency shift keying (FSK) and transmitted to the receiver [4].

10. At the receiver, in order to identify f_1 and f_2 , FSK signal is passed through a coupler which divides the corresponding signal into two parts. These two parts contain frequencies both frequencies f_1 and f_2 [21].
11. The resulting two signals are passed through two different resonating cavities of frequencies of f_1 and f_2 to identify them and then summed up using circuit to get the original speech signal.
12. RSA decryption algorithm or ElGamal decryption algorithm (whichever is applied in Step 8) is applied on analog signal for encryption and get back original speech signal which is authenticated in Step 6.
13. Stop

4 Results and Discussions

The relevant spectrograms obtained from the voice signal preceded by the plotting of wave plots and recognition and implementation of RSA algorithm on this sound signal are shown, respectively, in Fig. 8 and Fig. 9. It shows two such samples used for

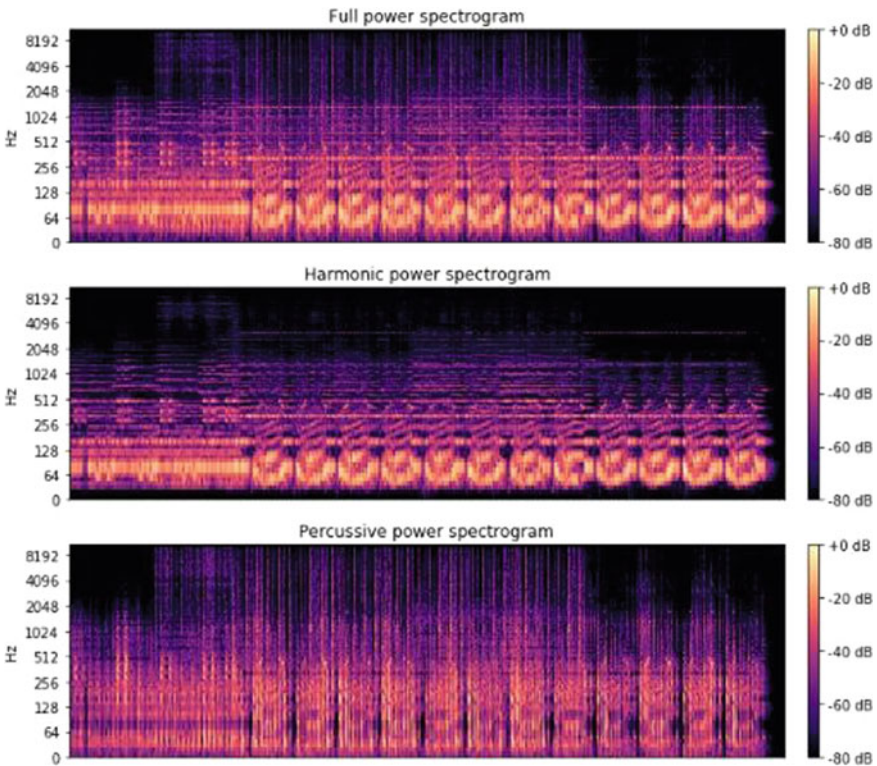


Fig. 8 The three corresponding spectrograms

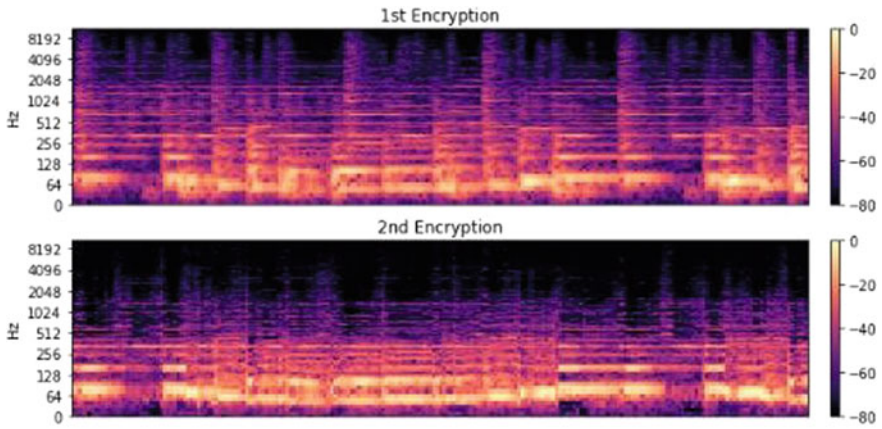


Fig. 9 Encrypted waves

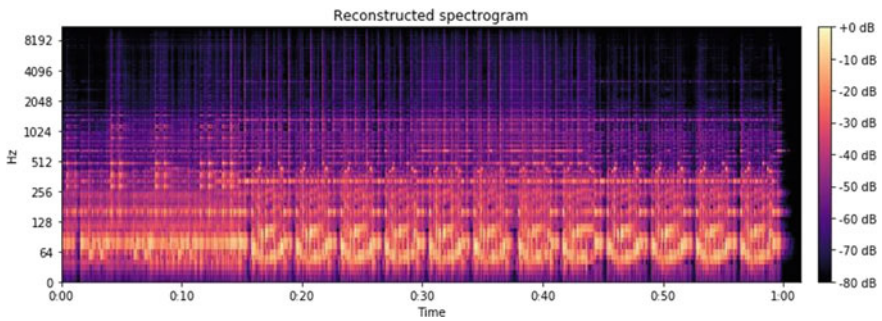


Fig. 10 Reconstructed spectrogram

recognition. The sound wave is further processed to plot its harmonic, percussive, and full power spectrogram. This is depicted in Fig. 11. These breakdowns are suitable when sound analysis is done at higher levels of processing. The ultimate stage in the encryption stage involves implementing RSA algorithm. The waveforms are shown in Fig. 9. The recovered waveform after applying the decryption algorithm is shown in Fig. 10. The graph for klystron characteristics using external modulation is shown in Fig. 11 [4, 21].

4.1 Comments on Results

The results were obtained using Python with the help of librosa, matplotlib, numpy, PIL, and glob. As the results are based on software simulation hence the original and

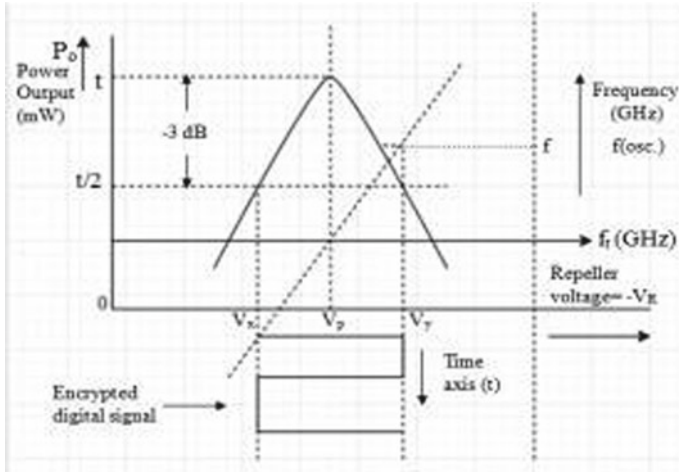


Fig. 11 Klystron characteristics using external modulation

Fig. 12 Setup of klystron characteristics using external modulation [21]



reconstructed spectrograms shows almost match. The DAC was implemented, and the output was brought into effect in the form of a LED signal.

5 Conclusion

In addition to RSA or ElGamal encryption techniques, voice-based authentication of authentic person's gives additional security. Therefore, only authentic persons data would be taken for encryption and finally for transmission. The methodology can be used in highly secured environment like in defense applications. High

power microwave devices like reflex klystron is used for generation of FSK modulated signals. The signals are reconstructed after FSK demodulation and decryption process. Similarity index of their constructed signal is very high with respect to transmitted signal. Furthermore, here the results show that the performance of RSA algorithm is faster than its ElGamal counterpart. The reason behind this is the logarithmic time complexity of the ElGamal is slower than the RSA, where prime number division arithmetic is taken into consideration. Also security of ElGamal is more challenging in comparison to RSA for the time taken to compute the discrete logarithms.

References

1. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **21**, 120–126 (1978)
2. Elgamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Trans. Inf. Theory* **31**(4), 469–472 (1985)
3. Middlestead, R.W.: Frequency shift keying (FSK) modulation, demodulation, and performance. In: *Digital Communications with Emphasis on Data Modems: Theory, Analysis, Design, Simulation, Testing, and Applications*, pp. 207–225. Wiley (2017)
4. Chakraborty, M., Mallick, A.: AES encrypted FSK generation at X-band frequency using a single reflex klystron. In: *Wireless Communication over ZigBee for Automotive Inclination Measurement*. China Communications, vol. 7, pp. 1–9 (2010)
5. Bae, H., Lee, H., Lee, S.: Voice recognition based on adaptive MFCC and deep learning. In: *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1542–1546. Hefei (2016)
6. Pak, J., Kim, M.: Convolutional neural network approach for aircraft noise detection. In: *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pp. 430–434. Okinawa, Japan (2019)
7. Tan, T.: The effect of voice disguise on automatic speaker recognition. In: *3rd International Congress on Image and Signal Processing*, pp. 3538–3541. Yantai (2010)
8. Yeo, C.Y.Y., Al-Haddad, S.A.R., Ng, C.K.: Animal voice recognition for identification (ID) detection system. In: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pp. 198–201. Penang (2011)
9. Bui, N.C., Monbaron, J.J., Michel, J.: An integrated voice recognition system. In: *ESSCIRC '82: Eighth European Solid-State Circuits Conference*, pp. 158–161. Brussels (1982)
10. Lau, Y.-K., Chan, C.-K.: Speech recognition based on zero crossing rate and energy. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 1, pp. 320–323; Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford digital library metadata architecture. *Int. J. Digit. Libr.* **1**, 108–121 (1997)
11. Wenndt, S.J., Mitchell, R.L.: Machine recognition versus human recognition of voices. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4245–4248. Kyoto (2012)
12. Yamazaki, Y., Tamaki, M., Premachandra, C., Perera, C.J., Sumathipala, S., Sudantha, B.H.: Victim detection using UAV with on-board voice recognition system. In: *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pp. 555–559. Naples, Italy (2019)
13. AlShu'eili, H., Gupta, G.S., Mukhopadhyay, S.: Voice recognition based wireless home automation system. In: *2011 4th International Conference on Mechatronics (ICOM)*, pp. 1–6. Kuala Lumpur (2011)

14. Aktar, N., Jaharr, I., Lala, B.: Voice recognition based intelligent wheelchair and GPS tracking system. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–6. Cox's Bazar, Bangladesh (2019)
15. Sharifuddin, M.S.I., Nordin, S., Ali, A.M.: Voice control intelligent wheelchair movement using CNNs. In: 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), pp. 40–43. Ipoh, Perak, Malaysia (2019)
16. Pleshkova, S., Zahariev, Z., Bekiarski, A.: Development of speech recognition algorithm and lab view model for voice command control of mobile robot motion. In: 2018 International Conference on High Technology for Sustainable Development (HiTech), pp. 1–4. Sofia (2018)
17. Obin, N., Roebel, A.: Similarity search of acted voices for automatic voice casting. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(9), 1642–1651 (2016)
18. Rashid, R.A., Mahalin, N.H., Sarijariand, M.A., Abdul Aziz, A.: Security system using biometric technology: design and implementation of voice recognition system (VRS). In: 2008 International Conference on Computer and Communication Engineering, pp. 898–902. Kuala Lumpur (2008)
19. Tahir, A.: Design and Implementation of the RSA algorithm using FPGA. *Int. J. Comput. Technol.* **14**, 6361–6367 (2015)
20. Prodeus, Kukharicheva, K.: Automatic speech recognition performance for training on noised speech. In: 2017 2nd International Conference on Advanced Information and Communication Technologies (AICT), pp. 71–74. Lviv (2017)
21. Pal, P., Sahana, B.C., Poray, J., Mallick, A.K.: Generation of encrypted FSK RF signals for secured communication inspired with high frequency technique. In: International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications (ICAISC-2020)

UD-RMM: A Remote Monitoring and Management Tool Using PowerShell Universal Dashboard for Universities



Pranav Nagarajan and Jayavignesh Thyagarajan

Abstract A Remote Monitoring and Management Tool is a software that performs multiple operations like monitoring a group of computers and observing the behavior, installing software remotely, troubleshoot the computers from a single location. Nevertheless, remotely monitoring a network of more than 500 computers is a cumbersome task. This project aims at implementing a hassle-free solution, particularly for computer laboratories in a University by developing a GUI (Graphical User Interface) based tool with a Software-as-a-Service approach. This tool is built entirely out of Microsoft PowerShell (PoSH) right from scripting the functions of each feature to designing the tool's GUI, while making use of many Networking Protocols and Standards wherever necessary. With PoSH being available as a cross-platform tool, this tool easily supports the implementation in both Windows and Linux operating systems.

Keywords Remote monitoring and management · Cross-platform tool · Powershell universal dashboard · Software-as-a-service

1 Introduction

Network Administration is still one of the areas where the majority of the Universities tend to not prioritize for daily usage. Private Universities have the luxury to invest heavily in software packages for a long term cycle [1]. A few among them land up in a dilemma of the trade-off between using more than two software to cover many features as possible by trying to minimize their expenses. Others resort to using a pay-as-you-go model, which is still an expensive solution. Though many software

P. Nagarajan (✉) · J. Thyagarajan
School of Electronics Engineering, Vellore Institute of Technology,
Chennai 600127, Tamil Nadu, India
e-mail: pranav.n2016@vitstudent.ac.in

J. Thyagarajan
e-mail: jayavignesh.t@vit.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_33

401

caters to our needs, they either lack a variety of features, or the pricing is too high. Therefore with existing open-sourced tools on the market, an attempt to create an RMM (Remote Monitoring and Management of the network) to be explicitly used in Universities has been made here. It can save a lot of workforce and time required to perform routine tasks periodically and manually on all computers [2]. Also, this tool has been made open-source (see Appendix). Hence a lot of new features can be merged into this tool by fellow open-source enthusiasts, matching the performance of expensive RMM tools on the market.

1.1 Desired Features of an RMM Tool in Universities

On a day to day basis, some of the highly demanded features by laboratory technicians are to

1. Power-on the remote computers and put computers to sleep when not in use to conserve energy, from one location.
2. Monitor screens of remote computers, produce warnings on detecting any unprofessional behavior of users, and use the remote desktop connection to a particular computer to troubleshoot any specific issue, rather than having to access it physically.
3. Restrict usage of applications in situations like lab examinations and Block/Unblock USB Drive access.
4. Install/Uninstall required applications on more than one computer in one go. If possible, reset all computers with a single OS image (a feature used at the beginning of every semester).
5. Cluster a group of computers and perform group maintenance operations. (such as group policies, Firewall rules, changing the IP address on the whole.)

In Sect. 2, a survey of existing solutions has been briefed. Section 3 explains the proposed solution in detail, and Sect. 4 contains the sample demonstrations of the proposed tool. In Sect. 5, the paper concludes with inferences and possible future works.

2 Literature Survey

Following are the software packages whose list of features offered and the pricing are reviewed here: Action1 [3], CCleaner Cloud [4], Microsoft System Center Configuration Manager [5], Microsoft Windows Deployment Services [6], Microsoft PsExec [7], Bolt by Bolt Corporation [8], LogMeIn Central [9], Microsoft Windows Admin Center [10], PUD Admin Center: Paulo Di Maggio's attempted clone of Windows Admin Center [11].

2.1 Similarities in Existing Software

Applications run with (Windows Remote Management) WinRM [12] in the back-end. The majority of the application providers make use of PsExec. PsExec is a command-line tool that enables us to execute scripts on remote frameworks and redirect the output of console applications' to the host computer. The goal is that these applications seem, by all accounts, to be running locally.

2.2 Drawbacks in Existing Software

There is no GUI based remote endpoint configuration portal available either as free-ware or as Open Sourced version. Only Microsoft System Center Configuration Manager (SCCM) and "Bolt" by Puppet are compatible with all types of Operating Systems, namely Windows, Linux, and macOS. SCCM is very expensive for small scale network (not more than 100 computers) while "Bolt" require skills in PowerShell scripting. SCCM adds huge overhead by taking up more disk space in servers and endpoints systems. Microsoft's other solution named "Windows Deployment Services," only supports remote deployment of OS images. There are also many custom packages scripted in PowerShell by individual contributors for objectives, as mentioned above. However, as mentioned earlier, they require knowledge in PowerShell scripting. Most of the offered software deals only with OS image deployment and not a single application deployment. LogMeIn is very expensive for mid-tier and lower-tier user groups. Some of its services, which are an amalgamation of various native Windows services, are offered at a higher price, can be availed for 50% of the price from other vendors.

3 Proposed Solution

This section break downs the Proposed solution in terms of (1) User Interface design, (2) Networking Concepts used and (3) Tools used in reaching the objective.

3.1 UI/UX Design

Considering the desired features of an RMM tool, we end up with a list of features and the way the GUI can be developed and shipped for end users.

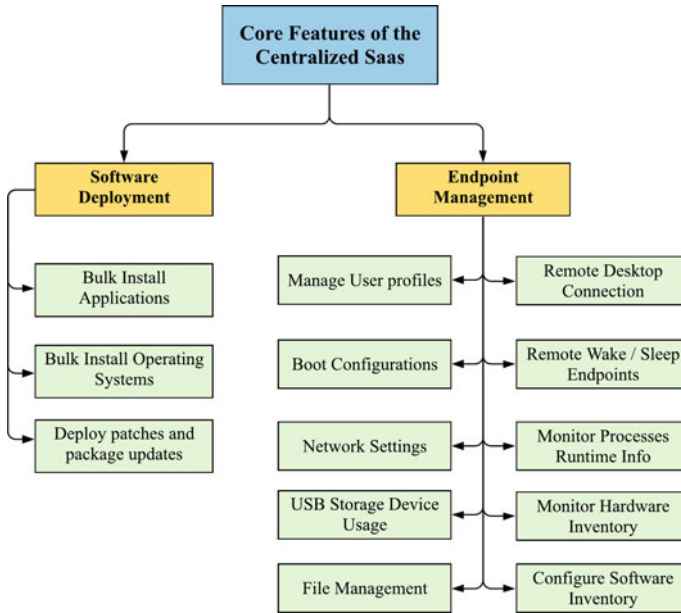


Fig. 1 Some of the core features offered by UD-RMM

The core features are represented in a block diagram, as shown in Fig. 1. The RMM tool can be considered as a centralized Software-as-a-Service (SaaS) model, as it can be invoked on any computer at any location in a network. Independent of the specifications of the computer.

3.1.1 Home Page

The RMM Tool’s home page should have standard functionalities (represented as block diagram as shown in Fig. 2) that apply to any computer in the network without any bias. The features offered on this page as seen on one computer remain the same when the page is opened on another computer.

3.1.2 Host Endpoint Page

Moving on to the layout of a host computer’s page—the contents of this page are shown in the block diagram, as seen in Fig. 3. This figure does an excellent job of covering core operations that can be performed directly on the same computer.

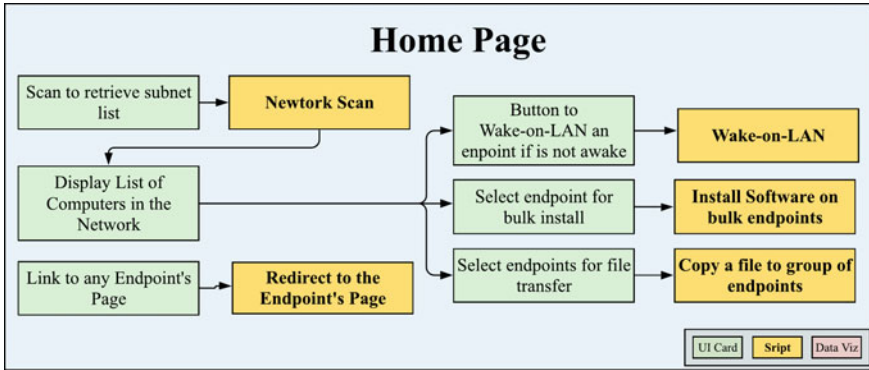


Fig. 2 Home page

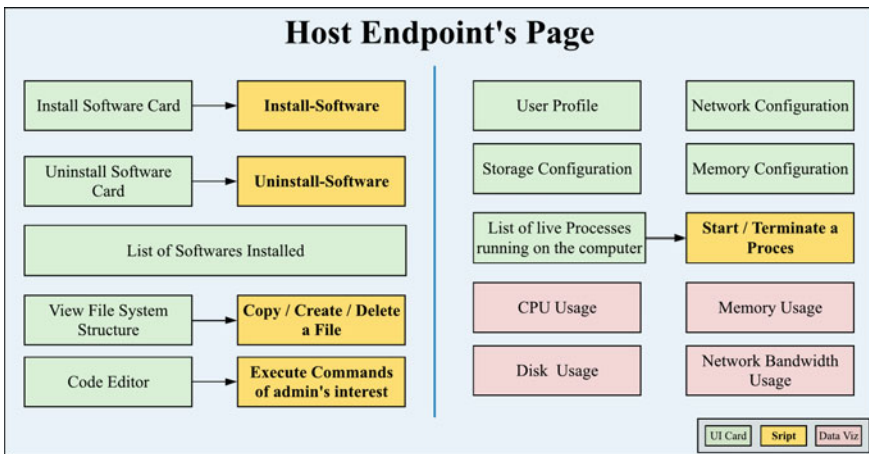


Fig. 3 Overview of features available on the Page pertaining to the host computer

3.1.3 Remote Endpoint Page

In contrast, Fig. 4 shows how the host computer's page can be modified to adjust operations to be performed on any remote computer available in the network. Some of the new features incorporated are: turning on/off the computer, establishing a remote desktop connection, restricting file system access to external storage devices (USB Drive, for example).

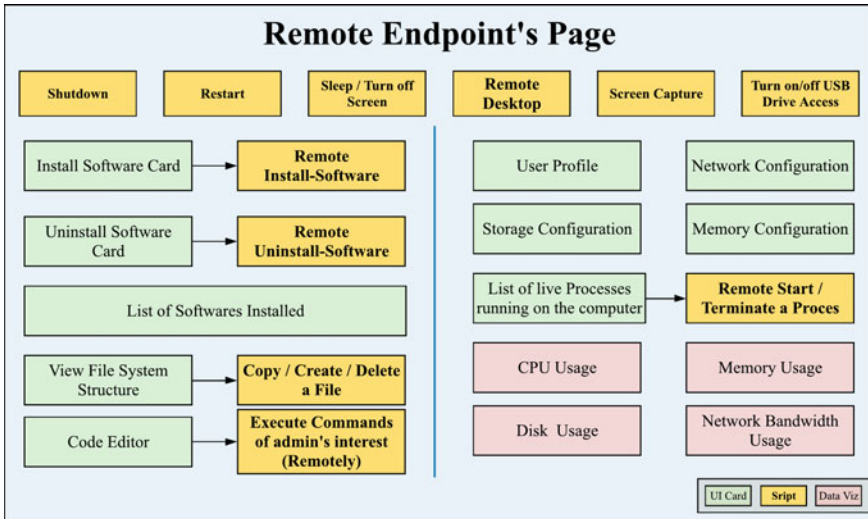


Fig. 4 Overview of features available on the page of any remote computer

3.2 Networking Concepts

3.2.1 WinRM

Windows Remote Management (WinRM) is a standard implemented by Microsoft to work on WS-Management Protocol. WS-Management Protocol is a Simple Object Access Protocol (SOAP). It allows us to enable communications hardware drivers and operating systems offered by various vendors, thereby achieving high interoperability (Fig. 5).

3.2.2 Wake-on-LAN

Wake-on-LAN (WoL) [13] is an industry-standard token-ring or Ethernet-based networking protocol that allows turning on remote computers by sending a message called “magic packet.” Recent advancements have enabled us to send a magic packet even from smartphones or, as a matter of fact, any device which is connected to the same network and has a Network Interface Card (NIC) of its own.

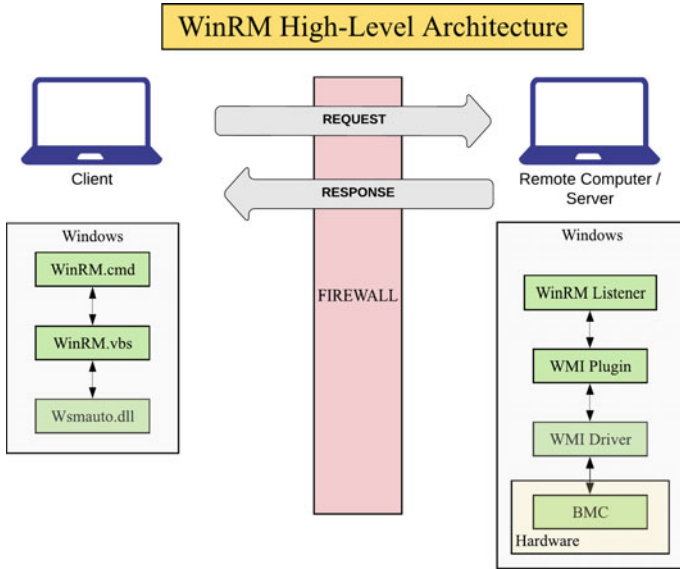


Fig. 5 Windows remote management architecture

3.3 Tools Used

3.3.1 Windows PowerShell

PowerShell is a CLI shell scripting language built on top of Microsoft Dot NET. PowerShell is object-oriented, which means every command produces an object as output. We can also ship the output object, through the pipeline, to another command as its input, which is one of the most demanding reasons to use PowerShell.

3.3.2 Universal Dashboard PowerShell Module

The “Universal Dashboard” [14] PowerShell module (also called as PoshUD) takes into consideration the production of browser-based dashboards. It makes a web server and site-dependent on PowerShell contents that we create. The client and server-side code for the Dashboard are created by only using PowerShell scripts—no compelling reason to learn JavaScript, HTML, or CSS.

4 Sample Demonstrations

Figures 6, 7 and 8 contain the demonstrations of UD-RMM.

As seen in Fig. 6, a network scan can be made either through in-built powershell script or by uploading a CSV file containing the IP Address, hostname, and password of every remote computer.

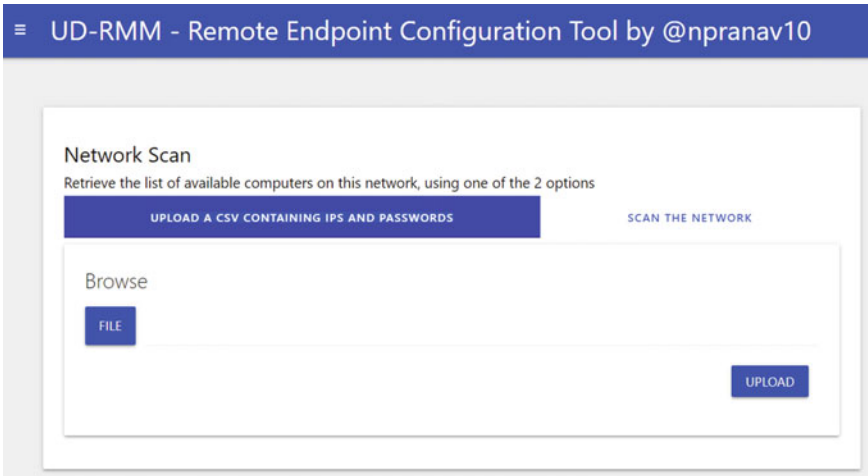


Fig. 6 “Network scan” feature as seen on homepage

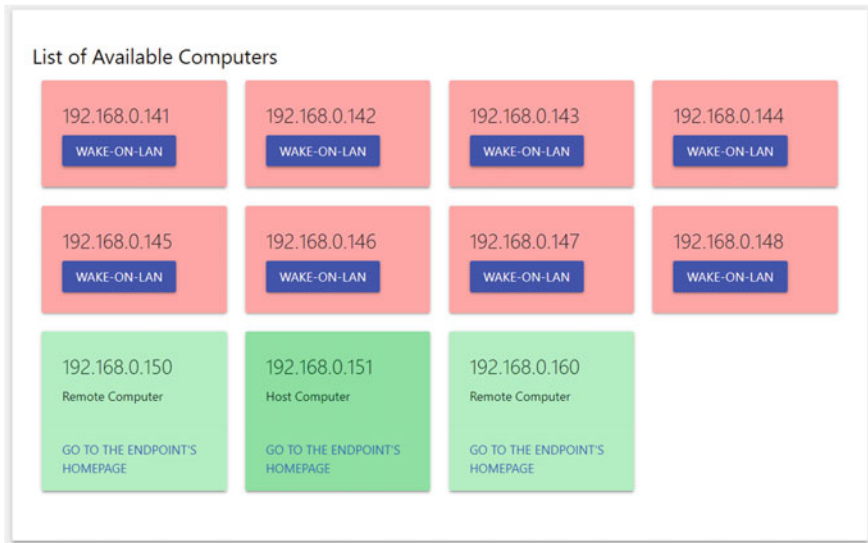


Fig. 7 “Network list” feature as seen on homepage

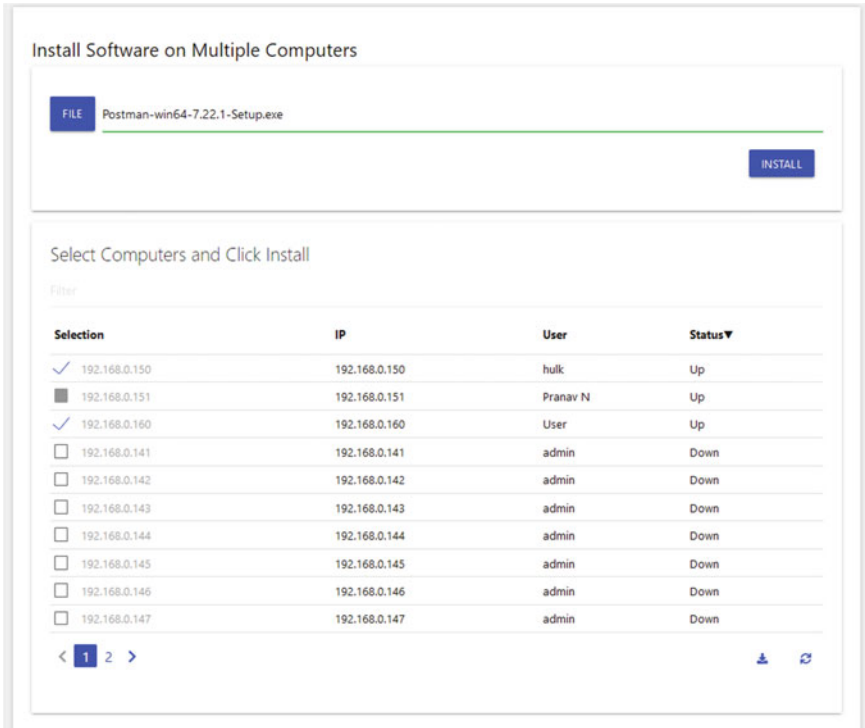


Fig. 8 Installing a software on multiple computers

Once the list is retrieved, the homepage’s “List of Available Computers” tab (Fig. 7) contains a card assigned to each one of the online computers and a link to its page. Computers that are in the list already but are powered off, can now be powered on by using the Wake-on-LAN button.

Installing a software on multiple remote computers is shown in Fig. 8. This tab allows you to select a list of computers using the checkbox aligned with computer’s IP address and later install a software on them, by browsing for an executable file on the host computer.

Few parts of the Remote computer’s page are shown in Fig. 9. Here the user can perform operations such as powering off the corresponding computer, restart it or put to sleep. One can also invoke Remote Desktop Service on host, to the remote computer and also be able to capture its screen as an image on the host computer’s clipboard.

Figure 10 portrays the Process Manager feature. One can filter a process name in the Process List grid and then click on “Terminate Filtered Processes” to stop the process on remote computer.

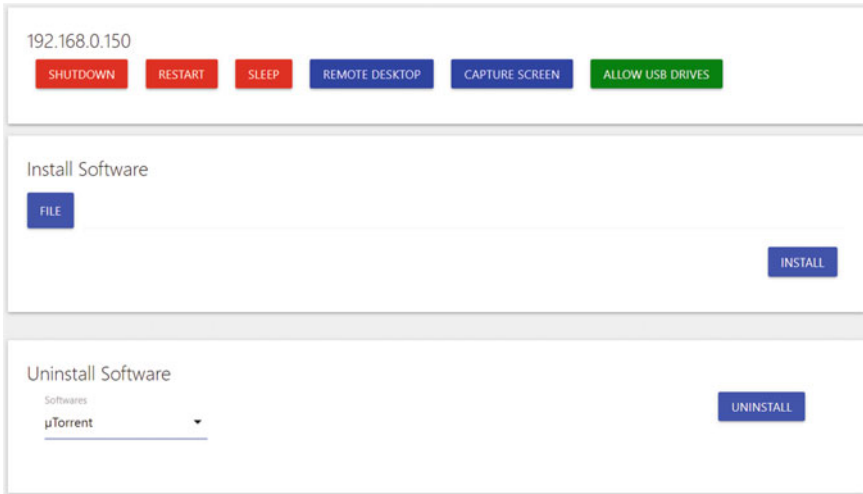


Fig. 9 A remote computer's page

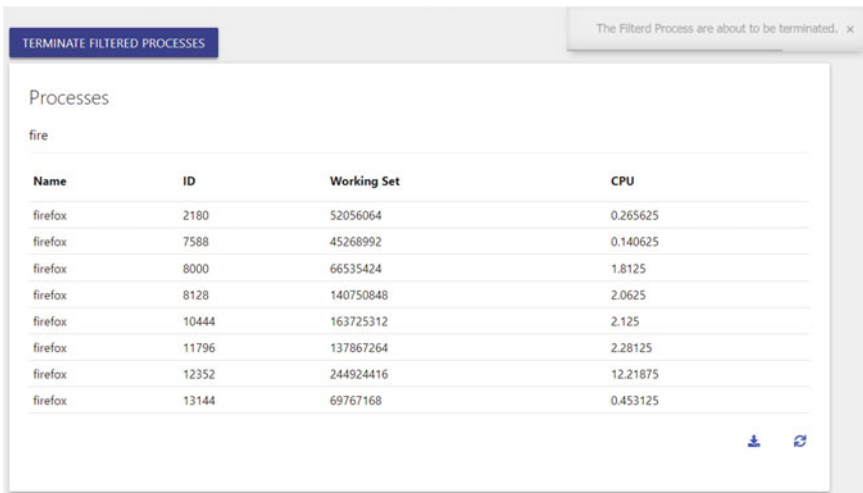


Fig. 10 Process manager on remote computer's page

5 Conclusion and Future Work

It is proven that a remote monitoring and management tool for use in Education institutes can be made free of cost from the above chapters. Also, by directly accessing the source code of this tool, we can quickly test and deploy many low-level features. These might incur extra costs and time on other vendors. It is straightforward to maintain, requires nearly zero scripting knowledge for daily use, and needs only a

team of 2–3 PowerShell developers to test and upgrade new features periodically. With the rise in the number of professional technical clubs in every University, it is only a matter of time before these projects are assigned to such groups. By doing so, Universities can now build the tool in-house.

Further works can also include moving the RMM tool ultimately to the cloud with no dependencies installed on any computer. Agreeing with authors at [15], the use of such tools becomes very apparent when one can access them outside the network.

Acknowledgements This whole project has been made possible because of the efforts by Adam Driscoll (creator of PowerShell Universal Dashboard) and Pal Sola for his guidance on designing the tool’s workflow.

Appendix: UD-RMM for public

The source code to build the UD-RMM tool can be found at <https://github.com/npranav10/udrmm>. The tool has been made open-source and is released under the Creative Commons License.

References

1. The Best Network Monitoring Tool for Universities, <https://www.helpsystems.com/resources/articles/best-network-monitoring-tool-universities>. Accessed 20 Dec 2019
2. The One IT Software Every School Should Budget for This Year, <https://www.helpsystems.com/intermapper/resources/articles/it-budgets-education>. Accessed 20 Dec 2019
3. Action1, <https://www.action1.com/help>. Accessed 20 Dec 2019
4. CCleaner Cloud, <https://www.ccleanercloud.com/>. Accessed 7 Jan 2020
5. Microsoft System Center Configuration Manager, https://en.wikipedia.org/wiki/Microsoft_System_Center_Configuration_Manager. Accessed 7 Jan 2020
6. Windows Deployment Services, https://en.wikipedia.org/wiki/Windows_Deployment_Services. Accessed 7 Jan 2020
7. Russinovich, M.: The PsExec utility, <https://docs.microsoft.com/en-us/sysinternals/downloads/psexec>. Accessed 7 Jan 2020
8. Bolt by Bolt Corporation, <https://puppet.com/docs/bolt/latest/bolt.html>. Accessed 7 Jan 2020
9. LogMeIn Central, <https://www.logmein.com/central/features>. Accessed 7 Jan 2020
10. Windows Admin Center, <https://docs.microsoft.com/en-us/windows-server/manage/windows-admin-center/overview>. Accessed 7 Jan 2020
11. PUDAdminCenterProtoype, Paulo di Maggio, <https://github.com/pldmgg/PUDAdminCenterPrototype>. Accessed 7 Jan 2020
12. Windows Remote Management (WinRM), Microsoft, <https://docs.microsoft.com/en-us/windows/win32/winrm/portal>. Accessed 7 Jan 2020
13. Wake-on-LAN, <https://en.wikipedia.org/wiki/Wake-on-LAN>. Accessed 20 Jan 2020
14. Powershell Universal Dashboard by Adam Driscoll, <https://adamdriscoll.gitbooks.io/powershell-universal-dashboard/>. Accessed 30 Jan 2020
15. Ismail, A., Hajjar, M., Hajjar, H.: Remote administration tools: a comparative study. *J. Theoret. Appl. Inf. Technol.* **140–148** (2005)

Performance and Resource-Aware Virtual Machine Selection using Fuzzy in Cloud Environment



Vikas Mongia and Anand Sharma

Abstract Cloud computing is an on-demand computing which provides elastic resources to users on pay-as-you-go basis. The ever-increasing resource demand of data-driven applications promulgated the deployment of data centers which results in high energy consumption. Therefore, efficient resource management is crucial to serve users' request. Live migration plays an important role in resource management. However, excessive migrations sometime degrade application performance. Thus, careful virtual machine selection must be done to ensure low migration count. This work develops a virtual machine selection policy: performance and resource-aware virtual machine selection using fuzzy (PRSF) that aims to utilize CPU resources to their maximum in order to reduce migration count. An attempt is also made to decrease migration time of a virtual machine by considering its main memory allocations. The policy implements Mamdani fuzzy controller to optimize the selection decision. The performance evaluation of proposed policy with benchmark algorithms has shown reduction of 32.78% and 81.7% in energy consumption levels and migration count, respectively.

Keywords VM selection · Fuzzy controller · Resource management · Service level agreement

1 Introduction

Cloud computing (CC) is a paradigm that offers utility-oriented IT services to the users [1]. The pay-as-you-go services of CC have revolutionized the industry which leads to proliferation of resource demand. To meet computational power demand, large-scale data centers are being deployed by cloud service providers (CSPs) that

V. Mongia (✉) · A. Sharma
UCCA, Guru Kashi University, Talwandi Sabo, Bathinda, Punjab, India
e-mail: vikasmongia@gmail.com

A. Sharma
e-mail: andz24@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_34

413

are inevitably responsible for high power consumption and increased CO₂ emissions into the environment. Therefore, efficient computing is foremost requirement of cloud computing industry to ensure committed service delivery keeping energy consumption (EC) and CO₂ emissions to their lowest. Infrastructure of cloud data center (CDC) implements virtualization and dynamic VM consolidation (DVMC) techniques that enables dynamic resource configuration for more efficient handling of scarce resources [2]. VM consolidation accommodates workload on limited number of servers by implementing live migration strategy. Virtual machines (VMs) are migrated between hosts with little performance disruptions, and the idle hosts are then put into sleep mode contributing toward energy-efficient cloud computing. However, aggressive migrations sometime cause poor quality of services consequently leading to service level agreement (SLA) violations. This is due to the fact that a VM during its migration demands CPU, memory, network bandwidth and other communication resources which lowers performance of running applications [3]. Therefore, selection of suitable VM for migration from overloaded host is a critical concern. This work contributes toward VM selection strategy based on fuzzy inference system that can select optimal VM for migration considering CPU and memory parameters in decision process. The main contributions of the paper are:

- Identification of an overloaded host by keeping record of resource usage by the host.
- Defining a power and performance-aware VM selection algorithm to handle trade-off between minimum migration time of VM and minimum number of migrations.
- Conducting extensive simulations on CloudSim toolkit and performance analysis between proposed policy and benchmark algorithms.

The work contributes toward energy performance trade-off that ensures minimum number of VM migrations ensuring minimum migration time along with. Rest of the paper is organized in the following manner: Sect. 2 discusses related work done in the area of VM selection decision. Sections 3 and 4 detail working of the proposed model and VM selection policy. The performance evaluation parameters are discussed in Sect. 5. Further Sects. 5 and 6 show experimental setup and results obtained after comparison of policies.

2 Literature Review

Numerous resource management techniques have been adopted by the researchers to attain the objective of high return on investments satisfying SLA constraints. The domain of static resource management is well known in cloud computing area. However, for dynamic workload environment, adaptive resource management policies are required. Nathuji and Schwan initiated power management concept on virtualized data centers [4]. They proposed the idea of resource profiling at local as well as global level, and resource reallocation decision is taken based upon the collected information. However, no specific policy is suggested for automatic resource reallocations.

To reallocate virtual resources, live migration technique is implemented in a data center that allows migration of VMs between hosts without execution suspension and with some downtime [5]. However, live migration has negative impact on application performance. Voorsluys et al. investigate the effect of migration and conclude that downtime depends upon number of pages needed to be updated during migration and it is approximately 10% of the CPU-MIPS utilization [6]. Therefore, selection of an optimal VM for migration is a crucial decision. To ensure low-performance degradation, a careful VM selection policy is needed, because a wrong selection sometimes increases reallocations and SLA violations as well.

Anton et al. suggested three policies for selection of a VM. The minimization of migration (MMT) policy considers RAM utilization levels, and the selection of VM is done that ensures least migration time. The random selection policy randomly selects a VM from an oversubscribed server. Selection of VM with highest CPU utilization is formulated by maximum utilization (MU) policy in order to ensure lesser migrations [7]. Verma et al. considered correlation between CPU utilization levels, and the VM having maximum value is selected for reallocation [8]. A modification in MC policy named maximum correlation extension (MCE) is suggested by [9]. In MCE, the VM is selected for migration having maximum positive correlation with other VMs with the belief that positive relationship is much prone to overloading of server than negative ones. However, correlation policies assume linear relationship between variables. So, consistency of proposed algorithms cannot be assured.

In another work, Anton et al. implement minimization of migration (MM) policy that selects minimum number of VMs from an overutilized server with an attempt to keep difference between upper threshold limit and new host CPU utilization levels to the minimum [10]. But this algorithm does not ensure maximum exploitation of resources [11]. So, a modification in this policy is suggested that implements backtracking algorithm to exploit entire search space and find globally optimal solution. The time and space involvement in backtracking algorithm are high which can be considered inadequacy of this algorithm.

In another approach, Zhou et al. suggest minimum product of both CPU utilization and memory size (MPCM) policy [12]. The policy considers CPU and memory as an important parameter for energy efficiency and selects a VM from an overutilized server having minimum product of these two parameters. However, prioritizing VM with least processing instructions sometime increases number of selections and thus contributes toward increased migration cost and SLA violations.

Another work suggested by Huixi et al. implements content-based VM selecting and placement (CVSP) policy. In CVSP policy, a cluster of VMs is selected based upon content similarity and is placed on same host to ensure minimum data transfer because common contents are transferred only once [13]. Another work proposes priority-based minimum migration time VM selection policy that classifies tasks into high-priority and low-priority classes. For high-priority tasks, performance is ensured even at some high energy consumption levels. But cost reduction is an important criterion for low-priority processes even sometimes by compromising SLA. The selection of the VM with lowest priority tasks for migration is done. The idea is to avoid downtime in execution of high-priority processes [14]. Masoumzadeh et al.

proposed max utilization policy that selects the VM having CPU utilization in order to quickly moderate the utilization levels of host [15]. However, migration time is not considered by the proposed strategy. Shidik et al. consider RAM and CPU parameters and implement Markov normal algorithm to optimize the selection decision [16].

The policies suggested in literature are centric toward reduction in either migration time or migration count only. Our proposed policy considers both aspects keeping performance parameters satisfied.

3 System Model

The proposed work is considered for IaaS environment where thousands of physical servers of a data center accommodates heterogeneous VMs with applications running over these machines. The environment has no knowledge of type of applications running on the VMs. Each single- or multi-core physical server is allocated CPU million instructions per second (MIPS), RAM and network bandwidth. These resources are further allocated to VMs for task execution. Figure 1 describes working of the system model. A profile engine keeps accounting of resource usage by every VM and an aggregated resource profile usage for each physical host. This profile usage is managed by hotspot detector to identify hotspots, where resource usage exceeds its upper limit. Numerous algorithms have been devised to set these threshold limits such as: median absolute deviation (MAD), local regression (LR) and interquartile range (IQR). These algorithms take into account past CPU utilizations and calculate threshold value for the physical server. On identification of an overloaded host, resource administrator is invoked for reallocation of virtual machines in order to moderate the resource usage on the host. The VM selection algorithm figures out the best VM for migration that can assure optimal resource usage keeping SLA infringement to the lowest. The selected VM is added to the migration list, and the algorithm is executed for more reallocations until the resource usage of host lies within limits. This work implements Mamdani fuzzy inference system (FIS) to optimize VM selection decision. The next subsection describes FIS briefly.

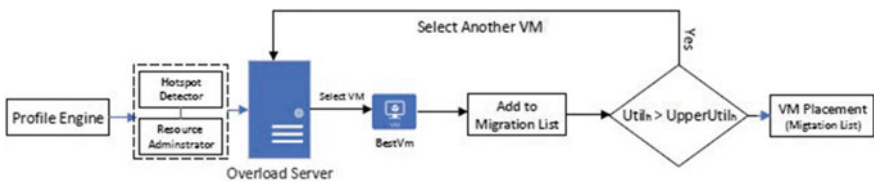


Fig. 1 System model

3.1 Fuzzy Inference System

FIS is a universal approximation system [17] suitable for environments where analytical functions are not sufficient. It implements mechanism of if-then rules to map input space or universe of discourse to an output space [18]. These rules form the knowledge base which is used by inference system. The expression of if-then rules is described by Eq. (1).

$$\text{If premise (antecedent), Then Conclusion (consequent)} \quad (1)$$

The knowledge base can have multiple rules with multiple antecedents with a consequent which is defined by r fuzzy rules. Equation (2) describes general form of these if-then propositions.

$$\text{If } \lambda_1 \text{ is } \bar{\alpha}_1^k \text{ and } \lambda_2 \text{ is } \bar{\alpha}_2^k \text{ and } \dots \lambda_n \text{ is } \bar{\alpha}_n^k \text{ Then } \delta^k \text{ is } \bar{\beta}^k \quad (2)$$

where $\lambda_1, \lambda_2 \dots \lambda_n$ represents the input variable such that $\lambda_k \in X$ (universe of discourse), $\bar{\alpha}_1^k, \bar{\alpha}_2^k, \bar{\alpha}_3^k, \dots, \bar{\alpha}_n^k$ and $\bar{\beta}^k$ are the fuzzy sets and δ^k is the consequent of k th pair. β^k is the output fuzzy set. The member function (μ) of fuzzy set $\bar{\beta}^k$ is defined in Eq. (3).

$$\mu_{\beta^k}(\delta^k) = \min[\mu_{\bar{\alpha}_1^k}(\lambda_1), \mu_{\bar{\alpha}_2^k}(\lambda_2), \mu_{\bar{\alpha}_3^k}(\lambda_3), \dots, \mu_{\bar{\alpha}_n^k}(\lambda_n)] \quad (3)$$

The value of each membership function is between 0 and 1: $\mu_{\bar{\alpha}}: \lambda \rightarrow [0, 1]$ and $(\lambda, \mu_{\bar{\alpha}}(\lambda)) \mid \lambda \in X$. The higher value of μ indicates higher membership.

After converting fuzzy inputs to fuzzy outputs using if-then rules, the rules are aggregated together to generate a single fuzzy output. This is mostly done using max operator described in (4):

$$\mu_{\beta^s}(\delta) = \max[\mu_{\beta^1}(\delta^1), \mu_{\beta^2}(\delta^2), \mu_{\beta^3}(\delta^3), \dots, \mu_{\beta^k}(\delta^k)] \quad (4)$$

After generating singleton fuzzy output, it is converted into crisp output by the process of defuzzification using some methods. This work implements center of gravity (CoG) defuzzification method.

3.2 Proposed VM Selection Policy (PRSF)

In a data center, a VM selection policy takes place when a hotspot is detected. A host is designated hotspot if its resource utilization rate exceeds maximum limits and consequently VM selection algorithm takes place to select some VMs for reallocations. This work is intended to achieve VM selection optimization in dynamic CC environment. The vision of proposed policy is two-folded. It ensures minimum

reallocations between hosts while keeping time involved in migration to the lowest. The selection decision considers CPU utilization levels and RAM allocations by a VM. Equation (5) details objective of the work.

$$R = \begin{cases} \left\{ \begin{array}{l} |S|S \in P(v_j), u_j - \sum_{v \in S} u_a(v) < T_u, \\ |S| \rightarrow \text{optimal and} \\ \sum_{v \in S} \frac{v_r}{v_b} < \sum_{v \in X} \frac{v_r}{v_b}, X \in P(v_j), X \neq S \quad \text{and} \quad |X| = |S| \\ v_j, \quad \text{if } u_j < t_l \end{array} \right. & \text{if } u_j > t_u \end{cases} \quad (5)$$

Here, S is the set of VMs selected for migration. $P(v_j)$ represents power set of all VMs on host j , and u_j represents CPU utilization of $host_j$. $u_a(v)$ is the allocated MIPS to VM_v . The variable T_u represents upper threshold limit of the host. The allocated RAM and network bandwidth to a VM are represented by v_r and v_b , respectively. If utilization of $host_j$ is greater than its upper threshold limit, then a subset of VMs (S) from the host is selected for migration in order to put utilization levels below upper threshold limits. The selection is done ensuring low migration count and migration time. Migration time is the ratio of RAM allocations to network bandwidth of a VM. A VM with high CPU utilization and lower RAM allocation is considered for migration. The detailed working of proposed policy: Performance and resource-aware VM selection using fuzzy (PRSF) is explained in Algorithm 1.

At first step, PRSF policy selects a VM from an oversubscribed host using the technique implemented in MM VM selection policy (up to line 18). The selection here is done ensuring minimum free leftover MIPS on the physical server to make better resource management and to reduce number of migrations. In order to reduce migration time as well, the selected VM and its K -nearest VMs are reconsidered based upon utilization of CPU-MIPS and allocated RAM (line 19).

The variable K can be defined as a total number of VMs whose CPU utilization ratio to selected VM (taken as τ) lies in the interval $0.75 \leq \tau \leq 1.25$. The inputs (RAM, τ) of selected VM are passed through fuzzy controller (FC)(line 20–23), and the VM that provides better trade-off between number of migrations and migration time is selected for migration. There are five and three input values for the linguistic variables MIPS(τ) and RAM, respectively, defined as: MIPS (τ): {Very High, High, Moderate, Low, Very Low} RAM: {Micro, Small, Large}

The fuzzy values for these variables are computed using triangular and trapezoidal membership functions as described in Fig. 2. The crisp output selection decision is observed by calculating degree of membership for each fuzzy input. Value of degree of membership ($\mu_a(x)$) lies between 0 and 1. The degree of membership is calculated for each fuzzy input value based upon strength of fuzzy rules. After computing the rule strength, all inputs are aggregated by the defuzzifier to generate one defuzzified output. The scheme used for defuzzification is center of gravity (CoG). It determines an aggregated crisp output considering all influences received from rules for particular input values defined as (6):

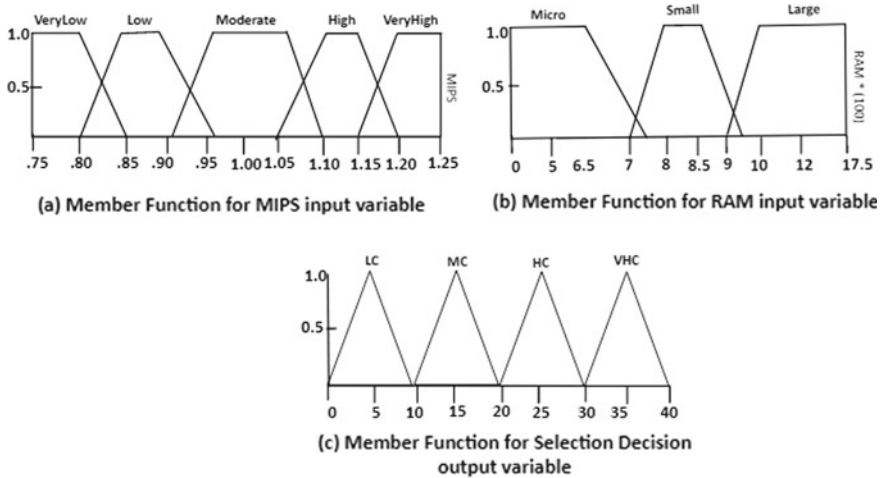


Fig. 2 Membership functions for input and output variables

Table 1 Rule base for inference engine

RAM\MIPS	Very low	Low	Moderate	High	Very high
Micro	MC	HC	VHC	HC	MC
Small	LC	MC	HC	MC	LC
Large	LC	LC	MC	LC	LC

$$\mu = \frac{\sum_{i=1}^m \beta_i \times \mu(\beta_i)}{\sum_{i=1}^m \mu(\beta_i)} \tag{6}$$

Here, β_i represents cell value and $\mu(\beta_i)$ represents degree of membership value for β_i . Fuzzy rule base applied in the research is summarized in Table 1. Working principle of FIS can be defined in (7).

$$\{x \in S \mid \forall y \in S, \text{DeFuzzified_output}(x) \geq \text{DeFuzzified_output}(y)\} \tag{7}$$

The VM having highest defuzzified value is selected for migration (line 26–28). The selected VM is added to migration list, and utilization of the host is again calculated (line 29–31). If it still exceeds the upper threshold limit, then the entire process is continued to select another VM. On detection of an underutilized host, then its all VMs are added to migration list and the host is put into sleep mode in order to attain EE (line 32–34).

Algorithm 1: Performance and Resource-aware VM Selection using Fuzzy (PRSF)

```

1 Input: HostList, List of overloaded hosts
2 Output: migrationList, List of selected VMs for migration
3 for ( $i = 1 : \text{HostList}$ ) do
4   vmList  $\leftarrow$  i.getVmList()
5   vmList.sortDecreasingUtilization()
6   hUtil  $\leftarrow$  i.getUtil()
7   bestFitUtil  $\leftarrow$  MAX
8   while ( $hUtil > \text{THRESH\_UP}$ ) do
9     foreach  $vm \in \text{vmList}$  do
10      if  $vm.getUtil() > hUtil - \text{THRESH\_UP}$  then
11         $t \leftarrow vm.getUtil() - hUtil + \text{THRESH\_UP}$ 
12        if  $t < \text{bestFitUtil}$  then
13          bestFitUtil  $\leftarrow$  t
14          bestFitVm  $\leftarrow$  vm
15      else
16        if  $\text{bestFitUtil} = \text{MAX}$  then
17          bestFitVm  $\leftarrow$  vm
18          break
19   nearestVmList  $\leftarrow$  getK_nearestVMs (bestFitVm)
20   foreach  $vm \in \text{nearestVmList}$  do
21     ratio  $\leftarrow$  vm.getUtil()/bestFitVm.getUtil()
22     setVariable("RAM", vm.getRam())
23     setVariable("MIPS", ratio)
24     FuzzyInferenceSystem.evaluate()
25     metric  $\leftarrow$  getVariable("Decision")
26     if  $\text{metric.getdefuzzifiedValue}() > \text{MIN}$  then
27       bestFitVm  $\leftarrow$  vm
28       MIN  $\leftarrow$  metric.getdefuzzifiedValue()
29   hUtil  $\leftarrow$  hUtil - bestFitVm.getUtil()
30   migrationList.add(bestFitVm)
31   vmList.remove(bestFitVm)
32   if  $hUtil < \text{THRESH\_LOWER}$  then
33     migrationList.add(i.getVmList())
34     vmList.remove(i.getVmList())
35 return migrationList

```

4 Experimental Setup

To test efficiency of proposed algorithm, a simulation is carried out for one day (24 h) on CloudSim toolkit [19].

4.1 Data Center Configuration

To carry out simulation, a data center with 800 hosts has been created with two different configurations: HP Proliant G4 (Xeon 3040 processor, dual core, 1860 MHz CPU frequency, 4 GB RAM) and HP Proliant G5 (Xeon 3075 processor, dual core, 2660 MHz CPU frequency, 4 GB RAM). Power utilization in a host at varying utilization levels is defined as per SPECpower [20] datasets as described in Table 2.

The host can accommodate 4 types of VMs: high CPU medium instance (CPU frequency—2500 MIPS, RAM—0.85 GB), extra large instance (CPU frequency—2000 MIPS, RAM—3.75 GB), small instance (CPU frequency—1000 MIPS, RAM—1.70 GB) and micro-instance (CPU frequency—500 MIPS, RAM—0.61 GB). In this experiment, it is assumed that each VM can execute one task at a time and the length of task is kept high so that execution can be continued till simulation ends. The utilization of VM is set according to CoMon project dataset [21]. It is a monitoring system that uses infrastructure of PlanetLab and monitors percentage of CPU-MIPS requested by VMs located at geographically dispersed locations. The utilization levels are observed every 5 min.

4.2 Performance Parameters

Providing QoS to customers is responsibility of every CSP. QoS is defined in terms of SLA between CSP and the client. It specifies minimum quality levels to be provided to the client and the penalties if the performance is not achieved. In this study, performance is measured in terms of energy consumption (EC), energy performance metric (EPM) and service level agreement violation (SLAV) parameters suggested by [7]. Next section briefly describes these parameters.

Table 2 Power utilization of hosts on varying utilization levels

Host	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Proliant G4	86	89.4	92.6	96	99.5	102	106	108	112	114	117
Proliant G5	93.7	97	101	105	110	116	121	125	129	133	135

4.2.1 Energy Consumption (EC)

This metric computes EC by all hosts involved in simulation. In this work, EC is computed considering CPU utilization level in a host and then mapping is done using SPECpower analytical model. This model shows EC by hosts at varying CPU utilization levels. Energy consumption can be defined in (8).

$$EC = \sum_{i=1}^h E(u_i) \quad (8)$$

Here, u_i represents CPU utilization of a $host_i$ in % and $E(u_i)$ is the EC for the utilization level according to SPECpower analysis.

4.2.2 SLA Violation (SLAV)

It measures the time for which desired QoS is not provided to the user. The violations occur when a VM could not get its requested resources. In a host, SLAV can be defined as product of service level violations per active host (SLATAH) which measures the ratio of time for which an active host experienced 100% utilization and performance degradation due to migration (PDM) which measures the performance disruptions caused due to involvement of extra resources in migration of a VM. It is assumed that a penalty is paid by CSP to user in case of SLAV.

4.2.3 Energy Performance Metric (EPM)

EPM is the combination of EC and SLAV defined as: $SLAV * EC$. There is trade-off between the two. The attempt for energy reduction results in increased SLA violations and vice versa. The algorithm is considered efficient if it ensures lower EC keeping SLA intact.

5 Results and Discussion

To evaluate the performance of suggested policy, an experiment on CloudSim toolkit has been conducted. It is a simulation platform for CC environment. Since on real clouds, repeatability of larger experiments is quite challenging so simulation tool is adopted for performance comparison of policies. The results of PRSF policy are compared with benchmark VM selection algorithms as shown in Table 3. The algorithms considered in the study are: maximum correlation (MC), minimum migration time (MMT), random selection (RS), maximum utilization (MU) and minimization of migration (MM).

Table 3 Comparative results of VM selection policies

Experiment	EC	Migrations	SLATAH	PDM $\times(10)$	SLAV	EPM $\times(10^{-2})$
MAD_MC	176.13	23,691	7.06	1.05	0.74	1.30
MAD_MMT	184.88	26,292	5.03	0.66	0.33	0.61
MAD_MU	200.40	30,051	7.57	0.67	0.51	1.02
MAD_RS	176.57	24,169	7.08	1.04	0.74	1.3
MAD_MM	105.86	3354	40.81	0.12	0.51	0.54
MAD_PRSF	108.17	3212	33.7	0.12	0.40	0.43
IQR_MC	177.1	23,035	6.9	1.02	0.70	1.24
IQR_MMT	188.86	26,476	4.96	0.63	0.31	0.59
IQR_MU	204.22	29,901	7.41	0.65	0.48	0.98
IQR_RS	179.59	23,765	6.95	1.02	0.71	1.27
IQR_MM	138.06	12,417	28.37	0.54	1.54	2.13
IQR_PRSF	107.93	2856	31.04	0.09	0.28	0.29
LR_MC	150.33	23,004	6.97	0.97	0.68	1.02
LR_MMT	163.15	27,632	5.84	0.79	0.46	0.76
LR_MU	174.24	29,555	8.18	0.72	0.59	1.03
LR_RS	150.36	23,064	7.20	0.99	0.72	1.08
LR_MM	137.62	14,986	23.62	0.62	1.46	2.01
LR_PRSF	111.49	2729	21.5	0.085	0.19	0.20

Three threshold policies: MAD, IQR and LR are considered to fix threshold limits for the host. If utilization of a host exceeds these threshold limits, then selection of some VMs is done by VM selection policies to migrate these VMs to other host. PABFD allocation policy is considered for VM placement. The number of migrations in PRSF policy is significantly reduced (Fig. 3a) due to the fact that the policy ensures maximum resource utilization in a host. The VMs are selected for migration that ensures minimum leftover MIPS on the host, thus contributing to reduced number of migrations. Table 4a shows % decrease in number of migrations in PRSF policy as compared to counterpart algorithms with different threshold policies. An average reduction of 87.4%, 89%, 90.2%, 87.6% and 54.3% in migrations as compared to MC, MMT, MU, RS and MM algorithms, respectively, has been observed.

The proposed policy outperforms other policies in terms of energy consumption as well (Fig. 3b). Average reduction in EC of PRSF in comparison with different selection policies ranges between 12.87% and 43.06% as described in Table 4b. The main cause of high EC is due to inefficient utilization of servers. A survey by Google on 5000 servers has been done for 6 months, and it showed that the servers are utilized only 10–50% and rest of their capacity is wasted [22]. According to another report, a completely idle server consumes 70% of energy [23]. So, in PRSF policy,

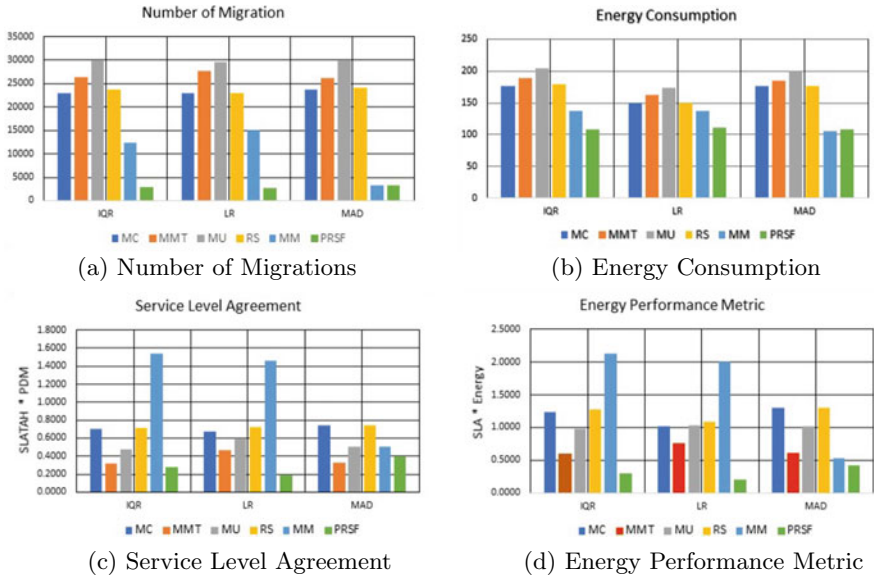


Fig. 3 Performance comparison of VM selection policies

Table 4 Result comparison of PRSF policy with benchmark policy

Threshold	% Reduction in Migrations				
Policies	MC	MMT	MU	RS	MM
IQR	87.6	89.2	90.4	87.9	77.0
LR	88.1	90.1	90.8	88.2	81.8
MAD	86.4	87.8	89.3	86.7	4.2

(a)

Threshold	% Reduction in EC				
Policies	MC	MMT	MU	RS	MM
IQR	39.1	42.9	47.2	39.9	21.8
LR	25.8	31.7	36.0	25.9	19.0
MAD	38.6	41.5	46.0	38.7	-2.2

(b)

Threshold	% Improvement in SLAV				
Policies	MC	MMT	MU	RS	MM
IQR	60.5	12.0	42.3	60.9	82.0
LR	72.2	59.4	68.2	73.8	87.1
MAD	46.3	-19.9	22.1	46.2	21.8

(c)

Threshold	% Improvement in EPM				
Policies	MC	MMT	MU	RS	MM
IQR	75.9	49.7	69.5	76.5	86.0
LR	79.9	73.0	80.2	81.1	89.8
MAD	67.0	29.9	58.0	67.0	20.1

(d)

an attempt is made to keep server utilization to its maximum so that workload can be accommodated on limited servers and remaining idle servers can be put to low power mode in order to attain energy efficiency.

Delivery of committed services is a liability of cloud service provider; otherwise, penalty per violation is to be paid to the customer. Therefore, SLA infringement is crucial to be considered by algorithms. Our proposed policy ensures minimum SLA

violations in contrast to other policies (Fig. 3c). This is due to the fact that an attempt is made to reduce migration time that results in lowered PDM with some increase in SLATAH. This is because an attempt is made to keep server maximum utilized in order to ensure lower energy consumption which increases risk of server overutilization. There is an average reduction of 59.67%, 17.16%, 44.21%, 60.31% and 63.63% in SLAV as compared to MC, MMT, MU, RS and MM policies, respectively (Table 4c).

The algorithm is termed efficient if it balances trade-off between EC and SLA violations. In a data center, with the growth in server utilization levels the EC reduces but with percentage of SLA violations. Thus, low EC by a server satisfying quality of services is highly desirable. The proposed policy PRSF efficiently handles the trade-off. The value of EPM is highly optimal as compared to its counterparts (Fig. 3d). There is an average reduction of 74.3%, 50.8%, 69.2%, 74.88% and 65.28% as compared to MC, MMT, MU, RS and MM policies, respectively (Table 4d).

6 Conclusion and Future Scope

This work plays a significant role in optimization of VM selection decision. It implements FIS that takes into account CPU usage and RAM allocations in decision taking. The objective of proposed policy is two-folded. Firstly, it ensures maximum resource utilization on a host in order to reduce migration count and to accommodate workload on limited servers to contribute toward EE. The experimental results have shown average 32.78% reduction in EC by proposed policy as compared to other benchmark policies. Secondly, the policy ensures reduced migration time of a VM that positively impacts SLA violations. The experiment shows reduction of average 48.99% in overall SLA violations. In the future, work is planned to consider more resources like network bandwidth in decision process and the evaluation will be done on more complex realistic datasets.

References

1. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. National Institute of Standards and Technology (2011)
2. Graubner, P., Schmidt, M., Freisleben, B.: Energy-efficient virtual machine consolidation. *IT Professional* **15**(2), 28–34. IEEE (2012)
3. Huang, Q., Gao, F., Wang, R., Qi, Z.: Power consumption of virtual machine live migration in clouds. In: Third International Conference on Communications and Mobile Computing, pp. 122–125. IEEE (2011)
4. Nathuji, R., Karsten, S.: Virtualpower: coordinated power management in virtualized enterprise systems. *ACM SIGOPS Oper. Syst. Rev.* **41**(6), 265–278 (2007)
5. Ye, K., Jiang, X., Huang, D., Chen, J., Wang, B.: Live migration of multiple virtual machines with resource reservation in cloud computing environments. In: 4th IEEE International Conference on Cloud Computing, pp. 267–274. IEEE (2011)

6. Voorsluys, W., Broberg, J., Venugopal, S., Buyya, R.: Cost of virtual machine live migration in clouds: a performance evaluation. In: IEEE International Conference on Cloud Computing, pp. 254–265. Springer, Berlin, Heidelberg (2011)
7. Beloglazov, A., Buyya, R.: Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency Comput. Pract. Exp.* **24**(13), 1397–1420 (2012)
8. Verma, A., Dasgupta, G., Nayak, T.K., De, P., Kothari, R.: Server workload analysis for power minimization using consolidation. In: Proceedings of the 2009 Conference on USENIX Annual Technical Conference pp. 28–28. USENIX Association (2009)
9. Cao, Z., Dong, S.: Dynamic VM consolidation for energy-aware and SLA violation reduction in cloud computing. In: 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, pp. 363–369. IEEE (2012)
10. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Gener. Comput. Syst.* **28**(5), 755–768 (2012)
11. Chien, N.K., Dong, V.S.G., Son, N.H. and Loc, H.D.: An efficient virtual machine migration algorithm based on minimization of migration in cloud computing. In: International Conference on Nature of Computation and Communication, pp. 62–71. Springer, Cham (2016)
12. Zhou, Z., Abawajy, J., Chowdhury, M., Hu, Z., Li, K., Cheng, H., Alelaiwi, A.A., Li, F.: Minimizing SLA violation and power consumption in cloud data centers using adaptive energy-aware algorithms. *Future Gener. Comput. Syst.* **86**, 836–850 (2018)
13. Li, H., Li, W., Wang, H., Wang, J.: An optimization of virtual machine selection and placement by using memory content similarity for server consolidation in cloud. *Future Gener. Comput. Syst.* **84**, 98–107 (2018)
14. Nadeem, H.A., Fadel, M.A.: Priority-aware virtual machine selection algorithm in dynamic consolidation. *Int. J. Adv. Comput. Sci. Appl.* **9**(11), 416–420 (2018)
15. Masoumzadeh, S.S., Hlavacs, H.: Dynamic virtual machine consolidation: a multi agent learning approach. In: IEEE International Conference on Autonomic Computing, pp. 161–162. IEEE (2015)
16. Shidik, G.F., Azhari, A., Mustofa, K.: Improvement of energy efficiency at cloud data center based on fuzzy Markov normal algorithm VM selection in dynamic VM consolidation. *Int. Rev. Comput. Softw. (IRECOS)* **11**(6), 511–520 (2016)
17. Precup, R.E., David, R.C.: *Nature-Inspired Optimization Algorithms for Fuzzy Controlled Servo Systems*. Butterworth-Heinemann (2019)
18. Ross, T.J.: *Fuzzy Logic with Engineering Applications*. Wiley (2015)
19. Buyya, R., Ranjan, R., Calheiros, R.N.: Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: challenges and opportunities. In: IEEE International Conference on High Performance Computing & Simulation, pp. 1–11. IEEE (2009)
20. Lange, K.D.: Identifying shades of green: the SPEC power benchmarks. *Computer* (3), 95–97. IEEE (2009)
21. Park, K., Pai, V.S.: CoMon: a mostly-scalable monitoring system for planetlab. *ACM SIGOPS Oper. Syst. Rev.* **40**(1), 65–74 (2006)
22. Barroso, L.A., Hölzle, U.: The case for energy-proportional computing. *Computer* **40**(12), 33–37 (2007)
23. Fan, X., Weber, W.D., Barroso, L.A.: Power provisioning for a warehouse-sized computer. *ACM SIGARCH Comput. Arch. News* **35**(2), 13–23 (2007)

Redefining Data Dimensionality Through Dynamic Linkages in Data-Space Continuum



Benard Alaka and Bernard Shibwabo Kasamani 

Abstract Dimensionality of data has been overly defined by scholars as the number of attributes a dataset holds. As such, high dimensional data is described as a curse to most analytics tools and algorithms; since, even a small increase in the number of attributes grows the predictive space significantly; so much so that the model is no longer as accurate as it ought to be. This study however seeks to redefine data dimensionality by first viewing data as a vector that exists within a data space continuum. Following the redefinition of dimensionality, this study proceeds to show the robustness and benefit of highly dimensional data in improving the quality of analytics. This is achieved in this study by introducing time as an important scalar in the data space continuum and showing the contribution of considering time during data analysis. This contribution explains the highly improved ability of the predictive space to not only being restricted to showing the how or to what extent attributes affect each other or the target output but also explains the pointer reasons why predicted events occur with relevance to a time series. This ability is thus deemed as a blessing to analysis following the redefinition of data dimensionality. The assumption made is that data points existing during different time stamps experience a variety of effects resulting from natural or artificial third parties.

Keywords Data vector · Data-space continuum · Data dimensionality · Timestamp

1 Introduction

In a world where data is treasured, there has arisen an age of massive data capturing for purposes of data analysis. [1–3] affirm that data is more and more easily acquired and stored, due to huge progresses in sensors and ways to collect data on one side, and in storage devices on the other side. The shadow effect of this has been that with

B. Alaka (✉) · B. S. Kasamani
Strathmore University, Nairobi, Kenya
e-mail: balaka@strathmore.edu

B. S. Kasamani
e-mail: bshibwabo@strathmore.edu

continuous gathering of data, there has been similar rise of organizing these data vectors, whose dimension correspond to the number of simultaneous measurements of the phenomenon of interest.

Most prediction algorithms have been very instrumental in the computation of either classification or regression problems. However, very few have gone a step further to rationalize the predictions gotten from the machine learning algorithms. With regards to the human problem, this is said to solve the problem only to some degree. Much is still to be explored yet as to the explanations regarding the causality of the different predictions.

It has become a preference to treat some of these machine learning algorithms as being black-boxes and thus giving leeway for little explanation of both the inner workings and the final output. This is bound to change especially with the advent of transparent deep learning, where humans are interested in knowing the inner workings as well as the reasons for the outputs of the different machine learning algorithms.

Another of the most popular theories used for purposes of vaguely explaining the inner workings of machine learning algorithms is dimensionality of data. This theory was developed by [3] and has since then been used as a measure for limiting the number of attributes allowable for an accurate machine learning model.

These next sections of this paper are Literature review, Methodology, Results and Discussion, and Conclusion.

2 Literature Review

The formal definition for dimensionality of the data as explained by [4] is the number of variables that are measured on each observation. It has been proven that dealing with high dimensional data during analysis poses a myriad of mathematical challenges. Additionally, having high dimensional data representations introduces redundancy posed by the inconveniences of variables correlation. This collinearity results into numerical instability whereby only a small sub-space of the original representation space is populated by the sample and by the underlying process.

The concept of the curse of dimensionality is later introduced as the problem of finding structure in data embedded in a high dimensional space [5]. This is emphasized by the fact that the more features a data set possesses, the more data points are needed to fill a space. This is analogized by representing data as a hypercube with length of one (1). A simulation is done to subdivide the hypercube into N possible pieces and thus the dimension for each hypercube becomes:

$$d = \left(\frac{1}{N} \right)^{1/n} \quad (1)$$

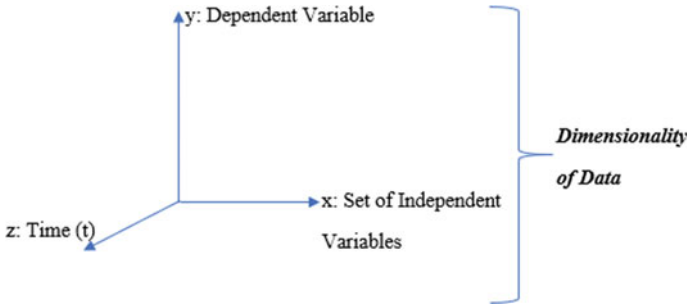


Fig. 1 Dimensionality of data

To perform a simple operation such as finding the volume (Euclidian volume for the hypercube with each measuring r units), we would have:

$$V = r^d \tag{2}$$

For a finite N , d converges to 1 when n goes to infinity. This illustrates the effect of a small alteration of dimension to a datapoint within a data space. In the worst-case scenario, it implies that the order of the Euclidian volume increases exponentially and similarly takes exponential time to solve. This problem has overly been alluded by classical researchers and modern alike as the curse of dimensionality.

A data dimension in this study is differently translated to physically imply the spatial orientation of a data point (as a vector) within a spatial data space continuum with respect to having the perspectives of *at least* the x , y and z co-ordinates. The x and y are treated as the independent variable and dependent variable, respectively (which are not collinear but rather possess a reasonable coefficient of correlation). Time as a scalar property is introduced as the z -axis (third dimension) that could yet be extrapolated further into a fourth dimension as discussed in subsequent sections of this paper. See Fig. 1.

2.1 Case: A Regression on Einstein’s Train Analogy

We allude to Einstein’s case of a moving train that is used to prove the theory of relativity with reference to Fig. 2.

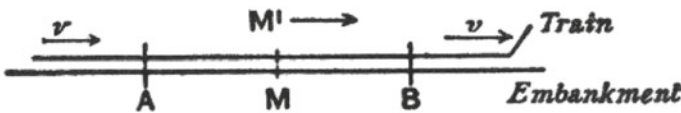


Fig. 2 Einstein’s relativity train [6]

We modify this case to suppose that a very long train travels along the rails with the oscillating velocities v_1, v_2, v_3 ; and in the direction indicated in Fig. 2. People travelling in this train use the train as a rigid reference-body (co-ordinate system); they regard all *events* in reference to the train. Then every *event* which takes place along the line also takes place at a particular point of the train.

We also have lightning strokes A and B , which are simultaneous with respect to the embankment. The rays of light emitted at the places A and B , where the lightning flash is experienced, meet each other at the mid-point M of the length $A \rightarrow B$ of the embankment. But the events A and B also correspond to positions A and B on the train. M' then becomes the mid-point of the distance $A \rightarrow B$ on the travelling train.

An observer sitting in the position M' in the train (considered with reference to the railway embankment) is hastening toward the beam of light coming from B , whilst he is riding on ahead of the beam of light coming from A . Hence, the observer will see the beam of light emitted from B earlier or later than he will see that emitted from A . This will depend on the oscillation point of the velocity as influenced by the events occurring between the points $A \rightarrow B$.

2.2 Kernel Trick

Other scholars have attempted to treat this curiosity about the hidden behavior of data points using the infamous Kernel trick originally formulated by [7, 8]. This approach works by mapping data points from a 2D input space into a 3D feature space.

The questions still remain as to whether the second dual purpose (b) has been met. Thus far, the Kernel trick has only served to explicitly map the dimensional different data points for a scenario that is obviously not linearly separable. As to why the alterations in the independent variables affect varies the dependent variable as it were still remaining remotely unknown.

Kernel trick in itself as an approach hence fails to comprehensively represent high dimensional spaces with respect to purpose (b). However, the application of the same to map to a higher feature space (third dimensional) is improved on in this study as discussed in Sect. 4.

3 Methodology

This study redefines the parameters of dimensional data thereby giving data dimensionality a whole new perspective that will aid in improving the predictive space for instances where data points in non-extrapolated spaces appear conjoined but are actually separable in higher dimensional spaces. An analogy is introduced as case study as an explanatory approach for deriving simple mathematical inferences for the purposes of redefining data dimensionality thus making this study a mixed research [9].

The data points are first truncated, and the logistic regression model of 82% prediction accuracy is fitted using the sigmoid function as in.

$$f_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{5}$$

A plot performed on the model yields Fig. 3. Of interest to this study was to obtain a detailed explanation that is dual in purpose. That is, an interpretation that would serve to explain:

- a. How the independent variables affect the dependent variable (*purpose (a)* here forth).
- b. Why the alterations in the independent variables affect vary the dependent variable as it were (*purpose (b)* here forth).

Table 2 shows the nature of relationship between the output of the measures (attributes) treated by a sigmoid to regress the classification problem of whether a delay in observing the lightning will occur (1) or not (-1).

Only the first purpose (a) is met through the coefficients which are log-odd units explaining the individual relationship between the attributes and the dependent variable as is outlined in Table 2. These basically translate into the measures (positive or

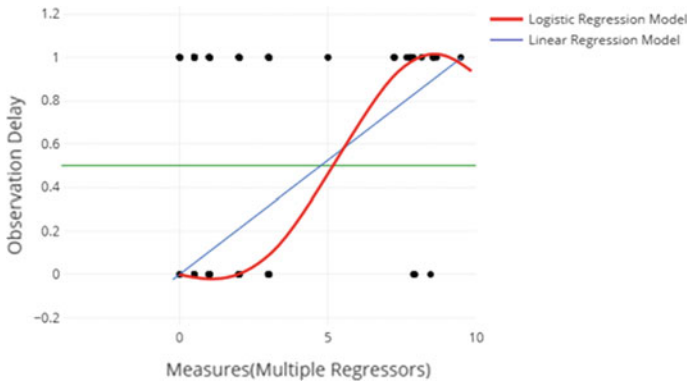


Fig. 3 Simple logistic regression (Einstein’s analogy)

Table 2 Coefficient of correlation (Einstein’s train regression)

Index	Attribute	Coefficient
0	v_1	-0.133024864
1	v_2	-0.453195398
2	v_3	0.342003721
3	cfreq	0.328363959
4	e_txtre	0.651475266
5	vsblty	-0.007765495

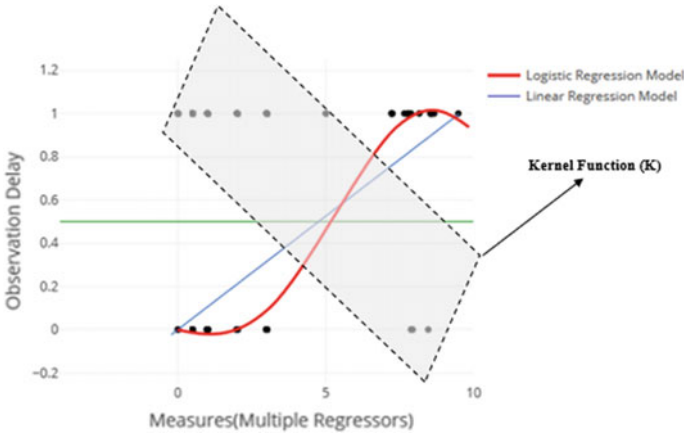


Fig. 4 Regression plotting with kernel function

negative) by which 1 log-odd unit change in the independent variable would influence the dependent variable.

In relation to our case, the data points which lie outside the regression lines could cheaply be considered as outliers to the model. However, this is not necessarily the case since the data considering the fact that our data points lie within a highly dimensional space and therefore a two-dimensional mapping would definitely leave out a whole other dimension(s). Thus, a Kernel function (K) is applied to the sigmoid activation function as:

$$K(f_0(x), y) = \sum \Phi(f_\theta(x))_i, \Phi(y)_i \tag{6}$$

Passing the x vectors and y through the dual function and plotting thus yields the plot shown in Fig. 4.

Barely any information is rendered with relation to the second purpose (b) as to even why the third oscillation (v_3) of the train’s velocity in contrary affects the delay in observation positively, unlike the first and second oscillations.

4.2 Experiment Results

Considering the Einstein’s Train Analogy, rather than viewing each attribute independently, dimensionality would have it that all the attributes have a correlation by virtue of experiencing a similar varying construct, *time*. As such, an informed assumption is made that as the oscillating velocities v_1, v_2 and v_3 approach an unknown critical oscillation range, simultaneity in observation might be achieved between the train boarder and the observer, which is nearly not likely.

Table 3 Preliminary dimensional dataset structure

v_1	...	$cfreq$	e_xtre	$vsblty$	T	$Obsv$
...

This is because the critical oscillation may never be achieved as a result of other variables that in turn affect the individual measures of each velocity thereby distorting further the probability of landing a critical oscillation range for the velocities. This is indeed an intriguing feature to consider with regards to achieving purpose (b).

To explain how this would be achieved, we add for each observation relative time (T) as a scalar which is computed using Lorentz factor [11] as:

$$T = \frac{T_o}{\sqrt{1 - (v/c)^2}} \quad (7)$$

where T is the observer's time taken and v is the absolute velocity for the given observation and C is a Kernel function applied on the activation function for the remaining x -values. And thus, ordinarily the dataset will have one more important attribute as indicated in Table 3.

At this point, we are yet to solve purpose (b) thus far since the relative time only serves to explain the degree to which the other attributes affect the oscillating velocities v_1 , v_2 and v_3 . And, therefore, upon ensuring the dependability of the variables by introducing a wholesomely reliable scalar which is the relative time.

Time stamping in data analysis as explained by [12] goes a long way to tag the absolute time onto the dataset. The absolute timestamp used is not a single unit, but rather a set of three additional variable, $tA1$, $tA2$, and $tA3$ which capture the pre-incidence, incidence, and post-incidence of the observed attributes. As such, having the relative time as the connector for the independent variables, a direct mapping can be achieved through observing the exact time an incident occurred and later on inferring other direct or indirect incidences that contributed to the observed effects.

This is to say, for instance, given the inclusion of both the absolute time and relative time, one should now be able specifically point out the point at which the oscillation of velocity differed (which resulted into different coefficients as Table 2) and proceed to explain purpose (b) by outlining other events (directly or indirectly linkable to the observation) that were happening at the particular timestamp.

Informed inference can thereafter be made from the set of identified activities as meriting a conclusive explanation to purpose (b). The plotting thus to a highly dimensional data space (also ultra-feature space) is as shown in Fig. 5 which is illustrative of a superimposed fourth dimension above the kernel's third dimension.

Having redefined data dimensionality, the following properties become the benchmark against which a data space would be checked in order to merit being highly dimensional.

A. An underlying third dimensional data (Mapping)

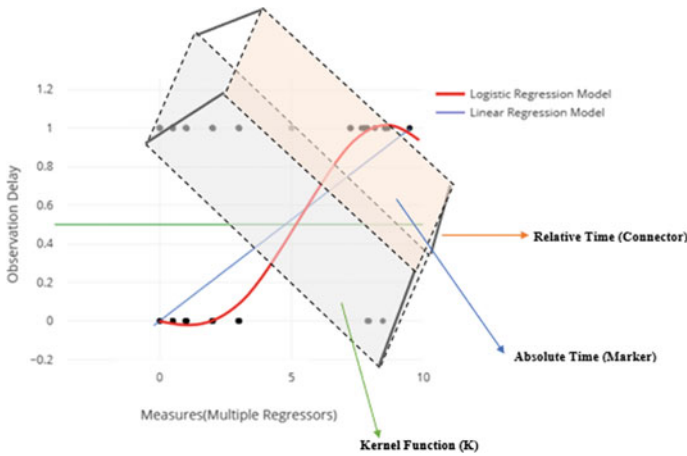


Fig. 5 Regression plotting on highly dimensional data space

That data must first be mapped into physical x , y , and z physical co-ordinates, where the third variable is not one of the independent variables but any suitable function that would clearly show dimensional separability where linear separability is not applicable or possible.

This is the very basic form of dimensional data; however, it serves least to draw any benefit for mapping data to higher feature spaces which is the explanatory capability.

B. Relative Time (Sets of dependability)

Within a highly feature space (dimensional data space), dynamic linkages are formed to allow the possibility of the independent variables to affect each other by virtue of how their relative time is altered by each other. As such, the data set is split into subsets, primary and secondary, where the secondary data subset alters the relative time of primary data subset.

C. An overlaying fourth dimension

Upon forming the linkages, a fourth dimension is formed by applying the third dimension mapping function to the relative time for purposes of a consistent continuum within the highly dimensional data space. This consistent continuum is instrumental in forging further the elasticity of the data subsets as they affect each other relative to time. The breakage of this consistency would dissociate the markers which point to the absolute time.

D. Explainability: Absolute time

Absolute time must be included for the data space continuum to be whole. This way, the position, weight, and direction of either individual data points or data subsets can be explained which relation to direct or third part relevant events associated with specific time stamps of observation.

5 Conclusion

Having outlined the benefit of mapping data up to a dimensional feature space, the following recommendations are made with reference to preparing data to be used for machine learning tasks more so regression tasks:

- i. That timestamps be capture either for the moment when the observation (tuple) was made or wholesomely for when the dataset was captured (snapshot) this provides the premise for using absolute time as addressed in the previous sections.
- ii. That data be viewed in subsets in relation to the dependent variable and a thorough correlation analysis be made in relation to obtaining the primary subset and the secondary subset.

These recommendations serve to build the premises upon which the properties of dimensional data would easily be applied.

References

1. Bansal, S., Kumar, D. (2020) IoT Ecosystem: a survey on devices, gateways, operating systems, middleware and communication. *Int J. Wirel. Inf. Netw.* <https://doi.org/https://doi.org/10.1007/s10776-020-00483-7>
2. Cavanillas J.M., Curry E., Wahlster W. (2016) The big data value opportunity. In: Cavanillas J., Curry E., Wahlster W. (eds.) *New Horizons for a Data-Driven Economy*. Springer, Cham
3. Verleysen M., François D.: *The Curse of Dimensionality in Data Mining and Time Series Prediction*. Springer, pp 758–550 (2005)
4. Sarveniazi, A.: An actual survey of dimensionality reduction. *Am. J. Comput. Mathem.* 55–72 (2014)
5. Rojas, R.: *The Curse of Dimensionality* (2015). https://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/dimensionality.pdf
6. Einstein, A.: *Relativity: The Special and General Theory*. Henry Holt & Company, New York (1920)
7. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer, New York (1995)
8. Vapnik, V.N.: *Statistical Learning Theory*. Springer, New York (1998)
9. Creswell J. (2003) *Qualitative, Quantitative and Mixed Methods Approaches*, 2nd edn. Sage Publishers, Thousand Oaks CA
10. Dennis A., Wixom B.H., Roth R.M.: *System Analysis and Design*. Wiley, Indiana (2012)
11. Steane, A.M.: *The Wonderful World of Relativity*. Oxford University Press, New York (2011)
12. Koen, R., Olivier M.S.: The use of timestamps in digital forensics. In: 7th International Workshop on Digital Forensics and Incident Analysis (WDFIA), Pretoria, South Africa (2012)

Implementation of Encryption Techniques in Secure Communication Model



Md. Sharif Hossen and Md. Shakhaowat Hossen

Abstract Wireless communication evolves the modern generation where communication has been possible to limit the distance without any physical connection between two parties. Basically, communication can be of two types namely wired and wireless where the second is preferred mostly than first. But we need a secure wireless communication in case of industries, companies, e-commerce and communication technologies, etc. In this research, we can transmit encrypted data wirelessly through the channel to provide data security using different encryption techniques. To do this we propose a model which comprises two parts namely transmitter and receiver. In the transmitter portion, the plaintext is encrypted and then after modulation, it is sent through the channel. The receiver at first receives the channel output and then demodulates and finally decrypts the encrypted data to get the original plaintext which is sent from the sender to the receiver. Here, we have considered two encryption techniques, i.e., Caesar and RSA, and compared them with no encryption in our proposed model. We implement our model using the MATLAB programming language. The overall investigation shows that the RSA algorithm deserves better performance than Caesar and no encryption techniques because RSA shows lower bit errors with signal-to-noise ratio compared to no encryption and Caesar cipher.

Keywords Encryption · Decryption · Key · Cryptography · Ciphertext · Plaintext · RSA · SNR · BER · Caesar

1 Introduction

With the increase of computer users, the information or data is now growing rapidly. This growing amount of data should be kept private or secure because there can be

Md. S. Hossen (✉) · Md. S. Hossen
Department of Information and Communication Technology, Comilla University, Comilla,
Bangladesh
e-mail: sharif5613@gmail.com

Md. S. Hossen
e-mail: mdshakhaowath@gmail.com

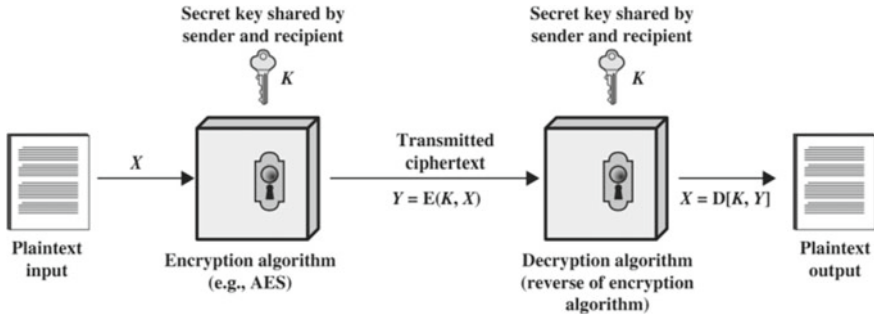


Fig. 1 General symmetric concept

some credential information like password, credit card details, etc. Third parties or hackers are trying to steal our data [1]. We can give an easy example, sometimes we get some messages in the spam folder in our email that encourages us to open the page or click on the link. Clicking or viewing the page may be a chance to hack or steal your data [2]. We should ensure the security of our data from those parties. And this can be possible through the concept of cryptography where the data or information is only readable by the authorized users and no other can read the data. This is termed as encryption which can be of two types namely symmetric and asymmetric [3, 4]. In the first type, a single private key is used to ensure the security which is available both to the sender and receiver. Another technique uses two different keys one is private and the other is public [5]. Figure 1 shows the general model of the symmetric technique [4].

From Fig. 1, we see that at first user data is encrypted using an encryption technique (e.g., RSA), and a shared key. Then, the encrypted data is transmitted through the media which is fed to the decryption process where the encrypted text is readable and sent to the receiver as the original text.

The remaining parts are arranged as follows. We use RSA and Caesar cipher encryption methods discussed in Sect. 2. Section 3 discusses the modulation scheme used. To ensure the security of data in the communication system, we propose a model described in Sec. 4. Result of analysis is discussed in Sect. 5. Section 6 concludes the research summary.

2 Encryption Techniques

There are several types of encryption techniques. In cryptography, generally, encryption techniques can be classified as symmetric and asymmetric [10]. In the first case, only the same key is treated to encrypt the original message and decrypt the ciphertext into the plaintext which is known both to the sender and receiver [11]. Examples of symmetric encryption are DES, RC4, Triple DES, Blowfish, and RC6 [12].

On the other hand, the other case (i.e., asymmetric) uses two keys, one is public and other is private where the first may be available to both parties but the second must be kept secret. The first key encrypts the plaintext into ciphertext while other key derives the plaintext by decrypting the ciphertext. Here, receiver’s public and private keys are used, respectively. The most common methods are RSA, DSA, Diffie-Hellman, etc. [13]. For our investigation, we utilize two techniques, namely Caesar and RSA irrespective of symmetric and asymmetric methods. Now, we will discuss here both methods.

2.1 RSA

RSA is an asymmetric key encryption technique consisting of six elements where encryption can be possible using either a public or private key. The following Fig. 2 shows the public key encryption scheme [4].

Bob wants to send his text to Alice. Using the above scheme, at first, Bob encrypts the message using Alice’s public key and the encrypted text is then fed to the decryption process where Alice decrypts the encrypted version using his private key to get the original text sent to him. The working principle of RSA is following in Fig. 3. Figure 4 shows an example of it.

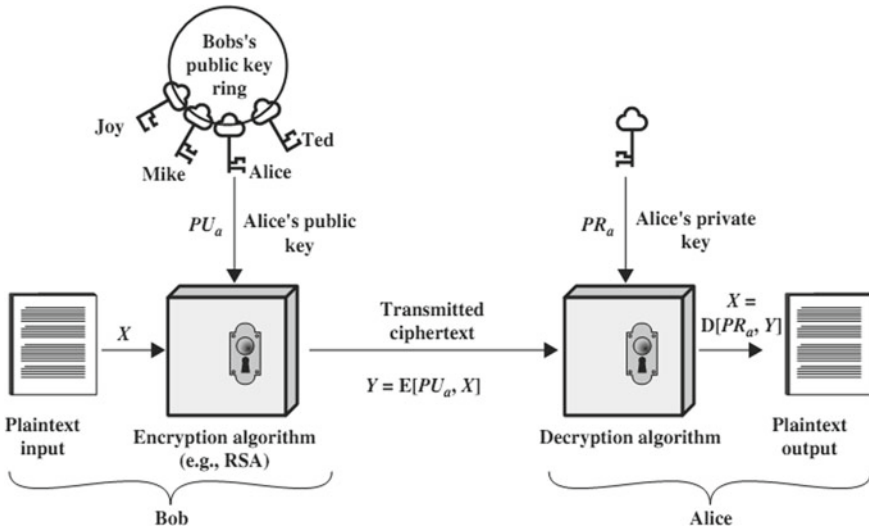
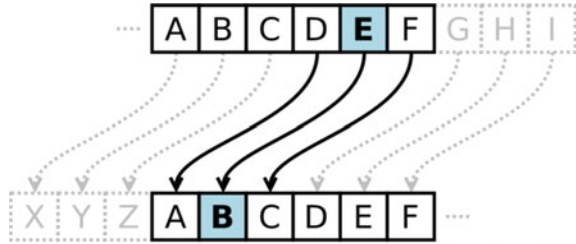


Fig. 2 Public key for encryption

Fig. 5 Shifts of each character



$$p = D(k, C) = (C - k) \text{ mod } 26$$

where C , p , and k indicate the ciphertext, plaintext, and shift of character (which value can be 1–25), respectively. We consider the following equivalent of character:

a	b	c	d	e	F	G	h	i	j	k	l	m	n	o
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
p	q	r	s	t	u	v	w	x	y	z				
15	16	17	18	19	20	21	22	23	24	25				

Plaintext, $p = \text{comilla}$.

For first character c , $p = c$, $k = 4$, then

$$\begin{aligned}
 C &= E(4, c) \text{ mod } 26 = E(4, 2) \text{ mod } 26 \\
 &= (4 + 2) \text{ mod } 26 \\
 &= 6 \text{ mod } 26 \\
 &= G
 \end{aligned}$$

Repeating this process, $C = \text{GSQMPPE}$ (using the above equivalent).

Now, with ciphertext, $C = \text{GSQMPPE}$, for first character G , $C = G$, $k = 4$, then,

$$\begin{aligned}
 p &= D(4, G) \text{ mod } 26 = D(4, 6) \text{ mod } 26 \\
 &= (6 - 4) \text{ mod } 26 \\
 &= 2 \text{ mod } 26 \\
 &= c
 \end{aligned}$$

Repeating this process, Plaintext, $p = \text{comilla}$ (using the above equivalent) is achieved.

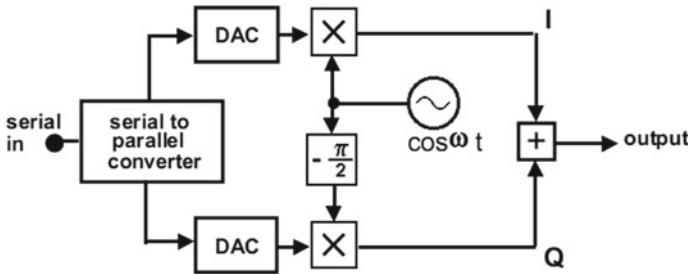
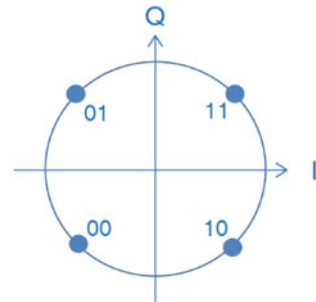


Fig. 6 Block diagram of QPSK

Fig. 7 Constellation diagram of QPSK



3 Modulation Techniques

Basically, digital modulation schemes are categorized into three kinds of shift keying depending on the phase, frequency, amplitude, and quadrature amplitude. In this paper, we have used phase-shift keying (PSK), actually, quadrature phase-shift keying (QPSK) also called 4PSK [13, 14] (Fig. 6).

Constellation of QPSK is shown in Fig. 7 where there are four points. Two bits are used to form a symbol.

4 Proposed Communication Model

This research motivates us to propose a model to ensure data security in the basic communication model shown in Fig. 8. It comprises two parts, transmitter and receiver [6]. In the first part, the user data is first encrypted using encryption techniques, for example, RSA, Caesar, etc., which is modulated using any scheme (e.g., BPSK). The encrypted modulated data is then sent through the channel where different kinds of noise can be added. Here, we have considered the AWGN channel [7, 8]. The output of it is then fed to the receiver which comprises two, namely demodulation and decryption where noisy data is demodulated and decrypted by the

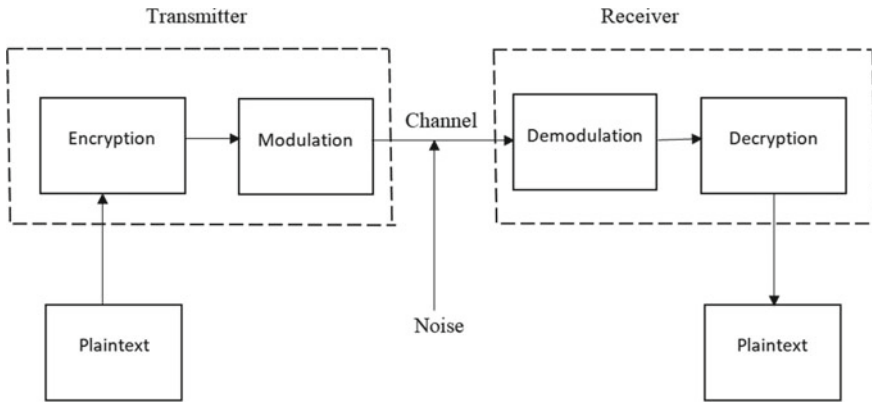


Fig. 8 Basic communication model with encryption

respective reverse operation of modulation and encryption technique. After that, the final outcome is found which is fed to another intended user [9].

5 Result and Discussion

We have used the AWGN channel in our proposed model. We use MATLAB as the programming tool to implement the coding. After implementing the model using MATLAB, we get the following comparison of bit error ratio (BER) with respect to signal-to-noise ratio (SNR) among no encryption, Caesar, and RSA encryption techniques.

5.1 No Encryption and Caesar Cipher

Comparison of no encryption and Caesar cipher is shown in Fig. 9. Comparing with no encryption and Caesar cipher, we see that Caesar cipher shows lower BER compared to No encryption.

5.2 No Encryption and RSA Algorithm

Comparison of no encryption and RSA encryption is shown in Fig. 10.

Comparing with no encryption and RSA encryption process, RSA encryption provides lower BER than no encryption.

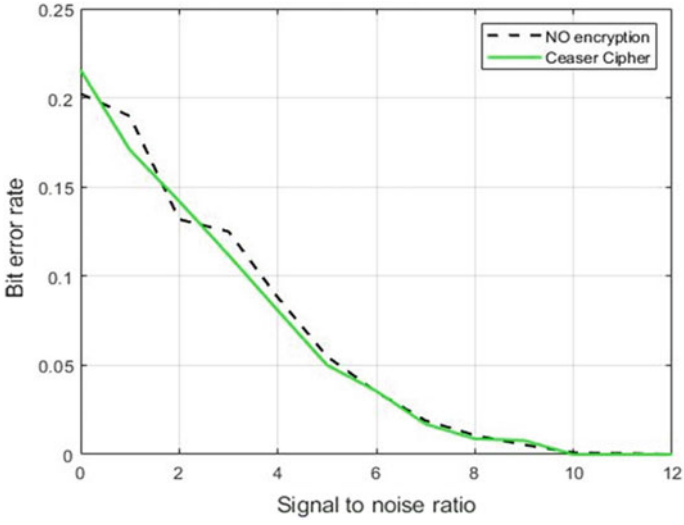


Fig. 9 Comparison of no encryption and Caesar cipher

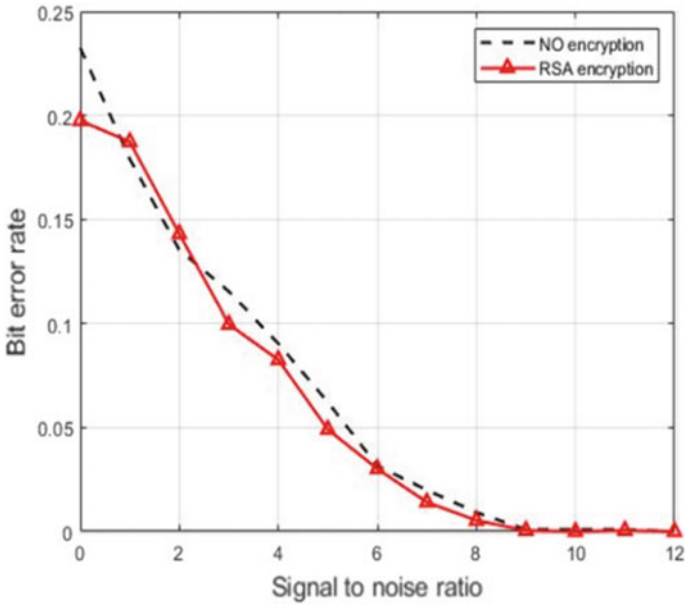


Fig. 10 Comparison of no encryption and RSA encryption

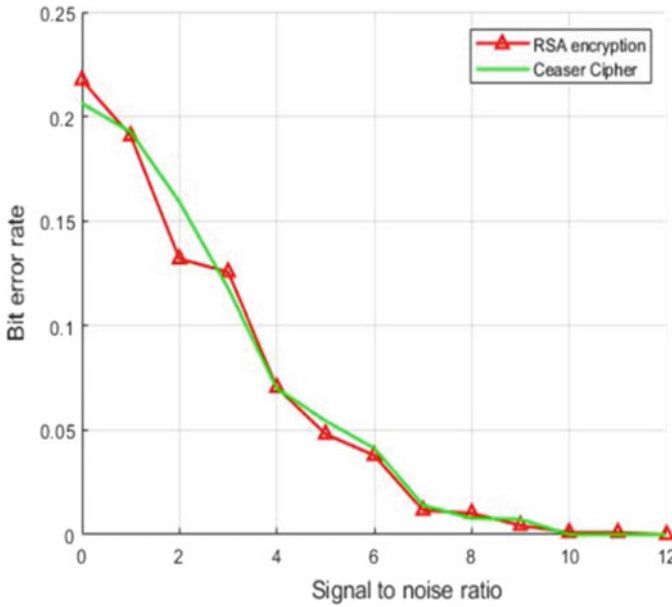


Fig. 11 Comparison of RSA and Caesar cipher

5.3 *RSA Encryption and Caesar Cipher*

A comparison of RSA and Caesar cipher is shown in Fig. 11. Figure 11 shows that RSA requires lower BER than Caesar cipher. So, the performance of RSA is better than Caesar.

5.4 *No Encryption, Caesar Cipher and RSA*

Comparison among no encryption, Caesar cipher, and RSA is shown in Fig. 12. Here, we see that RSA shows better results because both values of BER and SNR are less than others, i.e., no encryption and Caesar cipher. We can find a better signal in which the noise ratio is low.

6 Conclusion and Future Works

Wireless communication is the transfer of message, data, or information that is performed and delivered wirelessly. In the communication between the sender and the receiver, data security is very important which can be maintained through the use

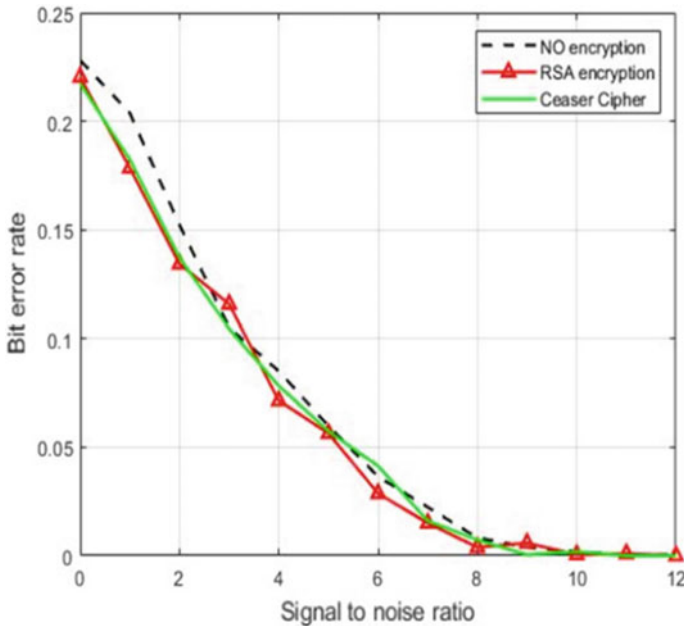


Fig. 12 Comparison among no encryption, Caesar cipher and RSA

of encryption. In this paper, we propose a model that consists of two major elements, namely transmitter and receiver. In the transmitter portion, data is encrypted and modulated for long-distance transmission which is then passed through the channel. While the receiver receives the signal, demodulates and decrypts it to get the original form. Here, we analyze the performance among no encryption, Caesar cipher, and RSA encryption algorithms in our proposed communication model. The good performance of a signal is achieved through a lower bit error ratio (BER) and signal-to-noise ratio (SNR). With the same input, we get different values of BER with respect to SNR for different encryption algorithms. Our research concludes that the RSA encryption process exhibits good performance by ensuring lower SNR and BER among Caesar and no encryption.

In the near future, we would like to implement and investigate the performance of other encryption techniques, namely DES, 3DES, AES, etc., in our proposed communication model. Moreover, we would like to propose another encryption technique.

References

1. Alsunbul, S., Le, P., Tan, J., Srinivasan, B.: A network defense system for detecting and preventing potential hacking attempts. In: International Conference on Information Networking (ICOIN), Kota Kinabalu, pp. 449–454 (2016)

2. Firth, A.: Communication. In: Practical Web Inclusion and Accessibility. Apress, Berkeley, CA (2019)
3. Buchmann, J.: Introduction to Cryptography. Springer, New York (2001)
4. William, S.: Cryptography and Network Security: Principles and Practice, 7th edn. Prentice Hall Press, USA (2019)
5. Hassan, N.A., Salminen, H.: Data Hiding Techniques in Windows OS: A Practical Approach to Investigation and Defense. Syngress, Cambridge, MA (2017)
6. Matin, M.A.: Communication Systems for electrical engineers. Cham, Switzerland: Springer.
7. Hossen, M.S., Ahmed, S.: Development of a Java Based Simulator for OFDM System. In: Proceedings of ICMEIE. RU, Bangladesh (2015)
8. Haykin, S.S.: Digital Communication Systems. Wiley, Hoboken, NJ (2014)
9. Cho, W.: Secure communications with asymptotically Gaussian compressed encryption. IEEE Sig. Process. Lett. **25**(1), 80–84 (2018). <https://doi.org/10.1109/LSP.2017.2773128>
10. Simmons, G.: Symmetric and asymmetric encryption. ACM Comput. Surv. (CSUR) **11**(4), 305–330 (1979). <https://doi.org/10.1145/356789.356793>
11. Kaur, E.M.: Data encryption using different techniques: a review. Int. J. Adv. Res. Comput. Sci. **8**(4), 4 (2017)
12. Maurer, U.: Confidentiality and integrity: a constructive perspective. Theory Cryptograp. **7194**, 209–229 (2012). https://doi.org/10.1007/978-3-642-28914-9_12
13. Manakshe, A.: Variants of public key cryptosystem RSA. Int. J. Adv. Res. Comput. Sci. **2**(3) (2011)
14. Haque, M.T.U., Hossen, M.S.: PAPR reduction in OFDM system using clipping and filtering methods based on CCDF. Int. J. Comput. Netw. Inf. Secur. (IJCNIS) **10**, 12–18 (2019)

A Group Decision Making Problem Involving Fuzzy TOPSIS Method



Prashanta Kumar Parida

Abstract Selection is a process to find out the best alternative solution result using the given alternatives, criteria and experts. The purpose of this manuscript is to development of fuzzy technique to group fuzzy technique method through FTOPSIS method. We introduce a literature survey in different models of fuzzy and that have been applied the field of decision making. In the multi-criteria decision technique, fuzzy TOPSIS is proposed for selection of four different projects by fuzzy TOPSIS software. Lastly, we determine the best project using group fuzzy TOPSIS methodology. To illustrate the sequel of the group ideal solution and have defend our replica to be structured and vigorous.

Keywords MCDM · Fuzzy TOPSIS · Group FTOPSIS · Relative closeness matrix

1 Introduction

Decision making (DM) is a process of finding solution in our day-to-day life. In every step, we are taking the decision by the help of human being or any technique or by soft computing process. In this paper, we are taking the decision through multi-criteria decision-making (MCDM) problems or multi-attribute decision-making (MADM) [1] using fuzzy technique for order performance by similarity to positive ideal solutions (FTOPSIS) techniques. This process is to define the ranking of all possible alternatives with respect to the goal and more than one criteria. There are several real-world applications of MCDM method; data are usually vague, ambiguous and/or unpredictable. The MCDM [2] problems credible and are excessively applied in many domains, such as different engineering sciences, management, mathematical sciences, economics, medical sciences and soon. The DM has to select, assess or rank these alternatives according to the weights of the criteria. The important branch of subject operation research is the MCDM technique [3–5] used in the last five decades.

P. K. Parida (✉)

Department of Mathematics, C.V. Raman Global University, Bhubaneswar, Odisha, India
e-mail: prashantakumarparida@cvrce.edu.in

In the real-life situations, the problems of DM are subjected to objectives, constraints and their consequences that are not meticulously known. The new decision theory is known today as fuzzy multi-criteria decision-making (FMCDM) [6, 7] is the condensation allying MCDM [8] and fuzzy set theory [9], where the DM models are deal with insufficient and undetermined intelligence and evidence. Many researchers have been preoccupied by decision-making (DM) problems [10–13] in fuzzy environments. To describe the subjective judgment of a DM in a quantitative manner, fuzzy numbers (FNs) most often used in triangular FN, trapezoidal FN.

In the TOPSIS method, the best alternative is one which is nearest to the PIS and at maximum length from the NIS. In the PIS, the benefit criteria get maximized and the cost criteria get minimized. In the NIS, the cost criteria get maximized and the benefit criteria get minimized. As a practical application of TOPSIS method, we can see [11, 12]. In this situation where the available information is vague, imprecise or uncertain, it is quite difficult to precisely asses the alternatives with respect to the criteria. The rating of every one alternative with respect to every one criterion can be described by fuzzy numbers [3].

The TOPSIS method has been broadened to handle MCDM with an unsettled DM with consequence in fuzzy TOPSIS [11–13], which has fortunately been used to solve different MCDM problems [13–18]. In this way, we obtain extensions of the TOPSIS method under fuzzy environment, i.e., fuzzy TOPSIS. The remnants of this paper are assembled into different segments having backdrop enlightenment about research methodology. In this paper, we organized as follows. In Section 2, we outlined the basic concepts of fuzzy set, fuzzy membership function, triangular fuzzy numbers, the TOPSIS method and FTOPSIS method. In Section 3, we suggested research methodology and proposed algorithm. Section 4 presents empirical studies and Sect. 5 concludes the study.

2 Basic Concepts

In this section, first, we briefly introduce some definitions and concepts related to fuzzy set, fuzzy membership function, triangular fuzzy numbers (TFN) and algorithm of TOPSIS method, fuzzy TOPSIS method by group decision-making method.

2.1 Definitions

Definition 1 (Fuzzy set). Let U be an universe of objects with an $u \in U$. A fuzzy set \bar{A} in U is characterized by $\mu_{\bar{A}}(u)$ membership function $u \in [0, 1]$ representing the grade of membership function of u in \bar{A} . Then.

$$\bar{A} = \{(u, \mu_{\bar{A}}(u)) : u \in U\}, \quad \text{where } \mu_{\bar{A}}(u) : U \rightarrow [0, 1] \quad (1)$$

Definition 2 (TFN). If \overline{Tr} is a TFN and $[\overline{t_n}]_\beta^l > 0$ and $[\overline{t_n}]_\beta^u \leq 1$ for $\beta \in (0, 1]$, so $\overline{t_n}$ is called a normalized TFN.

Definition 3 (Membership of TFN) Let \bar{a} be a fuzzy number which is defined by a triplet $\bar{a} = (a_1, a_2, a_3)$. Then the membership function is denoted as $\mu_{\bar{a}}(u)$, defined by.

$$\mu_{\bar{a}}(u) = \begin{cases} 0, & u < a_1 \\ (u - a_1)/(a_2 - a_1), & a_2 \geq u \geq a_1 \\ (u - a_2)/(a_3 - a_2), & a_3 \geq u \geq a_2 \\ 0, & u < a_3 \end{cases} \tag{2}$$

Definition 4 (Operation of TFN). Let $\bar{u} = (u_1, u_2, u_3)$ and $\bar{v} = (v_1, v_2, v_3)$ be two positive TFNs, then the operation with these fuzzy numbers is demarcated as follows.

$$\bar{u}(\mp)\bar{v} = (\bar{u}_1(\mp)\bar{v}_1, \bar{u}_2(\mp)\bar{v}_2, \bar{u}_3(\mp)\bar{v}_3) \tag{3}$$

$$\bar{u}(\times)\bar{v} = (\bar{u}_1(\times)\bar{v}_1, \bar{u}_2(\times)\bar{v}_2, \bar{u}_3(\times)\bar{v}_3) \tag{4}$$

$$\bar{u}(/)\bar{v} = (\bar{u}_1(/)\bar{v}_1, \bar{u}_2(/)\bar{v}_2, \bar{u}_3(/)\bar{v}_3) \tag{5}$$

$$k\bar{v} = (kv_1, kv_2, kv_3) \tag{6}$$

Definition 5 (Distance of TFN). Let $\bar{u} = (u_1, u_2, u_3)$ and $\bar{v} = (v_1, v_2, v_3)$ be two positive TFNs, then distance is computed by.

$$d(\bar{u}, \bar{v}) = \sqrt{\frac{1}{3}[(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2]} \tag{7}$$

Definition 6 (α -cut). The α -cut is a fuzzy set $\bar{A} \subset U$ and is defined by.

$$[\bar{A}]_\alpha = \{u | \mu_{\bar{A}}(u) \geq \alpha\}, \text{ where } \alpha \in [0, 1] \tag{8}$$

2.2 TOPSIS Method

Step 1 Choose decision matrix D is described by $D = A_i \begin{matrix} C_j \\ (u_{ij}) \end{matrix}$, where $A_i, i = 1, \dots, m$ are alternatives and $C_j, j = 1, \dots, n$ are criteria, u_{ij} are original scores express the grading of the alternative A_i with respect to criteria C_j . The weight

vector $w = (w_1, w_2, \dots, w_n)$ is collected the discrete weights $w_j (j = 1, 2, \dots, n)$ for every one criteria C_j .

Step 2 Construct normalized decision matrix N_{ij} , where $N_{ij} = u_{ij} / \sqrt{\sum u_{ij}^2}$ for $i = 1, \dots, m; j = 1, \dots, n$, where u_{ij} and N_{ij} are original and normalized matrix, respectively.

Step 3 The weighted normalized decision matrix $V_{ij} = w_j N_{ij}$, where w_j is the weight for j th criteria and $\sum w_j = 1$.

Step 4 The PIS and NIS are $A^+ = (v_1^+, v_2^+, \dots, v_n^+)$ and $A^- = (v_1^-, v_2^-, \dots, v_n^-)$, where $v_j^+ = \{\max_i V_{ij} | j \in J_1; \min_i V_{ij} | j \in J_2\}$ and $v_j^- = \{\min_i V_{ij} | j \in J_1; \max_i V_{ij} | j \in J_2\}$.

where J_1 represent benefit criteria and J_2 represent cost criteria.

Step 5 Compute the Euclidean lengths from the PIS A^+ and NIS A^- solutions for every one alternatives A_i :

$$\delta_i^+ = \sqrt{\sum_j (\Delta_{ij}^+)^2} \quad \text{and} \quad \delta_i^- = \sqrt{\sum_j (\Delta_{ij}^-)^2}$$

where $\Delta_{ij}^+ = (v_j^+ - V_{ij})$ and $\Delta_{ij}^- = (v_j^- - V_{ij})$ with $i = 1, \dots, m$

Step 6 Compute the relative closeness Ω_i for every one alternative A_i with respect to PIS A^+ as given by $\Omega_i = \delta_i^- / (\delta_i^- + \delta_i^+)$, where $i = 1, \dots, m$.

2.3 Fuzzy TOPSIS Method

Suppose there exists m possible alternatives u_1, u_2, \dots, u_m for which the decision maker (DM) has to choose on the basis on n attributes C_1, C_2, \dots, C_n both qualitative and quantitative A_i on a attribute C_j given by the decision maker is a triangular fuzzy number \bar{u}_{ij} , where $i = 1, 2, \dots, m, j = 1, 2, \dots, n$. The MADM problem can be expressed in the matrix form as

$$\bar{F} = \begin{matrix} & C_1 & C_2 & \cdots & C_n \\ \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} & \begin{pmatrix} \bar{u}_{11} & \bar{u}_{12} & \cdots & \bar{u}_{1n} \\ \bar{u}_{21} & \bar{u}_{22} & \cdots & \bar{u}_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{u}_{m1} & \bar{u}_{m2} & \cdots & \bar{u}_{mn} \end{pmatrix} \end{matrix}$$

2.3.1 Algorithm

- Step 1. Identify the evaluation criteria which may be expressed in linguistic variables.
- Step 2. Calculate every one alternatives in form of criteria.
- Step 3. Identify the weight of the criteria which may also be fuzzy in nature.
- Step 4. Establish the fuzzy decision matrix \bar{F} . In this matrix, every \bar{u}_{ij} is a triangular fuzzy number $\bar{u}_{ij} = (u_{ij}, \alpha_{ij}, \beta_{ij})$.
- Step 5. Establish the normalized fuzzy decision matrix \tilde{N}_{ij}

For every fuzzy number $\bar{u}_{ij} = (u_{ij}, \alpha_{ij}, \beta_{ij})$, we establish the set of α -cut as $\bar{u}_{ij} = ([\bar{u}_{ij}]_{\alpha}^l, [\bar{u}_{ij}]_{\alpha}^u)$, $\alpha \in [0, 1]$. Every one fuzzy number \bar{u}_{ij} is transformed into an interval. Now this interval is transformed into normalized interval

$$[\tilde{n}_{ij}]_{\alpha}^l = [\bar{u}_{ij}]_{\alpha}^l / \sum_{i=1}^m \left[([\bar{u}_{ij}]_{\alpha}^l)^2 + ([\bar{u}_{ij}]_{\alpha}^u)^2 \right], \quad j = 1, 2, \dots, n$$

$$[\tilde{n}_{ij}]_{\alpha}^u = [\bar{u}_{ij}]_{\alpha}^u / \sum_{i=1}^m \left[([\bar{u}_{ij}]_{\alpha}^l)^2 + ([\bar{u}_{ij}]_{\alpha}^u)^2 \right], \quad j = 1, 2, \dots, n$$

Now $([\tilde{n}_{ij}]_{\alpha}^l, [\tilde{n}_{ij}]_{\alpha}^u)$ is the normalized interval of $([\bar{u}_{ij}]_{\alpha}^l, [\bar{u}_{ij}]_{\alpha}^u)$ which is transformed into a fuzzy number $\bar{N}_{ij} = (n_{ij}, a_{ij}, b_{ij})$. According to setting the value of $\alpha = 1$, we have $[\tilde{n}_{ij}]_{\alpha=1}^l = [\tilde{n}_{ij}]_{\alpha=1}^u = n_{ij}$ and setting the value $\alpha = 0$, we have $[\tilde{n}_{ij}]_{\alpha=0}^l = n_{ij} - a_{ij}$ and $[\tilde{n}_{ij}]_{\alpha=0}^u = n_{ij} + b_{ij}$ then $a_{ij} = n_{ij} - [\tilde{n}_{ij}]_{\alpha=0}^l$ and $b_{ij} = [\tilde{n}_{ij}]_{\alpha=0}^u - n_{ij}$. Now $\bar{N}_{ij} = (n_{ij}, a_{ij}, b_{ij})$ is the fuzzy number of the normalized interval $([\tilde{n}_{ij}]_{\alpha}^l, [\tilde{n}_{ij}]_{\alpha}^u)$. This \bar{N}_{ij} be a normalized positive triangular fuzzy number.

- Step 6. Considering the every one criterion, we can construct the weighted normalized fuzzy decision matrix as $\bar{v}_{ij} = \bar{N}_{ij} \cdot \bar{w}_j$ where \bar{w}_j is the weight of the j th criterion.
- Step 7. Every one \bar{v}_{ij} is a normalized fuzzy number and their ranges belong to $[0, 1]$. So we identify the PIS $\bar{A}^+ = (\bar{v}_1^+, \bar{v}_2^+, \dots, \bar{v}_n^+)$ and the NIS $\bar{A}^- = (\bar{v}_1^-, \bar{v}_2^-, \dots, \bar{v}_n^-)$ where $\bar{v}_j^+ = (1, 1, 1)$ and $\bar{v}_j^- = (0, 0, 0)$, $j = 1, 2, \dots, n$ for every criteria.
- Step 8. Using the length definition, we calculate the length of every one alternatives from the PIS and NIS as $\bar{\delta}_i^+ = \sum_{j=1}^n d(\bar{v}_{ij}, \bar{v}_j^+)$ and $\bar{\delta}_i^- = \sum_{j=1}^n d(\bar{v}_{ij}, \bar{v}_j^-)$ $i = 1, 2, \dots, m$, respectively.
- Step 9. The relative closeness coefficients is $\bar{C}_i = \frac{\bar{\delta}_i^-}{(\bar{\delta}_i^+ + \bar{\delta}_i^-)}$, $i = 1, 2, 3, \dots, m$.

3 Research Methodology

Using the different steps to calculating the group of best alternatives is defined below:

The PIS A^+ (benefits) and NIS A^- (costs) for each group member $r = 1, 2, \dots, R$ as follows:

$${}^r A^+ = \left({}^r \bar{V}_1^+, {}^r \bar{V}_2^+, \dots, {}^r \bar{V}_m^+ \right) \text{ and } {}^r A^- = \left({}^r \bar{V}_1^-, {}^r \bar{V}_2^-, \dots, {}^r \bar{V}_m^- \right)$$

where ${}^r \bar{V}_j^+ = \left(\max_i {}^r \bar{V}_{ij}, j \in J_1; \min_i {}^r \bar{V}_{ij}, j \in J_2 \right)$.

and ${}^r \bar{V}_j^- = \left(\min_i {}^r \bar{V}_{ij}, j \in J_1; \max_i {}^r \bar{V}_{ij}, j \in J_2 \right)$.

where J_1 is criteria for benefit and J_2 is criteria for cost.

Evaluate the length of every one alternative for many members. The length of alternative A_i between the PIS and the NIS of the group members S_r , ${}^r \bar{D}_i^+$ and ${}^r \bar{D}_i^-$ is given with $i = 1, 2, \dots, m; r = 1, 2, \dots, R$ by:

$${}^r \bar{D}_i^+ = \sum_{j=1}^n D \left({}^r \bar{V}_{ij}, {}^r \bar{V}_j^+ \right) \text{ and } {}^r \bar{D}_i^- = \sum_{j=1}^n D \left({}^r \bar{V}_{ij}, {}^r \bar{V}_j^- \right).$$

where the lengths $D \left({}^r \bar{V}_{ij}, {}^r \bar{V}_j^+ \right)$ and $D \left({}^r \bar{V}_{ij}, {}^r \bar{V}_j^- \right)$ between two fuzzy numbers are calculated.

The relative closeness for every one alternative A_i of every one member r , $\bar{\Omega}^r(A_i)$ with respect to PIS as

$$\bar{\Omega}^r(A_i) = \frac{{}^r \bar{D}_i^-}{{}^r \bar{D}_i^+ + {}^r \bar{D}_i^-} \text{ with } i = 1, 2, \dots, m; r = 1, 2, \dots, R$$

Now, we calculate the $\bar{\Omega}^r(A_i)$ for every one member r we may form the relative closeness matrix as given by:

$$Q = \begin{pmatrix} \bar{\Omega}^1(A_1) & \bar{\Omega}^2(A_1) & \dots & \bar{\Omega}^R(A_1) \\ \bar{\Omega}^1(A_2) & \bar{\Omega}^2(A_2) & \dots & \bar{\Omega}^R(A_2) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\Omega}^1(A_m) & \bar{\Omega}^2(A_m) & \dots & \bar{\Omega}^R(A_m) \end{pmatrix}$$

The weighted RCM is given by:

$$Q\alpha = \begin{pmatrix} \alpha^1 \bar{\Omega}^1(A_1) & \alpha^2 \bar{\Omega}^2(A_1) & \dots & \alpha^R \bar{\Omega}^R(A_1) \\ \alpha^1 \bar{\Omega}^1(A_2) & \alpha^2 \bar{\Omega}^2(A_2) & \dots & \alpha^R \bar{\Omega}^R(A_2) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^1 \bar{\Omega}^1(A_m) & \alpha^2 \bar{\Omega}^2(A_m) & \dots & \alpha^R \bar{\Omega}^R(A_m) \end{pmatrix}$$

To establish the groups, PIS and NIS

$$A_G^+ = (V_{G1}^+, V_{G2}^+, \dots, V_{GR}^+)$$

$$= \left(\max_i \alpha^1 \Omega^1(A_i), \max_i \alpha^2 \Omega^2(A_i), \dots, \max_i \alpha^R \Omega^R(A_i) \right)$$

and $A_G^- = (V_{G1}^-, V_{G2}^-, \dots, V_{GR}^-)$

$$= \left(\min_i \alpha^1 \Omega^1(A_i), \min_i \alpha^2 \Omega^2(A_i), \dots, \min_i \alpha^R \Omega^R(A_i) \right)$$

Calculate to every one alternative A_i the lengths from the group positive and NISs A_G^+ and A_G^- , respectively, with $i = 1, 2, \dots, m$ as follows:

$$d_{Gi}^+ = \sqrt{\sum_{r=1}^R (\alpha^r \Omega^r(A_i) - V_{Gr}^+)^2} \text{ and } d_{Gi}^- = \sqrt{\sum_{r=1}^R (\alpha^r \Omega^r(A_i) - V_{Gr}^-)^2}$$

Construct the group relative closeness Ω_{Gi} for every one alternative A_i with respect to GIS (group ideal solution) as:

$$\Omega_G(A_i) = \frac{d_{Gi}^-}{d_{Gi}^- + d_{Gi}^+}$$

4 Computational Illustration

In this classification, we adduce one ciphering illustration to interpret the TOPSIS technique for DM problems with fuzzy data. Considering that, we have five alternatives Alt1, Alt2, Alt3, Alt4, Alt5 among which decision makers have to choose and evaluated by four experts or decision makers DM1, DM2, DM3, DM4 under fuzzy environment for behavior operational versus four benefit criteria Crt1, Crt2, Crt3, Crt4. The linguistic weights for performing the predominant of criteria are very little low (VLL), little low (LL), medium low (ML), medium little high (MLH), little high (LH), very little high (VLH), little excellent (LE) and excellent (E), with the following fuzzy numbers demarcated in Table 1.

Table 1 Linguistic terms

Linguistic terms	Fuzzy numbers
Very little low (VLL)	(0.0,0.0,0.12)
Little low (LL)	(0.0,0.12,0.24)
Medium little low (MLL)	(0.12,0.24,0.36)
Medium low (ML)	(0.24,0.36,0.48)
Medium little high (MLH)	(0.36,0.48,0.60)
Little high (LH)	(0.48,0.60,0.72)
Very little high (VLH)	(0.60,0.72,0.84)
Little excellent (LE)	(0.72,0.84,0.96)
Excellent (E)	(0.84,0.96,1.00)

Table 2 Performance of decision makers for alternatives and criteria

DM	Crt1	Crt2	Crt3	Crt4	Crt5	Crt6	Crt7
Alt1	MLL	VLL	LH	VG	E	VLH	ML
Alt2	G	LH	LE	LL	MLH	LP	LH
Alt3	LL	E	FLP	ML	VG	FLG	VLL
Alt4	LE	G	MLL	VLH	E	MG	ML
Alt5	MG	FLP	E	LH	LE	LH	LL
Alt6	LP	VG	MLH	LE	E	VLL	ML
Alt7	FLG	MG	MLL	FLP	G	VG	VLH
Alt8	LH	LE	G	VG	VLH	LL	LP
Alt9	VLL	MG	FLG	VG	MLH	E	LH

Based on the upper expansions, ourselves considering FTOPSIS for four decision matrices DM1, DM2, DM3 and DM4 with same appraises of weights with (0.250, 0.250, 0.250, 0.250). Evolved from this utility, we evaluated the DM, the NDM, the WNDM, fuzzy PIS, fuzzy NIS, the relative closeness coefficient for one after the other DM proportional to similar weights from Tables 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 and 18. From Tables 3, 4, 5 and 6, we established the appraise using first DM with NDM, propositional to the weights, the fuzzy PIS and fuzzy NIS, relative closeness coefficient with ranking order. From Tables 7, 8, 9 and 10, we established the appraise using second DM with NDM, propositional to

Table 3 Decision matrix for DM1

DM1	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.00,0.12)	(0.24,0.36,0.48)	(0.84,0.96,1.00)	(0.48,0.6,0.72)
Alt2	(0.24,0.36,0.48)	(0.72,0.84,0.96)	(0.00,0.12,0.24)	(0.6,0.72,0.84)
Alt3	(0.60,0.72,0.84)	(0.24,0.36,0.48)	(0.12,0.24,0.36)	(0.84,0.96,1.0)
Alt4	(0.84,0.96,1.00)	(0.60,0.72,0.84)	(0.48,0.60,0.72)	(0.24,0.36,0.48)
Alt5	(0.48,0.60,0.72)	(0.72,0.84,0.96)	(0.60,0.72,0.84)	(0.12,0.24,0.36)

Table 4 Normalized decision matrix for DM1

DM1	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.00,0.12)	(0.25,0.375,0.50)	(0.84,0.96,1.00)	(0.48,0.6,0.72)
Alt2	(0.24,0.36,0.48)	(0.75,0.875,1.00)	(0.00,0.12,0.24)	(0.6,0.72,0.84)
Alt3	(0.60,0.72,0.84)	(0.25,0.375,0.50)	(0.12,0.24,0.36)	(0.84,0.96,1.0)
Alt4	(0.84,0.96,1.00)	(0.625,0.75,0.875)	(0.48,0.60,0.72)	(0.24,0.36,0.48)
Alt5	(0.48,0.60,0.72)	(0.75,0.875,1.00)	(0.60,0.72,0.84)	(0.12,0.24,0.36)

Table 5 Weighted normalized decision matrix for DM1

DM1	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.00,0.04)	(0.083,0.125,0.167)	(0.28,0.32,0.333)	(0.16,0.20,0.24)
Alt2	(0.08,0.12,0.16)	(0.25,0.291,0.333)	(0.00,0.04,0.08)	(0.20,0.24,0.28)
Alt3	(0.20,0.24,0.28)	(0.083,0.125,0.167)	(0.04,0.08,0.12)	(0.28,0.32,0.333)
Alt4	(0.28,0.32,0.333)	(0.208,0.25,0.291)	(0.16,0.20,0.24)	(0.08,0.12,0.16)
Alt5	(0.16,0.20,0.24)	(0.25,0.291,0.333)	(0.20,0.24,0.28)	(0.04,0.08,0.12)

Table 6 Length from PIS, NIS and closeness with ranking order for DM1

DM1	D.P.I.S	D.N.I.S	Ci	Ri
Alt1	0.576	0.391	0.404	5
Alt2	0.535	0.435	0.448	4
Alt3	0.47	0.498	0.515	2
Alt4	0.345	0.623	0.644	1
Alt5	0.415	0.554	0.572	2

Table 7 Decision Matrix for DM2

DM2	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.0,0.12)	(0.12,0.24,0.36)	(0.36,0.48,0.60)	(0.6,0.72,0.84)
Alt2	(0.0,0.12,0.24)	(0.24,0.36,0.48)	(0.48,0.60,0.72)	(0.72,0.84,0.96)
Alt3	(0.12,0.24,0.36)	(0.36,0.48,0.60)	(0.60,0.72,0.84)	(0.84,0.96,1.0)
Alt4	(0.24,0.36,0.48)	(0.48,0.60,0.72)	(0.72,0.84,0.96)	(0.00,0.0,0.12)
Alt5	(0.36,0.48,0.6)	(0.60,0.72,0.84)	(0.84,0.96,1.00)	(0.12,0.24,0.36)

Table 8 Normalized decision matrix for DM2

DM2	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.0,0.20)	(0.143,0.286,0.429)	(0.36,0.48,0.6)	(0.60,0.72,84)
Alt2	(0.0,0.20,0.40)	(0.286,0.429,0.571)	(0.48,0.6,0.72)	(0.72,0.84,0.96)
Alt3	(0.20,0.4,0.60)	(0.429,0.571,0.714)	(0.6,0.72,0.84)	(0.84,0.96,1.0)
Alt4	(0.40,0.60,0.80)	(0.571,0.714,0.857)	(0.72,0.84,0.96)	(0.00,0.00,0.12)
Alt5	(0.60,0.80,1.0)	(0.714,0.857,1.0)	(0.84,0.96,1.0)	(0.12,0.24,0.36)

Table 9 Weighted normalized decision matrix for DM2

DM2	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.0,0.067)	(0.048,0.095,0.143)	(0.12,0.16,0.2)	(0.20,0.24,0.28)
Alt2	(0.0,0.067,0.133)	(0.095,0.143,0.19)	(0.60,0.2,0.24)	(0.24,0.28,0.32)
Alt3	(0.067,0.133,0.20)	(0.143,0.19,0.238)	(0.2,0.24,0.28)	(0.28,0.32,0.333)
Alt4	(0.133,0.20,0.266)	(0.19,0.238,0.285)	(0.24,0.28,0.32)	(0.00,0.00,0.04)
Alt5	(0.20,0.266,0.333)	(0.238,0.285,0.333)	(0.28,0.32,0.333)	(0.04,0.08,0.12)

Table 10 Length from PIS, NIS and closeness with ranking order for DM2

DM2	D.P.I.S	D.N.I.S	Ci	Ri
Alt1	0.660	0.227	0.256	5
Alt2	0.488	0.4095	0.456	4
Alt3	0.300	0.588	0.662	2
Alt4	0.446	0.443	0.498	3
Alt5	0.231	0.657	0.740	1

Table 11 Decision matrix for DM3

DM3	Crt1	Crt2	Crt3	Crt4
Alt1	(0.6,0.72,0.84)	(0.36,0.48,0.60)	(0.72,0.84,0.96)	(0.36,0.48,0.6)
Alt2	(0.36,0.48,0.6)	(0.12,0.24,0.36)	(0.48,0.60,0.72)	(0.24,0.36,0.48)
Alt3	(0.72,0.84,0.96)	(0.84,0.96,1.00)	(0.60,0.72,0.84)	(0.6,0.72,0.84)
Alt4	(0.60,0.72,0.84)	(0.72,0.84,0.96)	(0.48,0.60,0.72)	(0.12,0.24,0.36)
Alt5	(0.36,0.48,0.6)	(0.84,0.96,1.00)	(0.00,0.12,0.24)	(0.48,0.6,0.72)

Table 12 Normalized decision matrix for DM3

DM3	Crt1	Crt2	Crt3	Crt4
Alt1	(0.625,0.75,0.875)	(0.36,0.48,0.60)	(0.75,0.875,1.00)	(0.429,0.571,0.714)
Alt2	(0.375,0.50,0.625)	(0.12,0.24,0.36)	(0.50,0.625,0.75)	(0.286,0.429,0.571)
Alt3	(0.75,0.875,1.00)	(0.84,0.96,1.00)	(0.625,0.75,0.875)	(0.714,0.857,1.00)
Alt4	(0.50,0.625,0.75)	(0.72,0.84,0.96)	(0.50,0.625,0.75)	(0.143,0.286,0.429)
Alt5	(0.375,0.50,0.625)	(0.84,0.96,1.00)	(0.00,0.125,0.25)	(0.571,0.714,0.857)

Table 13 Weighted normalized decision matrix for DM3

DM3	Crt1	Crt2	Crt3	Crt4
Alt1	(0.208,0.25,0.291)	(0.12,0.16,0.20)	(0.25,0.291,0.333)	(0.143,0.19,0.238)
Alt2	(0.125,0.167,0.208)	(0.04,0.08,0.12)	(0.167,0.208,0.25)	(0.095,0.143,0.19)
Alt3	(0.25,0.291,0.333)	(0.28,0.32,0.333)	(0.208,0.25,0.291)	(0.238,0.285,0.333)
Alt4	(0.167,0.208,0.25)	(0.24,0.28,0.32)	(0.167,0.208,0.25)	(0.048,0.095,0.143)
Alt5	(0.125,0.167,0.208)	(0.28,0.32,0.333)	(0.00,0.042,0.083)	(0.19,0.238,0.285)

Table 14 Length from PIS, NIS and closeness with ranking order for DM3

DM3	D.P.I.S	D.N.I.S	Ci	Ri
Alt1	0.288	0.508	0.638	2
Alt2	0.582	0.214	0.269	5
Alt3	0.042	0.755	0.948	1
Alt4	0.390	0.408	0.511	3
Alt5	0.422	0.374	0.470	4

Table 15 Decision matrix for DM4

DM4	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.00,0.12)	(0.00,0.12,0.24)	(0.24,0.36,0.48)	(0.72,0.84,0.96)
Alt2	(0.12,0.24,0.36)	(0.00,0.00,0.12)	(0.36,0.48,0.60)	(0.00,0.00,0.12)
Alt3	(0.00,0.12,0.24)	(0.12,0.24,0.36)	(0.00,0.00,0.12)	(0.48,0.60,0.72)
Alt4	(0.24,0.36,0.48)	(0.00,0.12,0.24)	(0.00,0.00,0.12)	(0.60,0.72,0.84)
Alt5	(0.36,0.48,0.60)	(0.84,0.96,1.00)	(0.72,0.84,0.96)	(0.00,0.00,0.12)

Table 16 Normalized decision matrix for DM4

DM4	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.00,0.20)	(0.00,0.12,0.24)	(0.25,0.375,0.50)	(0.75,0.875,1.0)
Alt2	(0.20,0.40,0.60)	(0.00,0.00,0.12)	(0.375,0.5,0.625)	(0.00,0.0,0.125)
Alt3	(0.00,0.20,0.40)	(0.12,0.24,0.36)	(0.00,0.00,0.125)	(0.5,0.625,0.75)
Alt4	(0.40,0.60,0.80)	(0.00,0.12,0.24)	(0.00,0.00,0.125)	(0.625,0.75,0.875)
Alt5	(0.60,0.80,1.00)	(0.84,0.96,1.00)	(0.75,0.875,1.00)	(0.0,0.00,0.125)

Table 17 Weighted normalized decision matrix for DM4

DM4	Crt1	Crt2	Crt3	Crt4
Alt1	(0.00,0.00,0.067)	(0.00,0.04,0.08)	(0.083,0.125,0.167)	(0.25,0.291,0.333)
Alt2	(0.067,0.133,0.20)	(0.00,0.00,0.04)	(0.125,0.167,0.208)	(0.00,0.00,0.042)
Alt3	(0.00,0.067,0.133)	(0.04,0.08,0.12)	(0.00,0.00,0.042)	(0.167,0.208,0.25)
Alt4	(0.133,0.20,0.266)	(0.00,0.04,0.08)	(0.00,0.00,0.042)	(0.208,0.25,0.291)
Alt5	(0.20,0.266,0.333)	(0.28,0.32,0.333)	(0.25,0.291,0.333)	(0.00,0.00,0.042)

Table 18 Length from PIS, NIS and closeness with ranking order for DM4

DM4	D.P.I.S	D.N.I.S	Ci	Ri
Alt1	0.684	0.424	0.382	3
Alt2	0.834	0.269	0.244	5
Alt3	0.792	0.319	0.287	4
Alt4	0.658	0.450	0.406	2
Alt5	0.278	0.822	0.747	1

the weights, the fuzzy PIS and fuzzy NIS, relative closeness coefficient with ranking order. From Tables 11, 12, 13 and 14, we established the appraise using third DM with NDM, propositional to the weights, the fuzzy PIS and fuzzy NIS, relative closeness coefficient with ranking order. From Tables 15, 16, 17 and 18, we established the appraise using fourth DM with NDM, propositional to the weights, the fuzzy PIS and fuzzy NIS, relative closeness coefficient with ranking order. From Tables 19, 20 and 21, we computed group relative closeness DM, the weighted group relative closeness DM, group fuzzy PIS and group fuzzy NIS. Hence, the ranking order of all RCM in the GDM with five alternatives is Alt 2 < Alt 1 < Alt 4 < Alt 5 < Alt 3 and best optimal is Alt 3.

Table 19 Group relative closeness matrix

Alternatives	Relative closeness decision matrix			
	DM1Ci	DM2Ci	DM3Ci	DM4Ci
Alt1	0.404	0.256	0.638	0.382
Alt2	0.448	0.456	0.269	0.244
Alt3	0.515	0.662	0.948	0.287
Alt4	0.644	0.498	0.511	0.406
Alt5	0.572	0.740	0.470	0.747

Table 20 Weighted group relative closeness matrix

Alternatives	Weighted relative closeness decision matrix			
	WDM1Ci	WDM2Ci	WDM3Ci	WDM4Ci
Alt1	0.101	0.064	0.1595	0.0955
Alt2	0.112	0.114	0.06725	0.061
Alt3	0.12875	0.1655	0.237	0.07175
Alt4	0.161	0.1245	0.12775	0.1015
Alt5	0.143	0.185	0.1175	0.18675

Table 21 Group FPIS, FNIS and ranking order

Alternatives	GFPIS	GFNIS	GRCC	Rank
Alt1	0.180482	0.09849	0.3530464	4
Alt2	0.228189	0.051196	0.1832446	5
Alt3	0.121018	0.200007	0.623027	1
Alt4	0.151206	0.112075	0.4256847	3
Alt5	0.120848	0.186395	0.6066696	2

5 Conclusion

The MCDM method having various applications in fuzzy TOPSIS DM problems. In the present study, the outcomes controvert those four different DM with weights of the projected techniques. The algorithm was planned and tabulated values are calculated using fuzzy TOPSIS software. We believe that the projected techniques manifest value but, as a obstruction, it is tough and impenetrable to estimate subjectively the fuzzy information in a realistic way while the results of the research are dependent on the experts opinions and linguistic variables. We demonstrated a MCDM technique, with DM, comprising of the value of a fuzzy number greater than or equal to another fuzzy number, a new inter-space measure of one after another fuzzy number from FPIS as well as FNIS. This method yields the optimal solution.

However, some surveillances are obtained from the given illustration; we are assertive the consequence for numerous illustrations would give us similar resolutions. We quite reflect a scads of illustrations should be nominated for test in future studies. Each and every topic allied to group intercommunications would be an interesting one for group DM and will be left for future study.

Acknowledgements The content of this article has been prepared as a part of research work carried out in C.V. Raman Global University, Odisha, India. We would like to thank the reviewers and editor for helpful insights.

References

1. Hwang, C.L., Yoon, K.: Multiple Attribute Decision Making: Methods and Applications. Springer, New York (1981)
2. Chu, T.C., Lin, Y.C.: Improved extensions of the TOPSIS for group decision making under fuzzy environment. *J. Inf. Optim. Sci.* **23**, 273–286 (2013)
3. Yong, D.: Plant location selection based on fuzzy TOPSIS. *Int. J. Adv. Manuf. Technol.* **28**(7–8), 839–844 (2006)
4. Herrera, F., Herrera-Viedma, E., Verdegay, J.L.: A model of consensus in group decision making under linguistic assessments. *Fuzzy Sets Syst.* **78**, 73–87 (1996)
5. Ertugrul, I., Karakasoglu, N.: Performance evaluation of Turkish cement firms with fuzzy analytic hierarchy process and TOPSIS methods. *Expert Syst. Appl.* **36**, 702–715 (2009)
6. Wang, J., Liu, S.Y., Zhang, J.: An extension of TOPSIS for MCDM based on vague set theory. *J. Syst. Sci. Syst. Eng.* **14**, 73–84 (2005)
7. Wang, J.-W., Cheng, C.-H., Cheng, H.K.: Fuzzy hierarchical TOPSIS for supplier selection. *Appl. Soft Comput.* **9**(1), 377–386 (2009)
8. Parida, P.K.: Some Generalized Results on Multi-criteria decision making model using fuzzy TOPSIS technique. *Biol. Inspired Tech. Many-Criteria Decis. Making* **10**, 189–199 (2020)
9. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1995)
10. Kahraman, C., Otay, I., (eds.): Fuzzy multi-criteria decision-making using neutrosophic sets. In: *Studies in Fuzziness and Soft Computing*. Springer, Berlin (2019)
11. elgado, M., Verdegay, J.L., Vila, M.A.: Linguistic decision-making models. *Int. J. Intell. Syst.* **7**(5), 479–492 (1990)
12. Parida, P.K., Sahoo, S.K.: Fuzzy multiple attributes decision making models using TOPSIS technique. *Int. J. Appl. Eng. Res.* **10**(2), 2433–2442 (2015)

13. Parida, P.K.: A general view of TOPSIS method involving multi-attribute decision making problems. *Int. J. Innov. Technol. Exploring Eng.* **1.9**(2), 3205–3214 (2019)
14. Krohling, R.A., Andre, G.C.P.: A-TOPSIS-an approach based on TOPSIS for ranking evolutionary algorithms. *Precedia Comput. Sci.* **55**, 308–317 (2015)
15. Chen, C.T.: Extensions of the TOPSIS for group decision-making under fuzzy environment. *Fuzzy Sets Syst.* **114**, 01–09 (2000)
16. Wang, T.C., Lee, H.D.: Developing a fuzzy TOPSIS approach based on subjective weights and objective weights. *Expert Syst. Appl.* **36**, 8980–8985 (2009)
17. Wang, Y.M., Elhag, T.M.S.: Fuzzy TOPSIS method based on alpha level sets with an application to bridge risk assessment. *Expert Syst. Appl.* **31**(2), 309–319 (2006)
18. Wang, J., Liu, S.Y., Zhang, J.: An extension of TOPSIS for fuzzy MCDM based on vague set theory. *J. Syst. Sci. Syst. Eng.* **14**, 73–84 (2005)

IoT-Based Smart Intravenous Drip Monitoring System



Muskan Jindal, Nidhi Gajjar, and Nehal Patel

Abstract Health care being the most important aspect in heading toward a contented life also plays an essential role in India's progress. These days, automating the health monitoring devices leads to a drastic change in medical sphere as it ensures the safety of the patients and even helps in reducing the stress of doctors and nurses. In this field, intravenous remedy plays an important role as it is the system wherein the liquid substances are directly inserted into the patient's vein via an IV tube but it could also worsen the situations if not taken proper care. Thus, this paper emphasizes on the necessity to overcome such a consequence by introducing a solution to it. Hereby an automatic intravenous drip monitoring system is developed which directly sends an alert message to the assigned nurse when the fluid level of the bottle reaches a certain limit. This system measures the weight of the saline bottle with the help of a load cell and then using an automatic alerting and indicating device namely GSM sends the alert signal. This system would be a significant serve to build a different approach toward the intravenous therapy.

Keywords Load cell · HX711 sensor · Arduino microcontroller · GSM modem · Health care · IoT

1 Introduction

The Internet of things (IoT) and cloud technology play a dynamic role in the expansion of civilization as a brand novel tactical trade, and the modern technologies have

M. Jindal · N. Gajjar · N. Patel (✉)
K D Patel Department of Information Technology, Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology & Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India
e-mail: nehalpatel.it@charusat.ac.in

M. Jindal
e-mail: 18it040@charusat.edu.in

N. Gajjar
e-mail: 18it033@charusat.edu.in

a profound influence on the contemporary education organization [1–3]. With the increasing population around the globe, the demand of the healthcare [4, 5] units also increases. Also, it becomes the responsibility of the healthcare industry to provide best possible treatments to the patients at lower costs and maintaining their safety. Automation in health care is the step toward providing best services to patients in a cost efficient way. One of the challenges faced by the healthcare units is ensuring the safety of the patients during IV infusions. Intravenous infusions [6, 7] (commonly known as drips) are the process of infusing fluid substances directly into patients' vein. It is a typical method of treatment used for fluid volume replacement, to correct electrolyte imbalances in body, to send medicines, and for transfusion of blood or fluid injection. The task of keeping a constant watch on the level of the fluid in the fluid bottle is tedious and time consuming and restricts the efficiency of the hospital staff to do other tasks. Thus, this system helps in eliminating the constant manual task of keeping a regular watch on the level of fluid in the bottle.

In this paper, Section 2 defines background study of IoT-related sensors and health care. The proposed system is elucidated in Sec. 3. Section 4 indicates implementation and discussion of proposed system. Section 5 shows the result of the system. Finally, the last section shows summaries and future direction.

2 Existing System

Keerthi et al. [8] suggested that the system comprises of IR sensor which acts as a level sensor for monitoring saline level in the bottle and whenever the saline level reaches the predefined dangerous level of liquid alert message is sent through the internet to nurse. Along with that, buzzer alarm also starts ringing.

Anand [6] has proposed system, in which sending the message regarding the patients' health to the nurse is through GSM technology. The system automatically turns of the flow of liquid from intravenous bag using solenoid valve and also measures the pulse rate and blood pressure of the patient and display it on the LCD.

Bhavaasar et al. [7] developed a system where the level of the IV bag is monitored and checks if the level is dropped beyond the set level and senses the air bubbles or embolisms appearing in tube before it enters the patient's vein. The nurse station is also alerted about the same with help of alarm system so suitable action can be taken on time.

Arulious et al. [9] used the level of liquid light-dependent resistors (LDR) to address. The LDR and LED are fixed opposite to each other and once the level of liquid becomes low the conductivity of LDR sensor increases. The microcontroller is programmed in such a way that once the conductivity is increased it buzzes the buzzer in nurse's room and alerts the nurses.

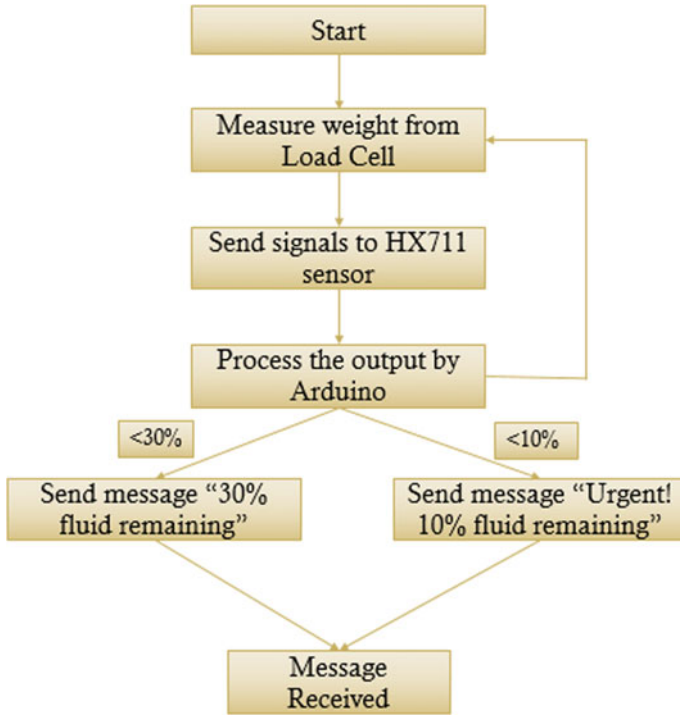


Fig. 1 Flow of entire system

3 Proposed System

This system aims at automating intravenous fluid monitoring system using the Arduino microcontroller. This project proposes GSM-based automatic alerting system where weight sensor is used as level sensor. Here, the intimation is given when the fluid reaches the certain fixed level, so that the nurse gets enough time to reach the room and replace the bottle. Figure 1 shows flowchart of how the system works. The load cell measures the weight and once the weight is less than the specified amount the GSM is used to send the message to the specified phone number.

4 Implementation

A. Connection of Load Cell and HX711 Sensor

A load cell is a sensor that detects changes in a physical stimulus (force, weight, or pressure) and then produces an output proportional to the physical stimulus. Load cell is basically used to sense the weight of the bottle and supply an electrical analog voltage to the HX711 load amplifier module. The HX711 is a

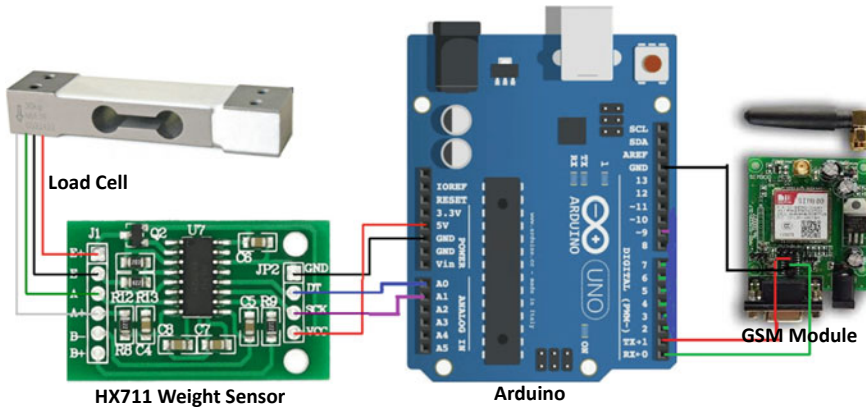


Fig. 2 Connections

load cell amplifier which is used to generate measurable data out from the load cell and strain gauge. The amplifier attached with the load cell helps in finding the actual weight of the saline bottle. Thus, by using the measured weight, the level of liquid present in the bottle can be calculated and is further passed onto the arduino.

B. Connection of HX711 Weight Sensor and Arduino

In this proposed system, an open-source electronics platform based on easy-to-use hardware and software is used namely Arduino. The HX711 load cell amplifier is connected to the Arduino and thus the software Arduino IDE is used to monitor the measured weight of the glucose bottle as predicted by the load cell and amplifier. It involves the task of calibrating the system for measuring correct weight. After calibration, weight measurement is done normally.

C. Sending Alert Message

Global system for mobile communications (GSM) is a mobile communication modem. A GSM module or a GPRS module is a chip or circuit that will be used to launch communication between a mobile device or a computing machine and a GSM or GPRS system [10]. After the Arduino receives the weight of the bottle in the form of voltage signals, it processes the signals and checks if the weight of the fluid is less than the specified amount. If the weight is less, then the GSM module through the serial communication sends the alert message to the nurse. Figures 2 and 3 indicates the working of entire system.

5 Results and Discussion

This automation system with Arduino and GSM is developed for the healthcare system of the human beings. This project provides the advantages for nurse and

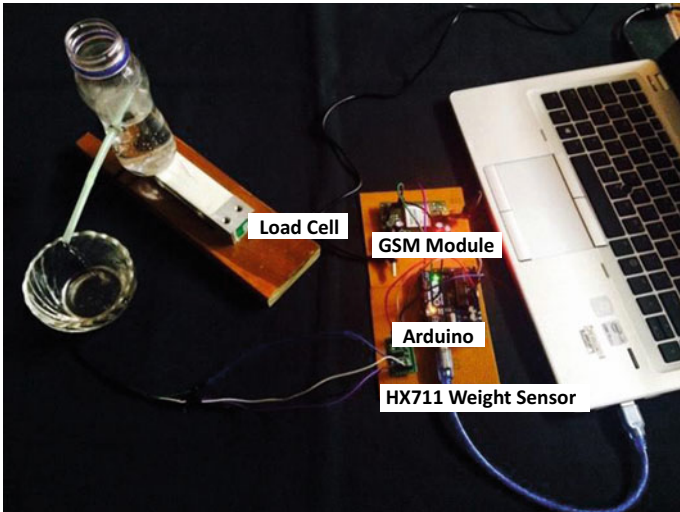


Fig. 3 Actual setup

assistants in healthcare system as it helps in eliminating the task of constantly monitoring the level of liquid in a bottle manually. The varying weight can be viewed on the serial monitor of the Arduino IDE. Following is the observed output of the load cell.

In Fig. 4, the alert is generated when the level has dropped below a certain limit. The alert message would be received in the form of a SMS on the registered phone number as shown in Fig. 5.

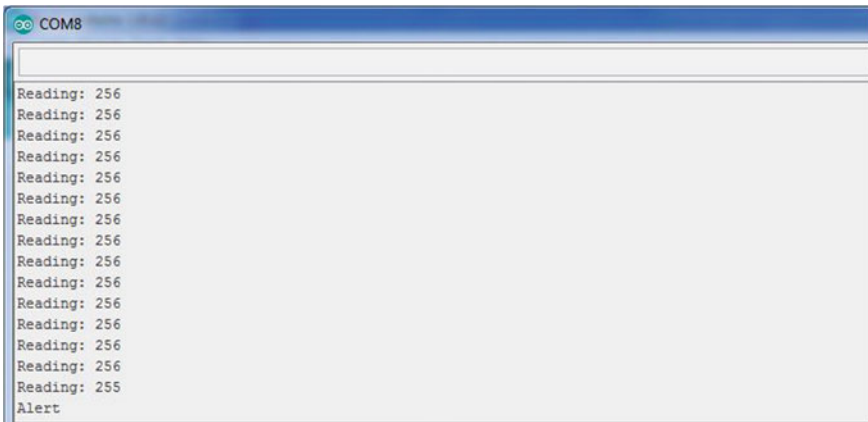
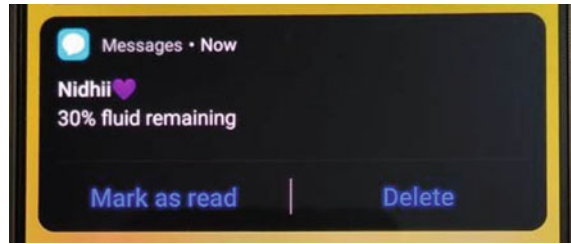


Fig. 4 Output on serial monitor

Fig. 5 SMS alert message received



6 Conclusion and Future Direction

To recapitulate, the constant task of manually monitoring the intravenous drip is eliminated. This system would be more beneficial at the nighttime to the hospital staff and patients. Also, implementing this system avoids a major risk of air bubbles entering the patient's bloodstream, which if happens can be fatal. This system can further be extended by including a call alert in it when the level of fluid is critical. Also, the system can be enhanced where along with the fluid-level message, the patient's body temperature, blood pressure level, and pulse rate are measured and sent to the hospital staff at regular intervals.

References

1. Singh, D., Pati, B., Panigrahi, C.R., Swagatika, S.: Security issues in IoT and their countermeasures in smart city applications. *Adv. Comput. Intell. Eng.* **1089**, 301–313 (2020)
2. Rath, M., Pati, B.: Security Assertion of IoT Devices Using Cloud of Things Perception. *Int. J. Interdisc. Telecommun. Netw.* **11**(4), 17–31 (2019)
3. Mishra, M., Choudhury, P., Pati, B.: Modified ride-NN optimizer for the IoT based plant disease detection. *J. Ambient. Intell. Human Comput.* (2020). <https://doi.org/10.1007/s12652-020-02051-6>
4. Khan, S.F.: Health care monitoring system in Internet of Things (IoT) by using RFID. In: 2017 6th International Conference on Industrial Technology and Management (ICITM), pp. 198–204. IEEE (2017)
5. Ahmed, M.U., Begum, S., Fasquel, J.B. (eds.): *Internet of Things (IoT) Technologies for HealthCare*. Springer, New York (2018)
6. Anand, M., Pradeep, M., Manoj, S., Marcel, L., Arockia Raj, P.: Thamaraiyani intravenous drip monitoring system. *Indo-Iranian J. Sci. Res. (IIJSR)* **2** (1), 106–113 (2018). (Article Published: 16 February 2018)
7. Bhavasaar, M.K., Nithya, M., Praveena, R., Bhuvanewari, N.S., Kalaiselvi, T.: Automated intravenous fluid monitoring and alerting system. In: 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), pp. 77–80. IEEE (2016)
8. Keerthi, N.S., Raju, A., Sowmya, N., Krishna, D.: Intravenous infusion monitoring system. *Int. J. Recent Develop. Sci. Technol.* **04**(03) (2020) March
9. Arulious Jora, A., Divya Laveena, A., Earlina, D., Nirmala, S.: Intravenous fluid level indicator. *Int. Res. J. Eng. Technol. (IRJET)* **05**(10) (2018)
10. <https://www.electronicsforu.com/>

Cryptanalysis of Lightweight Ciphers Using Metaheuristics



Seeven Amic, K. M. Sunjiv Soyjaudah, and Gianeshwar Ramsawock

Abstract There are several lightweight cryptographic ciphers (LWC) which have been proposed lately for constrained environments. When it comes to selecting one specific cipher for a particular application, the task becomes complicated. Several surveys have been conducted on hardware and software implementations of lightweight cryptographic primitives in order to evaluate and benchmark their characteristics such as throughput, ROM and RAM requirements, power consumption, and so on. But tackling the question about the strength of lightweight cryptographic primitives and their ability to withstand known and potential attacks is not a straightforward activity. The underlying design of the primitives or the key size of LWC is insufficient criteria to measure and compare their cryptographic strengths. This paper attempts to evaluate the relative strength of software implementation of lightweight block ciphers using metaheuristic algorithms and proposes a framework for selection of encryption algorithm. Experiments show that LWC algorithms are as strong as classical AES despite their simple design. Moreover, they execute faster with less resources than conventional AES. Finally, it is observed that the relative strength of LWC is independent of their underlying architectural design.

Keywords Lightweight cryptography · Cryptanalysis · Metaheuristics

S. Amic (✉)
Université Des Mascareignes, Rose Hill, Mauritius
e-mail: samic@udm.ac.mu

K. M. S. Soyjaudah · G. Ramsawock
University of Mauritius, Reduit, Mauritius
e-mail: ssoyjaudah@uom.ac.mu

G. Ramsawock
e-mail: gramsawock@uom.ac.mu

1 Introduction

Lightweight cryptography (LWC) is a relatively novel wave of cryptographic primitives aimed at providing security in constrained environments such as tiny devices in Internet of things, contactless smart cards, radio-frequency identification tags, and sensor networks. The constraints might be in terms of low-power microcontrollers, limited amount of ROM and RAM, lower gate equivalents (GEs), and higher throughput. Lightweightness of a cryptographic primitive is approximately proportional to the amount of resources required for it to execute in space and time. However, power consumption of the primitive is a significant factor that matters to both hardware and software implementations.

Lightweight cryptographic schemes fall under five categories: stream ciphers, block ciphers, authenticated ciphers, message authentication codes, and hash functions. This paper focuses on software-oriented lightweight symmetric block ciphers. It is observed that designs of symmetric block LWC favor the following trends in order to provide nonlinearity: lookup tables (LUT), ARX-based algorithms, bit-sliced S-box coupled with simple key schedules [1].

Lightweight block ciphers by design fall into three general categories: Feistel networks (FN), substitution–permutation networks (SPN), Add/Rotate/XOR (ARX) [2] with variations and combinations. SPN structures aim at providing confusion and diffusion by making use of substitution and permutation boxes, respectively. While confusion obfuscates the relationship between the ciphertext and the key, diffusion promotes the desirable avalanche effect. Figure 1 shows a broad classification of lightweight block ciphers based on their internal architecture.

Here are some examples of lightweight symmetric block ciphers designed by government, academia, and industry:

AES-like: AES [3], PRESENT [4], KLEIN [5], mCrypton [6], LED [7], Zorro [8];

Bit-sliced SPN: Fantomas/Robin [9], NOEKEON [10], Pride [11], RECTANGLE [12];

ARX-based SPN: SPARX [13];

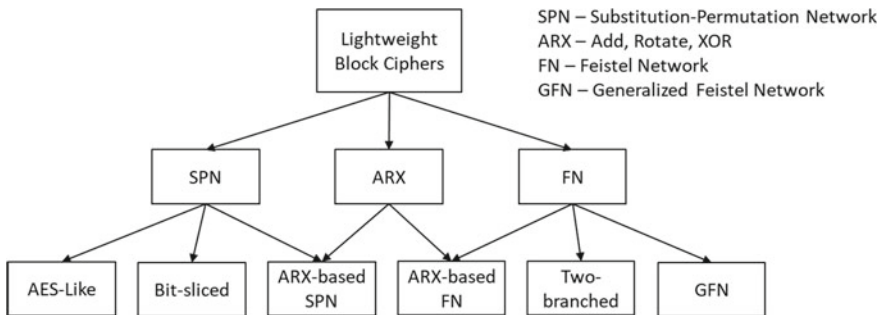


Fig. 1 Broad classification of lightweight cryptography

ARX-based FN: HIGHT [14], LEA [15], SIMON [16], Speck [17], SIMECK [18];

Two-branched FN: DESLX [19], GOST [20], MISTY [21], LBlock [22], Road-RunneR [23], SEA [24];

GFN: CLEFIA [25], TWINE [26], Piccolo [27].

The greatest challenge of LWC design is whether the cryptographic strength is compromised at the expense of reduced resources such as ROM, RAM, and number of cycles. This work attempts to comparatively study block LWC algorithms regardless of their internal structural design or contextual implementation, i.e., hardware or software. There is no straightforward method to evaluate the strength of a cryptographic primitive. In general, the strength of a cipher is equivalent to the effort required to find the key used for encryption. Some suggest that the key size is one of the factors to determine the strength of the cipher. The current recommended key size is 112 bits or higher so as to withstand bruteforce attack [28]. However, the resistance of a cipher against known cryptanalytic attacks should also be taken into consideration.

In this paper, we attempt to establish the relative strength of LWC algorithms by making use of metaheuristics as a cryptanalytic method of attack. Various authors have claimed to have cryptanalyzed Feistel ciphers [29, 30] using metaheuristic algorithms. Nevertheless, to the best of our knowledge, there are no reported publications of cryptanalytic attack of LWC using metaheuristic techniques.

The rest of the paper displays as follows: first, we give a short description of each of the selected ciphers under scrutiny. Second, we describe the metaheuristic algorithms we shall use for cryptanalysis. Then chapters on the Methodology, Results and Discussions, and finally, Conclusion and Future works are followed.

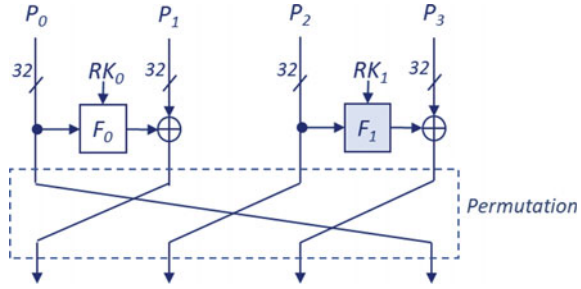
2 Short Specifications of Selected Ciphers

This section describes a few selected lightweight ciphers considered in this paper, namely AES (SPN), CLEFIA (GFN), SPARX (ARX-SPN), and Speck (ARX-FN). The ciphers have been selected such that they possess the same block size (i.e., 128 bits) and key size (i.e., 128 bits), but each from a different class of LWC. Though these ciphers may be also implemented in hardware, we consider only their software realization as their cryptanalysis depends only on the plaintext, key, and corresponding ciphertext and is independent of their internal structure or implementation medium.

2.1 AES

The Advanced Encryption Standard (AES) [31] is one of the most popular and robust 128-bit block ciphers since it became a standard in 2001. It offers three echelons of security based on the size of the key used. The key sizes are 128, 192, or 256 bits

Fig. 2 CLEFIA round



with number of rounds equal to 10, 12, and 14 correspondingly. AES has a classic substitution/permutation network structure (SPN). The 128-bit plaintext block, P , also known as the state, is stored in a 4×4 matrix of bytes.

2.2 CLEFIA

CLEFIA [32] is a 128-bit lightweight symmetrical block cipher designed by Sony Corporation in 2007. The authors of CLEFIA claim that the latter provides comparatively high level of security with good performance. CLEFIA permits the use of keys of size 128, 192, or 256 bits. CLEFIA has a generalized Feistel network (GFN) structure with a four-branch data line. CLEFIA uses a diffusion switching mechanism (DSM), which accounts for its robustness to both linear and differential cryptanalytic attacks by making use of two F —functions F_0 and F_1 . Due to the smaller size of the F functions, CLEFIA is more compact and executes with higher speed compared to ciphers in the same category.

Encryption is performed on the 128-bit plaintext data block, P , represented as four 32-bit words, as follows: $P = P_0|P_1|P_2|P_3$. One CLEFIA round is depicted in Fig. 2 whereas the CLEFIA encryption algorithm is shown in Fig. 3.

2.3 SPARX-N/k

SPARX is a relatively new lightweight block symmetric family of ciphers designed by Dinu et al. [33] in 2016 with different block size, n , and key size, k . SPARX has an ARX architecture based on SPN structure, ARX standing for Addition/Rotation/XOR, which are the primitive components of the cipher. The ARX box acts as a nonlinear function called Specky as shown in Fig. 4. Furthermore, SPARX uses the technique of long trail strategy (LTS), which accounts for higher immunity against linear and differential cryptanalysis, faster execution, and higher efficiency. Figure 5 depicts the round function used in SPARX, in which the Specky

Algorithm: CLEFIA Cipher

```

def  $P = P_0|P_1|P_2|P_3$  as the plaintext, 4 words of 4 bytes each
def  $N$  as number of rounds
def  $RK_{0,\dots,N}$  as round keys
def  $WK_{0,\dots,4}$  as whitening keys
 $P_1 \leftarrow P_1 \oplus WK_0$ 
 $P_3 \leftarrow P_3 \oplus WK_1$ 
for  $i \leftarrow 0$  to  $N - 1$  do
   $P_1 \leftarrow P_1 \oplus F_0(P_0, RK_{2i})$ 
   $P_3 \leftarrow P_3 \oplus F_1(P_2, RK_{2(i+1)})$ 
  Temp  $\leftarrow P_0$ 
   $P_0 \leftarrow P_1$ 
   $P_1 \leftarrow P_2$ 
   $P_2 \leftarrow P_3$ 
   $P_3 \leftarrow Temp$ 
end for
 $P_1 \leftarrow P_1 \oplus WK_2$ 
 $P_3 \leftarrow P_3 \oplus WK_3$ 
return  $P$ 

```

Fig. 3 CLEFIA cipher for encryption

Fig. 4 Speckey ARX box

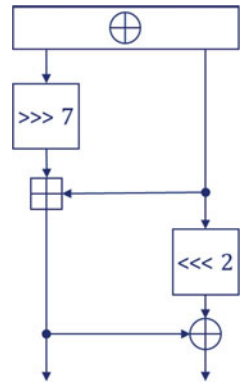
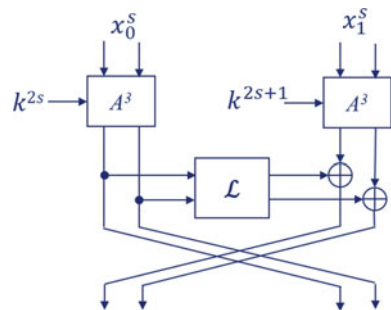


Fig. 5 SPARX round function



Algorithm: SPARX encryption

```

def  $P$  as the plaintext divided into two parts:  $P_0$  and  $P_1$ 
def  $N$  as number of rounds
def  $S$  as number of steps
def  $RK_{0,\dots,2NS}$  as round keys
for  $i \leftarrow 0$  to  $S - 1$  do
   $RKT \leftarrow RK_{2Ni}$ 
  for  $j \leftarrow 0$  to  $N - 1$  do
     $P_0 \leftarrow P_0 \oplus RKT_j$ 
     $Speckey(P_0)$ 
     $P_1 \leftarrow P_1 \oplus RKT_{j+N}$ 
     $Speckey(P_1)$ 
  end for
   $Temp \leftarrow P_0$ 
   $P_0 \leftarrow P_1 \oplus (P_0 \oplus (P_0 \lll 8) \oplus (P_0 \ggg 8))$ 
   $P_1 \leftarrow Temp$ 
end for
 $P_0 \leftarrow P_0 \oplus RK_{2NS}$ 
 $P_1 \leftarrow P_1 \oplus RK_{2NS+1}$ 
return  $P$ 

```

Fig. 6 SPARX algorithm for encryption [33]

function is denoted by A^3 . The components of the round function are used again in the key schedule with the purpose of minimizing the code size (Fig. 6).

2.4 Speck

The Speck family of ciphers is an ARX-based Feistel network proposed by Beaulieu et al. [16, 34] and endorsed by the National Security Agency (NSA) for a variety of applications. Speck is designed to support block sizes and various corresponding key lengths. The key size determines the number of rounds for encryption and decryption.

Speck makes use of the round function for key scheduling as well as the data encryption. As shown in Fig. 7, the Speck cipher has the typical Feistel structure whereby the data block is divided into the L and R halves, which are rotated α and β times to the left and right, respectively. The round function performs Addition (\boxplus), Rotation (\ggg , \lll) and XOR (\oplus) operations on them. The Speck key schedule is shown in Fig. 8 and the Speck algorithm is depicted in Fig. 9.

3 Metaheuristics Algorithms

Metaheuristics is a field of high-level strategies that are configured to solve combinatorial optimization problems, which cannot be tackled efficiently by traditional

Fig. 7 Speck round

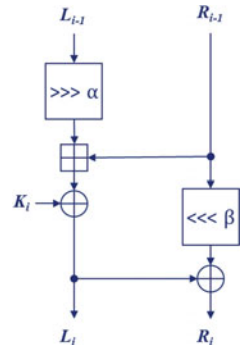
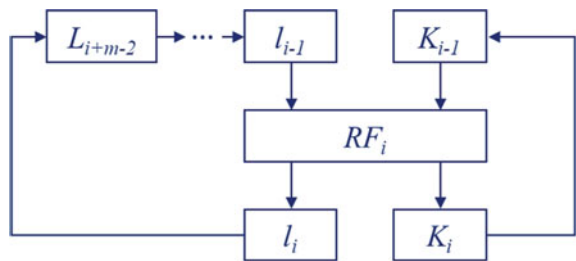


Fig. 8 Speck key schedule



Algorithm: Speck Cipher

```

def n ∈ {16, 24, 32, 48, 64}, word size in bits
def P ← L, R
def N as number of rounds
def RK0,...,N as round keys which are generated
α = 7 if n = 16, else 8
β = 2 if n = 16, else 3
for i ← 0 to N-1 do
    L ← (L >>> α + R) ⊕ RKi
    R ← (R <<< β) ⊕ L
end for
return P
    
```

Fig. 9 Speck cipher algorithm for encryption

deterministic approaches. Metaheuristics comes in handy in solving problems where information is imperfect or incomplete or computing resources are limited. There are a number of metaheuristic optimization models falling in different categories, such as population-based, combinatorial, model fitting, and nature-inspired algorithms. Specific metaheuristic techniques, namely genetic algorithm, firefly algorithm, and particle swarm optimization algorithm are described in the next section.

3.1 Genetic Algorithm

In the early 1970s, John Holland from the University of Michigan invented the genetic algorithm (GA) [35]. The idea of GA is to simulate the processes of natural evolution based on the principles of the work of Charles Darwin: “survival of the fittest” [36]. GA simulates the biological evolutionary process which constitutes of building a new population of chromosomes from an existing one by using the principle of “natural selection.” In order to achieve this goal, GA makes use of genetic transformations such as crossover and mutation. Each chromosome, also known as individual, is recorded as a bit string which represents a sequence of genes. The value of each gene represents an instance of a particular allele (e.g., 0 or 1).

The selection operation is the process during which chromosomes from the current population are picked to participate in the reproduction of the new population. Generally, the fitter chromosomes are expected to produce offsprings of higher quality or fitness than the inferior ones. An objective (or fitness) function is used to quantitatively establish the aptness of an individual for breeding. There are three main selection approaches: roulette wheel, tournament, and elitist. Crossover is a process of mixing bits and pieces of chromosomes from the parents to construct the offsprings. There are three traditional methods of doing crossover in vectors: one-point, two-point, and uniform crossover.

3.2 Firefly Algorithm

In 2008, Xin She Yang invents the firefly algorithm (FA) [37] based on the behavior of fireflies during courtship. Fireflies emit light of a given intensity and are attracted toward other fireflies of higher light intensity. The strength of attraction from one firefly toward another is directly proportional to the light intensity emitted by the latter firefly and varies inversely to the distance separating the two fireflies. When a firefly moves toward another firefly with higher brightness, it changes its own position as well as the intensity of its light. It is believed that the intensity of each firefly in the population will change until one or more of the fireflies reach maximum intensity after a certain number of generations of movement. The firefly with highest brightness in the population represents the solution to the problem determined by the position of the firefly in the D —dimensional space.

3.3 Particle Swarm Optimization Algorithm

The particle swarm algorithm (PSO) was developed by Kennedy and Eberhart [38] in 1995. In PSO, a particle represents a solution to a problem. At any time, t , the particle is at position, x , in a D —dimensional space and moves at a velocity, v . In

a fixed-size population of n particles, each particle moves and updates its position. This new position can be used to evaluate the fitness of the particle. The best particle at any given time is $pBest$, and the overall best particle is $gBest$, which eventually is the solution to the problem being solved after a number of iterations or the level of fitness required being reached.

4 Methodology

This work focuses on the cryptanalysis of block symmetric lightweight cryptosystems using selected metaheuristic algorithms, namely genetic algorithm, particle swarm optimization, and firefly algorithm using a known-text attack. Cryptanalysis of block ciphers may be considered as a search for a cryptographic key in a huge D —dimensional space. Cryptanalysis falls in the category of NP-hard combinatorial problems which can be addressed by metaheuristics because its time complexity increases exponentially with the size of the cryptographic key.

The cryptanalytic attack constitutes of searching for a key that matches the original key that was used for encryption. Experiments are set up such that an initial population of keys are generated and evaluated for fitness. For each generated key, K' , by the metaheuristic algorithm, the fitness of the key is obtained by using the formula:

$$\text{Fitness}(K') = \frac{\Delta C}{N} \quad (1)$$

where ΔC is the number of bits which match between the ciphertext obtained using the real key, K , and the ciphertext obtained with K' when encrypting the same plaintext. N is the data block size measured in number of bits. The higher the recorded fitness of K' , the higher the efficiency of the cryptanalysis using metaheuristics. For each iteration or generation, the fitness of the best key obtained is recorded. When the program is executed several times, the average fitness of the best keys obtained per generation is calculated.

The selected cryptographic algorithms under study are AES, CLEFIA, SPARX, and Speck. They have been selected so that they have similar data block size (i.e., 128 bits) and same key size (128 bits) but different internal architecture. The structure of the proposed system to tackle this problem is shown in Fig. 10.

As shown in the diagram, Cipher is an interface that allows to pick the cipher under scrutiny, and its appropriate key, to be executed in each experiment. The cryptographic key is chosen randomly, and for each execution of the program a different key is generated. Though AES has been optimized to work as a LWC on smaller processors, in our experiments, we have used the standard software version of AES in the Java Cryptography Extension. AES is used as the benchmark cipher. The Data Encryption Standard (DES) was also included among the ciphers in order to compare

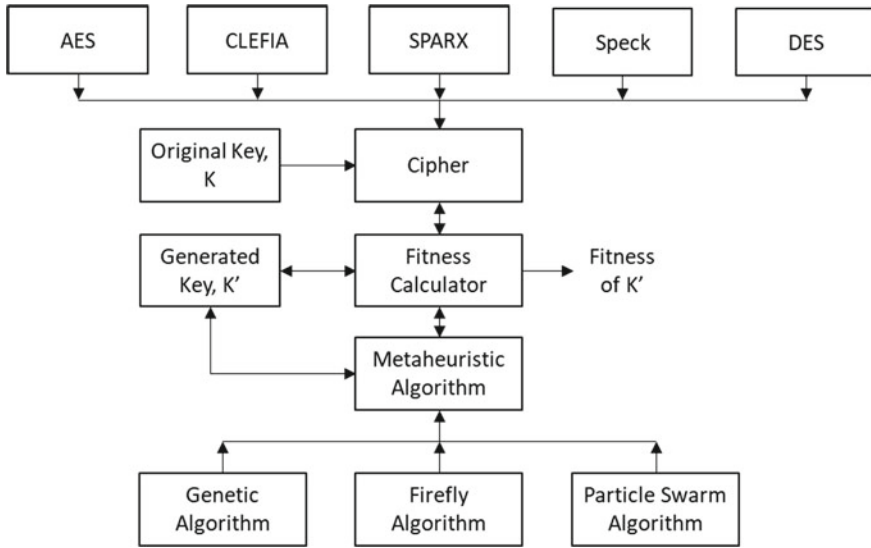


Fig. 10 Proposed framework for cryptanalysis of lightweight cryptographic algorithms using metaheuristics

its behavior when cryptanalyzed using the same metaheuristic algorithms. DES operates on data blocks of 64 bits with a key of 56 bits and has a Feistel network (FN) structure.

The ciphers were executed in Electronic Code Book Mode without padding. The system was implemented in the Java programming language due to its versatile object-oriented and polymorphism capabilities to easily plug-in different ciphers and metaheuristic algorithms under comparable parameters. The program was run on a CORE i7 computer with a 2.4 GHz clock and 16 GB of internal RAM.

For each execution, a cipher was taken and using a fixed key, a number of plaintext–ciphertext pairs were generated. Then, the metaheuristic algorithm was initialized with a fixed population of keys, usually as an XOR of plaintext and ciphertext pairs to generate initial random keys. The experiments were run 100 times and the average fitness of the best key generated for each run recorded at intervals of 100 between 100 and 2000 generations. The population size was fixed at 100 for all experiments with GA, FA, and PSO.

5 Results and Discussion

The results of the experiments are shown in Figs. 11, 12 and 13. The average fitness obtained through the cryptanalysis of the different ciphers by making use of particle swarm optimization, firefly, and genetic algorithms, respectively, is displayed. The

Fig. 11 Cryptanalysis using PSO

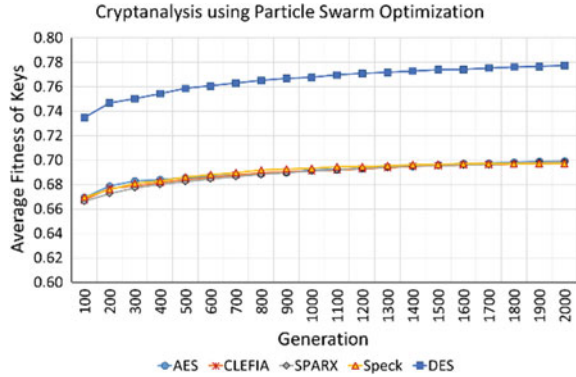


Fig. 12 Cryptanalysis using FA

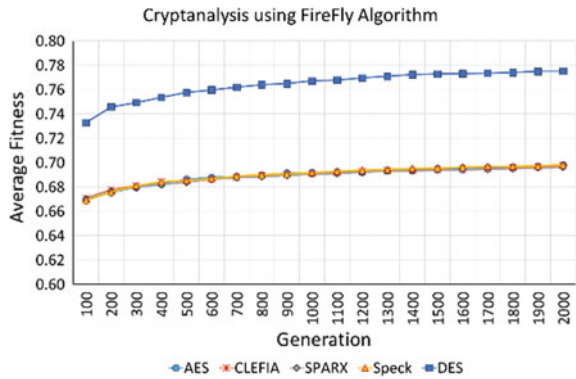
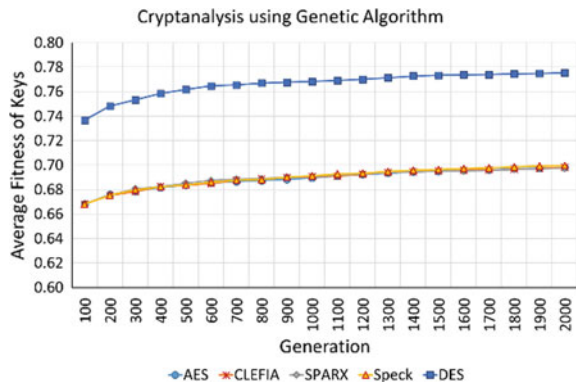


Fig. 13 Cryptanalysis using GA



corresponding root mean square difference (RMSD) between AES and the rest of the ciphers with respect to cryptanalysis using PSO, FA, and GA is shown in Tables 1, 2, and 3, respectively.

Table 1 RMSD for PSO

	AES
CLEFIA	0.0016
SPARX	0.0025
SPECK	0.0017
DES	0.0770

Table 2 RMSD for FA

	AES
CLEFIA	0.0012
SPARX	0.0014
SPECK	0.0009
DES	0.0768

Table 3 RMSD for GA

	AES
CLEFIA	0.0007
SPARX	0.0012
SPECK	0.0013
DES	0.0791

It is observed that the average fitness of keys generated by FA, PSO, and GA for the cryptanalysis of the LWC algorithms CLEFIA, SPARX, and Speck follows the same trend as AES with insignificant difference as proved by their respective RMSD. However, from the graphs, it is clear that the cryptanalysis of DES using FA, PSO, and GA generates keys with higher average fitness compared to the rest of the cryptographic algorithms.

Though the underlying architectures of the LWC studied are different, they are observed to yield similar behavior compared to AES cipher. It is fair to infer that CLEFIA, SPARX, and Speck show equivalent resistance to cryptanalysis using FA, PSO, and GA. On the other hand, DES shows to be more prone to cryptanalysis using the metaheuristic approach. This could be explained by its shorter key size (56 bits) compared to the other ciphers (128 bits). It means that the same metaheuristic algorithm generates keys with higher average fitness if the cipher is weaker.

6 Conclusion and Future Directions

From the trends shown in the results, it is observed that the selected lightweight ciphers behave in a similar manner when subjected to cryptanalysis using metaheuristics. In other words, CLEFIA, SPARX, and Speck yield approximately the

same average fitness of generated keys as AES. It can be deduced that CLEFIA, SPARX, and Speck are as strong as AES despite the fact that they are built on simpler and faster components than classical AES. The higher average fitness of keys obtained during the cryptanalysis of DES using the same metaheuristic algorithms confirms that it is weaker than the other ciphers under study. Furthermore, though lightweight cryptography was meant for constrained environments, it appears that they are strong enough to be used in other environments and applications.

Evaluating the strength of a cryptographic algorithm is a complex task. Using metaheuristics for cryptanalysis in a comparative manner may help to establish the relative strengths among ciphers. In future, metaheuristics can be exercised on other lightweight ciphers in order to make more conclusive deductions.

However, it is paramount to propose a more sophisticated and accurate fitness function for the cryptanalysis problem using metaheuristics. Furthermore, the use of other metaheuristic algorithms, like cuckoo search algorithm and ant colony algorithm, may be explored for the problem of cryptanalysis of LWC.

References

1. Biryukov, A., Perrin, L.: State of the Art in Lightweight Symmetric Cryptography. 1–40 (2017). <https://hdl.handle.net/10993/31319>
2. Biryukov, A., Perrin, L.: Lightweight Cryptography Lounge. CryptoLUX. (2015)
3. Bogdanov, A., Mendel, F., Regazzoni, F., Rijmen, V., Tischhauser, E.: ALE: AES-based lightweight authenticated encryption. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 8424 LNCS, 447–466 (2014). https://doi.org/10.1007/978-3-662-43933-3_23
4. Bogdanov, A., Knudsen, L.R., Leander, G., Paar, C., Poschmann, A.: PRESENT : An Ultra-Lightweight Block Cipher. 450–466 (2007)
5. Gong, Z., Nikova, S., Law, Y.W.: KLEIN : A New Family of Lightweight Block Ciphers
6. Lim, C.H., Korkishko, T.: MCrypton—a lightweight block cipher for security of low-cost RFID tags and sensors. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 3786 LNCS, 243–258 (2005). https://doi.org/10.1007/11604938_19
7. Guo, J., Peyrin, T., Poschmann, A., Robshaw, M.: The LED block cipher. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 6917 LNCS, 326–341 (2011). https://doi.org/10.1007/978-3-642-23951-9_22
8. Gérard, B., Grosso, V., Naya-Plasencia, M., Standaert, F.X.: Block ciphers that are easier to mask: How far can we go? *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 8086 LNCS, 383–399 (2013). <https://doi.org/10.1007/978-3-642-40349-1-22>
9. Grosso, V., Leurent, G., Standaert, F.X., Varici, K.: LS-designs: Bitslice encryption for efficient masked software implementations. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 8540, 18–37 (2015). https://doi.org/10.1007/978-3-662-46706-0_2
10. Daemen, J., Peeters, M., Assche, G. Van, Rijmen, V.: The NOEKEON Block Cipher. *First Open NESSIE Work*. 1–30 (2000)
11. Albrecht, M.R., Driessen, B., Kavun, E.B., Leander, G., Paar, C., Yalçin, T.: Block ciphers—focus on the linear layer (feat. PRIDE). *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 8616 LNCS, 57–76 (2014). https://doi.org/10.1007/978-3-662-44371-2_4

12. Zhang, W., Bao, Z., Lin, D., Rijmen, V., Yang, B., Verbauwhede, I.: RECTANGLE: A bit-slice lightweight block cipher suitable for multiple platforms. *Sci. China Inf. Sci.* **58**, 1–15 (2015). <https://doi.org/10.1007/s11432-015-5459-7>
13. Dinu, D., Perrin, L., Udovenko, A., Velichkov, V., Großschädl, J., Biryukov, A.: SPARX : A Family of ARX-based Lightweight Block Ciphers with Provable Bounds. (2016)
14. Hong, D., Sung, J., Hong, S., Lim, J., Lee, S., Koo, B., Lee, C., Chang, D., Lee, J., Jeong, K., Kim, H., Kim, J., Chee, S.: HIGHT cipher.pdf. 46–59
15. Hong, D., Lee, J.K., Kim, D.C., Kwon, D., Ryu, K.H., Lee, D.G.: LEA: A 128-bit block cipher for fast encryption on common processors. *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 8267 LNCS, 3–27 (2014). https://doi.org/10.1007/978-3-319-05149-9_1
16. Beaulieu, R., Shors, D., Smith, J., Treatman-Clark, S., Weeks, B., Wingers, L.: The SIMON and SPECK lightweight block ciphers. In: *Proceedings of the 52nd Annual Design Automation Conference 2015*. 1–6 (2015). <https://doi.org/10.1145/2744769.2747946>
17. Beaulieu, R., Shors, D., Smith, J., Treatman-Clark, S., Weeks, B., Wingers, L.: The SIMON and SPECK lightweight block ciphers. *Proc. 52nd Annu. Des. Autom. Conf. - DAC '15*. 1–6 (2015). <https://doi.org/10.1145/2744769.2747946>
18. Yang, G., Zhu, B., Suder, V., Aagaard, M.D., Gong, G.: The Simeck Family of Lightweight Block Ciphers. *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). **9293**, 307–329 (2015). <https://doi.org/10.1007/978-3-662-48324-4>
19. Leander, G., Paar, C., Poschmann, A., Schramm, K.: New Lightweight DES Variants. *Fast Softw. Encryption*. 196–210 (2007). https://doi.org/10.1007/978-3-540-74619-5_13
20. Poschmann, A., Ling, S., Wang, H.: 256 Bit standardized crypto for 650 GE - GOST revisited. *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6225 LNCS, 219–233 (2010). https://doi.org/10.1007/978-3-642-15031-9_15
21. Matsui, M.: New block encryption algorithm MISTY. *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 1267, 54–68 (1997). <https://doi.org/10.1007/bfb0052334>
22. Wu, W., Zhang, L.: LBlock: A lightweight block cipher. *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6715 LNCS, 327–344 (2011). https://doi.org/10.1007/978-3-642-21554-4_19
23. Baysal, A., Şahin, S.: RoadRunner: A Small And Fast Bitslice Block Cipher For Low Cost 8-bit Processors. 1–20
24. Standaert, F.-X., Piret, G., Gershenfeld, N., Quisquater, J.: SEA : A scalable encryption algorithm for small embedded applications. pp. 222–236 (2006)
25. Shirai, T., Shibutani, K., Akishita, T., Moriai, S., Iwata, T.: The 128-bit blockcipher CLEFIA. *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 4593 LNCS, 181–195 (2007)
26. Suzaki, T., Minematsu, K., Morioka, S., Kobayashi, E.: Twine: a lightweight, versatile block cipher. *ECRYPT Work. pn Light. Cryptogr. LC11*. 146–169 (2011)
27. Shibutani, K., Isobe, T., Hiwatari, H., Mitsuda, A., Akishita, T., Shirai, T.: Piccolo: An ultra-lightweight blockcipher. *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6917 LNCS, 342–357 (2011). https://doi.org/10.1007/978-3-642-23951-9_23
28. Barker, E., Roginsky, A., Locke, G., Gallagher, P.: Transitioning the Use of Cryptographic Algorithms and Key Lengths. *NIST Spec. Publ.* 800–131A Revis. **2**, 17–18 (2019)
29. Mekhaznia, T., Zidani, A.: Swarm intelligence algorithms in cryptanalysis of simple Feistel ciphers. *Int. J. Inf. Commun. Technol.* **13**, 114 (2018). <https://doi.org/10.1504/ijict.2018.090436>
30. Amic, S., Mohabeer, H., Soyjaudah, K.M.S., Ramsawock, G.: Cryptanalysis of DES-16 using binary firefly algorithm.
31. Daemen, J., Rijmen, V.: *The Design of Rijndael*, vol. 255. New York (2002). <https://doi.org/10.1007/978-3-662-04722-4>

32. Shirai, T., Shibutani, K., Akishita, T., Moriai, S., Iwata, T.: The 128-Bit Blockcipher CLEFIA (Extended Abstract). 181–195 (2007). https://doi.org/10.1007/978-3-540-74619-5_12
33. Dinu, D., Perrin, L., Udovenko, A., Velichkov, V., Großschädl, J., Biryukov, A.: Design strategies for ARX with provable bounds: SPARX and LAX. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. 10031 LNCS, 484–513 (2016). https://doi.org/10.1007/978-3-662-53887-6_18
34. Ray, B., Douglas, S., Jason, S., Stefan, T.-C., Bryan, W., Louis, W., Beaulieu, R., Shors, D., Smith, J., Treatman-clark, S.: The simon and speck families of lightweight block ciphers. *Cryptol. EPrint Arch.* **2013**, 1–42 (2013)
35. Holland, J.H.: *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence* (1975)
36. Darwin, C.R.: *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (1859)
37. Yang, X.: *Firefly Algorithm* (2010). <https://doi.org/10.1016/B978-0-12-416743-8.00005-1>
38. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: *Proceedings of ICNN'95-International Conference on Neural Networks*, vol. 4, pp. 1942–1948. IEEE. <https://doi.org/10.1109/ICNN.1995.488968>

Product Classification in E-Commerce Sites



Anannya Patra, V. Vivek, B. R. Shambhavi, K. Sindhu, and S. Balaji

Abstract Given a predefined catalog hierarchy where all the categories are defined, the task by which the catalog path of every product is automatically predicted is known as product classification. This paper explores some of the methods of product classification. With the help of supervised learning, we have presented various machine learning models to classify products into a set of known categories. With the information such as name and description of the product being provided, this model can accurately place it under a designated category. Machine learning models that have been deployed include decision trees, support vector machine (SVM), random forest, logistic regression, and Naïve Bayes, with logistic regression giving the highest accuracy of 91.55%. The proposed implementation has huge potential to be able to slowly but substantially increase the automation of the categorization of products.

Keywords E-commerce · Logistic regression · Machine learning (ML) · Supervised learning · Support vector machine (SVM)

A. Patra (✉) · V. Vivek · B. R. Shambhavi (✉) · K. Sindhu
Department of ISE, BMS College of Engineering, Bengaluru, India
e-mail: anannya.is17@bmsce.ac.in

B. R. Shambhavi
e-mail: shambhavibr.ise@bmsce.ac.in

V. Vivek
e-mail: vivek.is17@bmsce.ac.in

K. Sindhu
e-mail: ksindhu.ise@bmsce.ac.in

S. Balaji
Department of CSE, CIIRC, Jyothy Institute of Technology, Bengaluru, India
e-mail: drsbalaji@gmail.com

1 Introduction

In the past few years, the e-commerce industry has experienced an augmented rate. Due to the exponential development of transport and technology, all aspects of shopping have moved online. Naturally, this has increased the demand and production of goods made specifically for the online shopping portals. Large amount of data is being handled everyday by these e-commerce portals, and processing this data would be impossible for humans alone. Huge amount of small- and medium-sized businesses is flourishing and a lot of time is spent in classifying the enormous products into their appropriate categories. Customers browse through the product catalog to discover the products, and hence, accurate classification is a critical step for the e-commerce business. An improved user experience is facilitated by an efficient product classification model.

Classification is not a new concept in the world of machine learning. Classification has been performed with supervised and unsupervised data. However, classification is yet to be refined and used in the e-commerce industry [1]. Very few attempts have been made to classify products into their respective categories. The need for product classification is more than ever.

Some of the problems faced by various e-commerce sites are:

- Inability to classify huge numbers of products: When there are thousands of products and new data being included into the inventory every day, it is very common for the product to be classified in the wrong category.
- Ambiguity in classification: Most products tend to fall under more than one category. For example, an “Electric Razor” can be classified into the “Electronics” category as well as the “Body Care” categories.

The ability to classify the products into the correct category without human interference is still a big challenge. We propose a model which uses machine learning techniques to define the proper categories for the product based on the product description provided. The proposed model will be helpful for any business with a list of new products they aim to sell and aspire to have a system which would automatically classify their products based on the training data provided. Also, since our model takes into consideration the product description, the ambiguity in classification is decreased.

We were motivated to work on this problem keeping in mind the fact that this model could be really helpful to the small business enterprises.

This paper is organized as follows. Section 2 talks about the related work on product classification. Dataset used in the proposed work is discussed in Sect. 3. The initial methodology and modified approach are presented in Sects. 4 and 5. The results of the work are highlighted in Sect. 6 while Sect. 7 gives the concluding remarks.

2 Literature Survey

Kannan et. al. [2] worked on the topic of product classification using images and used the *confusion-driven probabilistic fusion (CDPF)* and *CDPF ++* algorithms. *CDPF ++* can be used for classification because it has to adapt to changing vocabulary without having to manually label the items every time the labels undergo some minuscule change. This helped in reducing manual labor. This research was done taking into account 17 categories related to computing, and the dataset consisted of 17,989 training instances and 10,026 testing instances. The small textual description and the product image together served as the input for the machine learning algorithm. *CDPF ++* was one of the most efficient classifiers with an accuracy of 86.8%.

Liu et al. [3] proposed *convolution neural networks* to address the problem of classifying products using images. Two datasets were used in this study. One was Caltech256 which consisted of 20 products. From this dataset for the purpose of training the model, 100 images of each product were chosen randomly and 50 images were chosen for testing. The second dataset consisted of 20 products, where each product had over 200 images. The data was augmented by increasing it fivefold. This was done by mirroring all the images both horizontally and vertically and all the images were rotated 90° and 180° increments. Finally, 20,000 product images were generated, out of which for training purpose 16,000 images were used and 4,000 product images were used for testing. As all the images were of different sizes, they were normalized into 256×256 pixels. They used 5 hidden layers and 100 epochs to achieve maximum accuracy without overfitting. The average accuracy achieved was 92.1%.

In yet another work by Cevahir et al. [4], the Rakuten product dataset was used, where 280 million products were processed. Products were assigned to 28,338 active categories. After removing all the duplicates by titles, only 170 million titles were left out of which, for training of the model, 90% of the products were randomly selected and the remaining 10% were selected to be the test data. Classification was first implemented on the first level and was able to predict 35 titles. For all of the two different data sources, i.e., titles and descriptions, *DBN* and *KNN* classifiers were used. The *probability distribution score of four different classifiers was averaged* and category predictions were made.

Yu Wenhui et al. [5] proposed the idea of merging test and training dataset to create word vectors. This would help in finding semantic similarity which can be further used to process new words in the dataset. This step would help in solving large-scale multi-class classification. The training dataset consists of 720,000 entries and a validation dataset is of size 80,000 with the *n*-parameter set to 2 and word window length set as 100.

Two models have been built, i.e., single label prediction where *Text-CNN*, *Fast-text*, *VDCNN*, *Text-RNN*, and *AbLSTM* algorithms have been used and *n multi-level label prediction* where hierarchical search tree and short path tree models have been constructed. *Fasttext* and *AbLSTM* gave the highest F1 score of 82%.

Another paper by Krishnan Abhinandan et al. [6] consisted of comparison between hierarchical models and flat models and also proposed an idea where structured attributes and their value have been used in an unstructured form along with convolutional format. The dataset used by them consisted of approximately 25 million labels, split into three subsets to generate train, validation, and test datasets. The models used were *multi-CNN* and *multi-LSTM*, where both performed equally well with the accuracy score of ~96.5% but, multi-CNN outdid by being more parallelizable hence, faster.

3 Data

The “Flipkart Dataset” available in Kaggle which lists about 20,000 products with various features was chosen for our study. The features of the initial dataset were product id, time stamp, product url, product name, product category, actual price, discounted price, product image, is_FK_Advantage_provided, product description, overall rating, brand, and product specification [7]. Several features were redundant and the dataset had to undergo thorough cleaning. The dataset was chosen because it was from one of India’s biggest online shopping portals.

3.1 Data Preprocessing

In supervised learning, the first step is dealing with a dataset. The first approach generally is to use “brute force” and use all the features provided. However, this proves to be a bad approach in tackling the problem. Therefore, feature engineering and selection are as important as the algorithm that is being applied on them [8].

Each product had a certain number of features of which all were not relevant to our models and just increased redundancy. Hence, the features which were pertinent to the problem were retained. The features chosen were product name, brand, and product description. The model would predict the product category as the result.

A restriction of machine learning (ML) models is that they cannot be applied directly to strings, so all the features had to be converted into a form where the algorithms can be successfully applied to the dataset. First step was to remove all the missing values because they lead to irregularity in the dataset and hence inaccurate model. Then, label encoding was applied to the brand column with an efficient algorithm to encode the categorical values into numeric values.

3.2 Splitting of the Existing Dataset into Training Data and Test Data

To train the machine learning model, we divided the given dataset into training and test data. The set of data whose observations form the experience that is used by the algorithm to learn is called a training dataset. On the other hand, the set of observations that is used to evaluate the performance of the model, on the basis of a chosen performance metric, is called test dataset. The dataset was finally split into:

- i. Training data: 70% of the entire dataset
- ii. Test data: 30% of the entire dataset.

4 Initial Methodology

After the preprocessing of the dataset and splitting of the entire dataset, into training data and test data, of 70% and 30%, respectively, a few machine learning algorithms, namely decision tree and support vector machine (SVM), were applied. It was found that SVM performed better than decision tree.

4.1 Decision Trees

Instances are sorted on the basis of features and classified in decision trees. In a decision tree, each node depicts a feature of the instance that is to be classified and each branch represents a value that can be assumed by the node. The classification starts from the root node, and the instances are tested and sorted based on the value of their features [11]. An online shopping portal contains many products. Thus, this makes the decision tree algorithm a bad way of tackling this problem. This was reflected in the result with an accuracy of 68% during testing.

4.2 Support Vector Machine

Support vector machine (SVM) is a linear model. A hyperplane or line is created and hence, all data are separated into classes. SVMs are based on structural risk minimization principle, from computational learning theory, which helps to find a hypotenuse h . This helps in guaranteeing the lowest true error [9]. The input vectors are mapped into high-dimensional space and the maximum margin supports are sparse, in other words, there are very few nonzero coefficients [10]. The learning generalization ability improves and SVMs also achieve minimization of confidence range in case of small samples. However, the sample size of the dataset considered was not small,

so it resulted in creation of a complex hyperplane and, hence, the result gave us a success rate of 72%.

5 Modified Approach

The reason for failure of using brand as a feature was that several brands had products spanning in multiple categories which caused many misclassified samples. On further research, the redundant nature of the product description and brand was observed; that is, the brand was almost always mentioned in the product description as well as the product title. To select the most appropriate feature, first the product description and the product title were used separately which gave good results. However, using both features together in the algorithm did not yield good accuracy. This was because of the misinformation and click bait nature of the product titles to attract customers. Hence, the product description was considered as the feature to train the model.

5.1 Preprocessing

The following steps were used to convert the product description into machine understandable data.

- **Tokenization:** The process of splitting a string into smaller strings or substrings is called tokenization. It was applied to the product description.
- **Normalization:** After tokenization, various words such as punctuation, stop words, and numbers were removed from the product description as they were unable to provide any useful data for the prediction of the product category [12, 13]. Stemming was used to convert tokenized words into more meaningful words.
- **TF-IDF vectorization:** The bag-of-words model is a common approach to extract features. The model takes the presence or frequency into consideration but the order of their occurrence can be ignored. Essentially, we calculate a measure called term frequency, inverse document frequency, for each term in a dataset. Term frequency, inverse document frequency is abbreviated as TF-IDF. It determines the frequency of a word in a document compared to the inverse proportion of that particular word over the entire document corpus. Basically, this calculation helps in determining the relevance of a particular word compared to the whole document.

In a single document, articles and prepositions have much lower TF-IDF number than certain words which are common. The encoding of TF-IDF is very simple and easy to use, and therefore, it forms a basis for more complicated algorithms [14]. For example, if “track pants” is mentioned 20 times in a document, it might be more important than if it was only mentioned once. The document frequency (the number of documents containing a given word) is used as a measure of how common the word

is. This minimizes the effect of domain-specific language that does not add much information, for example, words such as “Specifications” that might be present in most documents.

5.2 Algorithms Adopted

Logistic regression: A linear classifier is mostly similar to traditional linear regression but fits the output of the logistic function. Logistic regression is a binary classifier. For multi-class classification such as our problem in hand, there are two strategies:

- (a) One-versus-all (OvA): This method trains as many binary classifiers as there are classes.
- (b) One-versus-one (OvO): In this method, binary classifier is trained for every pair of classes. If there exists N classes, it is necessary to train $N*(N-1)/2$ classifiers.

In our model, we have considered logistic regression with the OvO strategy, that is, multinomial logistic regression.

Multinomial Naïve Bayes: This algorithm is a family of probabilistic algorithms. Given a class, the conditional probability of a particular word is estimated, because the relative frequency of term t in documents belonging to class c .

Conditional probability: The probability of one event occurring with some relationship to one or more events.

$$P(Z|Y) = P(Y \cap Z)/P(Y) \tag{1}$$

where Y and Z are two events.

Usually, in the classification problem, when multiple occurrences of the words matter, multinomial Naïve Bayes is used.

Random forest: Random forest consists of a huge number of decision trees and together they form an ensemble and every decision tree is trained on a random subset of the input features. The reason a random forest classifier works well is because of the principle it follows—a large number of relatively uncorrelated models outperforms the individual constituent trees.

6 Results

Table 1 gives the comparative results of the different algorithms experimented in this work.

Considering the results in Table 1, we notice that logistic regression gives us the best accuracy of 91.58%. Situations where large datasets are taken into consideration, machine learning algorithms like support vector machine (SVM) classifiers do not really scale well and fail to give the desired result. Logistic regression with the OvO

Table 1 Comparative results

Model	Accuracy (%)
Logistic regression	91.55
Naive Bayes	84.81
Support vector machine (SVM)	72
Decision tree classifier	68
Random forest classifier	50.84

strategy does better in these situations as it is much faster to train a lot of classifiers on a small dataset instead of training just one classifier on an extensive dataset [15]. Secondly, when the data in consideration is non-numeric data, then logistic regression is better than Naive Bayes.

Also, our dataset has low dimensions and has skewed data and hence logistic regression has better performance.

From Fig. 1, it is visible that there is an imbalance in the number of product descriptions per class label. The product descriptions are more biased toward “Jewellery” and “Clothing.” In some cases, like diabetes detection or spam message detection, it

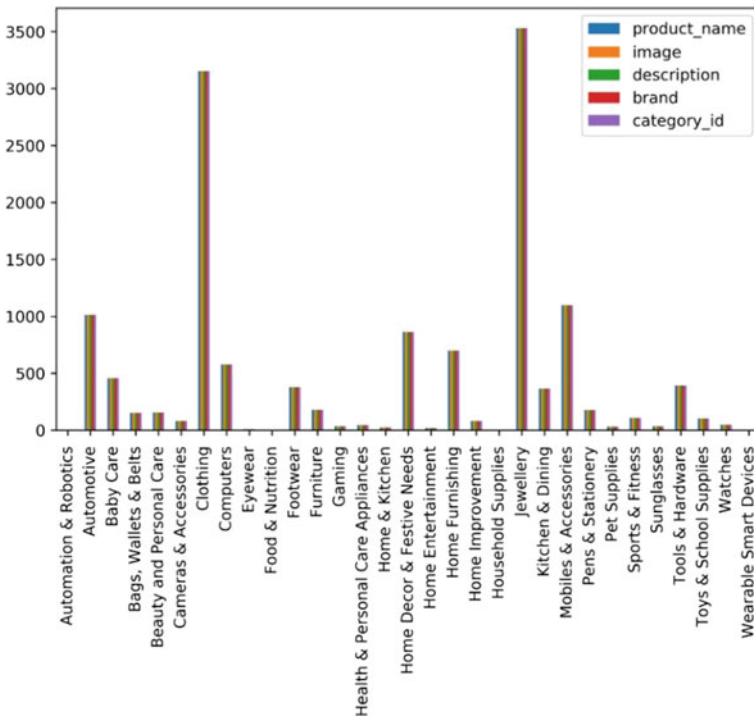


Fig. 1 Graph of class labels (x-axis) versus number of products in each class label (y-axis)

Fig. 2 Labels and the list of their *most related n-grams*

```
# ' Clothing ':  
  . Most correlated unigrams:  
    . shirt  
    . women  
    . bra  
  . Most correlated bigrams:  
    . branded clothes  
    . clothes flipkart  
    . india shop  
# ' Computers ':  
  . Most correlated unigrams:  
    . router  
    . laptop  
    . usb  
  . Most correlated bigrams:  
    . cell laptop  
    . laptop battery  
    . router rs
```

is extremely crucial to artificially balance the dataset, by either under sampling each class or oversampling each class.

But, in our problem, it is desirable to have reasonable accuracy for the minor classes and higher prediction accuracy for the majority class.

Figure 2 shows a small instance of the various labels present in our dataset and it tells us the terms that are most correlated with each of the output categories.

Figure 3 depicts an instance of the output window of our model. On giving the product description as “24 carat sparkling necklace,” the model correctly guesses the output label as “Jewellery.”

In Fig. 4, on entering the product description as “Red tshirt made out of pure cotton,” the model correctly predicts the category as “Clothing.”

There are still some shortcomings though:

1. Due to our limited dataset, and the imbalance in classes and number of products in our dataset, there are some cases where the output is not predicted correctly.
2. The accuracy can be even further increased if we include image as an attribute to classify the products.

7 Conclusion

The product classification is a very crucial field for today’s e-commerce business and has great potential for improvement. In this work, we have presented a classification method using product description. The proposed model has been trained with a dataset consisting of thousands of products and thirty-one classification categories. Different algorithms were used and the final accuracies are calculated out of which, the model which gave the best accuracy is logistic regression. If the dataset is augmented with product images and the model learns from state-of-the-art image processing algorithms, the system is expected to perform even better.

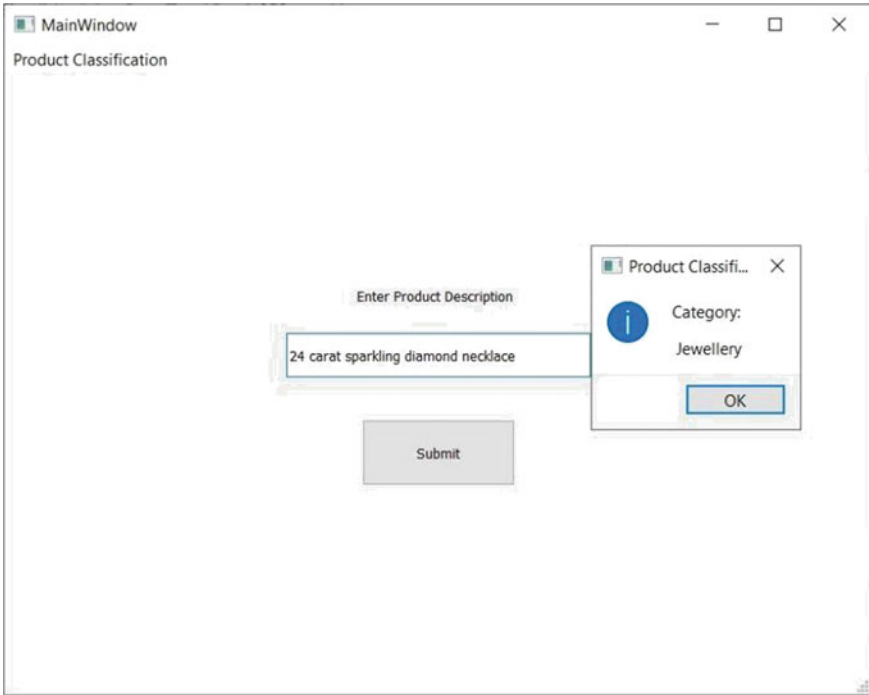


Fig. 3 Output instance 1

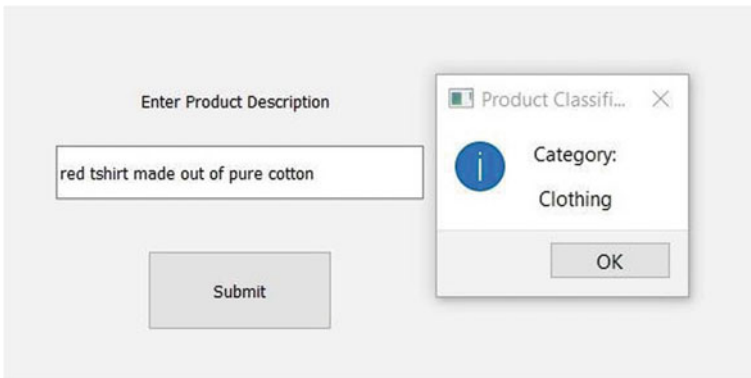


Fig. 4 Output instance 2

References

1. Pérez-Ortiz, M., S Jiménez-, Fernández, Gutiérrez, P.A., Alexandre, E., C , Hervás-Martínez, S Salcedo-, Sanz: A review of classification problems and algorithms in renewable energy applications. *Energies* **9**(8), 607 (2016)
2. Kannan, A., Talukdar, P.P., Rasiwasia, N., Ke, Q.: Improving product classification using images. In: 2011 IEEE 11th International Conference on Data Mining, pp. 310–319. IEEE (2011)
3. Liu, T., Wang, R., Chen, J., Han, S., Yang, J.: Fine-Grained Classification of Product Images Based on Convolutional Neural Networks. *Adv. Mol. Imaging* **8**(4), 69–87 (2018)
4. Cevahir, A., Murakami, K.: Large-scale multi-class and hierarchical product categorization for an E-commerce giant. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 525–535 (2016)
5. Yu, W., et al.: Multi-level deep learning based e-commerce product categorization. *eCOM@SIGIR* (2018)
6. Krishnan, A., Amarthaluri, A.: Large scale product categorization using structured and unstructured attributes. arXiv preprint [arXiv:1903.04254](https://arxiv.org/abs/1903.04254) (2019)
7. <https://www.kaggle.com/PromptCloudHQ/flipkart-products>
8. Mishra, V.K.: Comparative analysis: effective information retrieval using different learning approach.
9. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: European Conference on Machine Learning, pp. 137–142. Springer, Berlin, Heidelberg (1998)
10. Zhang, H., Zijun, S.: Product classification based on SVM and PHOG Descriptor. *Int. J. Comput. Sci. Netw. Secur. (IJCSNS)* **13**(9), 1 (2013)
11. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **160**, 3–24 (2007)
12. <https://techblog.commercetools.com/boosting-product-categorization-with-machine-learning-ad4dbd30b0e8>
13. Shankar, S., Lin, I.: Applying machine learning to product categorization. Stanford University, Department of Computer Science (2011)
14. Ramos, J.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the 1st Instructional Conference on Machine Learning, vol. 242, pp. 133–142 (2003)
15. <https://machinelearning-blog.com/2018/04/23/logistic-regression-101/>

Real-Time Detection of Inter-Frame Video Forgeries in Surveillance Videos



Kshitij Saluja and Naveen Aggarwal

Abstract Detection of video forgery is a critical requirement to ensure integrity of video data and with an increase in the use of cameras and dependability of surveillance systems, it becomes more critical. In this paper, video forgery detection is proposed based on the concept of exponential weighted moving average of optical flow variation factors. This mechanism relies on the fact that optical flow variation factor is almost continuous in an original video. However, discontinuity points in the optical flow variation factor are introduced in forged video which are not visible with the naked eye. The proposed mechanism detects inter-frame forgery process, i.e., frame deletion, insertion, and duplication in surveillance videos. The proposed detection mechanism is deployed on cloud server where a live video stream is received from different ATM vestibules. Upon detection, an alert message is sent to the concerned bank for immediate action. Experiments are performed on the videos captured from ATM vestibules of a bank. The accuracy and latency of proposed mechanism manifest pragmatism of the proposed idea.

Keywords Video forgery · Optical flow · Inter-frame forgery · Exponential weighted moving average

1 Introduction

The easy availability of low-cost cameras and their deployment in surveillance systems has led the generation of the huge amount of audio, video, and text, i.e., multimedia content. Further usage of online video conferencing platforms, YouTube videos, educational videos, and social networking platforms like Facebook, Instagram, etc., has given rise to manifold increase in especially video data. Videos can be generated, stored, and disseminated in digital format easily. Further, these videos can be edited easily with user-friendly editing tools. Even a novice user can do it with quite ease. Therefore, compromising authentication of a digital video for malign

K. Saluja (✉) · N. Aggarwal
Department of CSE, UIET, Panjab University, Chandigarh, India
e-mail: kshitijsaluja2018@gmail.com

purposes is not difficult. However, the authenticity of these digital videos is very critical as it is used as testimony in court of law.

Video forgery is a technique of modifying videos by adding or removing some content purposefully [1]. Thus, it questions the authenticity of digital videos. Video forgery detection is investigation of video data to check its authenticity [2]. It can be classified into two broad categories [3], viz. Active video forgery detection and passive video forgery detection. Active video forgery detection is based on watermark and digital signatures which are helpful to authentic content ownership and copyright violations. Since watermarks and signatures need to be embedded during the capturing phase, there is need of specially equipped digital cameras. Since these kinds of digital cameras are not widely used, it restricts the usage of this technique. Passive forgery detection techniques work in contrast to that of the active approach as no dedicated hardware is required to get information about videos. This approach is based on the fact that original videos have some consistent intrinsic characteristics or patterns which get disturbed in a forged video. These patterns are studied and analyzed by passive forgery detection techniques to detect video forgeries [4].

Passive detection methods [5] are used to detect any tampering performed either at intra-frame level or at an inter-frame level. Forgery performed at an inter-frame level includes simple deletion [6], insertion [7], or duplication [8] of frames in a video. Intra-frame forgery works at pixel level [9, 10] or at the entire frame level [11].

Inter-frame forgery can be detected using various techniques. One of the rationale used in detection of inter-frame forgeries is presence of double compression in manipulated videos. The video is normally compressed every time it is saved and whenever it is reconstructed it results in double compression [12, 13]. So, analysis of double MPEG compression has been utilized by some authors to detect inter-frame video forgeries. Double compression in MPEG-4 encoded videos was firstly detected in [14] by utilizing Markov statistics, which was dependent on quantization scale values of reconstructed video. These approaches result in a lot of false positives as legitimate videos also show signs of compression when these are uploaded, downloaded, and transmitted [3] so double compression cannot be considered as sign of inter-frame forgery.

Wang and Farid in [12] have analyzed periodicity in DCT coefficients of I-frames and prediction error of P-frames. However, empirical study is not provided. Some researchers examined Benford's law violation in quantized DCT coefficients by utilizing 36-D, 12-D, and 63-D feature vectors in [3, 13, 14], respectively. The accuracy of these approaches in comprehensive forgery scenarios is not up to the mark.

So, some other artifacts need to be tested to detect inter-frame forgeries. The authors in [3, 14, 15] detected frame tampering using optical flow consistency measure, but they validated their metric on a limited dataset of MPEG-2 encoded videos. A systematic threshold finalization criteria based on the anomaly score have been used in [15]. But, the anomaly score is computed on the basis of the mean and standard deviation values of the optical flow variation factor of all frames in a video. It makes the approach prone to false positives and forces to work in an offline manner.

Moreover, experimentally it is also found that the basic assumption of variation factor following Gaussian distribution in original videos was also wrong.

This work deeply investigates video forgery detection based on the concept of optical flow variation factor. The expected value of variation factor is computed using exponential weighted moving average EWMA [16]. The proposed approach relies on the fact that envelop of the optical flow sequence in an original video is approximately continuous. The corresponding optical flow variation sequence only fluctuates in a small range. However, the forgery processes will introduce discontinuity points in the optical flow sequence, even though there are no visible traces of tampering.

This paper is organized as follows: Sect. 2 defines the concepts of optical flow and optical flow variation factor as video forgery detection metrics. Section 3 describes the methodology. Section 4 discusses the experiment scenarios as well as results and finally Sect. 5 concludes the paper.

2 Optical Flow Variation Factor

The video is a stream of successive frames. Optical flow is a measure of relative transition in which each pixel experiences when we analyze successive frames of a video. As per [15], “Optical flow is the distribution of apparent velocities of movement of brightness patterns. Each pixel of the current frame is assigned a two-component velocity vector indicating the position of the same pixel in the reference frame.”

The proposed study is based on the fact that optical flow of successive frames has very less differences in case of legitimate videos whereas in case of forged videos it differs significantly at the place where inter-frame forgery has been performed.

For frame K , the optical flow velocity (u, v) at location (x, y) can be obtained. Here, we only consider the sum of the magnitudes of optical flow velocities of all pixels at (x, y) locations in K th frame while transforming to $(K + 1)$ th frame:

$$OF(K) = \sum_x \sum_y \sqrt{u^2 + v^2} \quad (1)$$

The optical flow $OF(K)$ alone cannot be used for detection of forgeries as in original videos there can be legitimate variations at various moments which results in change of $OF(K)$, resulting in false positives. Therefore, persistent differences in optical flow measured using optical flow variation factor promises to be more useful metric for detection of forgeries with minimum false positives.

Optical flow variation factor α helps to reveal the relative changes in the optical flow sequence. Here, K defines the frame number.

$$\alpha(K) = \frac{2 * OF(K)}{OF(K - 1) + OF(K + 1)} \quad (2)$$

3 Methodology

The value of $\alpha(K)$ varies in a narrow range for original videos. However, $\alpha(K)$ deviates appreciably for those values of K where inter-frame forgeries are present. In order to capture this variation, we propose to use exponential weighted moving average (EWMA) of optical flow variation factor β of each frame of any video to be tested. Let $\beta(K)$ is EWMA of K th frame and a is tunable parameter, the model to compute $\beta(K)$ can be given as below:

$$\alpha\beta(K) = a * \alpha(K) + (a - 1) * \beta(K - 1) \tag{3}$$

The value of $\beta(k)$ varies in a narrow range for original parts of video whereas tampered part of video will have unexpected value of $\beta(k)$. The error $\varepsilon(K)$ is difference of expected and actual value, i.e., $\varepsilon(K) = \beta(K) - \alpha(K)$. The value of error, i.e., $\varepsilon(K)$ determines whether the current frame is forged or not. The following figure explains the proposed methodology (Fig. 1).

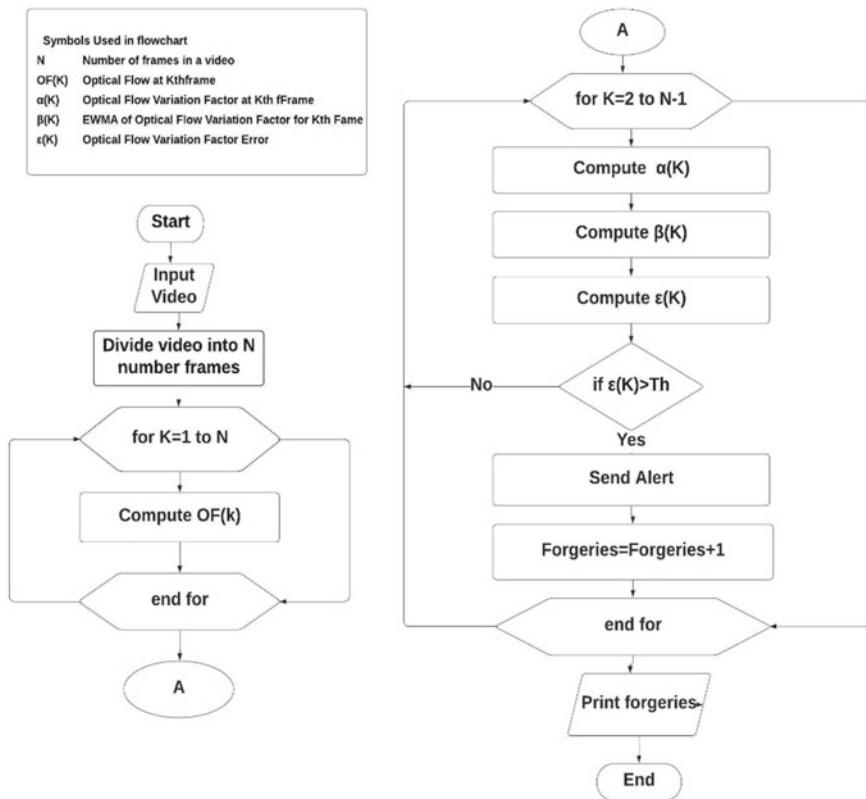


Fig. 1 Methodology for forgery detection

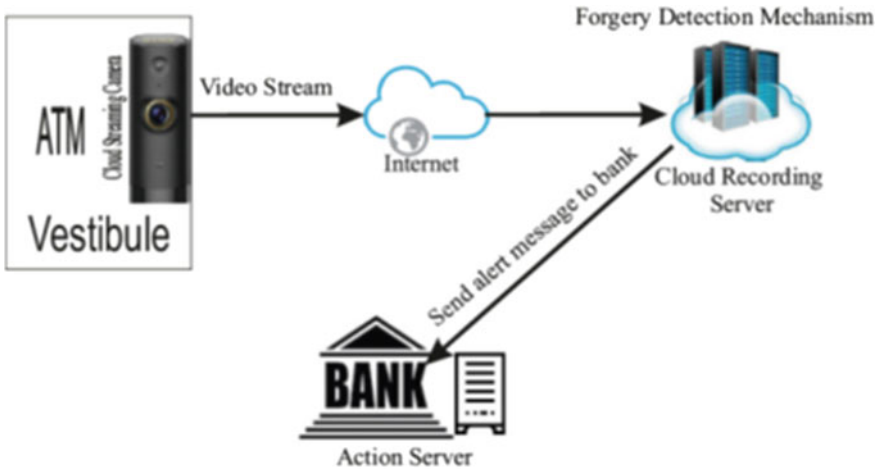


Fig. 2 Model of proposed forgery detection system

In this work, surveillance videos captured by static cameras are considered, so there will not be change in background schemes impacting optical flow sequence in original videos. Our main focus is on detection of forgery introduced by insertion, deletion, and duplication of frames in original videos.

First, original videos are converted into individual frames. Optical flow $O(K)$, optical flow variation factor $\alpha(K)$, and error $\varepsilon(K)$ are computed for different frames. The count of number of times $\varepsilon(K)$ crosses threshold yields number of forgeries in the video. The main advantage of this approach is detection of forgery in live video instantly without delays, whereas researchers in [15] computed mean and standard deviation of the complete video, afterward based on anomaly score forgeries were detected. The model of the proposed forgery detection system is depicted in Fig. 2.

Clearly, as per Fig. 2, live videos from the ATM vestibule are captured on cloud server and subsequently analyzed for forgery detection. In case, cloud streaming camera is operated maliciously by a masquerader in order to conceal facts or tamper videos to avoid detection, then on the fly proposed forgery detection mechanism can send alerts to concerned officials of the bank.

4 Experiment Scenarios and Results

In this work, surveillance video datasets of ATMs of a bank are used with due permissions of the authority. Sometimes customers claim that they have not visited the ATM and repudiate their withdrawals which were actually committed by them. In these cases, banks have to prove that they have not added the frames of the customers maliciously. Moreover, planned thefts of cash in ATMs can be avoided as deletion of some frames can be easily detected using the proposed approach. Finally, malevolent

intent of corrupt bank officials can be detected if there will be any trial of duplicating some frames in order to conceal facts related to deleted part of the videos. Figure 2a, b depicts the optical flow and variation factor for normal video (without forgery), whereas Fig. 3a, b shows same for the forged video.

4.1 Frame Deletion

Frame deletion is to delete some frames of the video. Clearly a real-life deletion forgery shall involve deletion of multiple frames, otherwise glimpses of malicious activity shall always be visible without in depth analysis. We deleted 10–30 frames in multiple videos for testing purposes.

4.2 Frame Insertion

Frame insertion is to insert one or more frames in the video. Frame insertion introduces two discontinuity points, i.e., at the beginning and ending of frame insertion. There are lots of free access tools available on the Internet for performing this type of forgery in digital videos.

4.3 Frame Duplication

Frame duplication is a scenario in which single frame is duplicated to show a stationary scene. Either the last or one of the previous frame is duplicated multiple times to maintain the size of the file.

4.4 Results

The optical flow and variation factors of original video are depicted in Figs. 3a, b. Clearly, optical flow values of frames 120–135 show large variation which can create false alarms. However, since the variation factor considers optical value of previous and next frames also, therefore, temporary variations have been neutralized and variation factor lies in narrow range from 0.5 to 1.5. However, for deletion forgery as depicted in Fig. 4a, b, supremacy of optical flow variation factor is quite obvious. But detection of deletion forgery on the fly requires computation of EWMA of variation factor. The expected values of variation factor lies in a narrow range as shown in Fig. 4b. Clearly, the difference between expected and actual values of

variation factor called variation factor error has the quantifiable capability to detect forgeries in real time with minimum false positives.

The graphs shown in Fig. 5a, b depict peaks at two places. The reason behind is that whenever a short video is inserted in the original video, variation factors get altered at the beginning as well as at the end. Finally, duplication of frames

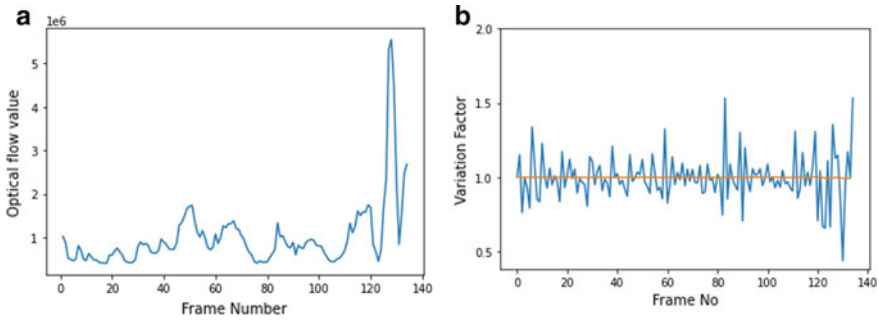


Fig. 3 a Optical flow of normal video. b Variation factor of normal video

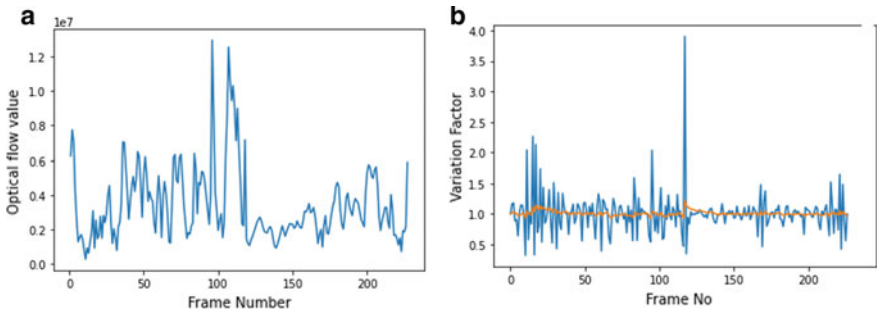


Fig. 4 a Optical flow for deletion forgery. b Variation factor for deletion forgery

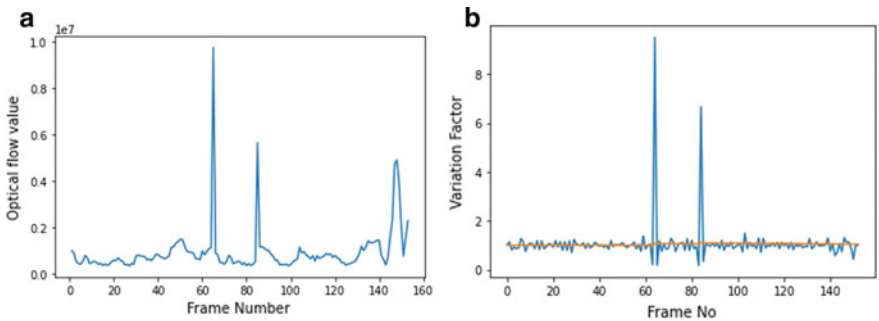


Fig. 5 a Optical flow for frame insertion forgery. b Variation factor for frame insertion forgery

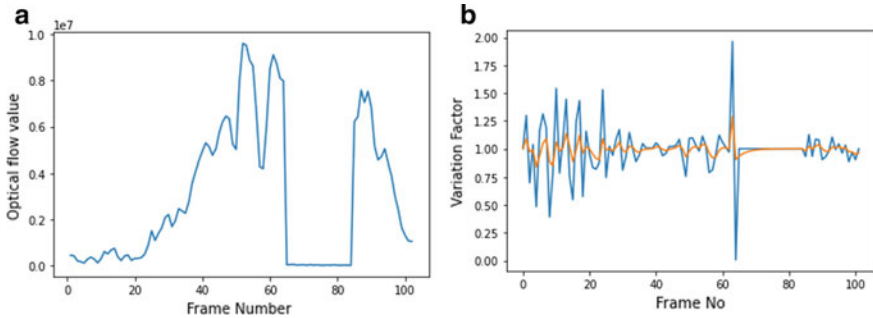


Fig. 6 **a** Optical flow for frame duplication forgery. **b** Variation factor for frame duplication forgery

results in a constant value of variation factor and hence very less value of error as depicted in Fig. 6a, b. The computation of error for each frame depends upon optical flow values of three neighboring frames. Hence, the detection of forgery can be done with minimum latency, accuracy, and in real time.

5 Conclusion

Surveillance videos have the same background. Therefore, successive frames in the video mainly depict movement of front objects. The forecasted value of optical flow variation factor using EWMA assists in finding the difference between expected and actual frame appearing in a video on the fly. In this proposed study, it has been found that optical flow variation in combination with EWMA is able to detect forgeries with accuracy and low latency of the proposed mechanism makes it pragmatic to be deployed in real time. In future, we are interested to explore more statistical metrics in combination with frame characteristics to detect different types of forgeries and other activities of interest.

References

1. Kingra, S., Aggarwal, N., Singh, R.D.: Video inter-frame forgery detection approach for surveillance and mobile recorded videos. *Int. J. Electr. Comput. Eng.* **7**(2), 2088–8708 (2017)
2. Farid, H.: Exposing digital forgeries in scientific images. In: *Proceedings of the 8th Workshop on Multimedia and Security (2006)* 29–36
3. Jin, H.: Research of blind forensics algorithm on digital image tampering. *Indonesian J. Electr. Eng. Comput. Sci.* **12**(7), 5399–5407 (2014)
4. Mizher, M.A., Ang, M.C., Mazhar, A.A., Mizher, M.A.: A review of video falsifying techniques and video forgery detection techniques. *Int. J. Electron. Secur. Digit. Forensics* **9**(3), 191–208 (2017)

5. Chao, J., Jiang, X., Sun, T.: A novel video inter-frame forgery model detection scheme based on optical flow consistency. In: International Workshop on Digital Watermarking, pp. 267–281. Springer, Cham (2012)
6. Shanableh, T.: Detection of frame deletion for digital video forensics. *Digit. Invest.* **10**(4), 350–360 (2013)
7. Zheng, L., Sun, T., Shi, Y.Q.: October. Inter-frame video forgery detection based on block-wise brightness variance descriptor. In: International Workshop on Digital Watermarking, pp. 18–30. Springer (2014)
8. Wang, W., Farid, H.: Exposing digital forgeries in video by detecting duplication. In: Proceedings of the 9th workshop on Multimedia and security, pp. 35–42 (2007)
9. Kobayashi, M., Okabe, T., Sato, Y.: Detecting video forgeries based on noise characteristics. In: Pacific-Rim Symposium on Image and Video Technology, pp. 306–317. Springer, Berlin, Heidelberg (2009)
10. Lin, C.S., Tsay, J.J.: A passive approach for effective detection and localization of region-level video forgery with spatio-temporal coherence analysis. *Digit. Invest.* **11**(2), 120–140 (2014)
11. Hyun, D.K., Ryu, S.J., Lee, H.Y., Lee, H.K.: Detection of upscale-crop and partial manipulation in surveillance video based on sensor pattern noise. *Sensors* **13**(9), 12605–12631 (2013)
12. Wang, W., Farid, H.: Exposing digital forgeries in video by detecting double MPEG compression. In: Proceedings of the 8th workshop on Multimedia and security, pp. 37–47 (2006)
13. Milani, S., Bestagini, P., Tagliasacchi, M., Tubaro, S.: Multiple compression detection for video sequences. In: 14th International Workshop on Multimedia Signal Processing (MMSP), pp. 112–117. IEEE (2012)
14. Jiang, X., Wang, W., Sun, T., Shi, Y.Q., Wang, S.: Detection of double compression in MPEG-4 videos based on Markov statistics. *IEEE Signal Process. Lett.* **20**(5), 447–450 (2013)
15. Wang, W., Jiang, X., Wang, S., Wan, M., Sun, T.: Identifying video forgery process using optical flow. In: International Workshop on Digital Watermarking, pp. 244–257. Springer, Berlin, Heidelberg (2013)
16. Raza, H., Prasad, G., Li, Y.: EWMA model based shift-detection methods for detecting covariate shifts in non-stationary environments. *Pattern Recogn.* **48**(3), 659–669 (2015)

An IoT-Based System Architecture for Environmental Monitoring



Binod Kumar Pattanayak, Deojeet Nohur, Sanjeev K. Cowlessur, and Rajani Kanta Mohanty

Abstract Internet of Things (IoT) has grown in popularity in all spheres of life. Environmental issues in the modern era present a major concern for researchers, scientists and government. In combating the environmental issues, monitoring, identifying and quantifying different contaminating parameters in the environment are necessary, and for this purpose, IoT technology can be extremely useful. In this paper, we propose an IoT-based environmental monitoring system that operates on a four-layer architecture wherein environmental data are captured by a sensor network, aggregated and then communicated to a web server via a cloud service. Then, these data are analyzed by various analytical IoT specific tools and then delivered to the applications for further processing. This system can be cost effective and easy to deploy considering that the underlying architecture is simpler than other existing similar systems.

Keywords IoT · Environment · Security · Sensor network

B. K. Pattanayak

Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India
e-mail: binodpattanayak@soa.ac.in

D. Nohur · S. K. Cowlessur

Department of Software Engineering, Faculty of Information and Communication Technology, Université Des Mascareignes, Beau Bassin-Rose Hill, Mauritius
e-mail: dnohur@udm.ac.mu

S. K. Cowlessur

e-mail: scowlessur@udm.ac.mu

R. K. Mohanty (✉)

Department of CSE, PSIT, Kanpur, Uttar Pradesh, India
e-mail: rkm.bbs@gmail.com

1 Introduction

Innovations in global Internet as a mass communication medium has resulted in Internet of Things (IoT) that has grown in popularity and has acquired almost all spheres of human life. As a complement to the traditional Internet, IoT has made it possible for various devices to communicate among themselves independent of any human intervention [1, 2]. As experts reveal, the number of devices to be connected to the global Internet by the year 2025 may rise up to a few billions [3]. It raises an apprehension over the security of the devices connected thereby [4]. Application of IoT has captured various fields such as healthcare, education, agriculture, environmental monitoring. [5, 6].

In this research article, a focus has been given on management of environmental monitoring using IoT technology. Issues related to environment have explored major concerns for government agencies, scientists, researchers, industries and various other enterprises in human society. Contamination in air resulting from high level of pollution in the atmosphere causes problems in breathing for humans, thereby leading to various critical health-related issues. Similarly, contamination in water has become another major concern that is driven by industrial waste products leading to various diseases even very critical ones. Hence, there arises a requirement for innovating methods for monitoring the level of pollution in the atmospheric air and the level of contamination in drinking water so as to facilitate necessary measures overcome these issues and protect people from health-related problems. Such measures require the appropriate devices to be deployed for recording the data from the environment and transfer it to a base station from where these data can be captured for analysis and finding solutions in order to remove the exacerbating properties of the environment. Different types of sensors can be useful for this purpose. In addition, various web applications are necessary for performing the necessary analytics of the captured data. In this article, we have proposed a four-layer IoT architecture for environmental monitoring wherein at the perception layer various devices such as sensors are proposed to record relevant to the environmental data for analysis. Then at network layer, the captured at perception layer data are communicated using the Internet services such as cloud computing. The middleware is responsible for storage and management of the retrieved data using various storage technologies such as relational database management system (RDBMS) along with different tools and software packages. Finally, the application layer performs the required processing, analysis and sharing of the analyzed data to the relevant services.

The rest of the paper is organized as follows. Section 2 covers an overview of IoT architecture. IoT security issues and challenges are discussed in Sect. 3. Related work from various sources on IoT in environment is detailed in Sect. 4. The proposed by us IoT architecture is included in Sect. 5, and Sect. 6 concludes the paper with probable extensions to the existing work.

2 IoT Architecture

In IoT, “things” are referred to as the devices that can be connected to the global Internet and that can communicate among themselves. IoT encompasses millions of small resource constrained radio frequency-driven smart devices that are capable of communicating independent of human intervention. IoT architecture supports a four-layer protocol stack [1, 3]. The perception layer incorporates the hardware components such as sensors for capturing data and actuators for executing necessary actions. The network layer is responsible for aggregating the captured by sensors data and communicating them to a cloud service which then can be stored for further analysis. The middleware layer incorporates the necessary tools for analysis of the received data and then the application layer comprises of various applications for processing after which the necessary actions can be executed.

3 IoT Security Issues and Challenges

As elaborated earlier, IoT encompasses millions of small resource constrained radio frequency-operated mobile devices, ensuring secure communication among such devices presents a challenge for implementers. Various layer-wise issues and challenges pertaining to IoT security are detailed below [7, 8].

Perception Layer: Various security issues at the perception layer of IoT protocol stack are unauthorized access to the RFID tags of IoT devices, cloning of RFID tags, eaves dropping, spoofing, radio frequency (RF) channel jamming as well as physical damaging to the IoT devices.

Network Layer: The issues related to the network layer of IoT protocol stack are sybil attack, sinkhole attack and malicious attack on the IoT devices.

Middleware Layer: Middleware layer for the IoT protocol stack can be susceptible to security issues such as data integrity violation, malfunctioning and privacy issues.

4 Related Work on IoT for Environment

Resolving environmental issues presents a major challenge for mankind as deterioration of environmental conditions and contamination of atmospheric air often tends to lead to severe health-related issues in people in rural as well as urban areas. The most challenging factor here is the monitoring the environmental characteristics and identification as well as quantification of environmental variables such as level of dusts, level of humidity, level of contaminating gases, level of noise, etc. IoT devices

can be effectively used for accurate measurement of the above environmental parameters for further analysis of the captured data in order to facilitate the methods of resolution of these issues.

Various authors have reported their research findings on implementation of IoT for resolving environmental issues. Implementation of IoT technology in recording and analyzing environmental variables has been addressed by authors in [9]. Here, the authors have proposed a IoT-based architecture that are integrated with different types of sensors that are associated with assessing various environmental variables in a smart city environment with respect to which the status of environment in a smart city can be accurately assessed, and further necessary actions can be executed in order to ease the environment to make it healthier for the inhabitants of the smart city. Implementation of IoT technology along with customization of existing IoT applications for sustainable environmental management in the conditions of South Africa has been discussed in [10]. In this paper, the authors focus on the question how can the IoT technology be made capable to leave an impact on the sustainable environmental management in South Africa. An integrated information system (IIS)-based prototype for environmental monitoring and management has been devised by the authors in [11]. This prototype relies on IoT technology for data collection from the environment, various web services along with supported applications such as cloud computing for data analytics, and as claimed by the authors here, the effectiveness of the environmental monitoring processes as well as decision making using this prototype has been improved significantly. A IoT-based system for atmospheric monitoring making use of long-term evolution (LTE) mobile communication network that addresses the issues of cost effectiveness and ease of installation of the measuring equipment for the purpose of monitoring the environment has been proposed by the authors in [12]. Here, the authors have devised a prototype that measures the level of dust and ozone contents in the atmosphere and transmits this information to the LTE network along with the status as well as location of the measuring equipment that further facilitates the process of analysis of the received environmental data. A IoT-based environmental monitoring system on wireless sensor network (WSN) platform has been proposed by the authors in [13] that relies on a GSM module for communication. Sensors are used for data collection that are subsequently stored onto a WAMP server and then dispatched to the monitoring device. Finally, after the necessary analysis of captured data, the results are delivered to the end user. As claimed by the authors in their research article, the proposed system is supported by a set of advantages such as low cost, ease of deployment, secure communication and quality output. An IoT architecture for green campus design has been addressed by the author in [14, 15]. This proposed system mostly focuses on energy saving for the computers as well as the air conditioners in the laboratories within a campus. In the process of designing this system, the author suggests to convert different objects into smart objects and then to integrate these smart objects. The sensor network is used here with an intention to save energy. An environmental monitoring system for smart cities based on IoT technology has been proposed by the authors in [16, 17] for monitoring temperature, levels of humidity and CO₂ in the atmosphere. Here, the data are transmitted from the transmitter to the receiver that are recorded at the

receiver in an Excel sheet on a computer using a graphical user interface (GUI), made on LabVIEW. AN android application then transfers the data from LabVIEW to a smartphone that makes it possible for monitoring the data remotely. A WSN application for environmental monitoring using IoT has been designed that mainly relies on two applications [18]. One of the applications here uses EPS8266 for sending data to the Internet directly and the second application uses Arduino and Xbee 802.15.4 in a multi-hop wireless network for collecting and aggregating the data from various nodes in the WSN and sending to a gateway which are further sent to a webserver.

5 Proposed IoT-Based Environmental Monitoring System

The architecture of our proposed environmental monitoring system based on IoT technology is depicted in Fig. 1. It comprises of four protocol layers: perception layer, network layer, middleware layer and application layer. The perception layer encompasses various sensors for recording environmental data. Such data include atmospheric temperature, level of humidity in the air, level of contaminating chemical components such as CO₂ and level of ozone. A specific type of sensor is used for each parameter listed here. The network layer that incorporates sensor networks is responsible for collecting the data from various sensors, aggregating them and communicate it to a web server via a cloud service. The middleware layer comprises of several analytical tools such as online analytical processing (OLAP) and artificial intelligent (AI)-based IoT tools which then carry out the necessary analysis of the captured data and then are delivered to the application layer. The application layer holds various applications for processing of the analyzed data and then provides a qualified output that indicates at the necessary action to be taken to eliminate the exacerbating components in the environment. The entire process is depicted in the functional block diagram as depicted in Fig. 2.

6 Conclusion and Future Work

In this research article, we have proposed an IoT-based environmental monitoring system that operates on a four-layer protocol stack. This system relies on a sensor network for data collection from the environment. Various environmental parameters can be used in this system where a specific sensor can be used for a specific environmental parameter. Collected from sensors data are aggregated and then communicated to a web server using a cloud service. Analytical tools such as OLAP can be used for analysis of the captured from environment data, and then, it can be processed in order to determine the necessary actions to be taken to ease the environment and to make it healthier. We look forward to practical implementation for this system and observe the accuracy of output. But, we assume that this system is simple enough

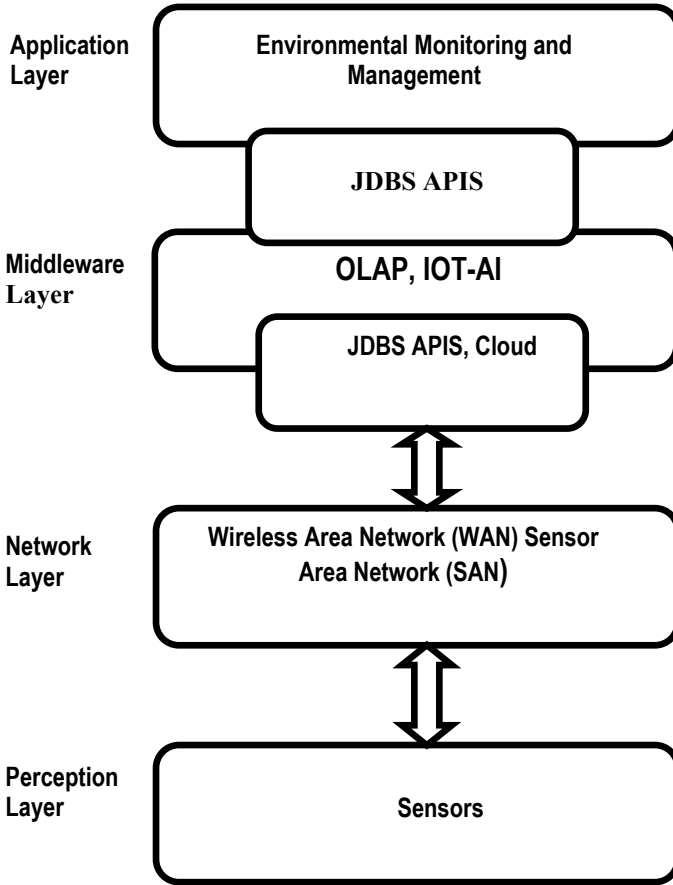


Fig. 1 IoT protocol architecture for proposed environmental monitoring system

to implement and cost effective as compared to other existing similar environmental monitoring systems.

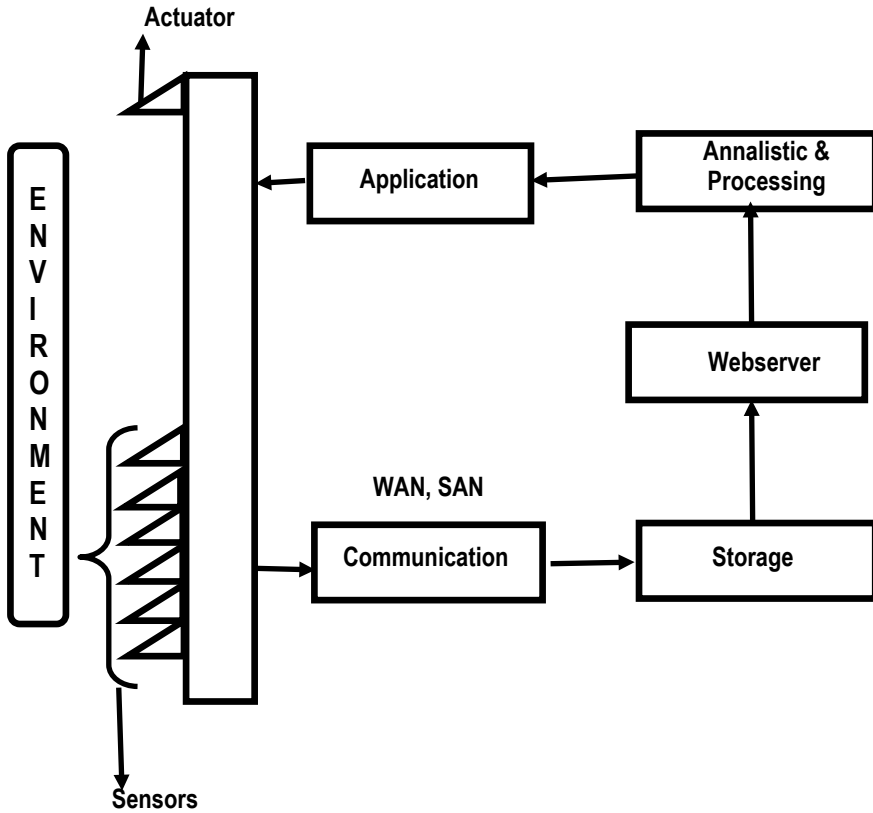


Fig. 2 Functional block diagram of proposed environmental monitoring system

References

1. Hosenkhan, R., Pattanayak, B.K.: A secured communication model for IoT. *Adv. Intell. Syst. Comput.* **863**, 187–193 (2019)
2. Singh, D., Pati, B., Panigrahi, C.R., Swagatika, S.: Security issues in IoT and their countermeasures in smart city applications. *Adv. Comput. Intell. Eng.* **1089**, 301–313 (2020)
3. Ramlowat, D.D., Pattanayak, B.K.: Exploring Internet of Things (IoT) in education: a review. *Adv. Intell. Syst. Comput.* **863**, 245–255 (2019)
4. Akila, K., Evanjaline, D.J.: Strengthening IoT-WSN architecture for environmental monitoring. *Int. J. Innovative Technol. Explor. Eng. (IJITEE)* **8**(2), 3339–3347 (2019)
5. Rahman, A., Hussein, H.: Internet of Things (IoT): research challenges and future applications. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **10**(6), 771–776 (2019)
6. Rath, M., Pattanayak, B.K.: Technological improvement in modern health care applications using Internet of Things (IoT) and proposal of novel health care approach. *Int. J. Human Rights Healthc.* **12**(2), 148–162 (2019)
7. Hosenkhan, M.R., Pattanayak, B.K.: Security issues in Internet of Things (IoT): a comprehensive review. *Adv. Intell. Syst. Comput.* **1030**, 359–369 (2020)
8. Pattanayak, B.K., Amic, S.: Modified lightweight aes based two level security model for communication on IoT. *TEST Eng. Manage.* **82**(1–2), 2323–2330 (2020)

9. Gomez, J.E., Marcillo, F.R., Triana, F.L., Gallo, V.T., Oviedo, V.W., Hernandez V.L.: IoT for environmental variables in urban areas. In: Proceedings of the 8th International Conference on Ambient Systems, Networks and Technologies (ANT 2017), pp. 67–74 (2017)
10. Dlodlo, N.: Adopting the Internet of Things technologies in environmental management in South Africa. In: Proceedings of the 2012 International Conference on Environment Science and Engineering, IPCBEE vol. **32**, pp. 45–55 (2012)
11. Fang, S., Xu, L.D.: An integrated system for regional environmental monitoring and management based on Internet of Things. *IEEE Trans. Indus. Informatics* **10**(2), 1596–1605 (2014)
12. Sasikumar, R., Anitha, M., Beebi, F., Abinaya, D.: Environmental Monitoring System Using IoT. *International Journal of Current Engineering and Scientific Research (IJCESR)* **5**(4), 64–68 (2018)
13. Sinde, S.A., Ghodke, V.N.: Environmental monitoring using IoT application for WSN platform. *Int. J. Innovative Res. Comput. Commun. Eng. (IJIRCCE)* **4**(7), 14451–14458 (2016)
14. Wang, H.: Constructing the green campus within the Internet of Things architecture. *Int. J. Distrib. Sens. Netw. (IJDSN)* **2014**, 1–8 (2014)
15. Rath, M., Pati, B.: Security assertion of IoT devices using cloud of things perception. *Int. J. Interdisc. Telecommun. Netw.* **11**(4), 17–31 (2019)
16. Shah, J., Mishra, B.: IoT Enabled environmental monitoring system for smart cities. In: Proceedings of the 2016 International Conference on Internet of Things and Applications (IOTA), pp. 383–388 (2016)
17. Mishra, M., Choudhury, P., Pati, B.: Modified ride-NN optimizer for the IoT based plant disease detection. *J. Ambient. Intell. Human Comput.* (2020). <https://doi.org/10.1007/s12652-020-02051-6>
18. Sastra, N.P., Wiharta, D.M.: Environmental monitoring as an IoT application in building smart campus of Universitas Udayana. In: Proceedings of the 2016 International Conference on Smart Green Technology in Electrical and Information Systems: Advancing Smart and Green Technology to Build Smart Society (ICSGTEIS 2016), pp. 85–88 (2016)

Performance Evaluation of VM Allocation Strategies on Heterogeneous Environments in Cloud Data Center



Rajni Garg, Indu Arora, and Anu Gupta

Abstract Cloud computing paradigm provides on demand utilities to the customers and involves high resource requirements. Virtual machine allocation plays a vital role in the optimization of resource usage in a data center. This approach consolidates workload on minimal number of servers to improve their resource utilization and thus contributes to energy efficiency. The literature evidences number of virtual machine allocation strategies that vary considerably in adopted approaches and evaluation environments. This research work aims to analyze performance of commonly used heuristic allocation policies. To ensure logistic comparison, the selection of policies is done that implements bin-packing approach in allocation. These algorithms are intensively compared to different workload data-sets and varying experimental setups. Performance of these algorithms with different threshold and virtual machine selection policies is also evaluated. The results conclude that the policies which consider power and computing capacity of the server perform better in almost all scenarios.

Keywords Virtual machine allocation · Workload traces · Server heterogeneity · Energy efficiency · Service level agreements

1 Introduction

Cloud computing (CC) is a paradigm that provides on demand elastic resources to the users based on pay-as-you-go basis [1]. Users can outsource their resource requirements to cloud and avoid high infrastructural cost. The increased trend of CC has led

R. Garg (✉) · A. Gupta
DCSA, Panjab University, Chandigarh, India
e-mail: 87rajnigarg@gmail.com

A. Gupta
e-mail: anugupta@pu.ac.in

I. Arora
Mehr Chand Mahajan DAV College for Women, Chandigarh, India
e-mail: indarora@yahoo.co.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_43

to proliferation of resource demand. To satisfy this demand, cloud service providers (CSPs) are setting up more data centers. The adoption of these data centers is driving the need of more power [2]. According to a survey, the energy consumption in a data center is estimated to reach 3000 Terawatt hour by year 2025 [3]. To reduce this high energy consumption, efficient utilization of scarce resources is a major challenge faced by cloud industry. Virtualization is a technique pervasively applied in a data center to manage resources effectively [4]. Multiple virtual machines (VMs) are accommodated on a physical server to increase its utilization. Since CPU utilization levels of VMs are not constant over time, therefore dynamic VM allocations are required to optimize resource utilization and reduce energy consumption (EC). The workload is dynamically allocated to limited servers using live migration strategy and the idle servers are put in lower power mode in order to reduce energy and attain resource efficient consolidation [5].

Empirical assessment and detailed evaluation of numerous policies devised in the literature for optimizing allocation of VMs are very important to guarantee their effectiveness. Performance comparison between these policies also plays a vital role to test their efficiency and adaptability to different scenarios of cloud. This research work aims to make detailed comparison between some of these allocation strategies that can be categorized on the basis of adopted VM allocation approach, objectives or experimental environment. To ensure better comparability, the algorithms that share common strategies and objectives are best to be considered. In this work, heuristic bin-packing algorithms that aim to reduce energy consumption and service level agreement are considered for comparison. The algorithms are intensively compared on different workload data-sets and varying experimental setups. Performance of these algorithms with different threshold and VM selection policies is also considered.

The structure of the paper comprises five sections. Section 2 describes existing VM allocation policies suggested in the literature. The performance evaluation parameters used to test the efficacy of algorithms are described in Sect. 3. Section 4 includes experimental setup which has been used for analyzing the performance and evaluating results followed by conclusion.

2 VM Allocation Policies

The VM allocation policies are devised to optimize VM placement decision in order to ensure better resource management. VM placement is a non-deterministic polynomial time (NP hard) problem [6] which can be dealt with bin-packing heuristic algorithms or meta-heuristic algorithms. Bin-packing algorithms consider physical servers as bins of different capacities, and the VM are to be allocated to the bins ensuring best resource utilization. The best fit (BF) algorithm selects a server for VM placement with least number of resources that are enough to accommodate a targeted VM. This technique ensures minimal residual resources of a physical server. Best fit decreasing (BFD) technique arranges VMs in decreasing order according to

their resource requirements, and then, BF algorithm is applied to ensure high mean utilization ratio [7]. For VM placement, power consumption is considered a major issue, and the algorithm power aware-best fit decreasing (PA-BFD) selects a machine for placement that shows minimum power difference before and after placement [8]. An algorithm, power and computing capacity-aware best fit decreasing (PCA-BFD), works to optimize energy consumption and network overheads incurred by VM migrations. The algorithm firstly calculates ratio of maximum power utilization of the server to its CPU capacity in order to calculate power efficiency of the servers. Then, the servers are arranged in increasing order based upon this ratio. The VMs are firstly allocated to the most power efficient server (the first server in the list), and then, allocations to further servers are made in increasing order. This algorithm attempts to utilize power efficient servers to attain energy efficiency [9]. Another variant of BFD, space-aware best fit decreasing (SA-BFD) algorithm, contributes to better resource utilization by placing a VM on the server that results in least free leftover MIPS after allocation [10]. The policy remaining utilization aware (RUA) considers average CPU usage by each VM and the remaining CPU capacity on the server in allocation decision. An attempt is made to accommodate workload on minimum number of servers to improve average CPU utilization rate while avoiding over-utilization of server by placing appropriate number of VMs on the server [11]. Power-ware next fit decreasing (PANFD) policy firstly sorts servers in increasing order according to available CPU MIPS, and then, all servers from last allocation to the end are examined to find the optimal server for VM allocation. The server is considered optimal that results in minimum power increase after allocation [12].

Since bin-packing algorithms do not always ensure global optimum solution, meta-heuristic algorithms for VM placement are preferred as they provide combination of randomization and local search and give optimal solution by adopting technique of exploration and exploitation [13]. Enhanced Cuckoo Search (ECS) algorithm for VM placement is recommended to ensure reduced energy consumption [14]. Its comparison with genetic algorithm, firefly search and ant colony algorithms has shown energy reduction of up to 25%, 27% and 26%, respectively, with local regression as overload detection algorithm and maximum utilization (MU) as VM selection algorithm. This algorithm does not ensure stability in the placement of VM. Sometimes varying workload or change in the need of resources of VM forces frequent need of re-allocation of VMs to other servers which leads to increased migration overheads and SLA violations. To resolve this problem and ensure high placement stability, three genetic algorithms—multi-objective genetic algorithm (MOGA), power-aware multi-objective genetic algorithm (pMOGA) and enhanced power-aware multi-objective genetic algorithm (EpMOGA)—have been proposed [15].

This work is focused towards in-depth performance analysis of VM allocation policies. To ensure logistic comparison, the policies are selected on the basis of common parameters like applied approach for allocation (heuristic), objectives of study (reduction in energy consumption and SLA), resource consideration (CPU) and implementation platform (CloudSim simulator). This work compares five heuristic bin-packing algorithms: PABFD, PCA-BFD, RUA, SABFD and PANFD. The

algorithms are tested on different scenarios such as varying workloads and physical machine (PM) heterogeneity. The performance is also tested with different VM selection and adaptive threshold policies.

3 Performance Evaluation Parameters

To analyze performance of algorithms, metrics suggested by Buyya et al. are considered [16]. These metrics are described as follows.

Energy Consumption (EC): EC measures total power consumption in a physical server. It is determined by processor and memory usage, disk storage and equipment used for cooling. It is difficult to construct an analytic model for power usage monitoring in the multi-core processing environment with large memory sizes. For analysis, real data-set of SpecPower [17] can be considered as it provides energy usage of different servers at varying workloads.

SLA Violations (SLAV): Service level agreement defines minimum service standards to be provided by CSP to the customers. The violation of SLA leads to penalties for service provider. A violation may occur due to over-utilization of server which is measured by SLA violations time per active host (SLATAH) or due to migration of VMs measured by performance degradation due to migrations (PDM). SLAV measures overall performance degradation caused during application execution and can be defined as $SLAV = SLATAH * PDM$.

Energy Performance Metric (EPM): An algorithm is considered efficient if it ensures minimum EC along with SLA compliance. EPM combines EC and SLAV and is described as $EPM = EC * SLAV$. Lower value of EPM ensures balanced performance of algorithm in terms of power consumption and SLA violations.

4 Experimental Setup and Result Evaluation

To cover every aspect that could affect performance of an algorithm, an extensive examination is done using CloudSim toolkit [18] and adopting different methodologies to test behavior of VM allocation policies. This study compares five VM allocation policies. An attempt is made to consider all time variant policies from year 2010 to year 2020. Performance comparison between these policies is done on the basis of different workload traces and physical machine heterogeneity to identify most consistent and robust strategy. These factors are described further.

4.1 Workload Traces

The rapid adoption of cloud data center (CDC) services exhibits variation in workload patterns. To continue adoption of these services, it is very important for a CSP to analyze the workload trends and implement strategies for better resource management. Performance testing of diverse resource management strategies on real workload traces is a key concept as it tests robustness and adaptability of algorithms to changing workload patterns. The current study implements selected algorithms on workload traces of Bitbrains CDC (fastStorage, Rnd) [19] and PlanetLab (CoMon) [20]. To test algorithms on vast possibilities, random workload is also considered for experiments. The next subsection briefly describes these workload traces.

Bitbrains Workload: Bitbrains as a CSP manages business computing for enterprises. Two workload traces—fastStorage and Rnd—are collected from its data center using infrastructure of VMWare. The information of these data-sets is provided in Table 1. The fastStorage data-set observes resource utilization of 1250 randomly selected VMs from the data center, whereas 500 VMs are observed by Rnd data-set. These data-sets record entries of seven performance parameters: number of cores, CPU frequency provisioned, CPU usage, provisioned memory capacity, actual memory usage, disk I/O throughput and network I/O throughput. Value of these parameters is observed every 5 min.

PlanetLab Workload: It is a global research network with 1353 geographically distributed computing nodes spanned over 48 countries. The CoMon project collected workload traces for 10 days from the randomly selected VMs. The project observes CPU capacity requested by a VM after every 5 min. Table 2 describes characteristics of these workload traces. This study considers workload traces of 3-3-2011 for the experiment.

Random Workload: For performance analysis of selected VM allocation strategies, these are also executed on artificial workload traces. These traces are randomly generated to test behavior of applications on vast possibilities. For experiment, this study considers stochastic workload.

The experimental results obtained by executing selected VM allocation policies on different data-sets are detailed in Tables 3, 4, 5 and 6. From the analysis, it is concluded that the algorithm PCA-BFD outperforms in terms of energy consumption on almost all workload traces. This is due to the fact that it ensures maximum power efficient allocations. The policy SA-BFD also gives optimal results for energy

Table 1 Configuration details of bitbrains workload traces

Trace name	# VM	Data collected for	Storage Tech.	T. Memory	# cores
fastStorage	1250	1 month	SAN	17729 GB	4057
Rnd	500	3 months	NAS, SAN	5485	1444

Table 2 Configuration details of planetlab workload traces

series Date	# VM	Mean (%)	Quartile 1 (%)	Quartile 3 (%)
03/03/2011	1052	12.31	2	15
06/03/2011	898	11.44	2	13
09/03/2011	1061	10.70	2	13
22/03/2011	1516	09.26	2	12
25/03/2011	1078	10.56	2	14
03/04/2011	1463	12.39	2	17
09/04/2011	1358	11.12	2	15
11/04/2011	1233	11.56	2	16
12/04/2011	1054	11.54	2	16
20/04/2011	1033	10.43	2	12

Table 3 Energy comparison among allocation policies on different workload traces

Workload	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
PlanetLab	184.88	111.95	185.09	128.59	180.65
fastStorage	247.82	186.59	237.41	189.04	270.98
Rnd	58.66	32.24	58.39	30.24	53.82
Random	386.91	308.76	430.70	311.57	440.17

Table 4 Migration comparison among allocation policies on different workload traces

Workload	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
PlanetLab	26,292	22,970	26,731	23,603	23,664
fastStorage	45,125	41,858	43,939	30,828	49,060
Rnd	12,876	7094	14,466	7176	17,170
Random	43,559	37,255	49,425	35,402	46,377

Table 5 SLAV * (10^{-2}) comparison among allocation policies on different workload traces

Workload	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
PlanetLab	0.33	0.25	0.35	0.39	0.30
fastStorage	0.65	0.15	0.59	0.41	0.62
Rnd	0.44	0.29	0.49	0.09	0.48
Random	2.66	3.16	1.88	3.69	1.72

Table 6 EPM comparison among allocation policies on different workload traces

Workload	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
PlanetLab	0.61	0.28	0.65	0.51	0.54
fastStorage	1.62	0.29	1.41	0.78	1.68
Rnd	0.26	0.09	0.28	0.03	0.26
Random	10.30	09.76	08.09	11.49	07.58

consumption and migration count as it attempts to utilize servers to their maximum capacity which results in lesser number of active servers. Hence, this policy contributes to reduction in energy consumption and number of migrations. Reduction in these parameters further optimizes SLA and EPM metrics. However, there is no significant comparison in results of PABFD, RUA and PANFD policies in term of energy consumption and migration count. The policy PANFD performs comparatively better in terms of SLAV. This is due to the quick allocation strategy adopted by the algorithm.

4.2 Physical Machine Heterogeneity

The applications in a CDC operate on heterogeneous physical servers, which are characterized based on CPU processing capacity, number of processing cores, RAM availability, storage and network bandwidth. An algorithm is termed efficient if it performs equally well for these heterogeneous scenarios. To test algorithms on these heterogeneous servers, two case studies have been developed.

Case Study 1: Implementation of algorithms on baseline servers—HP Proliant G4 and HP Proliant G—is done. Table 7 describes characteristics of these servers. Power utilization of these servers is observed according to SpecPower data-set.

Case Study 2: To keep data center upgraded with latest technology innovations, a CSP replaces baseline servers by latest energy efficient servers. But, complete transformation to upgraded technology is not possible due to high cost involvement and continuous innovations in technology. The hardware immediately starts to age after purchase. Therefore, data center operates with a bundle of traditional and latest servers. So, it becomes worthwhile to test adaptability of VM placement algorithms on these heterogeneous machines. Case study 2 implements four highly variant server configurations (HP Proliant G4, HP Proliant G5, PRIMERGYRX 1330 M3 and PRIMERGYRX 1330 M4). The variability is decided on the basis of performance parameters already discussed in Sect. 3. Table 7 describes detailed configuration of these servers.

The results obtained by executing allocation policies on heterogeneous servers are detailed in Tables 8, 9, 10, 11.

Table 7 Configuration details of Servers

Server	HP G4	HP G5	RX M3	RX M4
Processor (Xeon)	3040	3075	E3-1230	E-2176G
CPU Frequency (MHz)	1860	2660	3500	3700
# Cores	2	2	4	6
RAM (GB)	4	4	16	16
Min power (Watts)	86	93.7	13.3	14.9
Max power (Watts)	117	135	56.1	72.9

Table 8 Energy comparison among allocation policies on different server configurations

Configuration	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
Case study 1	184.88	111.95	185.08	128.59	180.65
Case study 2	15.07	11.49	14.85	118.35	12.22

Table 9 Migration comparison among allocation policies on different server configurations

Configuration	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
Case study 1	26,292	22,970	26,731	23,603	23,664
Case study 2	15,721	11,041	15,855	23,560	8849

- The comparison reveals that the utilization-aware algorithms such as SA-BFD do not perform well for heterogeneous physical servers as they do not consider power efficiency of the servers. The allocations are merely based on utilization of CPU processing instructions.
- The performance consistency has been observed in PCA-BFD algorithm. In contrast to other policies, results of this algorithm are highly optimal for baseline as well as heterogeneous servers.
- In PANFD policy, better results in terms of energy consumption, migration count and SLAV are observed on heterogeneous physical servers as compared to baseline servers since the central objective of PANFD is max utilization of power efficient servers like PABFD algorithm.
- The 91.84%, 89.73%, 91.97% and 93.23% energy reduction in PABFD, PCA-BFD, RUA and PANFD, respectively, has been observed on heterogeneous servers in contrast to baseline servers.
- The reduction of 84.0% and 73.3% in SLAV is also observed in PCA-BFD and PANFD policies, respectively. However, there is no significant improvement in other policies.

Table 10 SLAV * (10^{-2}) comparison among allocation policies on different server configurations

Configuration	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
Case study 1	0.33	0.25	0.35	0.39	0.30
Case study 2	0.32	0.04	0.32	0.33	0.08

Table 11 EPM comparison among allocation policies on different server configurations

Configuration	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
Case Study 1	0.612	0.282	0.646	0.506	0.540
Case Study 2	0.048	0.005	0.048	0.391	0.010

4.3 VM Selection Strategies

Careful selection of a VM from an oversubscribed server for migration is crucial because it influences overall execution. A wrong selection sometime leads to high energy consumption and SLA violations in a data center. In this work, performance evaluation of selected VM allocation policies in terms of energy consumption and SLAV is done with different VM selection policies. This study considers four benchmark VM selection strategies—Minimum migration time (MMT), maximum utilization (MU), random selection (RS) and maximum correlation (MC) for evaluation purposes. The MMT algorithm selects those VMs for migration that has minimum migration time. The migration time is the ratio of RAM utilization of the VM to the current available network bandwidth. In MU algorithm, a VM having highest CPU utilization is selected for migration in order to quickly moderate utilization level of the server. The random selection of a virtual machine is done in RS policy. The MC algorithm selects those VMs that have highest correlation of CPU utilization with other VMs because execution of highly correlated VMs on same server may lead to server overloading as they all demand high CPU-MIPS.

From the results, it is concluded that the selection policy impacts performance of allocation strategies. These policies behave differently with various VM selection policies.

- A consistency is observed in outcomes of PCA-BFD and SA-BFD allocation policies with all selection policies. These policies show 2.14% and 2.07% of respective variations in energy consumption levels, whereas the variation of 13.77%, 12.29% and 13.38% is observed in PA-BFD, RUA and PANFD policies, respectively, as shown in Table 12.
- The MMT policy contributes to lower SLA violations as represented by Table 13. The results are optimal with all VM allocation strategies, this is attributed to the fact that the policy considers migration time in selection decision, and VM with least value is selected that positively impacts SLATAH and PDM which collectively contribute to SLAV reduction.

Table 12 Energy comparison among allocation policies with different VM selection policies

Selection Policy	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
RS	176.58	112.51	174.61	127.77	172.62
MMT	184.88	111.95	185.09	128.59	180.62
MC	176.13	112.81	175.51	129.56	170.9
MU	200.4	110.44	196.08	130.42	193.77

Table 13 SLAV * (10⁻²) comparison among allocation policies with different VM selection policies

Selection Policy	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
RS	0.72	0.48	0.74	0.66	0.50
MMT	0.33	0.25	0.35	0.39	0.30
MC	0.74	0.45	0.72	0.63	0.47
MU	0.51	0.58	0.47	0.84	0.49

4.4 Adaptive Threshold Policies

Adaptive threshold policies identify oversubscribed servers by setting upper threshold limit for the server. On detection of such server, VM re-allocation strategies take place to migrate some of its running VMs. To study effect of adaptive threshold techniques on performance of VM allocation strategies, three adaptive threshold policies—Median absolute deviation (MAD), interquartile range (IQR) and local regression (LR)—are taken into account. MAD and IQR policies implement statistical techniques on historical data to calculate the threshold limit, whereas LR is based on regression analysis.

The selected allocation policies are executed with these threshold policies, and their observed impact on energy consumption and SLAV is described in Tables 14 and 15.

- The analysis concludes that the results of most of the VM allocation policies are better with LR threshold technique in regard to energy consumption metric except of PCA-BFD that works well with MAD allocation policy.

Table 14 Energy comparison among allocation policies with different threshold policies

Threshold Policy	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
MAD	184.88	111.95	185.09	128.59	180.65
LR	163.15	115.84	162.21	126.16	151.38
IQR	188.86	114.80	188.20	128.63	184.28

Table 15 SLAV * (10^{-2}) comparison among allocation policies with different threshold policies

Threshold Policy	PA-BFD	PCA-BFD	RUA	SA-BFD	PANFD
MAD	0.33	0.25	0.35	0.39	0.30
LR	0.46	0.10	0.45	0.13	0.14
IQR	0.31	0.24	0.32	0.34	0.32

- In regard to SLAV, results of PCA-BFD, SA-BFD and PANFD are optimal with LR adaptive threshold policy. However, PABFD and RUA policies perform better with IQR threshold policy.

5 Conclusion

This paper investigates performance and consistency of selected VM allocation policies on different adopted scenarios such as workload variations, PM heterogeneity, VM selection policies and adaptive threshold policies. From the analysis, the following observations have been made.

- Utilization-aware policies perform well in terms of energy consumption on different data-sets, but these policies lack in performance on latest physical servers. This is because power to computing ratio is not considered by these policies. So, these policies do not ensure better usage of power efficient servers.
- Performance consistency in PCA-BFD algorithm is observed on almost all scenarios because it attempts to reduce both energy consumption and network load of a data center.
- The policies such as PABFD, RUA and SABFD involve high allocation decision time because VM mapping is done after analyzing all available servers which makes these policies unsuitable for large-scale scenarios. Whereas PCA-BFD policy sorts all servers in increasing order according to ratio of power consumption and CPU capacity, and these allocations are made in order. Similarly, PANFD policy sorts servers in increasing order according to available CPU capacity, and the servers from last allocation are analyzed for further allocations. However, these policies include extra overhead of at least $O(n\log n)$ operations for server arrangements in sorting order.

From the analysis, it is further concluded that most of the policies consider single resource in allocation decision such as CPU-MIPS. But, multi-dimensional resource management policy is required because on complete utilization of one resource, the server becomes incapable to accommodate more VMs even when other resources are under-utilized.

References

1. Shang, P., Wang, J.: A novel power management for CMP systems in data-intensive environment. In: IEEE International Parallel & Distributed Processing Symposium, pp. 92–103. IEEE, New York (2011)
2. Beloglazov, A., Abawajy, J., Buyya, R.: Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Gener. Comput. Syst.* **28**(5), 755–768 (2012)
3. Barroso, L.A., Hözlze, U.: The case for energy-proportional computing. *Computer* **40**(12), 33–37 (2007)
4. Garg, S.K., Buyya, R.: Green cloud computing and environmental sustainability. *Harnessing Green IT: Principles Pract.* **2012**, 315–340 (2012)
5. Clark, C., Fraser, K., Hand, S., Hansen, J. G., Jul, E., Limpach, C., Warfield, A.: Live migration of virtual machines. In: *Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation*, vol. 2, pp. 273–286 (2005)
6. Calcevachia, N. M., Biran, O., Hadad, E., Moatti, Y.: VM placement strategies for cloud scenarios. In: *IEEE Fifth International Conference on Cloud Computing*, pp. 852–859. IEEE, New York (2012)
7. Toth, P., Martello, S.: *Knapsack problems: Algorithms and Computer Implementations*. Wiley (1990)
8. Beloglazov, A., Buyya, R.: Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency Comput.: Practice and Exp.* **24**(13), 1397–1420 (2012)
9. Tziritas, N., Xu, C. Z., Loukopoulos, T., Khan, S. U., Yu, Z.: Application-aware workload consolidation to minimize both energy consumption and network load in cloud environments. In: *42nd International Conference on Parallel Processing*, pp. 449–457. IEEE, New York (2013)
10. Wang, H., Tianfield, H.: Energy-aware dynamic virtual machine consolidation for cloud data-centers. *IEEE Access* **6**, 15259–15273 (2018)
11. Han, G., Que, W., Jia, G., Shu, L.: An efficient virtual machine consolidation scheme for multimedia cloud computing. *Sensors* **16**(2), 246–261 (2016)
12. Sengupta, J., Singh, P., Suri, P. K.: Energy aware next fit allocation approach for placement of VMs in cloud computing environment. In: *Future of Information and Communication Conference*, pp. 436–453. Springer, Cham (2020)
13. Blum, C., Roli, A.: Metaheuristics in combinatorial optimization: overview and conceptual comparison. *ACM Comput. Surveys (CSUR)* **35**(3), 268–308 (2003)
14. Barlaskar, E., Singh, Y.J., Issac, B.: Enhanced Cuckoo search algorithm for virtual machine placement in cloud data centres. *Int. J. Grid Util. Comput.* **9**(1), 1–17 (2018)
15. Deng, L., Li, Y., Yao, L., Jin, Y., Gu, J.: Power-aware resource reconfiguration using genetic algorithm in cloud computing. *Mob. Inform. Syst.* (2016)
16. Wu, L., Garg, S.K., Buyya, R.: SLA-based admission control for a software-as-a-service provider in cloud computing environments. *J. Comput. Syst. Sci.* **78**(5), 1280–1299 (2012)
17. Hsu, C. H., Poole, S. W.: Power signature analysis of the SPECpower_ssj2008 benchmark. In: *IEEE International Symposium on Performance Analysis of Systems and Software*, pp. 227–236. IEEE, New York (2011)
18. Calheiros, R.N., Ranjan, R., Beloglazov, A., De Rose, C.A., Buyya, R.: CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Pract. Exp.* **41**(1), 23–50 (2011)
19. Shen, S., van Beek, V., Iosup, A.: Statistical characterization of business-critical workloads hosted in cloud datacenters. In: *15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pp. 465–474. IEEE, New York (2015)
20. Park, K., Pai, V.S.: CoMon: A mostly-scalable monitoring system for PlanetLab. *ACM SIGOPS Oper. Syst. Rev.* **40**(1), 65–74 (2006)

QoSens: QoS-Aware Sensor Node Selection in Sensor-Cloud Architecture



Arijit Roy, Sudip Misra, and Aditya Kotasthane

Abstract In this paper, we propose a Quality-of-Service (QoS)-aware sensor node selection scheme, QoSens, for sensor-cloud architecture. In this architecture, a Sensor-Cloud Service Provider (SCSP) provisions Sensors-as-a-Service (Se-aaS) to the registered end-users. On the other hand, the end-users pay the charges for their availed services. This work has twofold objectives—first, we define the Service-Level Agreements (SLAs) in sensor-cloud to bind sensor owners, SCSP, and end-users together with certain contracts, and second, with the help of these SLAs, the proposed scheme provisions to select a suitable set of sensor nodes, based on the QoS value, to serve an application. The SLA between sensor owner and SCSP enforces the former to share the detailed specifications of his/her sensor nodes to the SCSP. On the other hand, the SLA between SCSP and the end-users enforces the SCSP to determine the optimal QoS of different available sets of sensor nodes and share with the end-users. We formulate the QoS of a sensor node with its specifications shared by the sensor owner. Further, we apply *Karush–Kuhn–Tucker (KKT)* conditions to obtain an optimal sensor node, based on the QoS value. Extensive experimental results depict that the total payable service price varies in the range 77.69–86.97% with the increase in the service price of SCSP from 500–1000 units. On the other hand, with the change in the price of sensor nodes from 500–1000 units, the total payable service price varies from 35.79–54.6%.

Keywords Service-level agreements (slas) · Sensor-cloud · Sensors-as-a-Service (Se-aaS) · Quality-of-Service (QoS) · Sensor node selection

A. Roy (✉) · S. Misra · A. Kotasthane
Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur, Kharagpur, India
e-mail: arijitroy@iitkgp.ac.in

S. Misra
e-mail: sudipm@iitkgp.ac.in

A. Kotasthane
e-mail: adityakotasthane@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_44

527

1 Introduction

Sensor-cloud is based on the service-oriented architecture (SOA), which consists of multiple actors such as sensor owners, end-users, and SCSP [1, 3]. This architecture provisions Sensors-as-a-Service (Se-aaS) for end-users using the concept of sensor *virtualization*. A sensor owner leases his/her sensor nodes and earns profit, depending upon the usage of the respective sensor nodes. On the other hand, an end-user pays rent to the SCSP for the services availed by him/her [2]. The SCSP acts as a centralized actor, who manages the sensor-cloud architecture along with the cash inflow and outflow in the system. As an end-user pays a significant amount of price for certain application, he/she expects for desirable Quality-of-Service (QoS). The proposed scheme, QSens, allows an end-user to select a sensor node for an application, based on its QoS. The service-level agreement (SLA) plays a crucial role in sensor-cloud for selecting a sensor node, depending upon its QoS, by the end-users. In sensor-cloud, SLAs are in the form of a certain commitment of services among SCSP, end-users, and sensor owners. The proposed scheme, QSens, comprises two SLAs— SLA_{SS} and SLA_{SE} . The SLA between SCSP and sensor owner is termed as SLA_{SS} , while the SLA between SCSP and end-users is known as SLA_{SE} . However, these SLAs may contain other service agreement, as per their requirement. The primary aim of this work is to minimize the price charged from an end-user and maximize the QoS of the sensor nodes.

In the existing literature of sensor-cloud, there is no such scheme which facilitates the end-users to select the sensor nodes, as per their requirements. This motivates us to propose a scheme, QSens, for allowing the end-users to select a suitable set of sensor nodes for serving an application. The specific *contributions* of this work are:

- The authors in the existing literature do not propose any work on SLA for the sensor-cloud architecture. Therefore, in this work, we introduce two different SLAs— SLA_{SS} and SLA_{SE} , which are specifically designed for the sensor-cloud. These SLAs bind sensor owners, SCSP, and end-users, legally with certain contracts.
- QSens allows the end-users to know the optimal value of QoS of the sensor nodes. On the other hand, the end-users are capable of selecting a suitable set of sensor nodes within their financial budget, considering the optimal QoS. In this work, we derive a function for computing an optimal QoS of the sensor nodes with the help of SLAs.
- We solve the problem by optimization function and proving it as convex. Further, we apply the *Langragian Multiplier* [4] and the *Karush–Kuhn–Tucker* (KKT) [5] conditions to derive the optimal value of QoS. Additionally, we analyze QSens with rigorous simulation.

2 Related Work

Existing literature reveals different research works on SLA for the traditional cloud architecture. On the other hand, in the literature, the authors explored the concept of sensor-cloud architecture, which replace the traditional WSNs. Considering all these aspects, we categorized the related works in two parts—SLA and sensor-cloud architecture.

For a SOA-based system, SLA plays an important role to bind the service providers and consumers. In the existing literature, the authors proposed several SLA-enabled schemes for different technologies and applications. Gaillard et al. [6] implemented SLA for WSNs. The authors discuss few important mechanisms—*SLA Observer*, *Service Registry*, *SLA Admitter*, *SLA Manager*, and *SLA Enforcer*—for ensuring QoS, in the context of WSNs. Similarly, Chieng et al. [7] proposed an SLA-driven scheme to facilitate the dynamic and flexible bandwidth reservation for a QoS-aware Internet. In order to discuss an SLA broker scenario, the authors used *Fujitsu's Phoenix Open Agent Mediator (OAM)*. Garcia et al. [8] modeled an SLA with Linked Unified Service Description Language (USDL) agreement. The authors utilized the benefit of the Web principle for incorporating the technical and business aspects in the SLA. The proposed model offers the necessary facilities for capturing the semantics of the agreements. Typically, for a cloud service, the SLA is proactive and difficult to dynamically modify. Considering the dynamic modification of SLA, Papatungan et al. [9] proposed a scheme for enabling dynamic negotiation in SLA for cloud architecture.

In the existing literature, the authors explored different works in the domain of sensor-cloud architecture. Yuriyama et al. [1, 10] proposed the concept of virtualization of sensor nodes. Further, Madira et al. [11] presented the Sensing-as-a-Service (Se-aaS) paradigm to offer a common service platform for multiple end-user. In this work, the authors also discussed the formation of the virtual sensor (VS) considering the resource-constrained environment of traditional WSNs. A VS comprises multiple physical sensor nodes and provisions multiple end-users to receive services, simultaneously. However, the composition of VS changes with time and the types of applications. In order to form the dynamic VSs, Roy et al. [12] designed a scheme for the sensor-cloud architecture. Typically, a sensor-cloud architecture is based on pay-per-use model, in which different actors are involved to receive certain benefits. Therefore, Chakraborty et al. [3] proposed a pricing scheme to manage the financial transactions among the actors of sensor-cloud while enforcing the trust among SCSPs.

3 Problem Scenario

We consider a sensor-cloud architecture, where sensor owners procure multiple heterogeneous sensor nodes and rent them to serve the end-user applications. The rent of these sensor nodes varies with *application type* and *duration* of their usage. Addi-

tionally, the rent of a sensor node depends on its QoS. However, the specification of the sensor node decides the QoS. The SLA plays a crucial role to provide the QoS of the sensor nodes to the end-user. We define two SLAs, which legally bind the sensor-owners with SCSPs and SCSPs with end-users, respectively.

Definition 1 The sensor owners are enforced, through a SLA, to share the detailed specifications of their respective sensor nodes to the SCSPs, such a SLA is known as SLA_{SS} .

Definition 2 The SCSPs are enforced, through a SLA, to share correct QoS of the sensor nodes with the total payable service price of the service to the end-users, such a SLA is known as SLA_{SE} .

The QoS of the sensor nodes is computed with the sensor node specifications, shared by the sensor owners, whereas the total payable service price of the service is derived using the service cost of SCSP, sensor node, and their QoS. SLA_{SS} and SLA_{SE} are the key enablers for providing the specifications of sensor nodes in sensor-cloud architecture. In this work, we propose a mechanism to compute the QoS and derive the total payable service price for the end-users.

Let $SO = \{SO_1, SO_2, SO_3, \dots, SO_p\}$ denote the set of sensor owners, where $SO_i \in SO$ represents any sensor owner and $1 \leq i \leq p$, such that p is maximum number of sensor owner present in the set. Any SO_i leases his/her respective sensor node to the sensor-cloud architecture and receives the rent as per the usage of the sensor nodes. The sensor node, j , owned by the i^{th} sensor owner is denoted as s_j^i . Further, we define the set of sensor nodes belonging to the i^{th} sensor owner as $S^i = \{s_1^i, s_2^i, s_3^i, \dots, s_q^i\}$. The maximum number of sensor nodes belonging to SO_i is denoted as q . In a sensor-cloud architecture, multiple SCSPs are present to provision Se-aaS to multiple end-users. Let $\mathcal{S} = \{S_1, S_2, S_3, \dots, S_r\}$ represent the set of available SCSPs in the system. The i^{th} sensor owner is legally binded to the k^{th} SCSP with a SLA, SLA_{SS}^{ik} . Also, we define the set of end-users as $EU = \{EU_1, EU_2, EU_3, \dots, EU_s\}$. Any SCSP, k , is legally binded with an end-user EU_l using a SLA, SLA_{SE}^{kl} . Figure 1 depicts the architecture of QSens. Additionally, we observe that Fig. 1 possesses a set of available sensor nodes with corresponding sensor owners. Among the available sensor nodes, a set of nodes are serves the end-user application. Using QSens, we facilitate the end user to choose the available set of nodes based on the QoS and its price. Figure 2 depicts the process flow of the proposed architecture of QSens. It shows the communications between three entities—sensor owner, SCSP and the end-user—to offer the sensor node with an optimal QoS at a minimized price.

4 QSens: The Proposed Scheme

In this section, we discuss the statistical variable used to categorize the various sensor nodes and formulation of QoS of these nodes.

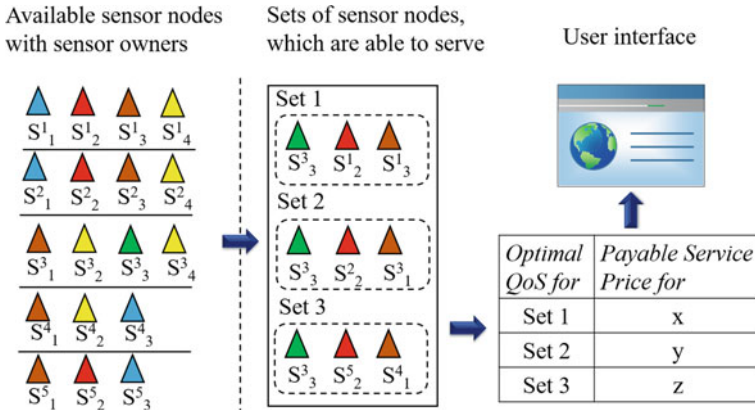
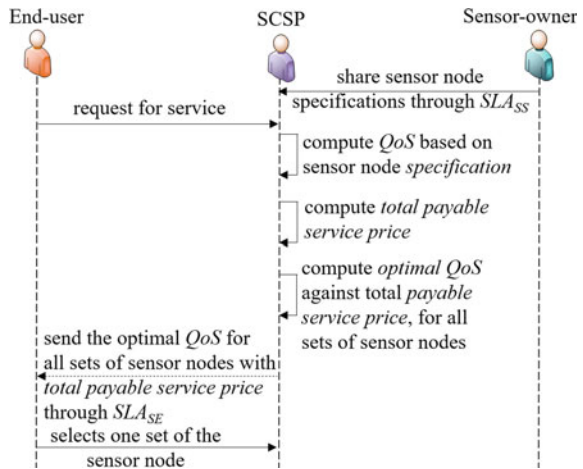


Fig. 1 Architecture of QSens

Fig. 2 Flow diagram of QSens



4.1 Statistical Variables for Sensor Node Specifications

SLA_{SS} enables the SCSP to receive different specifications of the sensor nodes. We compute the QoS of the sensor nodes using their respective specifications. However, depending on the types of sensor node, the specifications vary from one another. Further, we draw an analogy of the statistical variables—*dichotomous*, *continuous interval*, and *discrete ratio* [13] and map the specifications of sensor nodes to these variables.

Justification for using the statistical variables: Let there be a sensor owner (SO_i), owning a sensor node, s^i_j . The node, s^i_j , consists of a set of specifications such as technology support, sensor range, processing speed, energy consumption, temperature support, ADC resolution, interface support, debug support JTAG-SWD, Minimum

Energy Performance Standard (MEPS), ISO security compliance. We denote *security* and *MEPS* as dichotomous variable. Dichotomous variables are binary, i.e., they attain a value of 0 or 1. In our context, *security* and *MEPS* are either compliant or non-compliant with the sensor node. Therefore, if the sensor node is compliant with *security* and *MEPS*, we map these to the dichotomous variable with value 1; otherwise, with value 0. Similarly, the temperature of a sensor node varies within a minimum and maximum range. The minimum and the maximum temperature values are considered as range set. This minimum and sets of maximum range are brought under a predefined range of [0 1]. These new sets obtained contribute toward the specification, which has an interval range. We map it as continuous interval variable. The specifications such as communication technology support, sensor range, processing speed, energy consumption, ADC Resolution, interface support, debug support JTAG-SWD are provided with a sensor node. These technologies are provided corresponding to any sensor node, which are countable in nature and possess discrete countable values. Therefore, we categorize these specifications as discrete ratio variables.

4.2 QoS Formulation

Let the sets of dichotomous, continuous interval, and discrete ratio variables of a sensor node are denoted as $\mathbb{D} = \{d_1, d_2, d_3, \dots, d_t\}$, $\mathbb{I} = \{i_1, i_2, i_3, \dots, i_u\}$ and $\mathbb{R} = \{r_1, r_2, r_3, \dots, r_v\}$, respectively. We compute the QoS, Q_j , of the j^{th} sensor node with the help of these statistical variables. Moreover, Q_j depends on the effective dichotomous variable, effective continuous interval variable, and effective discrete ratio variable of sensor node, j . Therefore,

$$d_j^i = \begin{cases} 1, & \text{if a particular specification is present} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Definition 3 The effective dichotomous variable, $\mathcal{E}_j^{\mathbb{D}}$, of a sensor node, j , is the sum of all the dichotomous variables shared by its respective owner to the k^{th} SCSP, through SLA_{ss}^{ik} .

We compute the effective dichotomous variable $\mathcal{E}_j^{\mathbb{D}}$, as:

$$\mathcal{E}_j^{\mathbb{D}} = \sum_{i=1}^t d_i \quad (2)$$

The value of continuous intervals and discrete ratio variables lies in a range minimum and maximum ranges. However, the ranges of these variables are different from one another. Therefore, for both interval and ratio variables, we use *min-max normalization* technique [14] and bring them between in the range of $\{\sigma_{\min},$

σ_{\max} }. For simplicity, we consider $\sigma_{\min} = 0$ and $\sigma_{\max} = 1$. Then, any m^{th} element, i_m of the set of interval variables consist of its minimum value, i_m^{\min} and maximum value, i_m^{\max} , respectively. Therefore, the set of interval variables is represented as $\mathbb{I} = \{(i_1^{\min}, i_1^{\max}), (i_2^{\min}, i_2^{\max}), (i_3^{\min}, i_3^{\max}), \dots, (i_u^{\min}, i_u^{\max})\}$

Let the minimum and maximum values in set \mathbb{I} are denoted as ϵ_{\min} and ϵ_{\max} , respectively. Therefore, $\epsilon_{\min} = \text{Min}(\mathbb{I})$ and $\epsilon_{\max} = \text{Max}(\mathbb{I})$.

Also, ϵ denotes any values in set \mathbb{I} . The min-max normalization technique for the interval variable is represented as:

$$\mathcal{M} = \left(\left(\frac{\epsilon - \epsilon_{\min}}{\epsilon_{\max} - \epsilon_{\min}} \right) \times (\sigma_{\min} - \sigma_{\max}) \right) + \sigma_{\max} \tag{3}$$

where \mathcal{M} is the normalized value of i_m , such that $\sigma_{\min} \leq \mathcal{M} \leq \sigma_{\max}$. Therefore, we define the set of minimum and the maximum normalized values of the interval variables, for node j as:

$$\mathcal{I}_j^{\min} = \left\{ i_{(j,1)}^{\min}, i_{(j,2)}^{\min}, i_{(j,3)}^{\min}, \dots, i_{(j,u)}^{\min} \right\} \tag{4a}$$

$$\mathcal{I}_j^{\max} = \left\{ i_{(j,1)}^{\max}, i_{(j,2)}^{\max}, i_{(j,3)}^{\max}, \dots, i_{(j,u)}^{\max} \right\} \tag{4b}$$

We use Eq. (4) to define a parameter, *effective normalized interval variable* for deriving the QoS of any sensor node.

Definition 4 The effective normalized interval variable, $\mathcal{E}_j^{\mathbb{I}}$ of a sensor node, j , is the total sum of all the elements in sets \mathcal{I}_j^{\min} and \mathcal{I}_j^{\max} .

We derive the effective interval variable $\mathcal{E}_j^{\mathbb{I}}$, as:

$$\mathcal{E}_j^{\mathbb{I}} = \sum_{m=1}^u i_{(j,m)}^{\min} + \sum_{m=1}^u i_{(j,m)}^{\max} \tag{5}$$

The set of discrete ratio variables is denoted as $\mathbb{R} = \{r_j^1, r_j^2, r_j^3, \dots, r_j^v\}$, such that an element r_j^i represents the number of supporting technologies for a particular variable by a sensor node, j .

Definition 5 The effective ratio variable, $\mathcal{E}_j^{\mathbb{R}}$ of a sensor node, j , is the total sum of all the elements in set \mathbb{R} .

The effective ratio variable is derived $\mathcal{E}_j^{\mathbb{R}}$ as:

$$\mathcal{E}_j^{\mathbb{R}} = \sum_{p=1}^v r_j^p \tag{6}$$

We derive the QoS, Q_j , of a sensor node, j using Eqs. (2), (5), and (6). On the other hand, energy consumption is a crucial factor for a sensor node. Let the maximum energy consumption of any sensor node, among the available ones, be E_{\max} . Therefore, the effective energy consumption, E_{eff} , of the j th sensor node is computed as $E^{eff} = \frac{E_j}{E_{\max}}$.

Also, we consider the effective energy, E_j , of the j^{th} sensor node for determining its Q_j as:

$$Q_j = \frac{(\mathcal{E}_j^{\mathbb{D}} + \mathcal{E}_j^{\mathbb{I}} + \mathcal{E}_j^{\mathbb{R}})}{E^{eff}} \tag{7}$$

Similarly, we calculate the QoS of all the sensor nodes, available with the SCSP for certain application. The price of the j^{th} sensor node is denoted as, p_j , which depends on the QoS. Further, the effective energy, E^{eff} of a sensor node, j , influences the QoS, Q_j . However, Q_j must not dominate the total price, \mathcal{P}_j . Therefore, the total service price of the j^{th} sensor node is mathematically represented as $\mathcal{P}_j^S = \{Q_j\}^{1/\gamma} \times p_j$, where γ is a scaling factor with a positive constant value.

In Eq. (7), we notice that the effective energy E^{eff} influences the QoS of the sensor node. Let a value of effective energy be denoted as \mathcal{Y} . The value of \mathcal{Y} affects the QoS, Q_j . We say that \mathcal{Y} can affect the QoS as:

$$(Q_j) > 1, \exists, \text{ if } \{E^{eff} > \mathcal{Y}\} \quad (Q_j) \leq 1, \exists, \text{ if } \{E^{eff} < \mathcal{Y}\} \tag{8}$$

Proposition 1 *If n is the number of sensor nodes and the total effective energy of these nodes denoted as E_j^{eff} , then*

$$\sum_{j=1}^n (Q_j) \geq \left\{ \frac{n}{\sum_{j=1}^n \mathcal{Y}} \right\} \tag{9}$$

Justification: Let us assume that,

$$\sum_{j=1}^n Q_j < \frac{n}{\sum_{j=1}^n E_j^{eff}} \tag{10}$$

Also, we observe that $0 < E_j^{eff} \leq 1$. On the other hand, from Equation (7) we get:

$$\sum_{j=1}^n Q_j = \sum_{j=1}^n \frac{(\mathcal{E}_j^{\mathbb{D}} + \mathcal{E}_j^{\mathbb{I}} + \mathcal{E}_j^{\mathbb{R}})}{E_j^{eff}} \tag{11}$$

Therefore, from Eqs. (10) and (11), we obtain:

$$\sum_{j=1}^n \frac{(\mathcal{E}_j^{\mathbb{D}} + \mathcal{E}_j^{\mathbb{I}} + \mathcal{E}_j^{\mathbb{R}})}{E_j^{eff}} < \frac{n}{\sum_{j=1}^n E_j^{eff}} \tag{12}$$

The maximum possible effective energy, E_j^{eff} , for all the nodes is 1. Therefore, the maximum value of R.H.S. in Eq. (12) gives 1. On the other hand, the minimum value of L.H.S. is 1, which infer a contradiction to our assumption, as mentioned in Eq. (10). This concludes the justification of the proposition.

Further, the SCSP charges a certain amount from the end-users, which includes the maintenance charges of the sensor-cloud infrastructure for offering the Se-aaS, considering per unit price of the sensor node and the price charged by the sensor owner. We denote the charge of SCSP as \mathcal{P}^{SCSP} . Price, p_j , of the j^{th} sensor node consists of a maximum value, p_j^{max} . Similarly, maximum service price of the j^{th} sensor node is, $P_j^{SCSPmax}$. Thus,

$$\forall j \in \{j = 1, \dots, n\} \exists \mathcal{P}_j^{SCSP} \leq P_j^{SCSPmax} \tag{13a}$$

$$\forall j \in \{j = 1, \dots, n\} \exists p_j \leq p_j^{max} \tag{13b}$$

Further, the total payable service price, \mathbb{P} , to an end-user for the j^{th} sensor node is represented as: $\mathbb{P} = \mathcal{P}_j^{SCSP} + \mathcal{P}_j^S$. We compute the total payable service price, \mathbb{P}_{tot} , for a set of sensor nodes, which are eligible to serve the application, as:

$$\mathbb{P}_{tot} = \sum_{j=1}^n \mathcal{P}_{SCSP_j} + \sum_{j=1}^n \{(\mathcal{Q}_j)^{1/\gamma} \times p_j\} \tag{14}$$

The main aim of QoSens is to minimize \mathbb{P}_{tot} , while obtaining an optimal QoS. In order to achieve the minimum payable service price, we use $argmin_{\mathbb{P}_{tot}}$, for an optimal value of total QoS, \mathcal{Q} . Therefore, we represent Equation (14) as:

$$argmin_{\mathcal{Q}} \left(\sum_{j=1}^n \mathcal{P}_{SCSP_j} + \sum_{j=1}^n \{(\mathcal{Q}_j)^{1/\gamma} \times p_j\} \right) \tag{15}$$

Theorem 1 *The proposed function in Eq. (14) is convex, iff for each payable service price, $\mathbb{P}_{tot}^1, \mathbb{P}_{tot}^2 \in Z$, where Z is a non-empty open convex set.*

Proof Two service prices are denoted as \mathbb{P}_{tot}^1 and \mathbb{P}_{tot}^2 , such that:

$$\mathbb{P}_{tot}^1 = \sum_{j=1}^n \mathcal{P}_{SCSP_j}^1 + \sum_{j=1}^n \{(\mathcal{Q}_j^1)^{1/\gamma} \times p_j^1\} \text{ and } \mathbb{P}_{tot}^2 = \sum_{j=1}^n \mathcal{P}_{SCSP_j}^2 + \sum_{j=1}^n \{(\mathcal{Q}_j^2)^{1/\gamma} \times p_j^2\} \tag{16}$$

The respective first-order partial derivatives of \mathbb{P}_{tot}^1 with respect to $Q^{tot,1}$ is:

$$\frac{\partial \mathbb{P}_{tot}^1}{\partial (Q^{tot,1})} = \left\{ \frac{1}{\gamma} (Q^{tot,1})^{\frac{1-\gamma}{\gamma}} \times \sum_{j=1}^n p_j^1 \right\} \tag{17}$$

where, $\sum_{j=1}^n Q_j^1 = Q^{tot,1}$. Similarly, we obtain the first-order partial derivative of \mathbb{P}_{tot}^2 . From Equation (17), we obtain:

$$\left[\frac{\partial \mathbb{P}_{tot}^1}{\partial (Q^{tot,1})} - \frac{\partial \mathbb{P}_{tot}^2}{\partial (Q^{tot,2})} \right] (Q^{tot,1} - Q^{tot,2}) \geq 0 \tag{18}$$

Therefore, from Eq. (17), we infer that Eq. (14) is convex [5].

Corollary 1 *If the function for payable service price, in Eq. (14), \mathbb{P}_{tot} , is convex, then the function attains a minimum value.*

Proof In Theorem 1, we proved that the function derived in Equation (14) is convex. We apply the second-order partial derivatives on Equations (17) to attain minimum \mathbb{P}_{tot}^1 . Therefore, we obtain partial derivative of \mathbb{P}_{tot}^1 , as:

$$\frac{\partial^2 \mathbb{P}_{tot}^1}{\partial^2 (Q^{tot})} = \left\{ \frac{1}{\gamma} \times \frac{1-\gamma}{\gamma} (Q^{tot})^{\frac{1-\gamma}{\gamma}-1} \times \sum_{j=1}^n p_j \right\} \tag{19}$$

Further, Eq. (19) is equated to 0 and we get:

$$\frac{1}{\gamma} \times \frac{(1-\gamma)}{\gamma} \times ((Q^{tot})^{\frac{1}{\gamma}-1-1}) \times \sum_{j=1}^n p_j = 0 \tag{20}$$

As the value of γ is positive, $Q^{tot} > 0$, and $p_j > 0$, Equation (20) gives us a positive value. Therefore, we conclude that \mathbb{P}_{tot} attain a minimum value.

We apply the *Langragian multiplier* technique [4] on Eq. (15), using Equations (9) and (13):

$$L_j = \left\{ \sum_{j=1}^n \mathcal{P}_j^{SCSP} + \sum_{j=1}^n ((Q_j)^{1/\gamma}) \times \sum_{j=1}^n p_j - \mu_1 (P_j^{SCSPmax} - P_j^{SCSP}) - \mu_2 (p_j^{max} - p_j) + \mu_3 \left[\left(\frac{n}{y} \right) - \left(\sum_{j=1}^n (Q_j)^{1/\gamma} \right) \right] \right\} \tag{21}$$

Further, we apply the *Karush–Kuhn–Tucker (KKT)* [5] conditions to solve Eq. (21). We obtain the *dual feasibility* and *complementary slackness* conditions as follows:

$$\mu_i \geq 0, \forall i = \{1, 2, 3\} \text{ and } \mu_i X_i = 0 \tag{22}$$

where X_i are the constraints as mentioned in Eqs. (9) and (13). To obtain the optimal value of Q^{tot} , we use partial derivative on Eq. (21) with respect to Q^{tot} and equate to 0. Therefore,

$$\frac{\partial L_i}{\partial Q^{tot}} = \frac{1}{\gamma} \times ((Q^{tot})^{\frac{1}{\gamma}-1}) \times \sum_{j=1}^n p_j - \mu_3 \text{ and } \frac{1}{\gamma} \times ((Q^{tot})^{\frac{1}{\gamma}-1}) \times \sum_{j=1}^n p_j - \mu_3 = 0 \tag{23}$$

Theorem 2 *The function proposed by in Equation (23) is convex and attains a minimum on domain Z.*

Proof Let the variables of function, $f(x)$, be in domain Z, which are subset of real numbers R^n , and is twice differentiable over a domain, Z. We say that function f is convex, if its double differentiation, $f''(x) > 0, \forall x \in Z$. Equation (23) represents *Langragian* function, and the variable of the function is real numbers and is in domain Z. We apply double differentiation on Eq. (23) and obtain:

$$\frac{\partial^2 L_i}{\partial^2 Q^{tot}} = \frac{1}{\gamma} \times \frac{(1-\gamma)}{\gamma} \times ((Q^{tot})^{\frac{1}{\gamma}-1-1}) \times \sum_{j=1}^n p_j \tag{24}$$

From, Eq. (24) as we know that our domain Z conditions are γ is positive, $((Q^{tot})^{\frac{1}{\gamma}-1-1}) > 0$ and $\sum_{j=1}^n p_j > 0$. The double differentiation on the Lagrangian function gives a positive value, and therefore, the function is convex.

Finally, we obtained optimal $(Q^{tot})^*$, as:

$$(Q^{tot})^* = \left(\frac{\mu_3 \times \gamma}{\sum_{j=1}^n p_j} \right)^{\frac{\gamma}{1-\gamma}} \tag{25}$$

The multiple sets of sensor nodes are able to participate in serving an end-user application. However, the service charge depends on the quality of the sensor nodes. Therefore, the SCSP offers an optimal QoS for the sets of sensor node with the corresponding total payable service price to the end-users. Further, as per the requirement, the end-user selects one of the sets of sensor nodes, depending on the optimal QoS and total payable service price among the available ones.

5 Performance Analysis

In this section, we analyze the performance of our proposed scheme, QSens, with a detailed explanation of the results. In order to simulate the performance of QSens, we consider the presence of 100–1000 sensor nodes with a simulation area of $10 \times 10 \text{ km}^2$. The values of different simulation parameters are listed in Table 1.

Figure 3 represents the change in the value of QoS with the variations of different parameters such as \mathcal{E}^{D} , \mathcal{E}^{I} , \mathcal{E}^{R} , and E^{eff} . Figure 3a depicts the effect on QoS for increasing value of the dichotomous variables from 2 to 10. We observe that the general trend of the plot is increasing with the increase in the number of dichotomous variables. However, we also observe that the average QoS does not depend on the variation of the number of sensor nodes. Similarly, Fig. 3b depicts the variations in the QoS with a change in effective interval variables. Interestingly, we observe that the average QoS is lesser, when the total number of sensor nodes is 500, than that in the presence of 1000 nodes. We observe in Fig. 3c that with the increasing value of the effective ratio variables, the general trend of QoS is increasing. However, the presence of the number of nodes in the network does not affect the variations of QoS. We also evaluate the variations in the QoS with the change in effective energy consumption. In Fig. 3d, we notice a smooth decreasing pattern in the QoS with increasing value of effective energy consumption. We also observed for all the values of effective energy consumption that the average QoS value is higher when the total number of nodes in the network is 1000 as compared to that of 500.

We also examine the total payable service price for an end-user in Fig. 4. Figure 4 depicts the variations in the total payable service price with the price of the sensor nodes (p^j), considering the price of SCSP as 500 units. In this figure, we observe an increasing trend in the total payable service price with the increment in the price of the sensor nodes in the presence of 500 and 1000 sensor nodes. Similarly, in Fig. 4, we consider the price of the SCSP as 1000 and evaluate the effect on the total payable service price. In this figure, we notice an increasing trend in the total payable service

Table 1 Simulation parameters

Parameter	Value
Number of sensor nodes	100–1000
Deployment	Uniform random
Effective dichotomous variables (\mathcal{E}^{D})	0–10
Effective normalized interval variable (\mathcal{E}^{I})	0–10
Effective ratio variable (\mathcal{E}^{R})	1–10
Effective energy (E^{eff})	0–1
Scaling factor (γ)	1–5
Price for sensor nodes	200–1000
Service charges for SCSP	200–1000

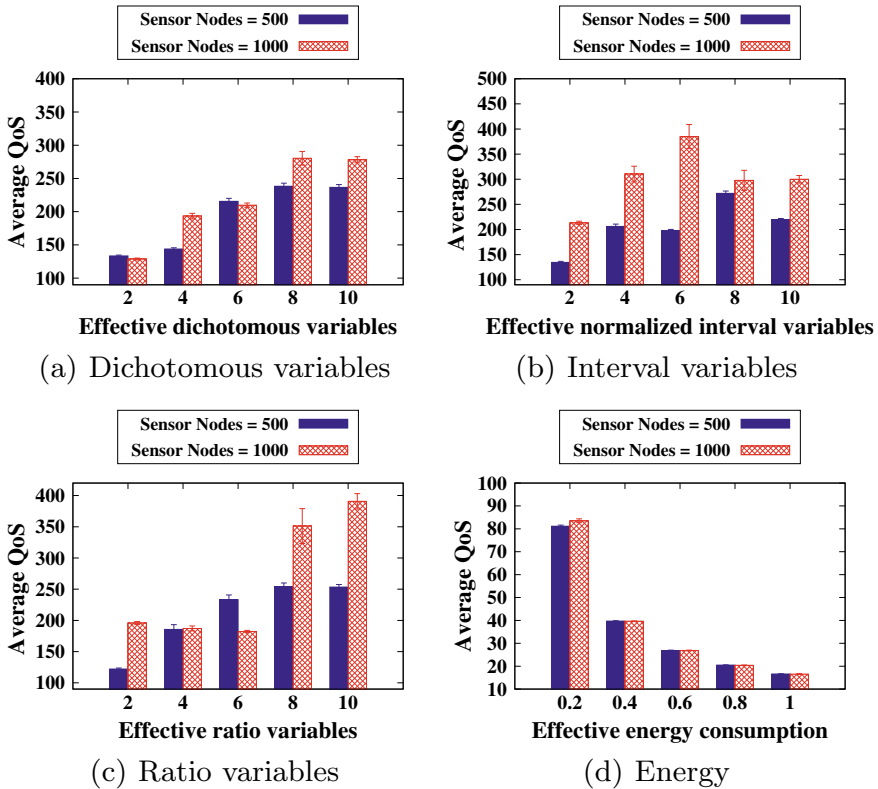
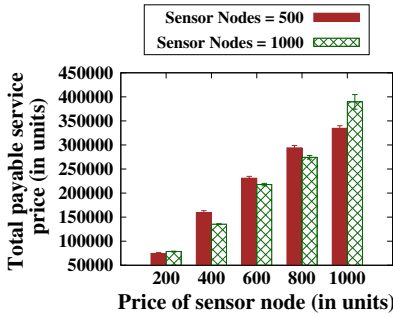


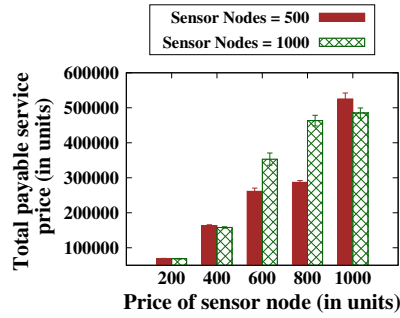
Fig. 3 Change in QoS

price with the increment of price of the sensor node from 200–1000. From Figs. 4 and 4, we infer that the total payable service price increases with the increase of the price of the sensor nodes, irrespective of the number of sensor nodes present in the network. We also evaluate the effect on total payable service price with the variation in the service charges of SCSP as depicted in Figs. 4 and 4. In Fig. 4, we vary the service charges of the SCSP from 200–1000 units and the price of the sensor nodes is fixed at 500 units. Similarly, Fig. 4 depicts the variations in the total payable service price with the increasing value of service charges of the SCSP from 200-1000 and the price of the sensor nodes is fixed at 1000 units. However, in both the Figs. 4 and 4, we do not find any specific standard trend in the plots. Therefore, we infer that the price of the sensor nodes has the primary effects on the total payable service price.

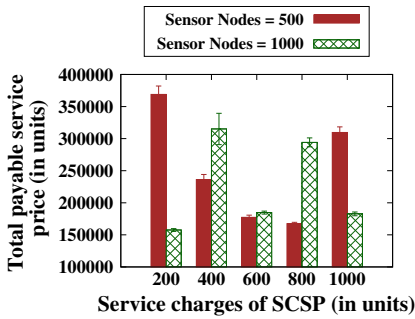
In Eq. (15), we use a variable γ as a scaling factor, which has significant effect on the total payable service price. Therefore, we analyze the variations in the total payable service price with change in the value of γ , as shown in Fig. 5a. For this analysis, we consider the presence of 200 nodes in the network. We observe that when the value of γ is 1, the total payable service price attains the maximum value



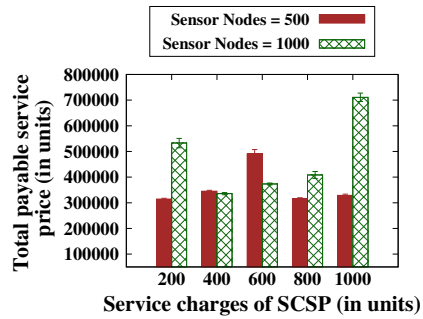
(a) Service price of SCSP = 500



(b) Service price of SCSP = 1000

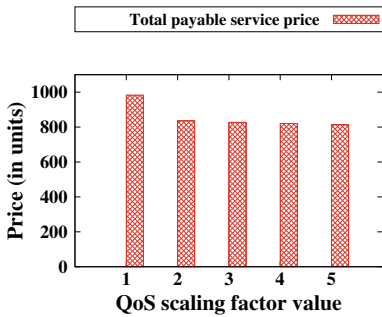


(c) Price of sensor nodes = 500

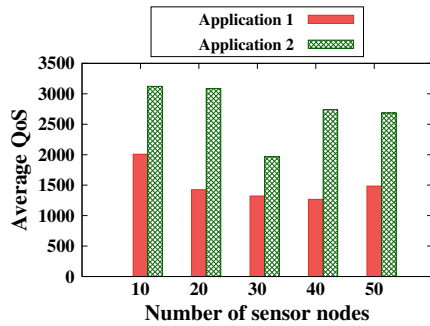


(d) Price of sensor nodes = 1000

Fig. 4 Change in total payable service price



(a) Price with scaling factor value



(b) QoS with number of nodes

Fig. 5 Change in pricing and average QoS

and decreases with the increasing value of γ . Figure 5b depicts the change in average QoS with the number of nodes in the networks in the presence of two applications. For this evaluation, we fixed the number of nodes to be 4 and 6 for the applications 1 and 2, respectively. We observe the average QoS of application 2 is higher as compared application 1. Therefore, we infer that the average QoS also depends on the total number of nodes used in an application.

6 Conclusion

In this work, we introduced the concept of SLAs for the sensor-cloud architecture for selecting a set of sensor nodes by the end-users. These SLAs provision the end-users to access the QoS of all the available sets of sensor nodes, which are suitable to serve an end-user application. We also designed a scheme, QoSens, which enables the end-users to select a suitable set of sensor nodes, based on the optimal QoS and the total payable service price, for serving their applications.

In future, we plan to extend the work by proposing a scheme for the QoS-based optimal resource allocation in sensor-cloud architecture and include as a component of SLA_{SS} . Further, we plan to design a data authentication scheme for the sensor-cloud architecture and include the same in SLA_{SE} .

References

1. Yuriyama, M., Kushida, T.: Sensor-cloud infrastructure - physical sensor management with virtualized sensors on cloud computing. In: Proceedings of the 13th International Conference on Network Based Information Systems (2010)
2. Roy, A., Misra, S., Dutta, P.: Dynamic pricing for sensor-cloud platform in the presence of dumb nodes. IEEE Trans. Cloud Comput. <https://doi.org/10.1109/TCC.2019.2950396>
3. Chakraborty, A., Mondal, A., Roy, A., Misra, S.: Dynamic trust enforcing pricing scheme for sensors-as-a-service in sensor-cloud infrastructure. IEEE Trans. Serv. Comp. p. 1 (2018)
4. Rockafellar, R.T. : Lagrange multipliers optimality. In: Society for Industrial and Applied Mathematics (SIAM) Review (1993)
5. Mokhtar, M.S., Bazaraa, S., Sherali, H.D.: Non Linear Programming Theory and Algorithms. Wiley (2006)
6. Gaillard, G., Barthel, D., Theoleyre, F., Valois, F.: Service level agreements for wireless sensor networks: A WSN operator's point of view. In: IEEE Network Operations and Management Symposium (2014)
7. Chieng, D., Marshall, A., Parr, G.: SLA brokering and bandwidth reservation negotiation schemes for QoS-aware Internet. IEEE Trans. Net. Serv. Manage. **2**(1), 39–49 (2005)
8. Garca, J.M., Fernandez, P., Pedrinaci, C., Resinas, M., Cardoso, J., Ruiz-Corts, A.: Modeling service level agreements with linked USDL agreement. IEEE Trans. Serv. Comp. **10**(1), 52–65 (2017)
9. Papatungan, I.V., Hani, A.F.M., Hassan, M.F., Asirvadam, V.S.: Real-time and proactive SLA renegotiation for a cloud-based system. IEEE Sys. J. **13**(1), 400–411 (2019)
10. Yuriyama, M., Kushida, T., Itakura, M.: A new model of accelerating service innovation with sensor-cloud infrastructure. In: Annual SRII Global Conference, pp. 308–314 (2011)

11. Madria, S.: Sensor cloud: Sensing-as-a-service paradigm. In: Proceedings of the 19th IEEE International Conference on Mobile Data Management, pp. 3–6 (2018)
12. Roy, C., Roy, A., Misra, S.: DIVISOR: Dynamic virtual sensor formation for overlapping region in IoT-based sensor-cloud. In: IEEE Wireless Communication and Network Conference, pp. 1–6 (2018)
13. Seltman, H.J.: Experimental Design and Analysis. stat.cmu.edu (2018)
14. Panda, S.K., Nag, S., Jana, P.K.: A smoothing based task scheduling algorithm for heterogeneous multi-cloud environment. In: International Conference on Parallel, Distributed and Grid Computing (2014)

Introduction to Adjacent Distance Array with Huffman Principle: A New Encoding and Decoding Technique for Transliteration Based Bengali Text Compression



Pranta Sarker and Mir Lutfur Rahman

Abstract Constructing a binary tree, the Huffman algorithm introduced the method of text compression that helps to reduce the size keeping the original message of the file. Nowadays, Huffman-based algorithm assessment can be measured in two ways; one in terms of space, another is decoding speed. The requirement of memory for a text file is going to be reasonable while the time effectiveness of Huffman decoding is being more significant. Meanwhile, this research is introducing the adjacent distance array with Huffman principle as a new data structure for encoding, and decoding the Bengali text compression using transliterated English text. Since the transliterated English text accommodates to reduce the unit of symbols accordingly, we transliterated the Bengali text into English and then applied the Huffman principle with adjacent distance array. By calculating the ASCII values, adjacent distance array is used to save the distances for each adjacent symbols. Apart from the regular Huffman algorithm, a codeword has produced by traversing the whole Huffman tree for a character in case, respectively adopting the threshold value and adjacent distance array can skip the lengthy codeword and perform the decoding manner to decode estimating the distances for all adjacent symbols except traversing the whole tree. Our findings have acquired 27.54% and 20.94% compression ratios for some specimen transliterated Bengali texts, as well as accomplished a significant ratio on different corpora.

Keywords Huffman · Data compression · Encoding · Decoding · Transliteration · Adjacent distance array

P. Sarker (✉) · M. L. Rahman

Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Kumargaon, Sylhet 3114, Bangladesh

e-mail: prantacse04@gmail.com

M. L. Rahman

e-mail: mirlutfur.rahman@gmail.com

1 Introduction

The research area to compress data or text has become more famous in new decades. Either computer science or the scope of data compression, the expeditious transference of data and less storage, the requirement of memory, and time efficiency are recognized as an indispensable part. Lossy and lossless are the two sorts of data compression while the Huffman principle is meant as a key methodology of lossless data compression domain. Appropriating the information of lossless data compression, [1] implemented a minimum redundant coding system that allows the variable code length to every input symbol, which again hooked into the occurrence of the symbols. Using this method, the modest code is generated for each maximum occurred symbol in a text, and therefore, the least occurred symbols get the highest code with a singular prefix for neglecting the ambiguity. There are many works associated with the enhancement of space efficiency in Huffman decoding, [2] represented such an array that requires $2n - 3$ memory with n number of symbols. Then, [3] modernized that array and decreased the size of memory to $\lceil 3n/2 \rceil + \lceil (n/2) \log n \rceil + 1$. Reducing the impact of sparsity occurs in the Huffman tree, which is grown on a particular side, [4] obtained the decoding time by $O(d)$. Lessening the memory from $O((n + d) \lceil \log 2n \rceil)$ bits to $O(2d \lceil \log 2n \rceil)$ bits, it also mitigated the necessity of additional memory offering by the ordering and clustering. Reference [5] reduced the memory requirement employing the corresponding data structure less than [4] significantly.

Later, the circular leaf nodes define a Huffman tree to diminish size of the memory to $\lceil 3n/2 \rceil$ as a new way introduced in [6]. Using the ternary-based adaptive Huffman tree, [7] reduced the memory space and accelerated the method to search a character in the Huffman tree. In terms to form, optimize the space and decoding time, there have plenty of researches supported on the lossless Huffman data compression method, many of the research executed good in space optimization and a certain amount of those well in performance of decoding time. By the use of mathematical representation to fulfill the Huffman tree based on recursion, [8] investigated the decoding performance of a Huffman tree where it decodes more symbols at a time that helps to enhance the decoding speed. However, it outperforms for small block sizes, while the requirement of a large memory was the drawback of this. In order to reach a balance between usage of memory and decoding time, [9] proposed a quaternary tree rather than a usual binary tree-based Huffman code. The tree has four branches constructed with 00 for left branch, 11 for right branch, 01 for left mid branch, and 10 for right mid branch. This methodology performed $O(\log_4 n)$ rather than $O(\log_2 n)$ for decoding in Huffman based algorithm. The tree generated the optimal codeword that helps to search in the Huffman tree effectively.

In terms of Bengali text, compression is not rich as well. Reference [10] given a proposal for dictionary-based Bengali text compression corpus name Ekushe-Khul considering the concept of Type to Token ratio (TTR) and Compression Ratio for only short and middle-sized files. Using Huffman principle, [11] made a balance between compression and decoding time for the Bengali short text message system.

They have implemented a static compression technique for such a device, i.e., a mobile, which has small memory and relatively low processing power to compress the short message of a text. Reference [12] proposed a new technique of compression for a more symbolic Bengali language using the shortest one English. They have used the transliteration mechanism to convert Bengali to English, which is a less symbolic language. On the transliterated language, they have shown applying the Huffman principle that the transliteration based methodology offers significantly by performing a compression ratio than the extant any conventional technique. Transliteration is the process of converting a text from the letters of one language to another that not a concern with designing the phonemics of genuine rather endeavors to express the characters correctly is called a graphemic conversion. Reference [13] has applied our approach only on any English texts and gain a significant improvement in decoding time, however, our research is to represent the adjacent distance array and Huffman principle with maintaining the stability of the compression ratio for Bengali text by combining transliterated English text compression.

2 Architecture

2.1 Process of Transliteration

In the transliteration approach, first, it represents each Bengali alphabet considering the graphemic conversion of the English alphabets. The major advantage of replacing more symbolic language into smaller ones is to enhance the compression ratio found in [12]. For example, we need a single lower or upper case English character to express every Bengali character (e.g. we may represent the Bengali letter অ by English letter o, or ঙ by I). On the other hand, we may also represent a Bengali character composed by one more than English letters (e.g. Bengali letter ঙ is represented by English letters Th, ঙ by ng, or ঞ by rri). Sometimes, we need a set of characters from the ASCII character set (e.g. ঁ is represented by ^) to express a Bengali letter. However, it is strengthened that we will have required the incorporation of English and ASCII characters set to design any Bengali document. Therefore, expressing 65 Bengali symbols, it will use 39 some ASCII, or English characters, or their aggregation. This is reflected in the Avro Phonetic Layout in Fig. 1. This method will achieve more efficiency by improving the frequency of occurrences for each English symbol comparing the Bengali.

The requirements of the bit are calculated by the approach [14] in terms of a total of 65 Bengali characters:

$$N_T = N_I + N_E = 64 + 65 = 129$$

Here N_T used to measure the whole amount of nodes that can be calculated from the sum of internal nodes N_I and external nodes N_E . The subsequent formula is used to

ক	k	ট	T	প	p	স	s	অ	o	ও	ৌ	OU	০	0
খ	kh	ঠ	Th	ফ	ph,f	হ	h	আ	।	ব	(ফলা)	w	১	1
গ	g	ড	D	ব	b	ড়	R	ই	ি	ঈ	- য ফলা	(c)y, Z	২	2
ঘ	gh	ঢ	Dh	ভ	bh,v	ঢ়	Rh	ঋ	ী	ঌ	- র ফলা	(c)r	৩	3
ঙ	Ng	ণ	N	ম	m	য়	y,Y	উ	ু	ঊ	- রেফ	(v)rr(c)	৪	4
চ	c	ত	t	য	z	ৎ	t`	ঊ	ু	ঋ	- হসন্ত	,,	৫	5
ছ	ch	থ	th	র	r	ং	ng	ঋ	ু	ঌ	- দাড়ি	.	৬	6
জ	j	দ	d	ল	l	ঃ	:	এ	ৈ	চ	- টাকা	\$	৭	7
ঝ	jh	ধ	dh	শ	sh,S	ৎ	^	ঐ	ৈ	.	- ডট	(NumPad)	৮	8
ঞ	NG	ন	n	ষ	Sh	জ	J	ও	ৌ	:	(কোলন)	:	৯	9

Fig. 1 Avro phonetic layout

measure the depth of the tree:

$$Depth = \text{Floor}(\log_2(N_T + 1)) = \text{Floor}(\log_2(129 + 1)) = 8$$

Therefore, representing 65 symbols in Bengali text required the number of bits that can be measured as following:

$$N_B = 65 \times \text{Largest Level} = 65 \times (Depth - 1) = 65 \times (8 - 1) = 455 \text{ bits}$$

On the other hand, for 39 English symbols, we can calculate the requirement of bits correspondingly as before:

$$N_T = N_I + N_E = 38 + 39 = 77$$

$$Depth = \text{Floor}(\log_2(N_T + 1)) = \text{Floor}(\log_2(77 + 1)) = 7$$

$$N_E = 39 \times \text{Largest Level} = 39 \times (Depth - 1) = 39 \times (7 - 1) = 234 \text{ bits}$$

Hence, the resulting equation is used to calculate the compression ratio (r):

$$r = F(N_B, N_E) = \left(\frac{N_B - N_E}{N_B} \right) \times 100\% = \frac{455 - 234}{455} \times 100\% = 48.57\%$$

Most of the Huffman trees are belonging to a 2-tree (a tree, which has 0 or 2 child) or an extended tree of binary, though the Huffman tree remains not a complete binary tree ever. Nevertheless, if we want to represent the equivalent Bengali text into English, the reduction may happen for the minimization of a certain amount of characters by $26(65 - 39 = 26)$.

2.2 *Compression Using Huffman Principle and Adjacent Distance Array*

After transliterating Bengali text into English, this research used a modified compression technique based on the Huffman principle with adjacent distance array that is applied on [13] instead of the traditional Huffman approach for any English text to optimize the decoding time during the decoding process.

The regular Huffman technique generates the shortest code for the foremost recurrent symbols, whenever the codeword for the smallest recurring symbols expand sequentially. It often takes ample time to see the encoded file for being comprehensive in the decoding manner. This research has employed a threshold value and an adjacent distance array to recognize the adjacent characters and put those distances, respectively. The adjacent symbols will have $S_{i+1} + S_{i+2} + \dots + S_{i+m}$ for a certain symbol S_i , if the distance between S_{i+1} and S_i is equal or less than the threshold value T .

$$|S_{i+1} - S_i| \leq T \quad (1)$$

If the condition is satisfied by S_i , the encoded file and adjacent distance array will save the Huffman generated code and the distances for all satisfied adjacent symbols $S_{i+1} + S_{i+2} + \dots + S_{i+m}$ respectively. However, an appropriate symbol will be considered as the latest symbol in the encoded file for any dissatisfied condition, and this process is recommended for all the upcoming adjacent symbols. A separator bit '0' will have existed on the encoded file for a corresponding symbol specifically, in the adjacent distance array.

The distances placed on the adjacent distance array are interpreted by a distinct coding pattern that depends upon the threshold value T . Every code can generate the two different types of bit models.

First of the type is constructed by the pattern of 2 bits:

- There always will be the first bit is '1' means the starting code bit.
- The next code bit will be '0' or '1' used to state whether the value of distances is negative or positive.

The second type is responsible to represent the distances in binary, which may acquire the identical unit of bits, and it is equal to the highest bit illustration of the threshold value T . This is computed from this conception as:

$$T = 2^x - 1 \quad (2)$$

Here, x is indicating the number of bits are expected to present the threshold value T . Moreover, the equation is always performed by the threshold value according to the method. Thus, the encoded file demanding the memory is:

$$\sigma_1 = \sum_{i=1}^N (F_i - A_i) \times C_i \quad (3)$$

Here,

F_i = total frequency of occurrence for a symbol.

A_i = the frequency is lessened from the encoded file into adjacent distance array.

C_i = code bits from Huffman coding scheme.

N = total number of symbols in the encoded file.

The depiction of memory for an adjacent distance array is:

$$\sigma_2 = \sum_{i=1}^M (A_i \times x) \quad (4)$$

Here,

x = quantity of bits needed to interpret the distance.

M = the entire number of distances for adjacent symbol.

Hence, there is expected memory to cache the full information is:

$$\sigma = \sigma_1 + \sigma_2 + H_T + S_N \quad (5)$$

Here,

H_T = the header of Huffman tree.

S_N = measurement the whole separators in adjacent distances.

In decoding time, at first, this method is used to decode the first character from the particular encoded file. By evaluating the distances corresponding to the adjacent distance array, the subsequent adjacent symbol can be decoded, accordingly, this process will have run until obtaining any separator bit or leads the end of the array of adjacent distance. The process literally bypasses the traversing entire Huffman code list for any adjacent symbols.

Assume an encoded string of text composed by the alphabet of k symbols with the size of n length. To decode any encoded symbol there is needed to traverse the entire binary tree according to the regular Huffman principle, which process may obtain the time complexity is $O(n \log_2 k)$, if the tree includes k nodes on medium. However, following the proposed approach has gained the complexity $O(((n - a) \log_2 n) + a)$ by performing only arithmetic progression that gets $O(1)$ rather not visiting the all nodes of adjacent symbols for decoding. This is to be noted, a is used as the code length of whole symbols of adjacent to calculate the complexity.

3 Implementation

We have given a conceptualize analysis for the process of transliteration of Bengali known as a more symbolic language to less symbolic language English, as well as

discussed a qualified compression approach utilizing the adjacent distance array with Huffman principle. With the reduction of encoding symbol, the method is applied to decode the characters having the symbols of adjacent in consideration of threshold value (T), while it does not visit the entire code list or Huffman tree. The implementation of the introduced methodology has two processes:

- The process of Encoding, and
- The process of Decoding.

3.1 Explanation of Encoding Process

To describe the process of encoding, the proposed approach allows the Huffman generated code list (L) along with adjacent distance array (*adjacent*), threshold value (T), and the transliterated English text (E) for corresponding Bengali.

Beginning of the encoding process, a Bengali text (B) is taken as an input and transliterates into corresponding graphemic English text (E). It inserts a threshold value (T) and forms a code list (L) of symbols relating the Huffman principle in the next step. Accordingly, the process of encoding will save a code that generated from the Huffman into the encoded file from the transliterated English text for any particular symbol (S_i) to check the distances of adjacent symbols considering the threshold value (T) for that Huffman provided code. Besides, this process also helps to take decision between Huffman code length and adjacent distance whether the code length for adjacent distance is minimum or not. Fulfillment of the mentioned cases, the distances of code will be saved into the adjacent distance array. Whatever the mismatching any situation an adjacent symbol will be considered as a fresh symbol on the encoded file, and there will be the same occurrence by reforming the system. For a particular symbol, the adjacent distance array uses a separator bit '0' after accumulating all the distances of adjacent symbols. The process will be continued until to move the edge of the English text. We will get two coded files after terminating the encoding process:

- Encoded codes file, and
- Distance array of adjacent codes.

Encoding Algorithm:

```

1: Take a Bengali text ( $B$ ) as an input
2: Transliterate the corresponding Bengali into English text ( $E$ )
3: Set threshold value ( $T$ ) and generate the Huffman code list ( $L$ )
4: for  $i \leftarrow 0$  to  $size(E)$  do
5:      $S_i \leftarrow E[i]$ 
6:     if  $i = 0$  then
7:          $encoded \leftarrow encoded \cup \{S_i\}$ 
8:          $previous \leftarrow S_i$ 
9:     else
10:        if  $distance(S_i, previous) \leq T$  and  $size(L_i) \geq size(T)$  then
11:             $adjacent \leftarrow adjacent \cup \{(distance(S_i, previous))\}$ 
12:        else
13:             $encoded \leftarrow encoded \cup \{S_i\}$ 
14:             $previous \leftarrow S_i$ 
15:        end if
16:    end if
17: end for
18: exit

```

► reach the size of Encoded file

► distance between present and past symbol will be saved into *adjacent*

3.2 Explanation of Decoding Process

The desired encoded file (*encoded*) along with Huffman created a code list (L), and adjacent distance array (*adjacent*) will be taken to the beginning of the decoding process. This process is implemented to decode all of the adjacent symbols evaluating the distances and either unless it reaches the final range of the adjacent distance array or receives a separator bit and so on. Hereafter, decoding the English text will convert the graphemic Bengali using the transliteration process.

Decoding Algorithm:

```

1: Input encoded, adjacent and Huffman code list (L)
2: for i ← 0 to size(encoded) do
3:     e ← encoded[i]
4:     if search(e, L) = true then
5:         decoded ← decoded ∪ {L.symbol}
6:         for j ← 0 to size(adjacent) do
7:             if adjacent[j] = 0 then
8:                 break
9:             else
10:                adjs ← adjs ∪ {(char) distance(L.symbol, adjacent[j])}
11:                decoded ← decoded ∪ {adjs}
12:            end if
13:        end for
14:    else
15:        i ← i + 1
16:        e ← e.concat(encoded[i])
17:    end if
18: end for
19: exit

```

▶ search *e* on the code list *L*
▶ the corresponding symbol from code list *L* will be written into decoded file

4 Analysis and Discussion

Here, we have applied the transliteration process in sample string 1 and 2 stated in Table 1.

We observed that the Bengali string requires more bits considering the symbols of the Unicode coding scheme, whereas the English string requires a less significant amount of bits in terms of ASCII. Therefore, representation of both Bengali strings requires 1122 and 2965 Huffman bits and transliterated English strings 813 and 2344 bits. Table 2 has recorded the result of our review. Besides, from Table 2, it is observable that approving the transliterated method, we may attain a much better compression than any other regular Huffman technique at present. The ratio of compress data will be enhanced when the size of the data constantly expands. From that analysis, we can also observe that regular Huffman has enough ability to compress a notable volume of data, though the most reliable performance can be managed by applying the Huffman principle on the transliterated string after transliteration of the Bengali to English string.

In Table 2, we have observed the number of Huffman bits for transliterated English considering the traditional Huffman technique, but using the modified Huffman-based technique with adjacent distance array performs well in terms of compression ratio.

We have compared both in an experiment that streamed on different some popular English corpora. The target is to maintain the compressed file size as well as the

Table 1 Sample strings

String	Language	Sample text	Total symbols	Distinct symbols	Unicode/ASCII bits
1	Bengali	আমার সোনার বাংলা, আমি তোমায় ভালবাসি। চিরদিন তোমার আকাশ, তোমার বাতাস, ও মা আমার প্রানে বাজায় বাঁশি।	98	28	$98 \times 16 = 1568$
	English	amar sOnar bangla, ami tOmay valobasi. cirdin tOmar akaS, tOmar batas, O ma amar prane bajay ba^Si	98	25	$98 \times 7 = 686$
2	Bengali	আমার সোনার বাংলা ও মা, আমি নয়ন জলে ভাসি।	404	42	$404 \times 16 = 6464$
	English	amar sOnar bangla O ma, ami noyon jole vasi	431	34	$431 \times 7 = 3017$

Table 2 Overall compression using transliteration

String	Number of Unicode bits (original Bengali string)	Number of Huffman bits (Bengali string)	Number of Huffman bits (Transliterated English string)	Compression ratio (%) = $\left(\frac{HB-T HB}{HB}\right) \times 100$
1	1536	1122	813	27.54%
2	6464	2965	2344	20.94%

compression ratio for English text in consideration of the original file size. The following environment was related to stream the experiment: Language: C++, OS: Linux (Ubuntu 14.04 LTS) 64-bit, Graphics: Gallium 0.4 on llvmpipe (LLVM 3.4, 256bits), IDE: CodeBlocks 16.01, Processor: Intel(R) Core(TM) i5-6500 CPU @ 3.20 GHz \times 4, and Memory: 7.7GiB. The approach of adjacent distance array with Huffman principle, here, two different types of threshold values have used for the experiment on different corpora. Threshold value $T = 7$ and $T = 15$.

Table 3 presents the experimental result of compressed file size and the ratio of compression for the Canterbury Corpus. The size of the Canterbury Corpus [15] is 1158.08 Kilobytes (KB) and it contains 87 distinct English characters.

The result of Table 3 indicates that the execution of compression for threshold value $T = 7$ is comparably better than the other one $T = 15$.

For the Supara Corpus [16], Table 4 is shown that the Huffman-based method with adjacent distance array that has the threshold value, $T = 7$ outperformed than $T = 15$ considering the compression ratio and it much compressed the original file.

Table 3 Compression performance using proposed method on Canterbury Corpus (1158.08 KB)

Source file	Original size (OS) in Kilobytes (KB)	Proposed approach/algorithm	Compressed size (CS) in Kilobytes (KB)	Compression ratio (%) = $\left(\frac{OS-CS}{OS}\right) \times 100$
The Canterbury Corpus [15]	1158.08	The modified approach with (T = 7)	788.18	31.94%
		The modified approach with (T = 15)	796.97	31.18%

Table 4 Compression performance using proposed method on Supara Corpus (1464.21 KB)

Source file	Original size (OS) in Kilobytes (KB)	Proposed approach/algorithm	Compressed size (CS) in Kilobytes (KB)	Compression ratio (%) = $\left(\frac{OS-CS}{OS}\right) \times 100$
The Supara Corpus [16]	1464.21	The modified approach with (T = 7)	989.04	32.45%
		The modified approach with (T = 15)	1002.35	31.54%

Table 5 is introduced the Brown Corpus [17] that is 6040.63 Kilobytes (KB) and contains 95 distinct English characters.

Therefore, the result of Table 5 indicating the threshold value T = 7 has more capability to achieve an impressive compression ratio by compressing the size of original source file than the threshold value T = 15.

All of the experimental outcomes along with Fig. 2 determines that the execution of compression based on adjacent distance array with the Huffman law is comparatively work better when we pick the minimum threshold value (T = 7); that indicating a specific amount of adjacent symbols are not existing as much. However, the size

Table 5 Compression performance using proposed method on Brown Corpus (6040.63 KB)

Source file	Original size (OS) in Kilobytes (KB)	Proposed approach/algorithm	Compressed size (CS) in Kilobytes (KB)	Compression ratio (%) = $\left(\frac{OS-CS}{OS}\right) \times 100$
The Brown Corpus [17]	6040.63	The modified approach with (T = 7)	4037.61	33.15%
		The modified approach with (T = 15)	4093.39	32.23%

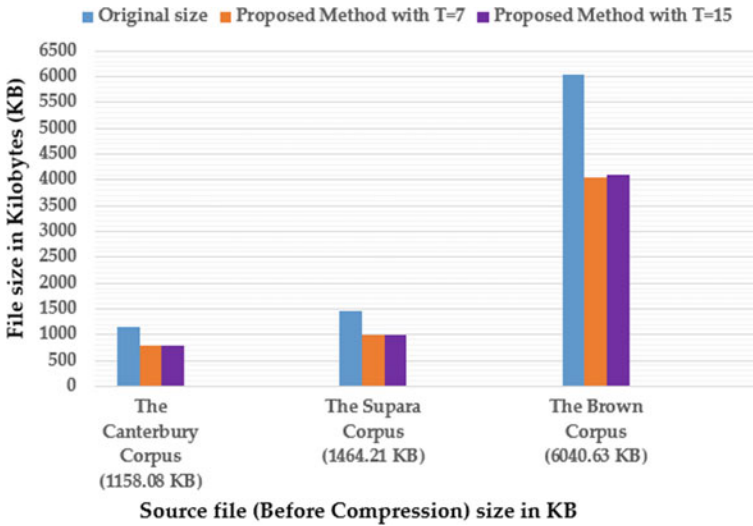


Fig. 2 Chart of compression performance using the proposed method on different corpora

of the encoded file may be enlarged if there may exist a less amount of adjacent symbols. Now, we may decide from that analysis, altering the threshold value, it may create a significant effect on the performance of data compression. Finally, it is remarkable that in the case of more symbolic language Bengali combining the process of transliterated English text and Huffman principle using adjacent distance array can improve a notable amount in compression ratio by compressing the size of original source file.

5 Conclusion

In this research, we have concentrated on the Huffman based lossless data compression designs, more specifically the Bengali language that is much symbolic. Here, we have combined the transliteration approach and Huffman based method with an adjacent distance array. We performed the transliteration approach upon Bengali text to get corresponding English and hence implement the modified Huffman based method on the transliterated text. The performance of the proposed approach is shown in consider of compression ratios. In the experiment, we observed that our proposed approach achieved significant results in the enhancement of compression by applying the modified Huffman based adjacent distance array method on the transliterated text. Our proposed strategy has been explained and implemented to all Bengali texts exclusively, though the basic idea could be used to inspect for all any other languages, moreover, the language which has more characters compared to the English language.

References

1. Huffman, D.A.: A method for the construction of minimum redundancy codes. *Proc. IRE* **40**(9), 1098–1101 (1952)
2. Chung, K.L., Lin, Y.K.: A novel memory-efficient Huffman decoding algorithm and its implementation. *Signal Process.* **62**(2), 207–213 (1997)
3. Chen, H.C., Wang, Y.L., Lan, Y.F.: A memory efficient and fast Huffman decoding algorithm. *Inf. Process. Lett.* **69**(3), 119–122 (1999)
4. Hashemian, R.: Memory efficient and high-speed search Huffman coding. *IEEE Trans. Commun.* **43**(10), 2576–2581 (1995)
5. Lin, Y.K., Chung, K.L.: A space-efficient Huffman decoding algorithm and its parallelism. *Theor. Comput. Sci.* **246**(1–2), 227–238 (2000)
6. Chowdhury, R.A., Kaykobad, M., King, L.: An efficient decoding technique for Huffman codes. *Inf. Process. Lett.* **81**(6), 305–308 (2002)
7. Suri, P., Goel, M.: Ternary Tree and memory-efficient Huffman decoding algorithm. *Int. J. Comput. Sci. Issues* **8**(1), 483–489 (2011)
8. Lin, Y.K., Huang, S.C., Yang, C.H.: A fast algorithm for Huffman decoding based on a recursion Huffman tree. *J. Syst. Softw.* **85**(4), 974–980 (2012)
9. Habib, A., Rahman, M.S.: Balancing decoding speed and memory usage for Huffman codes using quaternary tree. *Appl. Inf.* **4**(1), 1–15 (2017)
10. Islam, M.R., Rajon, S.A.: On the design of an effective corpus for evaluation of Bengali text compression schemes. In: 11th International Conference on Computer and Information Technology, pp. 236–241. *IEEE Xplore*, Khulna, Bangladesh (2008)
11. Arif, A.S.M., Mahamud, A., Islam, R.: An enhanced static data compression scheme of Bengali short message. *Int. J. Comput. Sci. Inf. Secur.* **4**(1–2), 97–103 (2009)
12. Hossain, M.M., Habib, A., Rahman, M.S.: Transliteration based bengali text compression using huffman principle. In: 2014 International Conference on Informatics, Electronics and Vision, pp. 1–6. *IEEE Xplore*, Dhaka, Bangladesh (2014)
13. Rahman, M.L., Sarker, P., Habib, A.: A faster decoding technique for Huffman codes using adjacent distance array. In: Proceedings of International Joint Conference on Computational Intelligence, pp. 309–316. *Algorithms for Intelligent Systems*, Dhaka, Bangladesh (2019)
14. Lipschutz, S.: *Data Structures with C*, 1st edn. McGraw-Hill, New York (2010)
15. The Canterbury Corpus, <https://corpus.canterbury.ac.nz/resources/cantrbry.zip>, last accessed 2020/01/25
16. The Supara Corpus, <https://github.com/mir1234/Supara-Corpus/>, last accessed 2020/01/25
17. The Brown Corpus, <https://ia800306.us.archive.org/21/items/BrownCorpus/brown.zip>, last accessed 2020/01/25

BSAT: A New Tool for Analyzing Cryptographic Strength of Boolean Function and S-Box of Symmetric Cryptosystem



Pratap Kumar Behera and Sugata Gangopadhyay

Abstract The cryptographic primitives such as Boolean function and S-Box are used as a building block to design stream cipher and block cipher, respectively. However, it is important to evaluate the cryptographic strength to measure the security of such cryptosystem resistance against important cryptanalytic attacks. We develop a new tool called Boolean function and S-Box analysis tool (BSAT) to evaluate the cryptographic strength of the Boolean function and S-Box. As the size of S-Box increases, it takes a lot of computational effort to evaluate all the required cryptographic properties. In this paper, we develop a tool to evaluate important cryptographic properties of both Boolean function and S-Box to minimize the computational time. The computational time of BSAT is lesser than the SET tool in terms of calculating the algebraic normal form and algebraic degree. Our BSAT takes only 0.404 s to evaluate the AES properties faster than the SET, which takes 650 millisecond (0.65 s).

Keywords Stream cipher · Block cipher · Boolean function · S-Box · Nonlinearity

1 Introduction

The security of symmetric cryptosystem such as stream cipher and block cipher depends upon the cryptographic strength of the primitives used in the cipher. The cryptographic primitives, which are used as a building block of the stream cipher and block cipher, are Boolean functions and S-Boxes, respectively [1, 2]. To measure the security of such cipher, the cryptographic properties need to be evaluated against a various attacks. The two most important cryptanalytic attacks on block

P. K. Behera (✉) · S. Gangopadhyay
Department of Computer Science and Engineering,
Indian Institute of Technology Roorkee, Roorkee, India
e-mail: pbehera@cs.iitr.ac.in

S. Gangopadhyay
e-mail: sugata.gangopadhyay@cs.iitr.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_46

557

cipher are linear approximation attack [3] and differential attack [4]. The cryptographic properties resistance to both these attacks are nonlinearity and differential uniformity, respectively [1, 5]. Apart from these two properties, there are other cryptographic properties, which also play a major role for the security of block cipher. It is necessary to evaluate the cryptographic strength of Boolean functions and S-Boxes to measure the security of stream cipher and block cipher, respectively. Since the Boolean function and S-Box used as a cryptographic primitives, we need to evaluate the cryptographic properties of such cryptographic primitives to measure security against various cryptanalytic attacks. There are various tools available in public to analyze the cryptographic strength of Boolean function and S-Box. These tools are explained in a detailed manner as follows:

Boolfun package in R: The *Boolfun* package is developed using R language, and the important cryptographic properties are evaluated in this paper [6, 7]. The *Boolfun* package does not have a support for evaluating S-box properties.

Boolean function and S-Box in Sage: Sage is a free and open source mathematical tool, which is developed on top of many existing open source Python package such as NumPy, SciPy, Sympy, Maxima and many more [8]. There are two different modules in sage, i.e., Boolean function and S-Box. The *Boolean function* module evaluates all the required cryptographic properties of the Boolean functions. For S-Box, the corresponding module called *SBox* evaluates the cryptographic properties resistance to the linear approximation attack and differential attack. The module *SBox* takes the input in integer form and calls the corresponding function for evaluation. There are different optimized approaches for evaluating properties to reduce the computational complexity. The sage tool has no freedom to see the exact implementation details, and the user can only call to the predefined function.

SET: The SET tool is developed using ANSI C programming language [9]. This tool evaluates the cryptographic properties of the Boolean function and S-Box resistance to linear approximation attack, differential attack and side channel attack [10–12]. This tool has limited number of functionalities for evaluating the S-Box properties and also takes more time for evaluation.

In this paper, we develop a new tool called Boolean function and S-Box analysis tool (BSAT) to analyze the cryptographic strength of Boolean function and S-Box. Since the construction of such cryptographic primitives can be formulated as an combinatorial optimization problem [13–15], researcher are actively involved to find the optimal Boolean functions and S-Boxes using heuristic search [16–18]. Since the optimization problems take a lot of computational effort, the performance of the algorithm depends upon the time taken to evaluate the fitness function. The fitness function is designed based upon the cryptographic properties to be optimized. So, we develop a tool using C++ with optimized version of some original approach for evaluating the cryptographic properties. Our tools cover a wide range of cryptographic properties, and it takes less computational time than the existing tools. Our BSAT tool is developed using C++ STL, and it is publicly available in GitHub for cryptographic researcher. Since BSAT is an open source tool, the researcher can have complete freedom to customize the code in a more optimized way or can add new

functionalities. Our BSAT is different and more user friendly than the existing tools in the following scenario:

- (1) The state-of-the-art algorithm is implemented in a precise manner.
- (2) The BSAT tool is available on GitHub repository (<https://github.com/pratapiitr/BSAT>), which can be further customized by adding new properties or optimizing existing algorithm.
- (3) Since the BSAT tool is on public domain, so the user has complete freedom to know the exact implementation details of all the functionalities.

The remainder of this paper is organized as follows: In Sect. 4, we discuss about the Boolean function and S-Box definition, its representation and cryptographic properties. In Sect. 5, we present the practical implementation of mentioned cryptographic using optimized approach. In Sect. 6, we explain the BSAT execution, analyze the performance of the tool and future work. Finally, Sect. 7 presents the conclusion.

2 Preliminaries

In this section, we describe the definition of Boolean function and S-Box, its representation and cryptographic properties resistance to various cryptanalytic attacks.

2.1 Boolean Function

The Boolean function is defined as mapping of an n -bit input vector to one bit output vector defined over the finite field \mathbb{F}_2 . Mathematically, it is defined as follows:

$$f : \mathbb{F}_2^n \rightarrow \mathbb{F}_2$$

There are three ways to represent the Boolean function such as truth table form, algebraic normal form and Walsh–Hadamard transform. The truth table is represented using (0, 1) sequence.

$$(f(x_0), f(x_1), \dots, f(x_{2^n-1}))$$

The ANF of any Boolean function f is represented as follows:

$$f(x_1, \dots, x_n) = \bigoplus_{u \in \mathbb{F}_2^n} \lambda_u \left(\prod_{i=1}^n x_i^{u_i} \right) \quad (1)$$

where $\lambda_u \in \mathbb{F}_2$ and $u = (u_1, u_2, \dots, u_n)$.

The Walsh–Hadamard transform of Boolean function f is a real-valued function, i.e., $W_f : \mathbb{F}_2^n \mapsto \mathbb{R}$, and it is defined as follows [19]:

$$W_f(u) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x)+u \cdot x} \quad (2)$$

where $u \in \mathbb{F}_2^n$ and $u \cdot x$ is the dot product of two input vector. The dot product is defined as $u \cdot x = u_1x_1 \oplus u_2x_2 \oplus \dots \oplus u_nx_n$.

2.2 Cryptographic Properties of Boolean Function

The cryptographic strength of Boolean function denotes the resistance against corresponding cryptanalytic attack [20, 21] on the stream cipher. We list out only important cryptographic properties in this paper.

Balancedness: The balancedness property can be checked either calculating the number of zeros and ones present in the entire sequence or through the Walsh–Hadamard transform. If the number of zeros and ones is equal, then the function is said to be balanced.

Nonlinearity: The nonlinearity properties are used to measure the security against the linear approximation attack. The higher nonlinearity denotes the more secure is stream cipher. The minimum hamming distance of the Boolean function to the set of all affine functions defined over the finite field \mathbb{F}_2 is said to be nonlinearity of the Boolean function. Mathematically, it is denoted as NL_f and defined as follows:

$$NL_f = \min_{g \in \mathcal{A}_n} d_H(f, g) \quad (3)$$

where \mathcal{A}_n is the set of all affine function defined over \mathbb{F}_2^n and $d_H(f, g)$ is the Hamming distance between f and g . The Hamming distance between two Boolean functions f and g is defined as follows:

$$d_H(f, g) = \#\{x \in \mathbb{F}_2^n | f(x) \neq g(x)\}$$

Correlation Immunity: A Boolean function f is correlation immune of order m ($CI(m)$), if its output bits are statistically independent of any combination of m input bits [21, 22]. Mathematically, if $W_f(u) = 0$, $1 \leq wt(u) \leq m$ and $u \in \mathbb{F}_2^n$, then the Boolean function satisfies the correlation immunity of order m . The Boolean function which is both balanced and satisfies the correlation immunity of order m is called m -resilient Boolean function.

Autocorrelation: The autocorrelation of a Boolean function f denoted as AC_f is defined as

$$AC_f = \max_{\alpha \in \mathbb{F}_2^n \setminus \{0\}} |C_f(\alpha)| \quad (4)$$

where

$$C_f(\alpha) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x)+f(x+\alpha)}$$

and $\alpha \in \mathbb{F}_2^n$.

Algebraic Degree: The polynomial representation of a Boolean function f is called algebraic normal form (ANF) [22], and it is defined as follows:

$$f(x) = \sum_{\tau \in \Omega_f} \left(\prod_{i=1}^n x_i + \tau_i + 1 \right)$$

Where, $\tau = (\tau_1, \tau_2, \dots, \tau_n)$ The algebraic degree of Boolean function f is the number of variables of highest monomial having nonzero coefficient in algebraic normal form.

Strict Avalanche Criteria Whenever a single input bit is complemented, and if the output bit changes with a probability of one half, then the Boolean function satisfies the strict avalanche criteria [23].

$$\sum_{x \in \mathbb{F}_2^n} f(x) + f(x + a) = 2^{n-1} \tag{5}$$

For all $a \in \mathbb{F}_2^n$ such as $wt(a) = 1$.

2.3 S-Box Definition and Representations

The mapping of an n -bit input vector to m -bit output vector defined over the finite field \mathbb{F}_2^n is called vectorial Boolean function or (n, m) -function or $n \times m$ S-Box [1]. Mathematically,

$$F : \mathbb{F}_2^n \rightarrow \mathbb{F}_2^m$$

Let $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_m)$ and $F(x_1, x_2, \dots, x_n) = (y_1, y_2, \dots, y_m)$ then we can write $F(x) = y$, where each $y_i = f_i(x_1, x_2, \dots, x_n)$ ($1 \leq i \leq m$). Each $f_i : \mathbb{F}_2^n \rightarrow \mathbb{F}_2$ is a single output Boolean function. The sequence of m single output Boolean function is also known as vectorial Boolean function, and it is represented as $F = (f_1, f_2, \dots, f_m)$. The single output Boolean function f_1, f_2, \dots, f_m are called coordinate function, and their nonzero linear combination is called component function.

The Walsh–Hadamard transform of an (n, m) -function is a real-valued function, and it is defined as follows:

$$W_F(u, v) = \sum_{x \in \mathbb{F}_2^n} (-1)^{v \cdot F(x) + u \cdot x} \tag{6}$$

For all $u \in \mathbb{F}_2^n$ and $v \in \mathbb{F}_2^{m*}$, the Walsh–Hadamard spectrum is a $2^m - 1 \times 2^n$ matrix, and each value of the matrix lies between $[-2^n, 2^n]$.

2.4 Cryptographic Properties of S-Box

Bijjective/Balancedness: An (n, m) -function is called bijective, when all the component functions satisfy balancedness property, i.e., if $w_t(v \cdot F(x)) = 2^{n-1}$, for all $v \in \mathbb{F}_2^{m*}$. The balancedness property can also be checked by computing the Walsh–Hadamard transform. If $W_F(0, v) = 0$, for all $v \in \mathbb{F}_2^{m*}$, then the S-Box is bijective.

Nonlinearity: The nonlinearity of an $n \times m$ S-Box F denoted as NL_F is defined as the minimum Hamming distance of all component functions to the set of all affine function defined over \mathbb{F}_2^n . Mathematically,

$$NL_F = \min_{v \in \mathbb{F}_2^{m*}, g \in \mathcal{A}_n} d_H(v \cdot F, g) \tag{7}$$

Differential Uniformity: The differential uniformity of an $n \times m$ S-Box F , denoted as δ_F , is defined as

$$\delta_F = \max_{a \in \mathbb{F}_2^{m*}, b \in \mathbb{F}_2^m} \delta_F(a, b) \tag{8}$$

where a , and b are called input and output difference, respectively, and $\delta_F(a, b)$ is defined as

$$\delta_F(a, b) = \#\{x \in \mathbb{F}_2^n | F(x) + F(x + a) = b\}$$

The differential uniformity(δ_F) should be minimized to resist the differential cryptanalysis. The δ_F is always even, so $\delta_F \geq 2$. If $\delta_F = 2$, then the S-Box F is called almost perfect nonlinear (APN). The given S-Box is called differentially δ uniform, if for every $a \in \mathbb{F}_2^{m*}, b \in \mathbb{F}_2^m$,

$$\#\{x \in \mathbb{F}_2^n | F(x + a) + F(x) = b\} \leq \delta$$

Algebraic Degree The algebraic degree of an $n \times m$ S-Box F , denoted as deg_F , is the minimum degree among all of its component functions. Mathematically,

$$Deg_F = \min_{b \in \mathbb{F}_2^{m*}} deg(b \cdot F) \tag{9}$$

The higher the degree of an $n \times m$ S-Box is more resistance to algebraic attack.

Fixed Points/Opposite Fixed Points The presence of fixed points (FP) and opposite fixed points (OFP) of an S-Box is vulnerable to statistical attack. A fixed point occurs when the input of an S-Box F maps to the same output value ($F(x) = x$), and the opposite fixed point occurs when the output of F is the complement of its input value ($F(x) = x'$)

3 Practical Implementation

In this section, we will explain about the efficient implementation of all the cryptographic properties of the Boolean function and S-Box. There are some cryptographic properties which are directly implemented as per theoretical definition, and some cryptographic properties need some optimized approach to reduce both time and computational complexity. In this paper, we discuss about the practical implementation of some cryptographic properties which are optimized version of its theoretical definition.

3.1 *Balancedness*

To check the balancedness property of Boolean function and S-Box, we use Walsh–Hadamard transform for the zero input vector. For Boolean function f , if $W_f(0) = 0$, then the function f is balanced. In case of S-Box F , if $W_F(0, v) = 0$ for all $v \in \mathbb{F}_2^{m*}$, then the S-Box is balanced.

3.2 *Nonlinearity*

Basically, the nonlinearity of Boolean function is evaluated using Walsh–Hadamard transform or Hadamard matrix using polarity truth table form. The polarity form of Boolean function f , denoted as \hat{f} , is defined as $(-1)^f$. The product of Hadamard matrix and its transpose becomes an identity matrix. Mathematically,

$$HH^t = nI_n$$

The Hadamard matrix is defined as follows:

$$H_0 = (1) \quad H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad H_n = \begin{pmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{pmatrix}$$

This Hadamard matrix for an n -variable Boolean function is the polarity form of the linear function $u \cdot x$. The u^{th} rows of the Hadamard matrix are the polarity truth

Fig. 1 Fast Walsh–Hadamard transform using butterfly approach

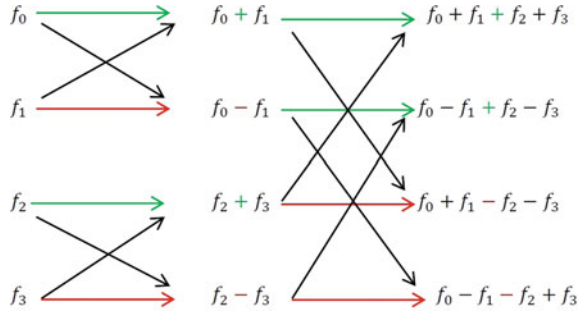


table of the linear function $u \cdot x$ for any $u \in \mathbb{F}_2^n$. The Walsh–Hadamard spectrum now can be expressed in terms of Hadamard matrix and polarity truth table of Boolean function [21] as follows:

$$[W_f] = H_n[\hat{f}] \tag{10}$$

For $n = 2$, the Walsh–Hadamard spectrum can be represented as

$$W_f = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} (-1)^{f_0} \\ (-1)^{f_1} \\ (-1)^{f_2} \\ (-1)^{f_3} \end{bmatrix}$$

Since the size of the Hadamard matrix is $2^n \times 2^n$, the number of operation required for this matrix multiplication is $O(2^{2n})$ [21]. As the size of n increases, it is not an efficient way to calculate the nonlinearity using this approach. To reduce the time complexity, we apply the fast Walsh–Hadamard transform as shown in Fig. 1.

Fast Walsh–Hadamard Transform: In fast Walsh–Hadamard transform, we do not need to calculate and store any matrix, and we store only Boolean function. So, less memory is needed in this approach. For n -variable Boolean function, the total number of operation at each step is 2^n and repeated for n times. The total number of operation in fast Walsh–Hadamard transform is $O(n \cdot 2^n)$. The algorithm of fast Walsh–Hadamard transform is explained in Algorithm 1.

For vectorial Boolean function or S-Box F , the expression becomes $[W_F] = H_n \cdot [\hat{F}]$, where \hat{F} is the polarity truth table of all component function.

3.3 Algebraic Degree

For practical implementation, we calculate the algebraic degree from the algebraic normal form of Boolean function specified by its support [22]. Mathematically,

$$f(x) = \sum_{\tau \in \Omega_f} \left(\prod_{i=1}^n x_i + \tau_i + 1 \right)$$

Algorithm 1: Algorithm for computing Fast Walsh–Hadamard Spectrum

```

1 Initialize Boolean function( $n$ -var) into  $f$  in terms of truth table (0,1) form ;
2 for  $i \leftarrow 0$  to  $2^{n-1}$  do
3   |  $polarity\_tt = (-1)^{f[i]}$  ;
4 end
5  $Walsh\_spec = polarity\_tt$  ;
6 for  $i \leftarrow 0$  to  $n$  do
7   |  $skip = 2^i$  ;
8   |  $pair = 2^{i+1}$  ;
9   | for  $j \leftarrow 0$  to  $2^n - 1$  by  $j+pair$  do
10  |   | for  $k \leftarrow j$  to  $j + skip$  do
11  |   |   |  $temp = Walsh\_spec[k]$  ;
12  |   |   |  $Walsh\_spec[k] = temp + Walsh\_spec[k + skip]$  ;
13  |   |   |  $Walsh\_spec[k + skip] = temp - Walsh\_spec[k + skip]$  ;
14  |   | end
15  | end
16 end
17 Output  $Walsh\_spec$  ;

```

Where $\tau = (\tau_1, \tau_2, \dots, \tau_n)$. But this approach is not efficient to calculate the algebraic degree as the size of n increases. For efficient implementation, we use the *Transeunt triangle* which is being proved by Shafer et al. [24]. In this method, it is easy to convert ANF to truth table form and vice versa. For calculating algebraic normal form, the truth table of Boolean function is placed in a row. In an inverted Pascal’s triangle fashion, consecutive values of the rows are performed Xor operation and placed in the next higher row in between two values on which it is performed. This operation is continued in this way, and we are left with a row having only one value. The left side of this triangle is (0, 1)-sequence of ANF coefficient. For a three-variable Boolean function as defined in Table 1, we calculate the algebraic normal form and algebraic degree using Transuent triangle representation.

Figure 2 demonstrates converting Boolean function truth table form to ANF coefficient.

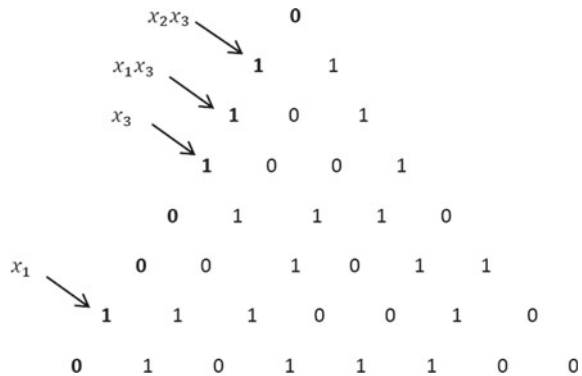
In Fig. 2, the ANF coefficient value 1 represents the algebraic normal form. In position 1, 4, 5 and 6 have values 1, and its ANF representation is $x_1 \oplus x_3 \oplus x_1x_3 \oplus x_2x_3$.

Similarly, the algebraic degree of an S-Box F is the minimum algebraic degree of all component function of F .

Table 1 Truth table of Boolean function

x_3	x_2	x_1	f
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	0
1	1	1	0

Fig. 2 ANF using Transeunt triangle representation



4 BSAT: Execution and Performance Analysis

In this section, we will discuss about our implementation details, execution and performance analysis of BSAT tool.

Boolean function and S-Box analysis tool (BSAT) analyze the cryptographic properties of Boolean function and S-Box, which is developed using C++ code. All the important cryptographic properties resistance to existing cryptanalytic attacks defined in section 2 are evaluated using BSAT. This tool is developed using most efficient and optimized approach to reduce the time and memory complexity.

Our BSAT is open source, and the source code is available in GitHub repository (<https://github.com/pratapiitr/BSAT>). The user can further customize the program according to their requirements and can add new cryptographic properties. The execution of BSAT can be done in two ways, first by using stand-alone tool using C++ library to execute entire program, and second approach is graphical user interface form. In stand-alone tool, user has no choice to call a particular function, instead all the cryptographic properties are evaluated. In this paper, we use our BSAT as a stand-alone tool to evaluate all the cryptographic properties. In future, we can convert this tool that can call an individual function, but it takes more computational effort.

Table 2 BSAT execution results for Boolean function

Boolean function	Size	Balanced	NL	AC	Deg	CIDEV	SAC Satisfied	Time (sec)
f_1	8	YES	114	24	7	20	NO	0.001
f_2	10	YES	486	48	9	44	NO	0.019
f_3	12	YES	2002	88	11	88	NO	0.045

Table 3 BSAT execution results for S-Box

S-Box	Size	Bijjective	NL	DU	Deg	FP/OFP	Time [Sec]
PRESENT	4×4	YES	4	4	2	0/1	0.007
GIFT	4×4	YES	4	6	2	0/1	0.008
AES	8×8	YES	112	4	7	0/0	0.404

4.1 Practical Results and Execution Time

In this section, we analyze the performance of BSAT in terms of execution speed by taking the input of an n -variable Boolean function and different size of $n \times m$ S-Boxes used in well-known block ciphers (PRESENT, GIFT, AES). We also present the practical results of some important cryptographic properties defined in Sect. 2.

The execution time of BSAT depends upon two things: firstly, the number of cryptographic properties to be evaluated. Secondly, it depends upon the system configuration upon which BSAT is being executed. We execute our BSAT tool on a system having Xeon processor 3.70GHz computational power and 32 GB internal memory. The experiment is conducted on Windows 10 operating system. Since we have more number of cryptographic properties and the system configuration is totally different, it is not practically feasible to compare the performance of BSAT with the existing tools. In this paper, we present our results of different size of Boolean functions and S-Boxes in Tables 2 and Table 3, respectively.

5 Conclusion and Future Work

Boolean functions and S-Boxes are two most important cryptographic primitives of symmetric cryptosystem. So, it is necessary to evaluate the cryptographic strength of such primitives to measure the resistance against cryptanalytic attacks. In this paper, we develop a tool called BSAT, which uses optimized approach for calculating Walsh–Hadamard spectrum and algebraic normal form for evaluating nonlinearity and algebraic degree in a less computational effort. We evaluate the performance of BSAT by experimenting on different size of Boolean functions and S-Boxes and

present the practical results along with execution time. This tool considered all the important cryptographic properties, and it is available in public domain which is open source for further modification and adding the new cryptographic properties.

There are several aspects to be considered for further development of BSAT. The most important aspect is to add new cryptographic properties resistance to different cryptanalytic attacks such as side channel attack, algebraic immunity and k -normality of Boolean functions. The another aspect is to enhance the performance of BSAT in terms of execution time, when it is applied for larger size S-Boxes up to 16×16 . The third approach is to develop a graphical user interface (GUI) version of BSAT to make it more user convenient.

References

1. Stinson, D.R.: *Cryptography: Theory and Practice*. Chapman and Hall/CRC (2005)
2. Carlet, C., Crama, Y., Hammer, P.L.: Boolean functions for cryptography and error correcting codes. *Boolean Models Methods Math. Comput. Sci. Eng.* **2**, 257–397 (2010)
3. Matsui, M.: Linear cryptanalysis method for DES cipher. In: *Workshop on the Theory and Application of Cryptographic Techniques*, pp 386–397 (1993)
4. Biham, E., Shamir, A.: Differential cryptanalysis of DES-like cryptosystems. *J. cryptol.* **4**(1), 3–72 (1991)
5. Heys, H.M.: A tutorial on linear and differential cryptanalysis. *Cryptologia* **26**(3), 189–221 (2002)
6. Lafitte, F.: The boolfun Package: Cryptographic Properties of Boolean Functions
7. Lafitte, F., Van Heule, D.: Cryptographic Boolean functions with R. *Methods* **1**(1), 1 (2011)
8. Stein, W.A., et al.: *Sage Mathematics Software (Version 5.10)*. The Sage Development Team, <http://www.sagemath.org> (2013)
9. Picek, S., Batina, L., Jakobovi, D., Ege, B., Golub, M.: S-box, SET, match: a toolbox for S-box analysis. In: *IFIP International Workshop on Information Security Theory and Practice*, pp. 140–149 (2014)
10. Alvarez-Cubero, J.A., Zufiria, P.J.: A C++ class for analyzing vector boolean functions from a cryptographic perspective. In: *2010 International Conference on Security and Cryptography (SECRYPT)*, pp. 1–9 (2010)
11. Prouff, E.: DPA attacks and S-boxes. In: *International Workshop on Fast Software Encryption*, pp. 424–441 (2005)
12. Lerman, L., Veshchikov, N., Picek, S., Markowitch, O.: On the construction of side-channel attack resilient s-boxes. In: *International Workshop on Constructive Side-Channel Analysis and Secure Design*, pp. 102–119 (2017)
13. Daemen, J., Rijmen, V.: *The Design of Rijndael: AES-the Advanced Encryption Standard*. Springer Science & Business Media (2013)
14. Bogdanov, A., Knudsen, L.R., Leander, G., Paar, C., Poschmann, A., Robshaw, M.J., Seurin, Y., Vikkelsoe, C.: PRESENT: An ultra-lightweight block cipher. In: *International Workshop on Cryptographic Hardware and Embedded Systems*, pp. 450–466 (2007)
15. Sarkar, S., Syed, H.: Bounds on differential and linear branch number of permutations. In: *Australasian Conference on Information Security and Privacy*, pp. 207–224 (2018)
16. Burnett, L.D.: *Heuristic optimization of Boolean functions and substitution boxes for cryptography (Doctoral dissertation, Queensland University of Technology)* (2005)
17. Clark, J.A., Jacob, J.L., Stepney, S.: The design of S-boxes by simulated annealing. *New Gener. Comput.* **23**(3), 219–31 (2005)

18. Millan, W., Clark, A., Dawson, E.: Heuristic design of cryptographically strong balanced Boolean functions. In: International Conference on the Theory and Applications of Cryptographic Techniques, pp. 489–499 (1998)
19. Banik, S., Pandey, S.K., Peyrin, T., Sasaki, Y., Sim, S.M., Todo, Y.: GIFT: a small present. In: International Conference on Cryptographic Hardware and Embedded Systems, pp. 321–345 (2017)
20. Behera, P.K., Gangopadhyay, S.: Analysis of cost function using genetic algorithm to construct balanced Boolean function. In: TENCN 2018-2018 IEEE Region 10 Conference, pp. 1445–1450 (2018)
21. Cusick, T.W., Stanica, P.: Cryptographic Boolean Functions and Applications. Academic Press (2017)
22. Fukuzawa, M.B.: The spectra of DES S-Boxes. NAVAL POSTGRADUATE SCHOOL MONTEREY CA (2014)
23. Kim, K., Matsumoto, T., Imai, H.: A recursive construction method of S-boxes satisfying strict avalanche criterion. In: Conference on the Theory and Application of Cryptography, pp. 565–574 (1990)
24. Shafer, J.L.: An analysis of bent function properties using the Transeunt triangle and the SRC-6 reconfigurable computer. NAVAL POSTGRADUATE SCHOOL MONTEREY CA (2009)

Sorted Galloping Prevention Mechanisms Against Denial of Service Attacks in SIP-Based Systems



Sheeba Armoogum and Nawaz Mohamudally

Abstract The IP telephony service is gaining its popularity over the past fifteen years. Statistical reports show that there has been a hike in the number of attacks on VoIP systems over the last five years. Among the many existing threats, Denial of Service (DoS) flood attack is considered as the worst to VoIP environment. In this paper, we present three statistical models to mitigate flood attacks. The models can detect distributed attacks for two behaviors. They also show very high accuracy and very small false positive alarms. The proposed models can train the VoIP system sequentially every 500 ms to reduce false positive cases to zero. Based on all criteria considered in this paper, we notice that the q-SGP prevention model is better than the two other models.

Keywords Denial of service (DoS) · Distributed denial of service (DDoS) · Flood attacks · Session initial protocol (SIP) · Detection rate · False alarm

1 Introduction

The Internet has ironed telecommunication suppliers and service providers to transmit voice using IP. Voice over IP (VoIP) uses a network for voice, hence benefiting many advantages [18] while giving trendy and state-of-the-art telecommunication services. The Session Initiation Protocol (SIP) [14] is one of the two protocols to enable this service [8, 21]. The SIP, which is an application layer protocol, is indeed very easy to implement and is based on ASCII codes. Because of its simplicity, working at layer 7, illegitimate users find ways and techniques to attack VoIP systems and their nodes. The taxonomy [19] report gives a technical awareness to communities, namely technical designers community, service providers, vendors, software

S. Armoogum (✉)
University of Mauritius, Reduit, Mauritius
e-mail: s.armoogum@uom.ac.mu

N. Mohamudally
University of Technology Mauritius, La Tour Koenig, Mauritius
e-mail: alimohamudally@umail.utm.ac.mu

developers, and researchers by providing them with a detailed structure of technical vulnerabilities.

The most common DoS attack is when an attacker sends a massive amount of messages to a client or server which leads to congestion, thereby paralyzing the processing capabilities [20], and hence rendering it inoperable. Therefore, the attackers exploit this issue by using, for example, a SIP generator tool to send a huge number of messages in a few seconds to exhaust the system. Since the messages are not malformed or tampered ones, they are considered valid ones, hence, they are very hard to detect [6, 11, 17]. Ahmad and Singh [1] state that, “a denial of service attack can be initiated by exhausting the three main resources: memory, CPU, and network bandwidth”. For example, a hacker can send a huge number of INVITE messages using fake addresses to disestablish a user or the server. Coulibaly and Liu [5] argue that “many attacks are mitigated by implementing the VoIP Defense System consisting of powerful hardware with a high level of authentication mechanism using parallel processing”. However, as quoted by Ehlert et al. [6], “sophisticated attacks can still crack these systems, as they are generally undetectable by the proxy servers or any monitoring nodes”. The second type of attack under the flooding category is the Distributed DoS (DDoS) one, which is an advanced version of the SIP DoS flood attack and is even trickier than the simple flood attack. The setting up of such a threat is complex but achievable by the attackers. Ahmad and Singh [1] simulate a DDoS attack test by setting up a testbed using several computers to exhaust several users. The authors added that the attacks penetrated the system effectively and significantly and exploited the resources of several PCs. An example for setting up of this attack is to use the mastermind attacker to search for unsecured devices (also known as the handlers and zombies) inside and outside the VoIP system to plant malicious codes. As mentioned by Semerci et al. [15], it may be part of strategic plans to exhaust the network and to disrupt service which leads to severe financial losses to the service providers.

Researchers have tackled over the years; many researchers have addressed diverse attacks and their solutions over the years. Network designers have termed the network solution as SIP Protectors or SIP Defenders or VoIP Defense systems which imply the same definition and tasks.

This paper presents a defense system where three new statistical flood attack mitigation techniques that focus on signature-based and anomaly-based intrusions, will be tested. The architecture proposed has an application firewall that can detect some known intrusion signature-based anomalies. Compared to several researchers [4, 13] who use the established threshold-based rate-limiting technique, a method introduced by Iancu [10], our method will address the processing issue in a more efficient manner using our modified hybrid algorithms, known as the Sorted Galloping Prevention (SGP) algorithms. Furthermore, more fields of the INVITE messages will be considered to determine the behavior and patterns of random attacks and hence to obtain additional statistical results. Therefore, for example, flood patterns that consist of one message at a periodic time interval to trick the SIP server will be caught by the proposed algorithms. In the end, a performance analysis of the three mitigation techniques will be conducted.

The remaining structure of the paper will be organized as follows: Sect. 2 will deal with a review of DoS flood occurrences against VoIP systems. Section 3 will present the techniques and the proposed architecture. The methodology with the experimental setup will be discussed in Sect. 4. The findings and performance with concerns will be elaborated in Sect. 5. The paper will wrap up in Sect. 6 with a conclusion and future work.

2 Related Work

In this section, a list of studies, which focuses on security issues, security systems set up and the prevention algorithms to countermeasure SIP-based DoS, will be reviewed.

Xin et al. [20] present multiple source flood attacks using novel detection algorithms. The system is a two-layer defense mechanism that is capable to detect and defend against the distributed attacks. The first layer is in charge of the extraction of the only required information of the incoming messages to preserve users' integrity and the second layer will take the needed action. The results indicate that the objectives concerning performance are met.

Another effort by Bansal and Pais [3] is made to address the problem of flooding against the VoIP system. The authors considered different scenarios and metrics. The experiment is conducted for a very high attack rate only and for INVITE and BYE message attacks. According to them, the results are much better than existing defense systems.

Ahmad and Singh [1] propose a VoIP security system to mitigate attacks against DDoS flood attacks (invite, register messages). This system uses an enhanced SIP proxy server and an enhanced application layer stateless firewall. The main idea of this system is to maintain users' IP addresses at the level of the firewall and the SIP server for priority handling. Ahmad and Singh [1] claim that the SIP server has one more specific task to train the firewall by updating it with the IP addresses of legitimate users and alerts the firewall when a legitimate user's IP address expires and needs to be removed from the list so that it can adjust its rules according to the information fed. In addition to the two devices, one more server is required between the SIP server and the callees to ensure proper balancing concerning network bandwidth. Unfortunately, the authors did not provide any quantitative results.

Semerci et al. [15] propose a two-tier intelligent system that can monitor DDoS attacks together with a unit that can "discriminate" users in the system from malicious ones. In their paper, the authors introduce a novel real-time online intrusion detection and prevention system for communication networks by developing a change detection algorithm that monitors the network traffic intensities at the server side for analyzing the traffic pattern [15].

For over fifteen years, despite that several methods have been initiated and executed (hardware-based or software-based), the attackers still enjoy conducting

malicious activities and, therefore, have led us to join hands to continue battling against them.

3 The Mitigation Methods

In this section, we propose three statistical algorithms (SGP models) to mitigate the DoS and DDoS flood attacks. Algorithms are the practical tools in computer science and networks for problem solving. The sorting and searching algorithms have remained a priority in many technological fields. The goal to counteract attacks is basically to search and take action (e.g., to quarantine, to take legal actions, etc.). In this context, sorting, searching, and machine learning algorithms are appropriate methods that are commonly used in today's research. Concerning sorting algorithms, "despite several new techniques like three-way radix quick sort, LSD radix sort, Tim sort and multi-key quick sort being introduced" [16], researchers still rely on the three stable and "pin to" quick sorting, merge sorting, and heap sorting algorithms due to poor performance displayed by the former mentioned.

Similarly, in computer science, searching algorithms are techniques that are used to retrieve information located in a fixed storage space (e.g., list of addresses in the Iptables utility of a firewall) or incoming and outgoing information (e.g., packets in a network) whether structured or unstructured information. On the other hand, we have noted that the galloping search algorithm has not been considerably explored in-depth in research. Since our work also involves memory utilization and CPU usage, this search will be improved to address this issue besides the main objectives described in Sect. 1.

Therefore, the three statistical defense mechanisms (also known as the m-SGP, q-SGP, and the h-SGP) will be devised by using three types of sorting algorithms, combining with a modified galloping algorithm. The defense structure will be embedded in the testbed for verification (Fig. 2). The INVITE message is considered since it is the most predominant one among attackers by analyzing its source IP address, a destination address, and its "info" and time parameters and by analyzing the behavior and the attacking patterns.

The flowchart in Fig. 1 illustrates a fragment of the mitigation system using the galloping algorithm. The galloping structure is an exponential binary search method to locate the position of the target which consists of an array (technically, known as a list) of elements [Source, Destination, and Info] in a given sorted array (list) SorceSIPList, where the list can be of unbounded size.

Refer to Fig. 1, the algorithm is implemented using the recursive method, which involves two main steps as follows:

- (i) To loop SorceSIPList from zero to the size of the SorceSIPList by checking the condition as follows: if the loop index variable is less than the size of the list and also the "Source" element at the same index is less than the "Source" element of the target, then the loop index variable is increased by raising it to the

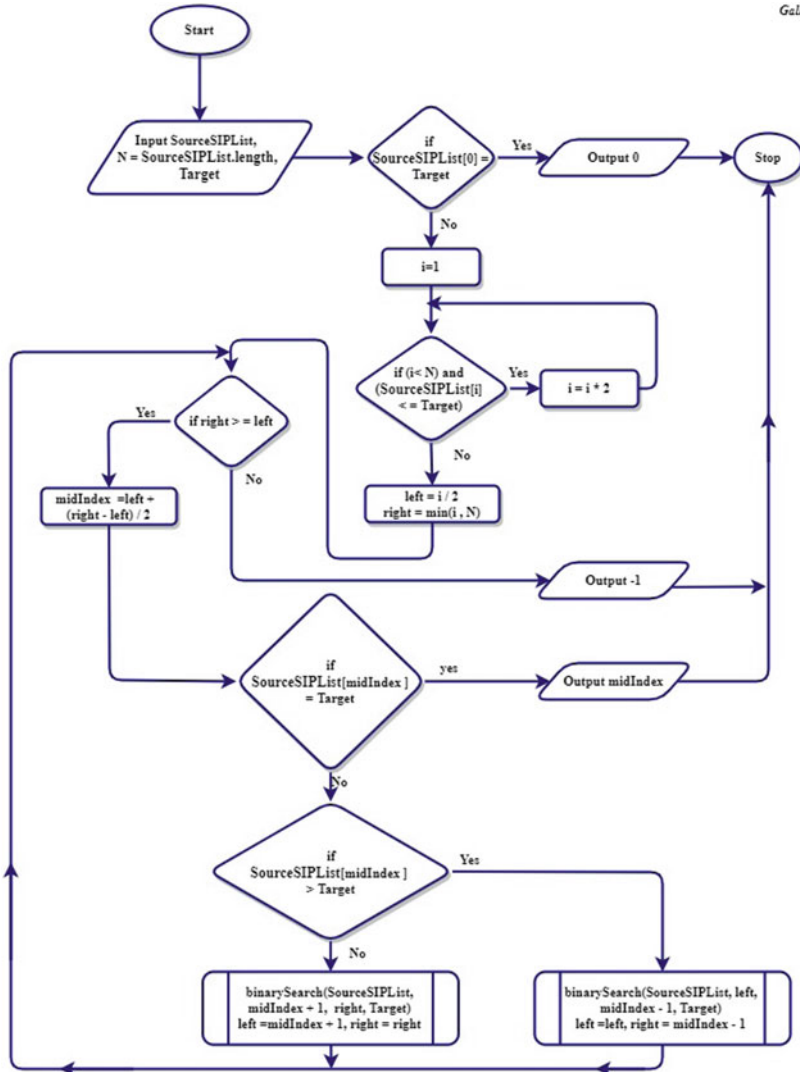


Fig. 1 Galloping algorithmic segment of the SGP

power of 2. After the completion of the loop, then we conduct a binary search test on the SourceSIPList in the range between half the resultant index and that index value itself.

- (ii) To perform a binary search by dividing the list to find its mid-position followed by recursively searching the first half of the list and second half of the list until no further recursion can be applied. The position of the target is then returned to the called function.

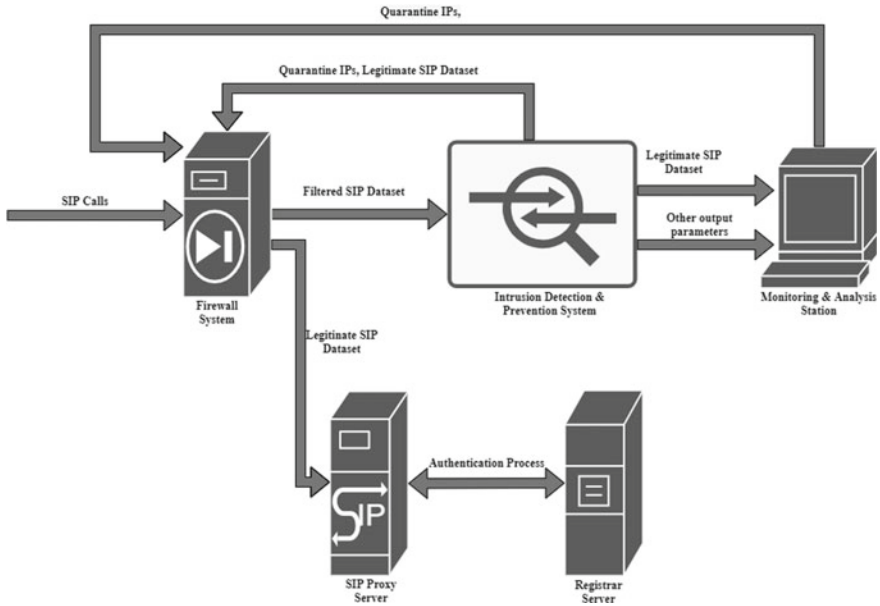


Fig. 2 Defense system

The technique, which is developed using the Python programming platform and libraries, demonstrates how flood attacks can be detected and prevented using the three prevention mechanisms (m-SGP, q-SGP, h-SGP) and the next step is to validate them in a defense system.

4 The VoIP Defense System Architecture

In this section, the proposed testbed, the dataset, and the attack scenarios are explained.

4.1 Proposed Architecture and Methodology

Figure 2 depicts the defense mechanism which has been mounted to test the mitigation algorithms. It depicts our proposed defense system using a firewall and an intrusion detection and prevention system (IDPS) for an attempt to mitigate DoS attacks.

The experiment is conducted in a private network due to safety reasons. We make several users initiate calls using the softphones (LinPhone, Zoiper, and MiniSipPhone 7.3) randomly within 16 s. At the same time, an attack tool is used (SIPp) and is

launched at different attack rates during a fixed duration of 16 s. The computers and a server are interconnected using a switch router of a maximum speed of 20 Mbps. The SIP server (MiniSIPServer) is mounted on Linux Ubuntu 18.04 and an Intel Quad Core i7 processor machine at a processor frequency of 3.3 GHz, supporting Gigabit Ethernet connection with installed RAM of 8 GB. The remaining computers are of the following specifications: Intel Core i5 processor of 4 GHz and an installed RAM of 16 GB (firewall and IDPS); Intel Core processor of 4 GHz, two laptops, and two mobile devices (where softphones and the attack tool are installed).

As explained by many researchers [6, 7], “the IDPS detects and mitigates ongoing attacks using proactive security measures, which is either hardware or a software solution that analyses both internal and external network traffic for malicious activities”. As shown in Fig. 2, the users communicate with the SIP Server through a security server consisting of an application-level firewall and an IDPS. The firewall consolidates the received SIP call request messages from the users/attackers into a CSV file with the five tuples: time, Source, Destination, Protocol, and Info. The consolidated CSV file is then forwarded to the IDPS for malicious activity detection. At the level of IDPS, the analysis is carried out through different evaluation behaviors on source addresses, destination addresses, time stamps, pattern matching, and malformed packets. Upon detection of malicious activities, the IDPS prevents the attack by quarantining the malicious IP addresses. The quarantined IP addresses are then sent to the firewall for blocking. After a complete analysis, the firewall sends the legitimate dataset to the SIP Server for further processing. Once the server receives the legitimate datasets, the server proceeds further with the registration process. Every SIP user agent sends REGISTER messages to the SIP Server to associate the SIP URI with the public IP address. The registrar records the association to the location service database.

4.2 Dataset and Attack Scenarios

As it is often reported by many researchers, due to privacy concerns, except for the list of suspected IP addresses provided by Bad Packets LLC [2], there are no real datasets available for testing. In this context, we have generated a list of INVITE messages using a flood message generator which are captured by the firewall and which returns the information in CSV format. The flood generator produces good and bad INVITE messages. To simulate real data, the consolidated dataset (refer to Fig. 3) in CSV format will consist of the following:

- (i) Genuine real network traffic at several instances of a day for five days using different computers and IP addresses.
- (ii) Non-genuine network traffic data similar to attack traffic generated using the penetration testing and message generator tool. This will hence characterize different scenarios for a VoIP network and will make the dataset more representative.

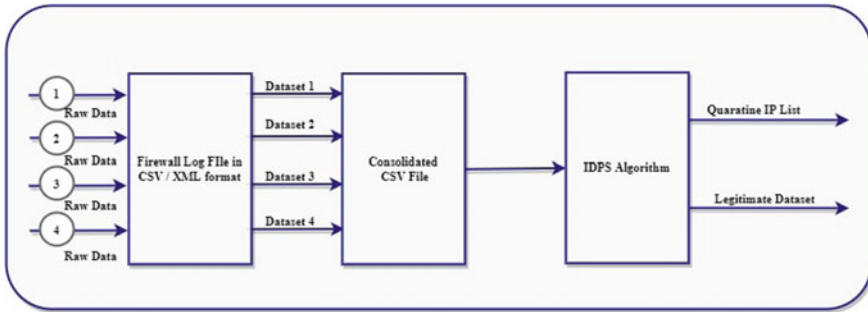


Fig. 3 Consolidated dataset

- (iii) Other traffic data are been created manually and using the list from Bad Packets LLC [2]. The traffic data hence consists of simulated genuine and non-genuine traffic.

According to SIP specifications [14], response time for an invite message varies in the order of 500, 1000, 2000, 4000, 8000, 16,000, 20,000, 24,000, 28,000, and lastly 32,000 ms which is the timeout response time. The testing will be conducted at various attack rates varying from 20 messages per second (mps) to 200 mps for 16 s. However, compared to other efforts conducted by researchers, the window size will not be equal to the response time set at the firewall. Instead, our algorithmic-based IDPS will calculate the window size based on the difference between the time of the first record and the last record of the CSV file. This may enable faster processing time for each slot.

The three models make provision for single flood attacks and distributed attacks as per the regular definitions given by the research community [9]. Furthermore, we have created two additional patterns as follows: firstly, by sending one message per slot by one attacker or by zombies and secondly, by sending one message every alternate slot during the attack time.

5 Experimental Analysis and Discussion

In this section, the results are presented in a tabulated form for further analysis. The experimental testing is conducted for the duration of attacks of 16 s. The classifications of results for conducting technical comparative analysis among the SGP group of statistical algorithms are explained in the next section.

5.1 *Experimental Metrics Used*

Measurements are conducted every 500 ms for 16 s (32 slots). The system is therefore been trained consecutively every 500 ms to detect the maximum number of false positive, false negative cases, and illegitimate messages, that is, the training is achieved every $1/32^{\text{th}} = 3.125\%$ of the total messages injected.

To evaluate the performance of the prevention mechanisms, several scores are used to determine the accuracy and precision of the IDPS which can be computed using the four classifiers true positive (TP), false positive (FP), true negative (TN), and the false negative (FN). Therefore, to simplify the analysis of the performance of the mitigation mechanisms, the following parameters are taken into consideration:

- (i) Attack detection rate is the proportion of the total number of illegitimate messages detected by the IDPS to the total number of illegitimate messages injected into the IDPS for testing.

$$\text{Detection Rate (DR)} = \frac{\text{Total attacks detected}}{\text{Total attacks injected}} \quad (1)$$

- (ii) False positive rate which is the false alarm rate is the ratio of the total number of false positive to the total number of negative events, that is, the probability that our IDPS will falsely detect legitimate messages as attackers.

$$\text{False Alarm rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (2)$$

- (iii) System accuracy is the ratio for which the IDPS correctly achieves predictions to the total number of cases examined.

$$\text{System Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

5.2 *Performance Evaluation*

In this section, we will evaluate the three models based on the measurements taken and as per the four classifiers and the three parameters explained in the previous section.

Experiments were conducted using the cross-sectional design and the before-and-after design approach. Therefore, initially, a specific dataset was tested (cross-sectional), and later during the day, the same dataset was tested again (the before-and-after). In addition to the two types of approach, for a specific attack rate, three different populations of the same size were tested, while keeping the number of genuine calls the same (that is, around 30). The experiment is repeated for different

Table 1 Attack scenarios

Attack rates	Number of genuine messages	Number of attacks injected	Total number of messages
20	28	320	348
40	28	640	668
80	29	1280	1309
200	30	3200	3230

attack rates. The table below (Table 1) shows the total messages injected for different attack rates.

The table (Table 2) and bar chart (Fig. 4) show the variation of the attack detection rate for the three models. It is observed that all three methods have the same outcome.

We can distinguish that our model can detect 100% precision when the attack rate is greater than 80 messages per second for q-SGP and h-SGP. With a very low attack rate, all models can have an accuracy of at least 83%. The first result indicates that all three models are not similar as depicted in the table and figure (Fig. 4).

It is noticed that all three models produce a false positive rate of zero at the end of 16 s, inclusive of the zombies-users. Regarding system accuracy, it is again observed that all three models show a good performance of at least 95.5% for an attack rate above 80 mps (Table 3). The findings also prove that q-SGP is the most efficient when considering the accuracy of 100% above 40 mps.

Table 2 Attack detection rate (%)

Attack rates	m-SGP	q-SGP	h-SGP
20	83.33	91.67	91.67
40	95.83	100.00	95.83
80	91.67	100.00	100.00
200	95.83	100.00	100.00

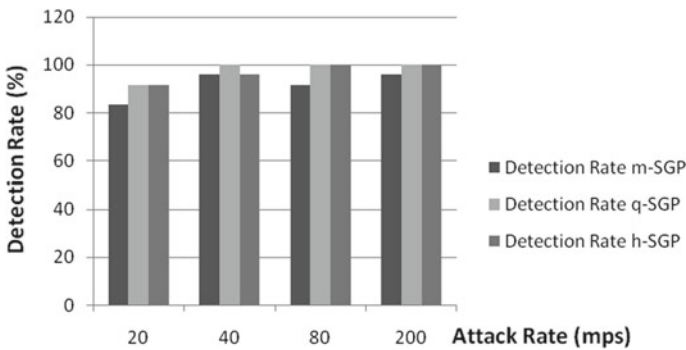


Fig. 4 Detection performance

Table 3 System accuracy (%)

Attack rates	m-SGP	q-SGP	h-SGP
20	91.67	95.83	95.83
40	97.92	100.00	97.92
80	95.83	100.00	100.0
200	97.92	100.00	100.0

Table 4 Processing time (milliseconds)

Total number of messages	Attack rates	m-SGP	q-SGP	h-SGP
348	20	6.20	5.86	5.86
668	40	6.35	5.80	5.84
1309	80	6.49	5.85	5.87
3230	200	6.70	6.05	6.12

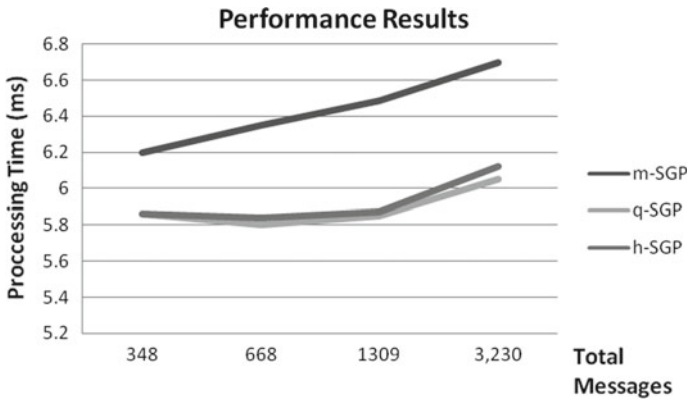


Fig. 5 Variation of processing time

Finally, however, when analyzing the variation of the processing time for each IDPS (Table 4 and Fig. 5) against the total number of messages injected, we note that the q-SGP model gives better processing time than the other two models.

6 Conclusion

In this paper, we present an intrusion detection and prevention system to mitigate DoS and DDoS attacks against VoIP systems that can focus on signature-based and anomaly based intrusions by three statistical algorithms. It provides three additional efficient techniques (or tools) that can be integrated into a VoIP system to further protect a SIP server against DoS flood attacks.

The experiment shows that all IDPS support both single flood and distributed flood attacks. They also cater for two behavioral patterns that an attacker can use to deviate and trick the VoIP system. The results show that the two models (q-SGP & h-SGP) give optimum results when the attack rate is greater than 40 mps. However, concerning the overall performance by considering all criteria, the q-SGP is leading.

As it is always said that flood attack is considered to be the most difficult threat which SIP providers need to challenge, we believe that our proposed techniques can be enhanced further by considering additional behaviors. The flood detection could be addressed using other techniques such as the application of machine learning techniques with linear vector machine classifiers which our colleagues and researchers have conducted [12]. In our future work, we will further develop our system to improve the attack prevention rate and to maintain the false positive and false negative rates close to zero. We are currently conducting experiments to verify the behavior of our IDPS for very high attack rates. The final result obtained in this paper will be used to confirm the findings of another research work in the field of deep learning.

References

1. Ahmad, W., Singh, D.: VoIP Security: A model proposed to mitigate DDoS attacks on SIP based VoIP network. In: A Multi-Disciplinary Research Book, pp. 37–48. Researchgate (2018)
2. Bad Packets, L.: Meaningful Intelligence for an Evolving Cybersecurity Landscape. Available at <https://badpackets.net/> (2020). Accessed 4 June 2017
3. Bansal, A., Pais, A.R.: Mitigation of flooding based denial of service attack against session initiation protocol based VoIP system. In: International Conference on Computational Intelligence & Communication Technology, pp. 391–396. IEEE (2015)
4. Bouzida, Y., Mangin, C.: A framework for detecting anomalies in VoIP networks. In: IEEE Third International Conference on Availability, Reliability, and Security. (2008)
5. Coulibaly, E., Liu, L.H.: Security of VoIP networks. In: 2nd International Conference on Computer Engineering and Technology (ICCET). (2010)
6. Ehlert, S., Wang, C., Magedanz, T., Sisalem, D.: Specification-based denial-of-service detection for SIP voice-over-ip networks. In: The Third International Conference on Internet Monitoring and Protection. IEEE (2008)
7. Fielder, J., et al.: VoIP defender: highly scalable SIP-based security architecture. In: International Conference on Principles, Systems, and Applications of IP Telecommunications. ACM (2007)
8. Geneiatakis, D., Kambourakis, G., Lambrinouidakis, C., Dagiuklas, T., Gritzalis, S.: SIP message tampering: the SQL code injection attack. In: 13th International Conference on Software, Telecommunications and Computer Networks (2005)
9. Hussain, I., Djahel, S., Zhang, Z., Naït-Abdesselam, F.: A comprehensive study of flooding attack consequences and countermeasures in the session initiation protocol (SIP). Secur. Commun. Netw. 4436–4451 (2015)
10. Iancu, M.: SER PIKE excessive traffic monitoring module (2003). <http://www.iptel.org/ser/doc/modules/pike>. Accessed 4 June 2017
11. Lahmadi, A., Festor, O.: A framework for automated exploit prevention from known vulnerabilities in voice over IP services. IEEE Trans. Network Serv. Manage. 114–127 (2012)
12. Nazih, W., et al.: Efficient detection of attacks in SIP based VoIP networks using linear L1-SVM classifier. Int. J. Comput. Commun. Control **14**(4), 518–529 (2012)

13. Ormazabal, G., Nagpal, S., Yardeni, E., Schulzrinne, H.: Secure SIP: a scalable prevention mechanism for dos attacks on SIP based VoIP systems. In: Principles, Systems, and Applications of IP Telecommunications Conference. (2008)
14. Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A.: SIP: Session Initiation Protocol, RFC 3261. IETF Network Working Group, IETF (2002)
15. Semerci, M., Cemgil, A.T., Sankur, B.: An intelligent cybersecurity system against DDoS attacks in SIP networks. *J. Comput. Netw.* **136**, 37–154 (2018)
16. Solehria, S.F., Jadoon, S.: Multiple Pivot Sort Algorithm is Faster than Quick Sort. *Int. J. Electr. Comput. Sci.* **11**(3), 14–18 (2011)
17. Tang, J., Cheng, Y., Yong, H.: Detection and prevention of SIP flooding attacks in voice over IP networks. In: INFOCOM conference, pp. 1161–1169. IEEE (2012)
18. Tzvetkov, V., Zuleger, H.: Service provider implementation of SIP regarding security. In: International Conference on Advanced Information Networking and Applications Workshops. (2007)
19. VOIPSA: VoIP security, and privacy threat taxonomy. In: Security and Threats (Official Doc.). (2005)
20. Xin, L., Hu, L., Hongbin, L., Xiongwei, X.: Distributed intrusion prevention system for SIP DDoS attack. *J. Chin. Comput. Syst.* **34**, 2095–2099 (2013)
21. Zhang, L., Tang, S. & Zhu, S.: An energy-efficient authenticated key agreement protocol for SIP-based green VoIP networks. *J. Netw. Comput. Appl.* 126–133 (2016)

Performance Enhancement and Reduce Energy Consumption with Load Balancing Strategy in Green Cloud Computing



Hitesh A. Bheda, Chirag S. Thaker, and Darshan B. Choksi

Abstract Cloud computing is an innovative technique which provides on-demand access of computing resources like processing power, memory, storage, network, etc. Cloud providers always suffering from the proper allocation of the resources for the execution of requested tasks. Because of this conflict or improper matching of tasks with the computing resources, they lead with QoS and performance degradation, and at end result, it may have bigger issue with greater energy consumption. Hence, in cloud environment, load balancing is an important approach to allocate resources to execute requested tasks with utmost efficiency and with fully utilization of computing resources. With effective load balancing mechanism, we can utilize the cloud resources, enhance performance, achieve QoS, and can reduce significant amount of energy consumption. Proposed load balancing algorithm focuses on response time and processing time for the execution of requested tasks to better justify the results in Green Cloud Environment. By reducing response time and processing time of proposed system, we can enhance system performance and achieve better Quality of Service (QoS) compared with any existing load balancing policies like Round Robin, active monitoring, LFU, and Throttled in Cloud system. By cutting down processing time and response time would result in better energy efficient mechanism for Green cloud computing. By using proposed load balancing algorithm, we can tentatively reduce response time by 25% and can reduced data center processing time by nearly 20% with compared to Round Robin algorithm. With provided solution of effective load balancing strategy, there is a significant reduction in energy consumption by cloud environment and helps to manage Green Cloud Environment more robust.

H. A. Bheda (✉)

Computer/Information Technology Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India

e-mail: hitesh.a.bheda@gmail.com

C. S. Thaker

Computer Engineering Department, Government Engineering College, Rajkot, Gujarat, India

e-mail: chiragthaker@yahoo.com

D. B. Choksi

G. H. Patel Center of Computer Applications, V.V. Nagar, Gujarat, India

e-mail: dbchoksi@yahoo.com

Keywords Green cloud computing · Performance optimization · Energy consumption · Load balancing · Response time · Processing time

1 Introduction

Cloud computing is famous for its potential capabilities and also offers computing resources like CPU (Processing Power), memory, storage, network, and many more as a utility. According to NIST definition, cloud computing can be considered as an on-demand and convenient model for computing resources that can be easily and fast provisioned and offered with little interference of management or service provider interaction. According to its on-demand and pay-as-you-go utility, cloud computing offers its services online with the help of good Internet connectivity. Cloud computing also offers software utilities or any hardware instances on user-demand. Cloud offerings are available for individual use or for enterprise perspective.

Green cloud computing is the concept where cloud resources deal with energy parameter. Green cloud computing is the key solutions to all the crucial dynamic resources demand and scalability of IT industry. To meet massive and dynamic demand of cloud data centres will emit enormous amount of Carbon dioxide, which is the bigger issue now with regards to global warming [1]. Nowadays, energy consumption is the bigger challenge for the world, as multiple data centers burn countless energy [1]. There are various ways to reduce energy consumption. One of the ways suggested by Abualigah et al. [2] was using multi-objective task scheduling mechanism by minimize makespan with maximum resource utilization. Another straightforward approach is proposed by Chen et al. [3] with VM Placement, where over-utilized VMs are migrated to under-utilized VMs group. Sathya Sofia et al. [4] has proposed AI-based NSGA-II algorithm, where DVFS has been used to minimize energy consumption along with VMs are grouped and predicted based on the characteristics for either to be shutoff or to be migrated. Zhang et al. [5] have proposed VM selection and migration policy for better load optimization on Green cloud architecture. Masdari et al. [6] has proposed a VM Placement scheme which can classify into reactive and proactive groups, which works on future workload forecasting. Hamzaoui et al. [7] have compared scenario for SLA agreements and energy-efficient cloud resource scheduling to reduce energy consumption for cloud systems.

Cloud computing also offers “as-a-Service” utility which refer to how cloud services are available to clients or users. Cloud computing categorizes into three main service models: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS) [8]. Cloud computing also offers four deployment models based on its structure. Private cloud—cloud which operates specially for a dedicated organization and maintained by the organization itself or by any other third party. Public cloud—created for open usage for all users and managed by third party. Hybrid cloud—integration of private and public cloud. Community cloud—offers services shared by multiple organizations and available to those groups only. Any cloud provider may have their own infrastructure, or they can use any other

third-party infrastructure to offer cloud services. Usually, community cloud can be used for multiple organization which have similar requirements.

1.1 Virtualization

Abstraction and virtualization are the two concepts on which cloud computing works. Cloud computing hides the details of system implementation from end-users or developers. In cloud environment, whatever the applications run on physical systems or servers, they do not have the location awareness for their data storage or for their processing power requirements. Cloud computing contains one important component as virtualization which offers shared pool of IT resources amongst multiple applications in order to get better utilization [9]. As virtualization share same resources amongst multiple application, hence, we can utilize resources and simultaneously decrease cost for unutilized resources. Virtualization technology also focuses on migration policy and can migrate from one physical server to another. Virtualization also offers scalability and multi-tenancy.

Virtualization is an innovative technique which enhance resource utilization and with the help of migration, it minimizes energy consumption. Virtualization is important for cloud environment as it hides software layers from hardware and hence offers mechanism to fast reallocate applications and other tasks across server based on its resource requirements.

Hypervisor or virtual machine monitor (VMM) is a low-level program to provide access to system resources. Hypervisor can be of two types: Type 1 or native or Kernel hypervisor—that runs on bare metal. While Type 2 or hosted hypervisor are installed on host operating system. Type 1 and Type 2 hypervisors are display in Fig. 1 with multiple instances of operating system to parallel execute multiple tasks.

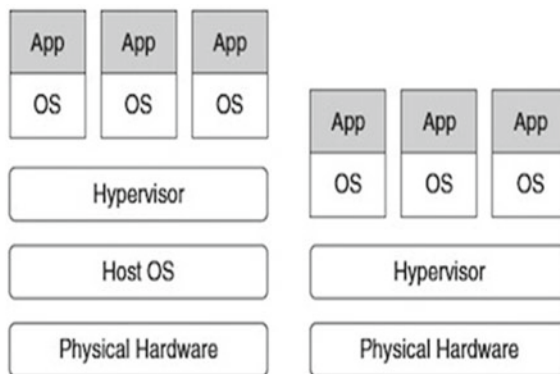


Fig. 1 Type I and type II hypervisors

1.2 Load Balancing

Load balancing is a strategy to assign or reassign tasks on individual parts of the system to make use of resources better and reduce system response time and processing time to execute tasks. As the greater number of tasks execution required, more effective load balancing strategies need to be implemented to have optimize performance [10].

The aim to use load balancing strategy to enhance performance by balancing load amongst available resources or underlying hardware to achieve optimal resource utilization, high throughput and avoiding overload [11]. By applying innovative load balancing approach, network and other resources can be utilized and hence we can minimize response time and enhance throughput.

By considering response time, there are two types of load balancing algorithms:

- *Cooperative load balancing algorithm:* These types of algorithms are used for the optimization of response time.
- *Non-cooperative load balancing algorithm:* All the tasks, which are currently in execution mode, are independent from each other and hence we can improve the total response time for the native task.

Based on resource scheduling and resource allocation, load balancing algorithm can be divided into three categories:

- *Sender-initiated algorithm:* The request message will be sent by the overloaded node until it find a suitable node which can be able to receive its load. This message request always originated by the sender node.
- *Receiver-initiated algorithm:* The request message will be sent by the under-loaded node till it gets a perfect overloaded node. This message request always initiated by receiver.
- *Symmetric algorithm:* Symmetric algorithm is a combination of sender-initiated algorithm and receiver-initiated algorithm.

At last, based on present system state, load balancing algorithms are categorized into two parts:

- *Static algorithms:* During the process of load balancing, these types of algorithms do not consider the previous states and performance of a node.
- *Dynamic algorithms:* During load balancing process, dynamic algorithms usually consider earlier states and performance of a node.

The rest of the paper is organized as follows. Section 2 describes the related study and works done in load balancing strategy for cloud computing in order to enhance performance. This is followed by existing system approaches for cloud environment in Sect. 3. Section 4 highlights the proposed system architecture. Section 5 contains Experimental setup in Cloud Analyst with result discussion. At the end Sect. 6 draws some important conclusion for proposed model of load balancing in cloud computing.

2 Related Study

Chen et al. [12] have provided user-priority guided min–min scheduling algorithm. Authors have provided two distinct techniques LBIMM and PALBIMM. If we compare the results of LBIMM and PALBIMM with min–min algorithm, then they are quite better to decrease tasks' completion time, proper load balancing of resources, and enhance overall system performance. This improved algorithm increases system performance over 20% for the utilization of resources and user services.

Behal et al. [13] have proposed load balancing technique for heterogeneous environment in cloud computing. Authors have proposed throttled technique for load balancing which can deal with optimized response time. This load balancing technique offers performance enhancement in distributed cloud environment using broker policy.

Garala et al. [14] have provided performance analysis using dynamic algorithm. Dynamic algorithm usually works on distinct parameters, like Response Time, Resources Utilization, Fault Tolerant, Waiting Time, Throughput, Turnaround Time, Process Migration, and Stability, and because of its dynamicity characteristics it always provides good performance compared to static algorithm.

Bhoi et al. have proposed a new algorithm with the modification in the Max–Min algorithm. This algorithm is used to identify tasks with average execution time [15]. Enhanced Max–Min task scheduling algorithm first selects the tasks with average or closest greater than average and then allocates to resources which requires minimum completion time.

Kruekaew et al. [10] have proposed method of artificial bee colony. ABC_LJF method provides efficient results compared to other methods and also provide high performance in terms of system scalability. In ABC algorithm, scheduling is executed with consideration of tasks' size and longest job first (LJF) to reduce the makespan of data processing time.

Malhotra et al. has proposed the adaptive load balancing algorithm. This adaptive load balancing algorithm optimizes the distributed system's performance and offers high scalability which is always a demand for Internet dependent computing and processing environment. With the consideration of response time and throughput, proposed algorithm provides very high performance compared to throttled and Round Robin algorithms.

3 Existing System

Load Balancing is an important approach in cloud architecture to distribute and assign load to existing resources in order to enhance performance. For cloud computing, few innovative load balancing strategies are famous for their advancements over overall system performance.

3.1 Round Robin Algorithm

Round Robin algorithm divides all the existing processes amongst all processors and then assign each process to the processor in Round Robin fashion. This algorithm distributes equal workload to existing processor [16]. Round Robin scheduling algorithm make effective use of time slice paradigm [17]. In this algorithm, total time splits into multiple partitions and each participating node will be assigned with particular time interval. Each node will have to execute its task or operation within assigned time quantum. If the user task will complete within assigned time quantum then user will not have to wait; otherwise, user will have to wait for the next round. Round Robin algorithm usually selects load randomly, for some cases server may be lightly loaded or it may be heavily loaded.

3.2 Equally Spread Current Execution (ESCE) Algorithm

In equally spread current execution (ESCE) or active monitoring algorithm [18], the load balancer actively monitors the load on all the VMs and distributes the load equally among them. The system load balancer preserves table of index for the VMs and the number of provisions allocated to individual VM. When a new job arrives, the data center controller requests the load balancer for the allotment of a new VM. The load balancer identifies index table for the VM which has minimum load. When the load balancer discovers the VM with minimum load, it provides the VM ID to the data center controller.

3.3 Throttled Algorithm

In Throttled algorithm, each VM processes only one job at a time; a new job can be processed only when the current job is completed successfully [19]. The load balancer entity manages an index table of all the VMs and their current states whether they are available or busy. As soon as the data center administrators query the load balancer for a VM allocation, the load balancer searches the index table for the available VM. If the load balancer finds an available VM at that time load balancer provides the VM ID back to the data center controller. Otherwise, the load balancer simply returns null. If the data center controller receives null from the load balancer, the request is queued until a VM becomes available.

3.4 Least Frequently Used (LFU) Algorithm

In least frequently used (LFU) algorithm, when task is submitted to a particular data center for the execution, the execution choice depends on the VM on which the tasks have been assigned very less number of times. Through this process, VMs remain busy for majority times for the purpose of processing of tasks.

4 Proposed System Model

By increasing load on the server in cloud environment with any existing approaches leads to some Quality of Service (QoS) issues and also degrade the overall performance. Our proposed system model works on response time of responses generated by server and data center processing time in order to utilize resource.

Proposed system model initially takes the user requests from the internet traffic as input in metatask. These user requests form cloudlets. Afterwards sorting is applied based on the length of the cloudlets. Initially, all VMs are available and have no requests. In this case, cloudlets are allocated to VMs on First Come First Serve basis. If VMs have requests to execute or it is busy then calculate the current load and processing capacity of VMs. If VM is underloaded then calculate the remaining load of VM and sort VMs based on remaining load. Allocate the cloudlets to VMs based on Firstfit. Delete the cloudlet from metatask and update VM allocation table.

4.1 Proposed Load Balancing Algorithm

VM load, processing capacity of VM, and Cloudlet execution time can be calculated using following Eqs. (1), (2) and (3) respectively:

$$\text{VM Load} = N * L / \text{MIPS of VM} \quad (1)$$

where, N = Number of Requests, L = Length of that request, MIPS = Millions of Instructions per second

$$\text{Processing Capacity of VM} = \text{PE}_{\text{num}} \times \text{PE}_{\text{mips}} + \text{VM}_{\text{bw}} \quad (2)$$

where PE_{num} = Processing elements number, PE_{mips} = Processing elements MIPS, VM_{bw} = Virtual machine Bandwidth.

Algorithm: Proposed Load Balancing Algorithm

Take Input as User requests (Cloudlets)
 Sort cloudlets based on length
 Initially All VMs are available and load is 0

For each VM
 Calculate the Current load of VMs
 Calculate the Processing capacity of VMs

 If VM load < VM capacity
 Calculate remaining Load of VMs
 Do Sorting Based on Remaining Load
 Allocate Cloudlet to Particular VM based on FirstFit
 Delete respective Cloudlet from metatask
 Update VM allocation Table
 End if

End for

$$\text{Cloudlet Execution Time} = \text{Filesize} / \text{MIPS of VM} \quad (3)$$

where, Filesize = File size of Cloudlet.

4.2 Flowchart of Proposed Load Balancing Algorithm

Figure 2 describes the flow of execution for the proposed load balancing algorithm to optimize resource and to enhance performance by considering all entities of Cloud Analyst architecture.

5 Experimental Setup

For simulation results, we use Cloud Analyst. Cloud Analyst provides graphical support for simulation and it is an extension to CloudSim [20]. Cloud Analyst provides separation of experimental simulation from programming aspects, as it is a need for the modeler to look for the complexities of simulations rather than spend time on coding and programming [21]. Cloud Analyst provides three Service Broker Policies [22] as given below:

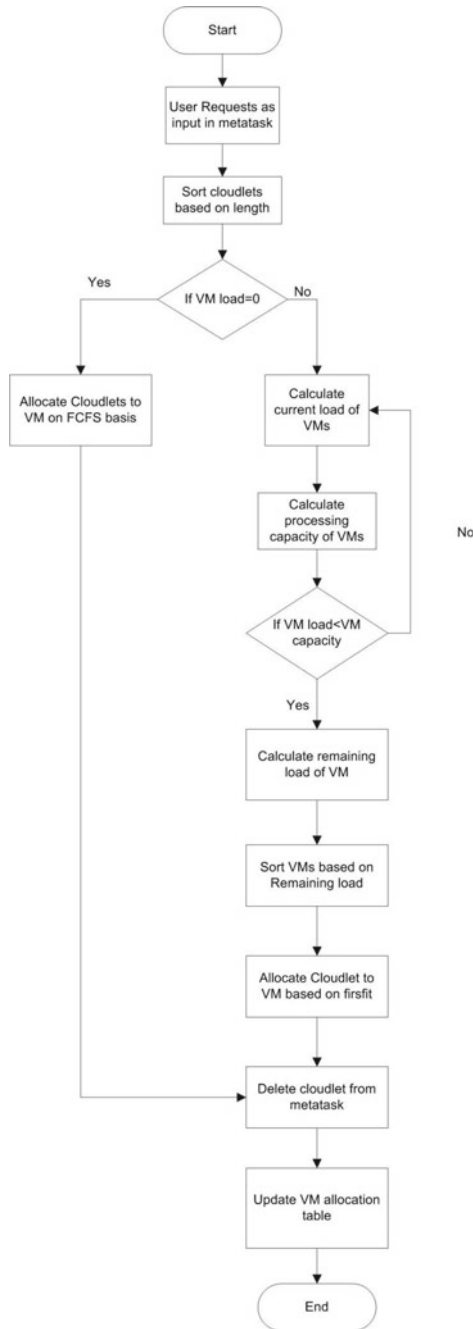


Fig. 2 Flowchart of proposed load balancing algorithm

- **Service Proximity-based Routing**

This routing policy identifies and selects the data center who has minimum network latency, it means the nearest data center. Whenever there are two or more nearest data centers available then selection should be done randomly. This routing policy does not consider load and cost during its working.

- **Performance Optimized Routing**

In performance optimized routing policy, performance of all the data centers are monitored by service broker and accordingly, it routes the traffic to the data centers based on the best response time. This routing policy never consider cost for its working.

- **Dynamically Reconfigurable Routing**

This routing policy is an extension to service proximity-based routing, in which routing logic remains common, but the service broker has extra task of scaling the application deployment according to the load. This routing policy may increase or decrease the VMs created in the existed data centers.

To implement proposed system architecture and to collect and compare results, we have configured Cloud Analyst with mentioned parameters and values: We have considered 10,000 MB VM image size, 512 MB VM Memory, 1000 MBPS VM Bandwidth, X86-based Data center—Architecture, Linux as Data center OS, Xen for Data center VMM, Number of VMs 20, 25 and 50 on DC1, DC2, and DC3, respectively, User base 10, User grouping factor 100, Request grouping factor 100 and Executable instruction length 100.

After setup above configuration in Cloud Analyst, we have collected simulation results for overall response time and data center processing time for service proximity-based routing, performance optimized routing, and dynamically reconfigurable routing.

Table 1 describes the collected results of overall response time in ms for Round Robin, equally spreaded current execution (ESCE), Throttled, least frequently used (LFU), and proposed load balancing technique.

Table 1 Comparative analysis of overall response time

	Proximity-based routing (ms)	Optimized response time (ms)	Reconfigure dynamically (ms)
Round Robin	201.42	152.65	158.49
Equally spreaded current execution (ESCE)	152.13	153.43	157.24
Throttled	152.04	152.67	156.92
Least frequently used (LFU)	151.56	151.66	155.8
Proposed load balancing policy	150.34	151.08	155.8

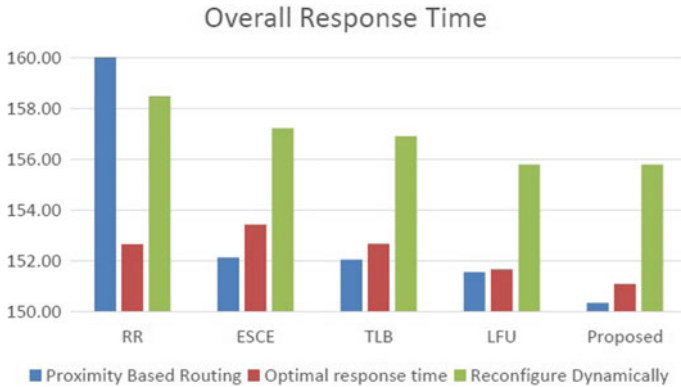


Fig. 3 Overall response time comparison

From Fig. 3, we can clearly say that our proposed algorithm for load balancing is quite better than Round Robin, ESCE, Throttled or LFU algorithm in all policies like service proximity-based routing, performance optimized routing, and dynamically reconfigurable routing.

Table 2 describes simulation results of data center processing time in ms for Round Robin, equally spreaded current execution (ESCE), Throttled, least frequently used (LFU), and proposed load balancing technique.

Figure 4 clearly concludes that proposed load balancing policy requires comparatively lower time compared to Round Robin, ESCE, Throttled, and LFU algorithm for all approaches like service proximity-based routing, performance optimized routing, and dynamically reconfigurable routing. Hence the performance is always be higher for proposed algorithm.

Table 2 Comparative analysis of data center processing time

	Proximity-based routing (ms)	Optimized response time (ms)	Reconfigure dynamically (ms)
Round Robin	1.96	1.44	7.02
Equally spreaded current execution (ESCE)	0.59	2.21	5.75
Throttled	0.43	1.46	5.43
Least frequently used (LFU)	0.40	1.23	5.36
Proposed load balancing policy	0.35	1.01	5.23

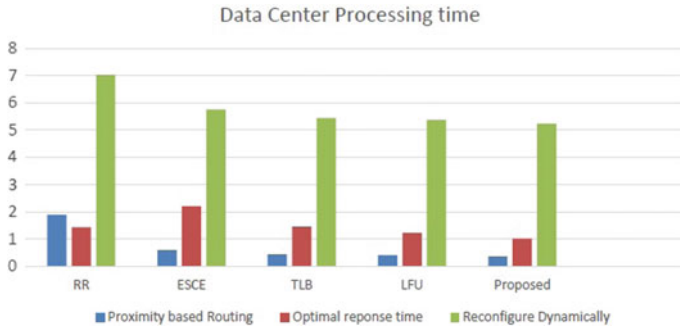


Fig. 4 Data center processing time comparison

6 Conclusion

Cloud computing always suffer to decrease response time and processing time for the execution of available tasks in order to increase cloud system performance, achieve good Quality of Service (QoS). Existing load balancing strategies like Equally Spread Current Execution Algorithm, Round Robin Algorithm, Throttled Algorithm or Least Frequently Used (LFU) Algorithm ideally work to decrease migration time, decrease processing time, decrease overhead, to improve utilization, etc. Decrease response time and processing time can be considered as the biggest challenges for any cloud provider or developer to offer the best throughput from available resource. Hence, we have implemented mechanism to increase the system efficiency with performance enhancement and QoS by considering response time and data center processing time. Various cases and collected results conclude that the proposed load balancing method has better development depending on the response time, processing of the data center and simultaneously optimize performance of cloud environment. By comparing proposed system with Round Robin, we have reduced overall response time by almost 25% and also decreased data center processing time by 20%. This all result in a greater advancement of Green cloud system by enhancing system performance and reducing response time hence reducing energy consumption.

References

1. Doshi, C., Verma, G., Chandrasekaran, K.: A green mechanism design approach to automate resource procurement in cloud. *Procedia Comput. Sci.* **54**, 108–117 (2015)
2. Abualigah, L., Diabat, A.: A novel hybrid antlion optimization algorithm for multi-objective task scheduling problems in cloud computing environments. *Cluster Comput.* (2020). <https://doi.org/10.1007/s10586-020-03075-5>
3. Chen, Y., Chen, X., Liu, W., Zhou, Y., Zomaya, A., Ranjan, R., Hu, S.: Stochastic scheduling for variation-aware virtual machine placement in a cloud computing CPS. *Future Gener. Comput. Syst.* **105**, 779–788 (2020)

4. Sathya Sofia, A., GaneshKumar, P.: Multi-objective task scheduling to minimize energy consumption and makespan of cloud computing using NSGA-II. *J Netw. Syst. Manage.* **26**, 463–485 (2018). <https://doi.org/10.1007/s10922-017-9425-0>
5. Zhang, K., Wu, T., Chen, S., Cai, L., Peng, C.: A new energy efficient VM scheduling algorithm for cloud computing based on dynamic programming. In: 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), pp. 249–254. New York, NY (2017). <https://doi.org/10.1109/CSCloud.2017.46>
6. Masdari, M., Zangakani, M.: Green cloud computing using proactive virtual machine placement: challenges and issues. *J. Grid Comput.* (2019)
7. Hamzaoui, I., Duthil, B., Courboulay, V., et al.: A survey on the current challenges of energy-efficient cloud resources management. *SN Comput. Sci.* **1**, 73 (2020). <https://doi.org/10.1007/s42979-020-0078-9>
8. Sosinsky, B.: *Cloud Computing Bible*. Wiley, Indianapolis (2011)
9. Malhotra, M., Singh, A.: Adaptive framework for load balancing to improve the performance of cloud environment. In: IEEE International Conference on Computational Intelligence and Communication Technology. (2015)
10. Kruekaew, B., Kimpan, W.: Virtual Machine Scheduling management on cloud computing using artificial Bee colony. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. (2014)
11. Mohapatra, S., Rekha, K.S., Mohanty, S.: A comparison of four popular heuristics for load balancing of virtual machines in cloud computing. *Int. J. Comput. Appl.* **68**(6), 0975–8887 (2013)
12. Chen, H., Wang, F., Akanmu, N.: User-Priority Guided Min-Min Scheduling Algorithm For Load Balancing in Cloud Computing. IEEE (2013)
13. Behal, V., Kumar, A.: Cloud computing: performance analysis of load balancing algorithms in cloud heterogeneous environment. In: Confluence The Next Generation Information Technology Summit (Confluence) 5th International Conference. (2014)
14. Garala, K., Goswami, N., Maheta, P.: A performance analysis of load balancing algorithms in cloud environment. In: IEEE, International Conference on Computer Communication and Informatics (ICCCI-2015). (2015)
15. Upendra, B., Ramanuj, P.: Enhanced max-min task scheduling algorithm in cloud computing. *Int. J. Appl. Innov. Eng. Manage.* **2**(4), 259–264 (2013)
16. Domanal, S., Reddy, G.: Load balancing in cloud computing using modified throttled algorithm. IEEE International Conference on Cloud Computing in Emerging Markets (CEEM), pp. 1–5. (2013)
17. Shoja, H., Nahid, H., Azizi, R.: A comparative survey on load balancing algorithms in cloud computing. In: 51th ICCNT. IEEE—33044 (2014)
18. Pasha, N., Agarwal, A., Rastogi, R.: Round Robin approach for VM load balancing algorithm in cloud computing environment. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* (2014)
19. Radi, M.: Efficient service broker policy for large-scale cloud environments. *IJCSI Int. J. Comput. Sci. Issues* **12**(1), (2015)
20. Buyya, R., Ranjan, R., Calheiros, R.: Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: challenges and opportunities. In: Proceedings of the 7th High Performance Computing and Simulation Conference (HPCS 09). IEEE Computer Society (2009)
21. Wickremasinghe, B.: Cloud analyst: a cloud-sim-based tool for modeling and analysis of large scale cloud computing environments. MEDC Project Report (2010)
22. Wickremasinghe, B., Calherios, R.: Cloud analyst: a cloud-sim-based visual modeler for analyzing cloud computing environments and applications. In: Proceedings of IEEE International Conference on Advance Information Networking and Application. (2010)

A Framework for Secure Communication on Internet of Things (IoT)



Mohammad Reza Hosenkhan and Binod Kumar Pattanayak

Abstract Internet of things (IoT) encompasses millions of resource constrained wirelessly connected devices that are very often prone to external malicious attacks. Hence, there arises a need for protecting these devices from such attacks. In this paper, we have proposed a secure communication scheme for resource constrained scheme where a device is identified by a key pair, public key and private key, and its unique identification (ID) that is assigned to it by Google Cloud IoT Core and the device stores the private key for communication with the server. Key generation uses a cryptographic approach. The communication can be facilitated by the device communicating its private key along with a cryptographically signed Java-based Web Token (JWT) to the MQTT bridge following which the connection can be established. The same JWT can be used for further communication as well.

Keywords IoT · Device · Security · Communication protocol

1 Introduction

Internet of things (IoT) has been an unprecedented and one of the most amazing revolutions in the recent years in the world of communications that has captured every field of human life. If traditional Internet could establish global communication among the people of the entire globe, it was further leveraged to IoT that could facilitate device-to-device (D2D) communication wherein the devices from any places of the world can communicate among themselves without any human intervention [1]. “Things” in IoT represent the devices connected to it. In the present scenario of IoT, millions of devices, mostly wireless, are connected to the global Internet, and as per

M. R. Hosenkhan

Department of Software Engineering, Faculty of Information and Communication Technology,
Universite Des Mascareignes, Beau Bassin-Rose Hill, Mauritius
e-mail: rhosenkhan@udm.ac.mu

B. K. Pattanayak (✉)

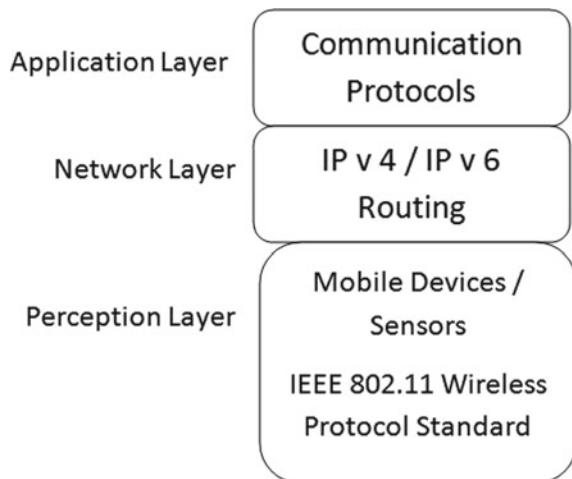
Department of Computer Science and Engineering, Institute of Technical Education and
Research, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India
e-mail: binodpattanayak@soa.ac.in

the experts, the number of devices to be connected to Internet may rise up to several billions by the end of the year 2025. The mobile devices connected to IoT are mostly small handheld ones operating with a radio frequency through a wireless channel supported by IEEE 802.11 wireless communication protocol and each such device is assigned with a radio frequency identification (RFID) tag for its unique identification. Such devices being small in dimensions operate with a limited memory space [2]. In addition, since these devices operate in wireless channels, very often they are susceptible to various attacks at each of the layers of IoT protocol stack. Hence, there arises a need of security measure at the level of IoT communication protocols for preventing the mobile devices from such attacks. IoT applications cover a wide spectrum of areas such as health care, education, agriculture, and environmental monitoring [3–5]. In this paper, we propose a security framework for communication of among the devices connected to an IoT environment.

2 IoT Architecture Protocol Stack

A simplified IoT architecture is depicted in Fig. 1. The perception layer that is identical to physical layer of OSI model protocol stack for communication incorporates various resource constrained devices, sensors for data collection and each mobile device is assigned with a RFID tag. Mobile devices operate on a radio frequency supported wireless channel that operates on standard IEEE 802.11 wireless protocol. These devices connect to the global Internet through gateways. Each device is uniquely identified by its RFID tag. The network layer operates on Internet protocol (IP) and that is equivalent to the network layer of OSI model of communication. It deals with routing functionality. The information routed between devices may

Fig. 1 Internet of things (IoT) protocol stack



be encrypted for security purposes. The application layer comprises of communication protocols, such as Message Queuing Transport Protocol (MQTT), Constrained Application Protocol (CoAP), that is responsible for communication of information between devices.

3 IoT Security Issues and Challenges

Since IoT environment incorporates huge number of resource constrained mobile devices, security concerns impose the major challenge for secure transmission of information between devices. The issues and challenges associated with IoT communication are detailed below.

3.1 Perception Layer

The issues and challenges pertaining to perception layer are as follows.

- (a) Unauthorized Access to RFID Tags: External malicious attackers attempt to modify the unique RFID tags of devices and as a result, the identification of devices becomes impossible. Lightweight cryptographic algorithms can be used to overcome this issue.
- (b) Tag Cloning: External attackers may tend to duplicate the RFID tags of devices as a consequence of which identification of the device becomes impossible for the receiver.
- (c) Eavesdropping: Malicious attackers may try to access the confidential information such as password which results in violation of data privacy that can be prevented using cryptographic systems like Rivest-Shakir0Adelman (RSA) and Data Encryption Standard (DES).
- (d) Spoofing: A device with a RFID tag may send fake information which may confuse the receiver if the information has been communicated by an authentic sender. This issue also can be resolved using the cryptosystems RSA and DES.
- (e) RF Jamming: The radio frequency channel along which the IoT devices communicate can get jammed due to presence of noise along the channel which can be resolved using dynamic risk assessment procedure.

3.2 Network Layer

The network layer security issues are listed below.

- (a) Sybil Attack: A device may tend to pretend its identity to other devices in the network that may lead to communication to an illegal device subsequently

resulting in loss of data. Strong authentication procedures must be used to protect from this type of attack.

- (b) Sinkhole Attack: Such an attack can bring harm to confidentiality as well as privacy of communicated data that can be resolved using hop-by-hop communication or source driven routing.
- (c) Malicious Attack: An external attacker may generate a malicious code to the network that subsequently may paralyze the functionality of the network which can be overcome using a strong authentication mechanism.

3.3 Application Layer

The issues related to IoT application layer security are:

- (a) Malicious code injection;
- (b) Denial-of-service (DoS) attack.

4 Related Work on IoT Communication Protocols

Application layer of IoT protocol stack comprises of the necessary communication protocols for performing the necessary communication between the IoT devices. A wide spectrum of research work has been carried out by various authors that are detailed below.

A MQTT protocol-based IoT communication system called SecT has been proposed by the authors in [6] that is thing-centered and secure lightweight wherein a device is capable of authenticating a legitimate user thereby preserving the user's privacy by virtue of end-to-end encrypted communication strategy. As claimed by the authors, this IoT communication system can be applicable to a wide range of IoT application domains. A comprehensive survey has been conducted by the authors in [7] for various protocols and standards implemented in IoT systems. Here, standards such as IEEE, IETF, and ITU are discussed in detail that significantly contribute to the growth of the IoT-based systems. A publish/subscribe secure communication protocol for IoT systems has been devised by the authors in [8] that operates with MQTT protocol. Three security levels have been provisioned here for communication on IoT environment. The first security level is meant for lightweight data communication. The second security level preserves the data privacy of sender and receiver and the third security level is implemented for robust long-term security between communicating entities. A secure lightweight identity-based cryptosystem for IoT devices has been developed by the authors in [10] that is capable facilitating secure authentication as well as data exchange among the devices on IoT where authors have used physically unclonable function (PUF) that is used for generation of public identities of IoT devices that can be used as the public key of the device in order for encryption of data. The proposed scheme as claimed by its authors is resilient against

active as well as passive attacks. A survey of communication protocols along with their potential for incorporation with fog and cloud computing has been addressed by the authors in [10]. This work can be very much useful designing integrated IoT-fog-cloud architecture-based systems. In smart city projects, IoT devices play a significant role in communicating the system-sensitive data about various urban infrastructure that inherently lack in data integrity which may create space for the malicious attackers to launch cyber-physical attacks on smart city projects. In order to address this issue, the authors in [11] have proposed a integrity-first IoT communication protocol using ethereum blockchain light client. A lightweight authentication protocol using key exchange procedure for IoT devices communicating via wireless channels has been proposed by the authors in [12] where two communicating devices operate with two unique keys for each of the devices, a master key along with a session key that is provided to the devices during the time of configuration that keeps on changing constantly. Here, the authors have used symmetric cryptography along with a hash message authentication code (HMAC)-based key derivation function (HKDF) in their implementation for the proposed scheme. In order to guarantee secure connectivity of resource constrained IoT devices operating in wireless channels, a lightweight proxy re-encryption scheme has been proposed by the authors in [13] and the authors claim to have successfully implemented the scheme for low-power and resource constrained devices on the IoT.

5 Proposed Security Framework

Security provisioning in an IoT device must take care of the following aspects.

- (a) Unauthorized network entities must not be able to access the data exchanged between an IoT device and the server;
- (b) Server should be capable of identifying each IoT device uniquely so as to detect outsiders sending fake data;
- (c) Security breach on one IoT device must not influence the security provisioning of other devices in the network.

The first issue can be resolved using Transport Layer Security (TLS) certificate which identifies an authorized IoT device.

While speaking about the second issue, using traditional user name and password pairs for unique identification of resource constrained wireless devices is not practicable. Rather a better solution to this is to use private key and public key pairs for unique identification of devices. As depicted in Fig. 2, an IoT device for the first time can register with Google Cloud IoT Core that creates a key pair along with a unique device ID. The private key along with the device ID is then stored with the device and device ID along with the public key is communicated to the device manager. As a result, the server can be capable of uniquely identifying each device. In the process of communication, when a device needs to send data, it attaches a cryptographically

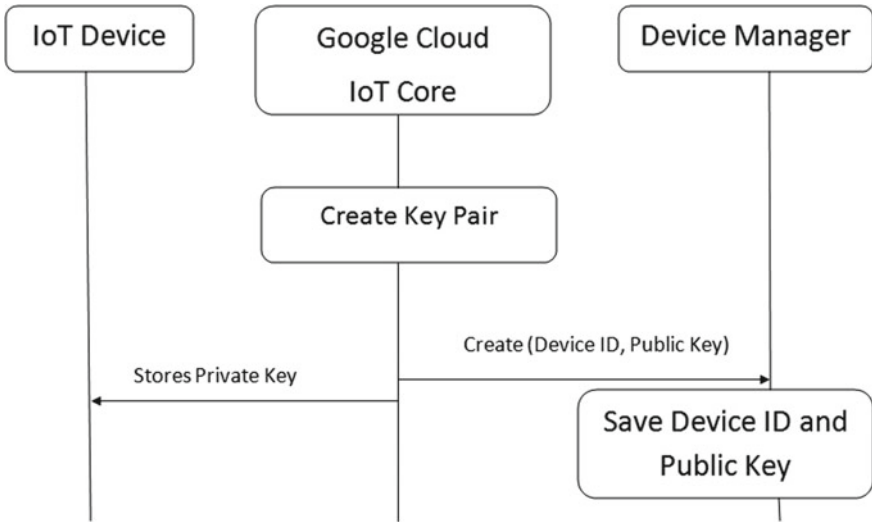
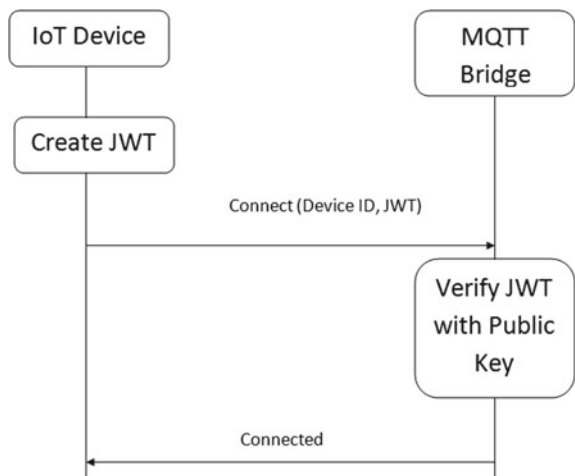


Fig. 2 Device registration and identification

signed Java-based Web Token (JWT) with the data. The server can verify the authenticity of the device by its device ID and key pair. Then, the device can be connected to the MQTT bridge after which the data communication can begin (Fig. 3). When JWT expires, the connection ends. During further communications, the same signed JWT can be used.

The third issue is resolved by the fact that each device uses a key pair which makes it difficult for a third party to compromise any device and as result, security breach on one device cannot affect the security provisioning of other devices.

Fig. 3 Device connection and session establishment



6 Conclusion and Future Work

Security issues in IoT environment present the major challenge for its implementers. In order to avoid security being compromised, the most important aspect is the unique identification of devices on IoT. Our proposed framework relies on assigning a key pair along with a unique device ID for unique identification of any device. This approach is more robust as compared to the traditional user name and password pair that can be easily compromised. This scheme can further be intensified using strong cryptographic algorithms for key generation. However, key exchange needs a strong authentication procedure which can be the future course of investigation.

References

1. Hosenkhan, R., Pattanayak, B.K.: A secured communication model for IoT. *Adv. Intell. Syst. Comput.* **863**, 187–193 (2019)
2. Pattanayak, B.K., Amic, S.: Modified lightweight aes based two level security model for communication on IoT. *TEST Eng. Manage.* **82**(1–2), 2323–2330 (2020)
3. Ramlowat, D.D., Pattanayak, B.K.: Exploring the internet of things (IoT) in education: a review. *Adv. Intell. Syst. Comput.* **863**, 245–255 (2019)
4. Mohapatra, S.K., Pattanayak, B.K., Pati, B.: A survey of IoT issues, methods and its implementation in higher education. *TEST Eng. Manage.* **83**, 5548–5568 (2020)
5. Rath, M., Pattanayak, B.K.: Technological improvement in modern health care applications using internet of things (IoT) and proposal of novel health care approach. *Int. J. Human Rights Healthc.* **12**(2), 148–162 (2019)
6. Gao, C., Ling, Z., Chen, B., Fu, X., Zhao, W.: SecT: a lightweight secure thing-centered IoT communication system. In: *Proceedings of the 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 1–9. (2018)
7. Salman, T., Jain, R.: A survey of protocols and standards for internet of things. *Adv. Comput. Commun.* **1**(1), 1–20 (2017)
8. Malina, L., Srivastava, G., Dzurenda, P., Hajny, J., Fujdiak, R.: A secure publish/subscribe protocol for internet of things. In: *Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES 2019)*, pp. 1–20. (2019)
9. Chatterjee, U., Chakraborty, R.S., Mukhopadhyay, D.: A PUF-based secure communication protocol for IoT. *ACM Trans. Embed. Comput. Syst.* **16**(3), 1–25 (2017)
10. Dizdarevic, J., Carpio, F., Jukan, A.: A survey of communication protocols or internet of things and related challenges of fog and cloud computing integration. *ACM Comput. Surv.* **51**(6), 1–29 (2019)
11. Reilly, E., Maloney, M., Siegel, M., Falco, G.: A smart city IoT integrity-first communication protocol via an Ethereum Blockchain light client. In: *Proceedings of the IEEE/ACM International Workshop on Software Engineering Research and Practices for the Internet of Things (SERP4IoT)*, pp. 53–56. (2019)
12. Rabiah, A.B., Ramakrishnan, K.K., Liri, E., Kar, K.: A lightweight authentication and key exchange protocol for IoT. In: *Proceedings of the Workshop on Decentralized IoT Security and Standards (DISS)*, pp. 1–6. (2018)
13. Ferretti, L., Marchetti, M., Colajanni, M.: Fog-based secure communications for low-power IoT devices. *ACM Trans. Internet Technol.* **19**(2), 1–21 (2019)

COVTrac: Covid-19 Tracker and Social Distancing App



Subhashish Das Mohapatra, Suwendu Chandan Nayak, Sasmita Parida, Chhabi Rani Panigrahi, and Bibudhendu Pati

Abstract In order to break the chain of SARS-CoV-2 virus infection, most countries are taking the help of mobile apps. The existing apps lack the necessary features to address the major issues of detection and alert of social distancing, detection and prevention of large gatherings, easy generation of travel permit during lock down, etc. The existing apps do not provide any solution for alerting user of the users social distance violation. Though many of these apps are being used for contact tracing, they do not say which places the user has visited recently which is a crucial parameter in contact tracing and preventing community spread. Smartphones empowered with the latest technology like Google geo-location and low energy Bluetooth (BLE) with such an app can complement a country's general Covid-19 control strategies—comprising testing, contact tracing, seclusion and social distancing. In this work, we develop a Covid-19 tracker and social distancing app named as *COVTrac*, an Android app that uses various capabilities of the smartphone to address these issues with the existing mobile app along with a number of other concerns such as privacy, accurate contact tracing and prevention of socio-physical interactions.

S. Das Mohapatra (✉)

Department of Computer Science & Information Technology, C. V. Raman Global University, Bhubaneswar, India
e-mail: subhashishdm@gmail.com

S. C. Nayak · S. Parida

Department of Computer Science & Engineering, Gandhi Institute for Technological Advancement, BPUT, Bhubaneswar, India
e-mail: suwendu2006@gmail.com

S. Parida

e-mail: sasmitaparida2004@gmail.com

C. R. Panigrahi · B. Pati

Department of Computer Science, Rama Devi Women's University, Bhubaneswar, India
e-mail: panigrahichhabi@gmail.com

B. Pati

e-mail: patibibudhendu@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_50

Keywords COVID19 · SARS-CoV-2 · Android · Bluetooth · BLE · GPS · Geo-Location · QRCode

1 Introduction

In December 2019, a global epidemic of an unidentified virus causing symptoms like pneumonia was exposed in the Chinese city of Wuhan [1]. The epidemic soon spreads throughout the world which imparted severe social and economic disturbance. This impressive spread of this global epidemic imposed confinement measures around the world.

Covid-19 often spreads when persons with the infection have close contact with non-infected individuals. This suggests spending just 15 min within two meters of a virus infested person is very likely to infect someone without infection [2]. Therefore, social distancing is the key to prevent the infection. The reduction of physical social interaction between people is the key in order to reduce the transmission of coronavirus (Covid-19). Physical social distancing is a non-pharmaceutical intervention [3] or measure intended to stop the spread of a contagion by keeping physical distance between individuals and decreasing the period of close contact with one another. Though preventive measures are taken in order to avoid close contact, it is still not possible to identify those who have symptoms or are contaminated by others merely inspecting at them from outside.

1.1 Motivation

A number of technological tools have been developed to monitor the spread of Covid-19 and tracing of individuals coming in contact with infected persons [4]. These tools also help in relaxing the confinement measures by tracking and keeping the pandemic local. These tools are often required to track people and their behavior by the authorities in order to restrict the spread of virus. In normal conditions, these tools would be considered to be violating privacy of an individual. Also, there is a need of energy efficient technique to be utilized [5], as many of these applications drain the energy of mobile devices in order to store and retrieve the user-specific data from their servers. Some research projects devised the way to trace vicinity while protecting a high level of privacy [6].

Android is a popular mobile platform which provides a lightweight framework for application development, installation and maintenance. Android open source has a large library to monitor various activities of the user using smartphone [7]. Major smartphones, wearable devices and IoT manufacturers support BLE technology [8], which is a short-range wireless standard. Android mobile devices equipped with Google geo-location services and BLE technology can be utilized to trace and track

the infected or contacted person, public gathering and clustering to maintain social distancing.

1.2 Contributions

The proposed mobile app uses a mechanism inspired by decentralized privacy-preserving proximity tracing (DP3T) system, originally proposed by the Pan-European privacy-preserving proximity tracing (PEPP-PT) project in order to track and alert users [9]. The application will also enable users to get latest updates of Covid-19 from reliable sources and allow users to call the emergency service numbers as per the requirement of the user.

The remaining parts of the paper are organized as follows. Section 2 presents the proposed system. The implementation and working of the app are described in detail in Sect. 3. Section 4 presents the experimental results of the proposed system, and Sect. 5 concludes the paper.

2 Proposed System

COVTrac is an Android-based mobile application which needs to be installed by all smartphone users. At the time of installation, it will authorize the user by validating the mobile phone by using OTP-based authentication system. At the initial configuration phase, the app will register the user details such as firstname, lastname, age, gender, address, district, block, city and e-mail. The data so collected are safely saved on a central repository using Firebase cloud services. In order to preserve privacy, the data of a user are not shared with any Government agency until a person is found infected by the SARS-CoV-2 virus. The home screen of the app has latest information about the contamination such as total number of infected people, number of active cases, number of recovered cases and number of deceased gets displayed along with the graphical view of the Covid-19 infection progress. Data are fetched from various authenticated sources by using Web scraping tools. There are a number of open source API available for Covid-19 data, which are volunteer-driven, crowd sourced databases for stats and patient tracing in India. This app uses one such API “COVID19- India API” [10].

Some clinical questionnaires must be filled every day by the user. The advantage of attending the daily questionnaires is that the user will be able to self-assess themselves for identifying possible SARS-CoV-2 virus infection. The user can generate a QR code-based daily pass, which is implemented in this app. When the user clears all the questionnaire, the app will generate a colored QR code. This code will allow the user to go outside and perform necessary tasks in lockdown situation, where restrictions are imposed. It removes the hectic task of the commoners where they are required to get pass authenticated by local authorities. The authorities have also a very tedious

Table 1 Comparison of *COVTrac* with other apps in India

Features	<i>COVTrac</i>	Arogya Setu	Corona Kavach	Odisha COVID Dashboard
Self diagnostic	YES	YES	NO	YES
Social distancing	YES	NO	NO	NO
Proximity tracing	YES	YES	NO	NO
Exposure notification	YES	NO	NO	NO
Information	YES	YES	YES	YES

task of verifying and allowing each person applying for a pass. The whole process is secure as the person who has obtained the pass only needs to show the QR code generated on the mobile itself, which can be verified by the personnel deployed using another mobile app. The tracking of suspected individuals becomes easy. The QR code can be used as an entry pass to any shop, mall or for any movement.

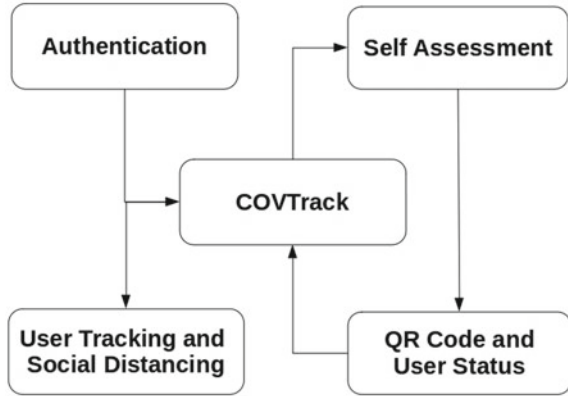
The Google geo-location which is the primary component of the tracker enables tracking of individuals in a certain geographic location [11]. The *COVTrac* tracks the movement of users, and when a person gets close to a certain area with another person having SARS-CoV-2 virus infection present or has certain symptoms of corona, then it alerts the user in a real time. It might happen that the user may disable location tracking, in that case, the status of that user and QR code will be changed to danger level, and the user is alerted.

BLE standard-based beacons are gradually getting preference as enablers for proximity-sensitive experiences [12]. Such beacons are used as proximity beacons and are used to enable location-based services and object-tracking beacons, which enable portable physical objects tracking. For keeping social distancing, both Google geo-location and Bluetooth of the device are utilized along with implementation of energy efficient mechanisms [13]. The grouping of 3 or 4 persons for a long time at a single location will create alerts in the mobile app. The distance between two persons are tracked with Bluetooth in the mobile app and an alert will be generated when the two come closer than two meters. Table 1 presents a comparative view of features [14] for a few Covid-19 apps with *COVTrac*.

3 Implementation and Working

The *COVTrac* is user-friendly and secure. It uses all security policies and supports Android Lollipop (API level 21) onward versions of Android.

Fig. 1 Android architectural framework and module development



3.1 System Architecture

The architecture of the proposed mobile app along with the different modules is shown in Fig. 1. The proposed app *COVTrac* consists of four modules namely:

- Authentication
- Self-Assessment
- QR Code and User Status
- User Tracking and Social Distancing.

Authentication: The authentication of a user is one of the important aspect of any user-centered mobile app. In *COVTrac*, authentication of the user is done by using one time password (OTP) mechanism with user’s mobile number. Further, this module also asks the user to allow a certain necessary Android features to be enabled, thereby providing a level of security to the Android device and the app. Following are the necessary features of Android which the user is required to give permission for:

1. *SMS Permissions:* This is required as the app will automatically read the received SMS when the user receives OTP on the Android mobile.
2. *LOCATION Permissions:* This is required as the app utilizes Google geo-location services for tracking the user.
3. *BLUETOOTH Permissions:* This is required as the app uses Bluetooth for tracking of social distancing status of the user
4. *CALL permissions:* This required permission allows the app to make phone calls on behalf of the user.

A number of user details are also registered from the user in this module, such as firstname, lastname, age, gender, street address, district name, block name, vil- lage/city name and e-mail (which is optional) used for sending occasional alert and other information to the inbox of user. The data furnished by the user are never shared

with any third party, unless it is absolutely necessary, such as in case of a user gets infected or comes in close contact with another person.

Self-Assessment: This module assesses a user based on the responses to a number of clinical questions asked and determines the status. This module has a total of eight clinical questions as per the recommendation of World Health Organization(WHO). Each question has a number of options to select. When a user exercises a choice, a value is assigned in each step of attending the questions. Table 2 represents the maximum weights assigned to each question in self-assessment module.

Based on Table 2, for answers chosen by the user, a score is assigned to each question, and the total is computed out of 14. The weigh is calculated for the assessment using Eq. 1, which gives the status of a user as per Table 3.

$$\text{weight} = \text{Round}((\text{total scores}/14) * 2) \quad (1)$$

The calculated weight value results in one of the values—0, 1 or 2.

QR Code and User Status: The QR code is used as a permission for the person to access various services and to step outside. A unique hash code is obtained from the user registered data from authentication module, and as per the result of self-assessment, the QR code is generated. The color of QR code and status of the user is obtained as specified in Table 3.

The user status is RED by default, which means the user is not permitted to go outside. In order to change the status code from RED to YELLOW or GREEN, the user has to take a self-assessment test on daily basis. The color of QR code and status may change again to RED based on a certain criteria which are described in Table 4.

User Tracking and Social Distancing: This module of the app utilizes Google geo-location services and BLE. Google geo-location service monitors user's movement by tracking the mobile device, returns a location and accuracy radius based on information about cell towers and Wi-Fi nodes that the mobile device can detect [15]. Using this information, users movement around crowded area and interaction with other app users can be determined.

The BLE is used as a beacon to determine the distance between two users with respect to each other. BLE uses RSSI measurements to measure the distance. Every advertisement package in BLE includes the RSSI value. Android OS broadcasts the RSSI system-wide on a regular basis while using Bluetooth or Wi-Fi. It represents the relative strength of the signal being received by the device. The distance is calculated based on the RSSI value on the mobile device, and the RSSI is obtained using the Eq. 2.

$$RSSI = Tx - 10 * n * \lg(d) \quad (2)$$

where $RSSI$ is received signal strength indicator, Tx is the transmission power of the device, n is a constant and d is distance of the device.

In free space, $n = 2$.

Thus, from Eq. 2, the formula for calculating distance can be derived as

$$d = 10^{((Tx - RSSI)/(10 * n))} \quad (3)$$

Table 2 Self-assessment questions and scores

S. No.	Category	Question	Max. Wt.	Score
1	Body temperature	Please let us know your current body temperature	2	0, 1, 2
2	Symptoms	Are you experiencing any of the following symptoms?	2	0, 1, 2
3	Additional symptoms	Do you have any additional symptoms?	2	0, 1, 2
4	Diseases/health condition	Do you have/ever had any of the following?	2	0, 1, 2
5	International travel history	Have you traveled anywhere internationally during last 14 days?	1	0, 1
6	Domestic travel history	Have you traveled anywhere outside Odisha within India during last 21 days? or have you returned from outside Odisha recently within 21 days?	1	0, 1
7	Contact tracing	Did you have a close contact with a confirmed Covid-19 patient in last 21 days?	2	0, 2
8	Contact tracing	Do you work in a care facility or have been in close contact with someone working in care facility?	2	0, 2

Table 3 Color codes

Color code	Weight	Status
RED	2	High risk
YELLOW	1	Medium risk
GREEN	0	Low risk

Table 4 Relationship of RSSI value with distance

Range of RSSI (dBm)	Distance (m)
RSSI > -51	0
-51 > RSSI > -52	1
-52 > RSSI > -53	2
-53 > RSSI > -54	3
-54 > RSSI > -56	4
-56 > RSSI > -58	5
-58 > RSSI > -60	6
-60 > RSSI > -65	7
-65 > RSSI > -69	8
-69 > RSSI > -76	9
-76 > RSSI	10 or more

Table 5 Criteria for change of status

S. No.	Criteria for change of status from YELLOW or GREEN to RED
1	If the user has not taken the self-assessment for the day (Self-assessment is mandatory each day)
2	If the user comes in close contact with other user who is infected by SARS-CoV-2 virus
3	If the user does not maintain social distancing norms for 15 minutes
4	If the user's age is above 60 or below 10

In order to provide privacy, each beacon uses a random id generated using the user's registration information. When two users are close to each other with less than two meters, this is considered as a violation of social distancing. The output of this module also results in change of status for the user. Table 5 presents various factors that affect the status of a user.

3.2 Working

Step 1: After installing the app when user starts the app for first time, the app requests user to permit the necessary four features of the mobile to be used. Once the user accepts to allow the above permissions, the app asks user to enter mobile number for generating and sending OTP for authentication (Fig. 2a). Once user enters the mobile number, a secure code of six digits is generated by the back end system which sends the code through SMS to the mobile.

Step 2: OTP is read by the application and automatically put in the field where it is required to be entered (Fig. 2b). User can manually type the OTP by reading the

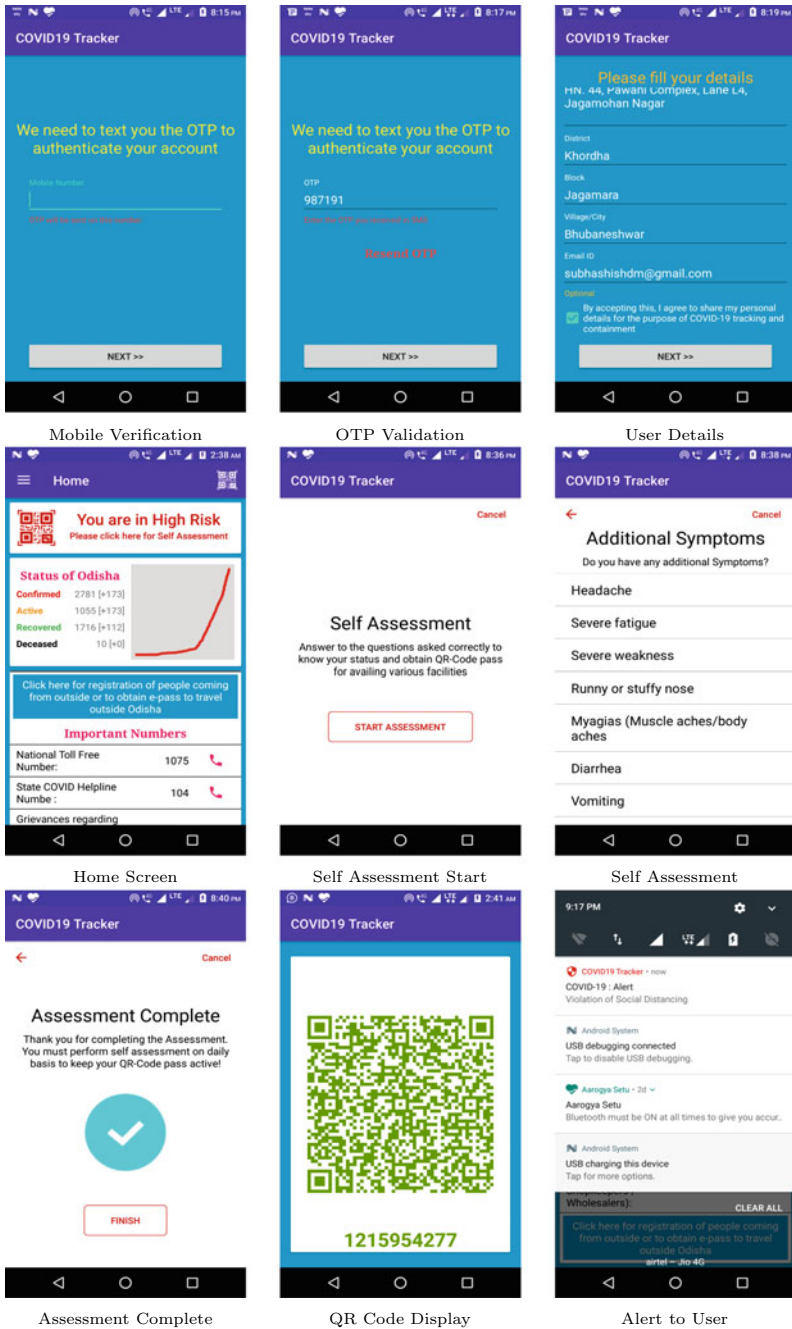


Fig. 2 Screenshots of various steps in COVTrac

OTP message also. The user gets verified, by matching the OTP received with the generated OTP in the back end, and user is taken to one time setup screen.

Step 3: In the one time setup screen, the user fills necessary information which are required for administration purposes only (Fig. 2c). Upon successful submission of information, the user is taken to home screen. This is the point where geo-location and BLE tracking are started.

Step 4: Home screen (Fig. 2d) is the primary screen where the user arrives every time the app runs. Components of this screen are self-assessment status, Covid-19 general status, important link and mobile numbers for emergency services. The “self-assessment status” shows colored user status for the day. The general status, which gets displayed below the self-assessment, displays following information up to current date.

- number of confirmed Covid-19 cases
- number of active Covid-19 cases
- number of recovered persons
- number of deceased

This section also has a graphical representation of change in Covid-19 confirmed cases, which gives an over all idea regarding the growth/decline of new cases. Below this section, there is a link to the portal where people coming from outside state/country can register their information. Along with it, a number of important contact numbers displayed in a list, which facilitate the user to contact the needed authorities directly from the mobile app itself.

Step 5: For taking the self-assessment test, the user has to click the status on homepage to open the self-assessment page. This consists of a number of questions (Table 2) for determining the status and generating QR code (Fig. 2e–g).

Step 6: For viewing QR code, user has to click on the small QR code icon on the top right corner in the home screen. It will display the colored QR code (Fig. 2h) according to the user status.

Step 7: Whenever user violates restrictions or social distancing norms, an alert (Fig. 2i) is shown on the mobile. This alert is displayed regardless of the fact that app is running or not. This works in the back end, and alert is created according to the user status based on geo-location and Bluetooth tracking.

4 Experimental Results

The *COVTrac* was developed and installed on a number of Android mobile devices. During several test runs, the data obtained were used to test the functionality and accuracy of the methods implemented in the app (Fig. 2).

Result of Self-Assessment: Figure 3 shows relationship between sum of individual score of each question and the final calculated weight based on the sum. It is evident from the plot that as per the options exercised by the user, the final weight of self-

Fig. 3 Self-assessment results

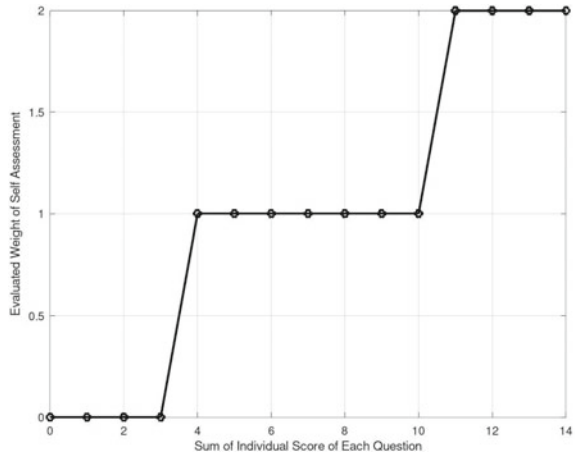
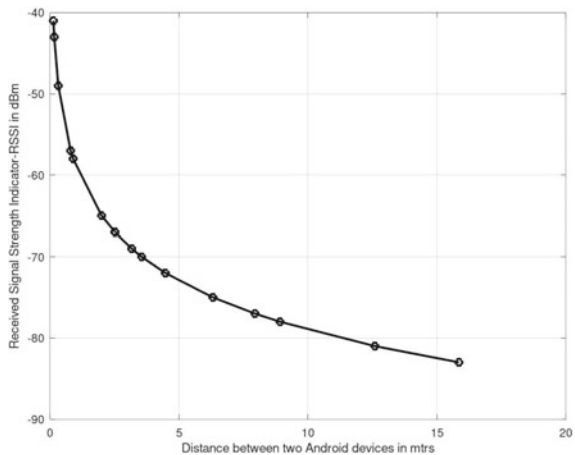


Fig. 4 RSSI versus distance curve

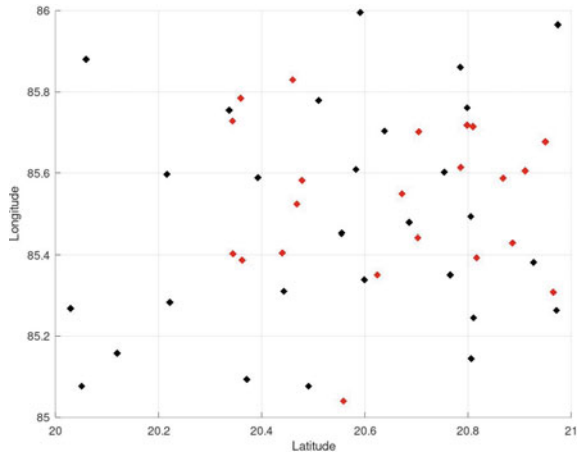


assessment has one of the three values 0, 1 or 2 as per Eq. 1. Based on this value, the predicted status of the user can be set as green, yellow or red.

Social Distance Tracking Using Bluetooth: RSSI value is easily obtained in Android without any special permission. A number of test data is collected, and it was found that as the RSSI value is increased, the distance is decreased. This means that the closer the two devices the greater the RSSI value. From Fig. 4, this fact is evident. Using Table 4 which represents the relation between RSSI and distance, the COVTrac determines proximity of two users and alerts the user if social distancing is not maintained.

GPS Tracking: A sample of 50 users were taken from the geo-location data obtained by using GPS tracking feature of theCOVTrac, and the data values are plotted as shown in Fig. 5. The red dots represent the users who are violating social

Fig. 5 Sampled location tracking data



distancing norms, and the black dots represent the normal users. This feature enables the authorities to track the violators and to take action against them.

5 Conclusion

It is evident that contact tracing and social distancing are the key to prevent the spread of SARS-CoV-19 virus. A better mobile app like *COVTrac* is the only thing which can be used as a weapon against this invisible enemy. In this work, we have developed a mobile app which has many features implemented to track social distancing, maintain privacy and security of users, etc. Though the app has most of the major features in built into it, a few feature enhancements can also be done. The app can be optimized for energy consumption. The app can include features like daily logging of other health parameters like blood pressure and temperature for easy prediction of infection. Geo-fencing can be added so that infected user and authorities can be alerted when moving out of containment. Integration of the app with external sensors through IoT technology can also be an added advantage so that the health parameters can be automatically recorded through the app.

Although there are many mobile apps are available and may be a number of more apps in development stage, *COVTrac* provides better features as compared to the existing apps. It not only prevents social gathering and close contact but with additional features like QR code-based tracking and permission system. It also addresses the problems related to privacy and security of users as present in existing apps.

References

1. WHO report on Pneumonia of unknown cause in China [Online], Available: <https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/>. Last accessed: 23 Apr 2020
2. Yuen, K., Ye, Z.W., Fung, S.Y. et al.: SARS-CoV-2 and COVID-19: The most important research questions, *Cell Biosci*, pp. 10–40 (2020)
3. Lakhani, H.V., Pillai, S.S., Zehra, M., Sharma, I., Sodhi, K.: Systematic review of clinical insights into novel coronavirus (CoVID-19) pandemic: Persisting Challenges in U.S. Rural Population. *Int. J. Environ. Res. Public Health* **17**, pp. 42–79 (2020)
4. COVID-19 Innovations in Healthcare Responds [Online], <https://www.innovationsinhealthcare.org/covid-19-innovations-in-healthcare-responds/>. Last accessed: 28 Mar 2020
5. Sarkar, J.L., Panigrahi, C.R., Pati, B., Trivedi, R., Debbarma S.: E2G: A game theory-based energy efficient transmission policy for mobile cloud computing. In: *Progress in Advanced Computing and Intelligent Engineering*, vol. 563. Springer, Berlin, pp. 677–684 (2018)
6. Vaudenay, S.: Analysis of DP3T: Between Scylla and Charybdis, *IACR Cryptol. ePrint Arch.*, vol. 2020, pp. 399–410 [Online]. Available: <https://eprint.iacr.org/2020/399>. Last accessed: 22 Apr 2020
7. Bluetooth low energy overview [Online], Available: <https://developer.android.com/guide/topics/connectivity/bluetooth-le>. Last accessed 5 Apr 2020
8. Bluetooth SIG, Specification of the Bluetooth System, Covered Core Package; Version 4.2; Bluetooth Special Interest Group (2014)
9. Troncoso, C. et al.: Decentralized Privacy-Preserving Proximity Tracing, PEPP-PT, v3 (2020) [Online], Available: <https://github.com/DP-3T/documents/blob/master/DP3T%20White%20Paper.pdf>. Last accessed: 08 Apr 2020
10. COVID19-India API. [Online], Available: <https://api.covid19india.org/>. Last accessed: 08 Apr 2020
11. Pati, B., Sarkar, J.L., Panigrahi, C.R., Debbarma S.: eCloud: An efficient transmission policy for mobile cloud computing in emergency areas. In: *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, vol. 719. Springer, Berlin, pp. 43–49 (2018)
12. Hassidim, A., Matias, Y., Yung, M., Ziv, A.: Ephemeral Identifiers: Mitigating Tracking & Spoofing Threats to BLE Beacons, 2016, [online] Available: <https://developers.google.com/beacons/eddystoneid-preprint.pdf>. Last accessed: 12 Apr 2020
13. Panigrahi, C.R., Sarkar, J.L., Pati, B., Bakshi, S.: E3M: An energy efficient emergency management system using mobile cloud computing. In: *IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, Bangalore, pp. 1–6 (2016)
14. COVID-19 Apps, Wikipedia [Online], Available: https://en.wikipedia.org/wiki/COVID-19_Apps. Last accessed: 12 Apr 2020
15. Pati, B., Sarkar, J.L., Panigrahi, C.R., Tiwary, M.: ECHSA: An energy-efficient cluster-head selection algorithm in wireless sensor networks. In: *Mining Intelligence and Knowledge Exploration, Lecture Notes in Computer Science*, vol. 94680. Springer, Berlin, pp. 184–193 (2015)

JOB-DCA: A Cost Minimizing Jaya Optimization-Based Data Center Allocation Policy for IaaS Cloud Model



Sasmita Parida, Bibudhendu Pati, Suwendu Chandan Nayak,
and Chhabi Rani Panigrahi

Abstract Cloud computing presents the new era of revolutionary computing of Internet to fulfill the user demands for computing that provides pool of services and resources that are globally distributed in different regions around the globe. The end user demands for the adequate resources are processed on pay-per-use. The end user computing is served through virtual machines (VMs), which are created and allocated at different data centers for the on-demand user request. Allocation of data centers is challenging one for the on-demand request. The data center allocation mechanism must consider different parameters of user data center and network. Most of the works address few of the parameters and applied optimization techniques. However, the implementation of new optimization techniques always opens a scope for the researchers in different fields. In this work, we discussed the Jaya optimization-based data center allocation (JOB-DCA) mechanism for infrastructure as a service (IaaS) cloud. The work considers selected parameters to find the optimized data center list and allocate to the on-demand resources with minimum cost. The presented JOB-DCA allocation technique tries to allocate the preferred data centers with reduced total VM cost. A new cost function is defined to minimize the cost and the existing benchmark cloud simulation tool known as CloudAnalyst is used to analyze the performance of the JOB-DCA technique. The proposed technique is compared with the benchmark mechanisms, and the results show 3.2% minimized cost in the JOB-DCA.

S. Parida (✉) · B. Pati · C. R. Panigrahi
Department of Computer Science, Rama Devi Women's University, Bhubaneswar, India
e-mail: sasmitamohanty5@gmail.com

B. Pati
e-mail: patibibudhendu@gmail.com

C. R. Panigrahi
e-mail: panigrahichhabi@gmail.com

S. Parida · S. C. Nayak
Department of Computer Science & Engineering, Gandhi Institute for Technological
Advancement, BPUT, Bhubaneswar, India
e-mail: suwendu2006@gmail.com

Keywords Jaya optimization · Allocation · Revenue · Routing · Cloud computing

1 Introduction

An assortment of topographical districts of data centers in cloud computing enables access of computing resources. As indicated by the client's request, the cloud offers types of service utilizing with targeting the proper resource utility and user satisfaction. To address the unlimited and high availability of sufficient resources, the inter-process correspondence among various data centers must be discussed so that the cloud permits imparting data to one another [1, 30]. However, the clients and cloud specialist co-ops need a center level for proficient mechanism. The middle person is liable for simple arrangement, distribution and the board of cloud administrations to the clients. Presently a day, we fundamentally center on the vicinity of end clients so specialist organizations guarantee greatest fulfillment of clients with all accessibility of assets. Because of the fast increment in various clients over cloud, the request of administrations is additionally expanded. The allocation and accessing of the on-demand resources, while considering the cost, revenue and utilization is a challenging one in cloud computing. Because of the immense interest of users, it is a provoking viewpoint to pick the most appropriate cost-effective data center [2, 29].

The prerequisites of various cloud clients differ as far as limitations are associated with least response time, cost, deadline time and so forth. Be that as it may, the test is to look through the best attainable data center to execute a specific client's solicitation with determined requirements [3]. Further, on account of clients having comparable sorts of limitations are designated with most fit data center and gets before more overburdened. For the most part, the server farms get overburden because of ill-suited determination which brings about non-uniform burden adjusting as a fast increment in various customers and their solicitations [4, 5]. Then again, the general execution will diminish as waiting and response time increment because of more dismissal of solicitations. So, an ideal portion of the most appropriate data center to customers depends on their various limitations and needs in a productive manner. Consequently, the broker assumes a job of interface among users and assigns the data center.

1.1 Motivation and Contributions

The proposed JOB-DCA algorithm is motivated by Jaya optimization technique which solves complex problems. The powerful computing resources need allocation in an optimum fashion to avoid mis-utilization of resources. The mis-allocation of powerful computing resources impacts high penalty such as increase of total cost, response time and processing time. However, for better data center allocation we apply Jaya to optimize the total cost.

In this work, we propose JOB-DCA allocation using population-based optimization technique. We believe that efficient allocation of data centers to the user request will result in cost minimization by considering the cost parameters such as VM cost and data transfer cost in cloud environment. The proposed JOB-DCA is studied and endorsed through broad investigations using ten scenarios in the cloud analyst. We have also proposed another intermediary strategy system to designate request to various data centers that are found topographically in various areas. The work limits the all-out expense and decreases the overall cost to maintain quality of service (QoS) parameter.

The rest of the work is organized as follows: Section 2 examines the current existing techniques related to allocation of VM. The proposed JOB-DCA mechanism with the expediting calculation and the solution implementing Jaya optimization approach are described in Sect. 3. Section 4 highlights the presentation and correlations of simulated results. Section 5 concludes the work along with highlighting the future extension.

2 Related Work

This section presents about current-related attempts to DC allotment and cost-based VM designation and intermediary strategies. Over the most recent couple of years, different allocation and VM distribution calculations have been proposed to choose proper data center dependent on client demand. VM allocation strategy aims to choose the most reasonable data center with effective consideration of numerous components along with cost [6]. Gaetano et al. [7] discussed QoS-based intermediary strategy utilizing the genetic methodology. It tackled the issue of overbooking issue in distributed computing. The authors in [8] proposed a value model to augment the income of cloud specialist co-ops. It utilized the dynamic conduct of cloud clients to expanding income. In [9], authors introduced a task planning system to asset the executives. It planned to oversee assets for the group subsidizing model. Manasrah et al. [10] proposed an upgraded administration representative approach for mist and cloud computing. The authors in this work talked about the new specialist strategy utilizing the differential advancement calculation to limit reaction time, preparing time and by and large expense in six distinct scenarios.

The role of optimization techniques is quite impressive among the researchers in engineering, industry, management and Internet computing. Many optimization models are suggested to solve complex problems in different domains. Jaya is one of the latest and simple optimization techniques to solve constrained and unconstrained optimization problems [11]. Researchers published more than 30 articles on Jaya optimization techniques in different reputed journals and claimed the performance is better than the benchmark optimization techniques [12, 13]. A few researchers also studied implementation of Jaya optimization technique in cloud environment to solve different issues. Mohanty et al. [14] discussed the implementation of Jaya optimization technique for load balancing in cloud computing. The authors claimed that it uses

Table 1 Related work on Jaya optimization technique in engineering domain

Article	Year	Application area	Objective
[12]	2018	Wireless network	Intelligent path selection for routing
[14]	2019	Cloud computing	Load balancing
[15]	2020	Cloud computing	Energy optimization VM placement
[16]	2019	Cloud computing	Workflow scheduling
[17]	2019	Cloud computing	Security to healthcare data
[18]	2020	Control system	Achieve optimal transient response
[19]	2016	Power distribution	Optimum power flow
[20]	2018	Clustering	Obtain optimal threshold in cluster
[21]	2016	Power system	Power quality improvement
[22]	2019	Wind power system	Online load frequency control

less control parameter and performs better than the other evolutionary approaches. Reddy et al. [15] proposed an energy optimization VM placement mechanism using Jaya algorithm in cloud computing and claimed the better result as compared to PSO mechanism. Gupta et al. [16] proposed a workflow scheduling using Jaya mechanism and claimed that Jaya performs better than other algorithms. Jaya optimization technique produces similar results and converges very quickly with minimum time. Sudha et al. [17] discussed a hybrid approach of Jaya and whale optimization algorithm for securing healthcare data in cloud computing. As a challenge, we motivated to implement Jaya optimization technique to optimize cost from the above study in cloud computing. The existing work of Jaya optimization techniques in different engineering domains is summarized in Table 1.

3 Proposed Work

Let $R = \{R_1, R_2, \dots, R_m\}$ be the set of regions, $DC = \{DC_1, DC_2, \dots, DC_k\}$ is the set of data centers and $\forall DC_i \in R_j$, where $i \in k, j \in m$. Each DC_i consists of set of physical machine units (PU): $DC_i = \{PU_1, PU_2, \dots, PU_p\}$, and PU_i represents with set of VMs: $PU_i = \{VM_1, VM_2, \dots, VM_v\}$. Consider set of user requests $U = \{U_1, U_2, \dots, U_N\} | U_{i=1 \dots N}$. We explore efficient resource allocation and propose the fitness function $C(VM_{total})$ to evaluate the on-demand resource allocation with minimized cost.

On-demand resource allocation associates with multiple parameters such as $\{R, AR, OS, VMM, VMP_{cost}, M_{cost}, ST_{cost}, DT_{cost}, PU\}$ where R = the region where the data center located, AR = architecture of the data center, OS = on-demand operating system to compute the user application, VMM = VM manager, VMP_{cost} = the cost per VM, M_{cost} = memory cost, ST_{cost} = storage cost, DT_{cost} = data transfer cost and PU = the physical hardware units. The existing fitness function involves

the cost and transfer time. It involves execution cost and data transfer cost, but does not discuss the cost associated with these. We derive the fitness function considering different associated parameters like cost and time as in Eq. (1)

$$C(\text{VM}_{\text{total}}) = C(\text{VM}_{\text{cost}}) + C(\text{DT}_{\text{cost}}) \tag{1}$$

The VM_{cost} is derived using VMP_{cost} , M_{cost} and ST_{cost} as in Eq. (2).

$$C(\text{VM}_{\text{cost}}) = C(\text{VMP}_{\text{cost}}) + \sum_{i=1}^x C(M_{\text{cost}}) + \sum_{i=1}^y C(\text{ST}_{\text{cost}}) \tag{2}$$

where x defines the number of required memory units with respect to MB of main memory and y signifies the required storage units with respect to X-MB [23, 24]. These values are defined by service provider during data center configuration using different pricing models [25].

Similarly, DT_{cost} is derived in Eq. (3) from available bandwidth ($\text{BW}_{\text{available}}$), D_R and the number of PU as:

$$C(\text{DT}_{\text{cost}}) = \sum_{i=1}^{\text{PU}} \left(\beta * \frac{D_R}{\text{BW}_{\text{available}}} \right) \tag{3}$$

where β defines the cost/GB in \$ during data center configuration.

We implement the Jaya optimization algorithm [13] for allocation of suitable VM belonging to a particular data center. The data center controller monitors the virtual machine activities. There is a mapping between a user request and the data center. Based on availability of VM in data center, the allocation is made along with consideration of cost factor associated. Jaya algorithm is a meta-heuristic optimization algorithm based on population size [14]. In each iteration, we consider the allocation of suitable DC based on minimized cost. Here, other DC parameters are involved to find the fitness value like data transfer cost and VM cost to execute the request, response time, execution time, network delay, etc. But, basically the performance of DC allocation mechanism varies due to the cost per VM and data transfer cost. Thus, in this proposed mechanism, we consider DT_{cost} and VM_{cost} as contributing parameters for calculation of fitness at each iteration. The fitness function $f(\text{DC})$ is to be minimized and is formulated as in Eq. (4):

$$f(\text{DC}) = C(\text{VM}_{\text{total}}) \text{ subject to minimized} \tag{4}$$

We derive the initial solution as the data center in the region with minimum VM cost and data transfer cost as the initial population and is formulated as in Eq. (5).

$$DC_{0,k}^1 = \min((C(\text{VM}_{\text{cost}}), C(\text{DT}_{\text{cost}})) \tag{5}$$

Similarly, we can consider m number of contributing problem design parameters and n number of possible candidate solutions where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$. From all the candidate solutions, we have to find the best DC to be allocated to the userbase. Here, $DC_{i,k}^j$ is the i th candidate of the population at j th iteration with k th problem design variable. For each iteration, the candidate gets updated according to Eq. (6).

$$DC_{i,k}^{j+1} = DC_{i,k}^j + r_{1,k}^j \left[\left(DC_{i,\text{best}}^j \right) - \left| DC_{i,k}^j \right| \right] - r_{2,k}^j \left[\left(DC_{i,\text{worst}}^j \right) - \left| DC_{i,k}^j \right| \right] \tag{6}$$

Here, $DC_{i,\text{best}}^j$ and $DC_{i,\text{worst}}^j$ are the best and worst candidate solutions for j th iteration. $r_{1,k}^j$ and $r_{2,k}^j$ are two random variables for j th iteration with k th problem variable. The cost of VM is varied as the number of requests increases. The DT_{cost} varies from DC to DC as depicted in different locations. The allocation of DC for each user request is changed. The $\text{Cost}_{\text{Total}}$ specifies the amount that has to pay against the usage of VM which is minimized, and henceforth the total cost is also minimized by Jaya algorithm. The allocation parameters are defined as given in Eq. (7).

$$\text{Cost}_{\text{Total}} = \sum_{\text{resource}^i} (C_i * t_i) \tag{7}$$

where C_i represents the cost of resource i per unit time and t_i represents the time utilization of resource i . The data center controller implements the new allocation policy which matches the suitable DC for the desired user request. The user requests are routed to the DCs which are in different geographical locations. The allocation of DC must consider the response time, data transfer cost and VM cost for different configurations.

Algorithm 1:JOB-DCA
 Inputs: $dcList=[1,2,\dots,d]$, $regionList=[1,2,\dots,r]$, where $r < d$
 Output: $dcName$

```

1  Begin
2  Initialization
    Set population size =  $n$ 
    Set Number of problem design parameters =  $m$ 
    Set Termination criteria
    Set Initial candidate value as Eq.5
3  do
4      Find Fitness function using Eq.4
5      Find best and worst candidate in the DC population for allocation
6      Update the Candidate using best and worst solution and random
       variables using Eq.6
7      if  $Cost_{total}[DC_{i,k}] < Cost_{total}[DC_{i,k}^*]$ 
           then
           Accept and replace the previous allocation
8      else
           Do not replace the previous solution
9  While (! termination criteria satisfied?)
10 End of do while
11 Report the optimized solution with minimum Cost
12 Find DC index
13 Return  $dcName$ 
14 Stop
    
```

The proposed Jaya optimization-based data center allocation (JOB-DCA policy) algorithm explains the computation of VM cost, DT cost and total cost as given in Algorithm 1. It optimizes these cost parameters through the formulation of different cost functions as given in Eqs. (1), (2) and (3). The derivation of these cost functions aims to reduce the value. JOB-DCA policy finds the best and worst solutions as Jaya optimization considers less parameter to compute. Here, the population size n is set as the number of data centers which are varied in simulation while considering different scenarios and the design parameter as the VM cost and DT cost. Most of the existing works consider the VM cost as fixed cost, but JOB-DCA policy derives it from per VM cost, memory cost and storage cost. The mechanism computes the VM and DT cost for all the data centers corresponding the user request. Finally, it provides the optimized data center name with index for allocation.

4 Simulation and Results

The proposed mechanism JOB-DCA is simulated using CloudAnalyst [23] an open-source toolkit which is integrated into CloudSim toolkit [23, 26]. This offers wide range of tools and methods for simulating the cloud and evaluating its service performance. Through service proximity-based routing, the broker allows to select the shortest path from the userbase (UB) to the DC by using the network latency. We consider ten different scenarios to evaluate the performance of the proposed mechanism. The performance is compared with the benchmark mechanism under closest DC policy and proposed allocation policy of the cloud using analyst simulator with six regions along with data centers from region code (0–5). During simulation, the

basic transmission characteristic is regular and necessary Internet bandwidths for different regions are considered as constant.

4.1 Performance

The work is reproduced with various situations to examine the performance efficiency of the proposed JOB-DCA policy. Every situation has a natural determination as for the quantity of UBs, various DCs and the quantity of districts took an interest. For this work, we assumed ten scenarios with client solicitations, DCs and DC regions. The detailed specification of UB is specified for every user in every hour ranging as follows: Request number 60–80, request size(1–500 KB), start peak hour 3 in GMT, end peak hour 9 in GMT and average peak users for userbase are 1000 throughout the experiment. Therefore, we organize the DCs with requisite determination as 204,800 MB main memory, 100,000,000 MB storage, 1,000,000 KB BW available and 4 numbers of processors. User gathering factor determines what number of users is treated as a group during reproduction. In this mechanism, we consider the current VM load adjusting plan like round-robin, FCFS and throttled.

The performance of JOB-DCA is evaluated compared with the benchmark allocation policies: closest data center policy (CDCP) and optimize response time (ORT) of CloudAnalyst [23, 26]. We compare the VM cost, data transfer (DT) cost and the total cost for different scenarios. Table 2 shows the comparison of different costs with respect to the benchmark mechanisms. In some scenarios, the proposed JOB-DCA minimizes the cost parameter as compared to CDCP and rest is minimized as compared to ORT benchmark mechanism. However, the overall cost is minimized as compared with the benchmark mechanisms. Among all the scenarios, the minimum VM cost as 69.76, DT cost 8.32 and total cost as 80.56 in dollar (\$) are noted which are quite less than the compared benchmark mechanisms.

4.2 Results and Discussion

In this section, we address the simulation results on VM cost, data transfer cost and total cost. Figure 1 presents the comparison of VM cost of proposed JOB-DCA mechanism with CDCP and ORT. In scenarios 5 and 6, the VM cost is minimized more as compared to other scenarios, but in all the scenarios the VM cost is minimized as compared to the benchmark mechanism CDCP and ORT. The computation comparison of DT cost is shown in Fig. 2. The DT cost is evaluated using Eq. (3) in the proposed JOB-DCA mechanism. As the data centers are wide spread over various regions, the data transfer cost changes from situation to situation. By and large, the data transfer cost is related to Internet boundaries like network, bandwidth and network delay. It is the covered-up or delicate expense for VM. We consider the bandwidth of different data centers that are provided in the cloud analyst. Under the

Table 2 Evaluation of VM cost, data transfer cost and total cost in \$

Scenario number	Policy specification	Cost		
		VM cost	DT cost	Total cost
Scenario 1	Closest DC policy	95.83	9.27	103.23
	Optimize response time	98.32	9.17	105.24
	JOB-DCA policy	94.78	8.42	102.11
Scenario 2	Closest DC policy	117.64	53.63	170.42
	Optimize response time	120.01	44.60	166.29
	JOB-DCA policy	116.76	42.83	162.33
Scenario 3	Closest DC policy	70.23	11.34	81.12
	Optimize response time	72.00	11.72	82.94
	JOB-DCA policy	69.76	10.84	80.56
Scenario 4	Closest DC policy	95.32	8.81	103.63
	Optimize response time	95.63	8.67	104.24
	JOB-DCA policy	94.71	8.32	102.78
Scenario 5	Closest DC policy	95.42	19.13	114.32
	Optimize response time	96.21	19.68	115.17
	JOB-DCA policy	91.33	19.02	112.64
Scenario 6	Closest DC policy	142.11	79.25	221.35
	Optimize response time	144.00	79.75	223.25
	JOB-DCA policy	141.34	78.65	220.19
Scenario 7	Closest DC policy	136.23	92.83	226.36
	Optimize response time	137.84	93.79	227.44
	JOB-DCA policy	135.92	92.87	225.78
Scenario 8	Closest DC policy	122.32	92.62	212.41
	Optimize response time	124.46	88.47	207.33
	JOB-DCA policy	121.87	87.25	205.87
Scenario 9	Closest DC policy	92.54	63.29	154.54
	Optimize response time	91.87	62.61	153.27
	JOB-DCA policy	91.43	62.33	152.87
Scenario 10	Closest DC policy	121.37	61.63	181.93
	Optimize response time	120.88	61.14	180.64
	JOB-DCA policy	120.24	60.35	179.23

standard bandwidth, we compare the result with the CDCP and ORT mechanisms. It is observed that the proposed JOB-DCA mechanism minimizes the DT cost as compared to the existing CDCP and ORT of cloud analyst. Note that the formulation function for DT cost reduces it and provides better data center allocation. In comparison with CDCP, the DT cost is quite impressive in scenarios 2 and 8. Overall, the DT cost is reduced in all ten scenarios as compared to CDCP and ORT mechanisms.

Fig. 1 Computation of VM cost in different scenarios

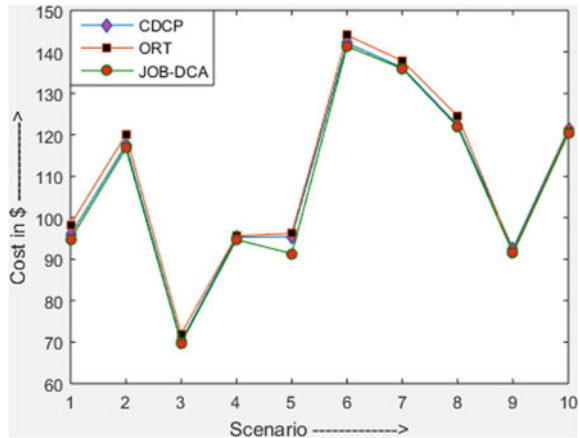
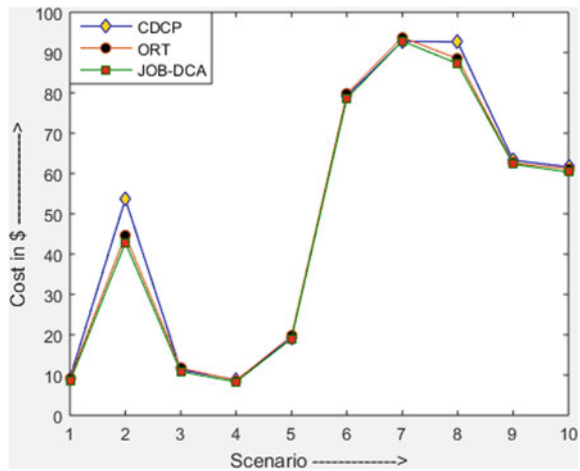


Fig. 2 Computation of DT cost in different scenarios



The total cost is calculated using Eq. (1) given as total cost of VM = VM cost + data transfer cost. The total cost is the actual cost that service providers charge for the VMs. Figure 3 depicts the evaluation of total cost in different scenarios with existing mechanisms CDCP and ORT of the cloud analyst. In most of the scenarios, the overall total cost is minimized with 3.2% as compared to other benchmark policies. As the work aims to optimize the cost, we compute the VM cost and DT cost and finally the total cost. We notice that in all the scenarios the computed costs are reduced. The comparison of total cost is presented with bar graph to differentiate in Fig. 4. In some scenarios, it is better than the CDCP and ORT. However, the total cost is reduced as compared to the benchmark mechanism. The experiments are done under different scenarios in cloud computing with respect to the user request and data center configurations. The QoS parameter states the availability of service with minimum

Fig. 3 Computation of total cost for different scenarios

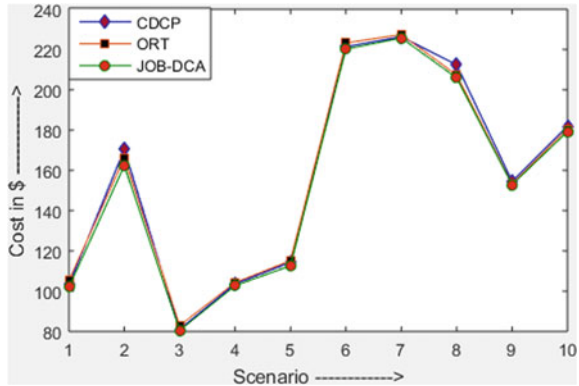
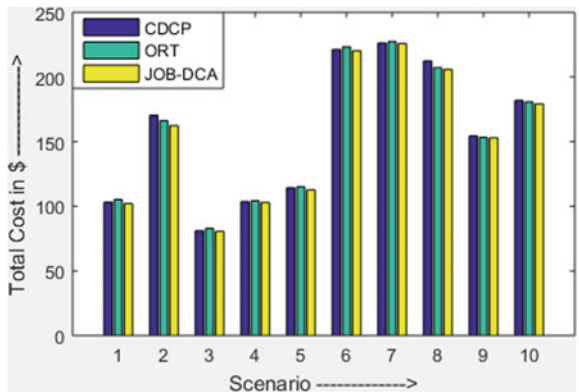


Fig. 4 Comparison of total cost



cost. It is essential to reduce the cost associated with computing in cloud computing. If the cost will be reduced, then the access usability will be increased, so the demand of resource also increased along with the revenue[27, 28]. So, the proposed JOB-DCA reduces the cost up to 3.2% less with respect to considered benchmark policies and could provide better data center allocation policy to manage the cloud resources.

5 Conclusion and Future Scope

Numerous mechanisms, algorithms and models for VM allocation are proposed by various analysts and researchers. It is found very fascinating and testing to propose a system to put VMs in an alternate data center in various regions for client applications, where the VM cost ought to be limited alongside least VM cost, DT cost and total cost. In this work, we propose JOB-DCA system to decrease distinctive cloud-related cost. The work is simulated over an open-source tool, *i.e.*, CloudAnalyst specified with the standard configuration. We examine with ten unique situations to watch the

exhibitions of the proposed JOB-DCA system. Finally, we performed comparison of the performance to other benchmark allocation mechanisms CDCP and ORT of cloud analyst with respect to cost. JOB-DCA implements new Jaya optimization technique to reduce the cost. The results obtained after simulation are found to be better as compared to the existing mechanisms. As the main aim of this work was to implement and reduce the cost using Jaya optimization technique, we have not considered other parameters such as response time and processing time. Moreover, a comparison with other optimization techniques can be taken as one of the future directions of this work.

References

1. Baker, T., Aldawsari, B., Asim, M., Tawfik, H., Maamar, Z., Buyya, R.: Cloud-SEnergy: a bin-packing based multi-cloud service broker for energy efficient composition and execution of data-intensive applications. *Sustain. Comput. Informatics Syst.* (2018)
2. Vakiliinia, S., Ali, M.M., Qiu, D.: Modeling of the resource allocation in cloud computing centers. *Comput. Netw.* **91**, 453–470 (2015)
3. Buyya, R., Garg, S.K., Calheiros, R.N.: LA-oriented resource provisioning for cloud computing: challenges, architecture, and solutions. In: *International Conference on Cloud and Service Computing*, pp. 1–10 (2011)
4. Nayak, S.C., Parida, S., Tripathy, C., Pati, B., Panigrahi, C.R.: Multicriteria decision making techniques for avoiding similar task scheduling conflict in cloud computing. *Int. J. Commun. Syst.* 1–31 (2019)
5. Parida, S., Nayak, S.C.: Truthful resource allocation detection mechanism for cloud computing. In: Indu, N. (ed.) *Third International Symposium on Women in Computing and Informatics (WCI'15)*, pp. 487–491. ACM (2015).
6. Nayak, S.C., Parida, S., Tripathy, C.: Modeling of task scheduling algorithm using petri-net in cloud computing. In: Saeed, K., Chaki, N., Pati, B., Bakshi, S., Mohapatra, D. (eds.) *Program Advance Computer Intelligence Engineering Advance Intelligence System Computer*, vol. 563, pp. 633–643. Springer, Singapore (2018)
7. Anastasi, G.F., Carlini, E., Coppola, M., Dazzi, P.: QoS-aware genetic cloud brokering. *Futur. Gener. Comput. Syst.* **75**, 1–13 (2017)
8. Do, C.T., Tran, N.H., Huh, E.N., Hong, C.S., Niyato, D., Han, Z.: Dynamics of service selection and provider pricing game in heterogeneous cloud market. *J. Netw. Comput. Appl.* **69**, 152–165 (2015)
9. Zhang, N., Yang, X., Zhang, M., Sun, Y., Long, K.: A genetic algorithm-based task scheduling for cloud resource crowd-funding model. *Int. J. Commun. Syst.* **31**(1), 1–10 (2018)
10. Manasrah, A.M., Gupta, A.B.B.: An optimized service broker routing policy based on differential evolution algorithm in fog/cloud environment. *Cluster Comput.* (2017)
11. Venkata Rao, R.: Jaya: a simple and new optimization algorithm for solving constrained and unconstrained optimization problems. *Int. J. Ind. Eng. Comput.* **7**(1), 19–34 (2016)
12. Sudhakar, T., Inbarani, H.H.: Intelligent path selection in wireless networks using jaya optimization. *Procedia Comput. Sci.* **133**, 976–983 (2018)
13. Pandey, H.M.: Jaya a novel optimization algorithm: what, how and why? In: *Proceedings of 2016 6th International Conference—Cloud System Big Data engineering confluence*, pp. 728–730 (2016)
14. Mohanty, S., Patra, P.K., Ray, M., Mohapatra, S.: An approach for load balancing in cloud computing using JAYA algorithm. *Int. J. Inf. Technol. Web Eng.* **14**(1), 27–41 (2019)
15. Reddy, M.A., Ravindranath, K.: Virtual Machine Placement Using JAYA Optimization Algorithm. *Appl. Artif. Intell.* **34**(1), 31–46 (2020)

16. Gupta, S., Agarwal, I., Singh, R.S.: Workflow scheduling using Jaya algorithm in cloud. *Concurr. Comput.* **31**(17), 1–13 (2019)
17. Sudha, I., Nedunchelian, r.: A secure data protection technique for healthcare data in the cloud using homomorphic encryption and Jaya-Whale optimization algorithm. *Int. J. Model. Simulation, Sci. Comput.* **10**(6), 1–22 (2019)
18. Jumani, T.A., et al.: Jaya optimization algorithm for transient response and stability enhancement of a fractional-order PID based automatic voltage regulator system. *Alexandria Eng. J.* (2020)
19. Warid, W., Hizam, H., Mariun, N., Abdul-Wahab, N.I.: Optimal power flow using the Jaya algorithm. *Energies* **9**(9) (2016)
20. Salih, A.H.A., Ali, A.H., Hashim, N.Y.: Jaya: an evolutionary optimization technique for obtaining the optimal dthr value of evolving clustering method (ECM). *Int. J. Eng. Res. Technol.* **11**(12), 1901–1912 (2018)
21. Mishra, S., Ray, P.K.: Power quality improvement using photovoltaic fed DSTATCOM based on JAYA optimization. *IEEE Trans. Sustain. Energy* **7**(4), 1672–1680 (2016)
22. Pradhan, C., Bhende, C.N.: Online load frequency control in wind integrated power systems using modified Jaya optimization. *Eng. Appl. Artif. Intell.* **77**, 212–228 (2019)
23. Wickremasinghe, B., Calheiros, R.N., Buyya, R.: CloudAnalyst: a cloudsim-based visual modeller for analysing cloud computing environments and applications. In: *Proceedings of International Conference on Advanced Information Networking and Applications*, pp. 446–452. AINA (2010)
24. Manasrah, A.M., Smadi, T., Almomani, A.: A Variable Service Broker Routing Policy for data center selection in cloud analyst. *J. King Saud Univ. Comput. Inf. Sci.* **29**(3), 365–377 (2017)
25. Huang, J., Kauffman, R.J., Ma, D.: Pricing strategy for cloud computing: a damaged services perspective. *Decis. Support Syst.* **78**, 80–92 (2015)
26. Magalhães, D., Calheiros, R.N., Buyya, R., Gomes, D.G.: Workload modeling for resource usage analysis and simulation in cloud computing. **47**, 69–81 (2015)
27. Parida, S., Pati, B.: A cost efficient service broker policy for data center allocation in IaaS cloud model. *Wirel. Pers. Commun.* 0123456789 (2020)
28. Nayak, S.C., Parida, S., Tripathy, C., Kumar, P.: An enhanced deadline constraint based task scheduling mechanism for cloud environment. *J. King Saud Univ. Comput. Inf. Sci.* (2018)
29. Panigrahi, C. R., Sarkar, J.L., Pati, B.: Transmission in mobile cloudlet systems with intermittent connectivity in emergency areas. *Digit. Commun. Netw.* **4**(1), 69–75 (2018)
30. Panigrahi, C.R., Sarkar, J.L., Tiwary, M., Pati, B., Mohapatra, P.: DATALET: an approach to manage big volume of data in cyber foraged environment. *J. Paral. Distrib. Comput.* **131**, 14–28 (2019)

Sustainable Computing and Engineering

Teaching and Learning Concepts of Audio Modulation Using Tangible User Interfaces



Ah-Kwet Rémi Wong Suk Hee, Hadija Ramadhani Halfani, and Girish Bekaroo

Abstract Digital technologies form an integral part of sound production. To meet industry requirements, audio processing and modulation are taught in computer science-related courses within modules such as multimedia, data communications and science, technology, engineering, and mathematics (STEM). However, due to the complexity of the subject, teaching and learning audio modulation within traditional settings involves challenges, thereby adversely impacting student engagement. The use of tangible user interfaces (TUI) can potentially improve student experience during teaching and learning, although this type of user interface has not been much explored in the area of sound processing and modulation. As such, this paper investigates the application of a novel TUI-based system to assist in teaching and learning concepts of audio modulation. Using an TUI evaluation framework, the proposed solution was evaluated in order to assess five key constructs, notably learnability, interaction, tangibility, enjoyment, and intention for future use. During the evaluation process, 29 students practically utilized the solution and provided feedback on the constructs assessed. Results revealed an inclination towards agreement for the different constructs investigated, although some limitations were identified. Based on these limitations, recommendations are provided towards improving design of such systems.

Keywords Audio modulation · Tangible user interfaces · Music production · Teaching and learning · Tangible interaction

1 Introduction

Music production is a field which is closely linked with computer science since digital technologies have been widely used by music composers to process their audio recordings. In order to address the needs of the industry, relevant concepts are integrated with modules including multimedia engineering, data communications as well

A.-K. R. Wong Suk Hee (✉) · H. R. Halfani · G. Bekaroo
School of Science and Technology, Middlesex University Mauritius, Flic en Flac, Mauritius
e-mail: remwon@hotmail.com

as science, technology, engineering, and mathematics (STEM). As such, students are taught a variety of concepts ranging from the mathematics of sound to manipulating and modifying associated properties [1]. Learning audio synthesizing can be challenging due to the mathematical concepts involved [2] and the use of traditional techniques to teach such concepts can result in lack of engagement and motivation from students during the learning process.

A potential solution to address such learning issues is through the use of tangible user interface (TUI)-based systems [3]. Tangible user interface (TUI) extends capabilities of graphical user interfaces (GUI) to enable a user to interact with a system by using physical objects instead of the conventional mouse and keyboard. This technology has often been applied in educational setups to facilitate student learning experience, while involving physical objects which stimulate the human brain whereby visually helping students to better understand concepts and captivate attention [4]. Even though this technology can potentially help to simplify concepts of audio modulation during teaching and learning, limited work has been done in this area. This paper complements literature by investigating the application of a TUI-based system to assist in teaching and learning concepts of audio modulation and proposes recommendations towards improving design of such systems.

This paper is organized as follows: In the next section, related works in relation to the application of TUI for teaching sound processing are reviewed. Then, in order to achieve the purpose of this paper, the design and implementation of a TUI-based system is presented. In the fourth and fifth sections, evaluation of the prototype is provided where the methodology used, data collection procedures and results are discussed. Recommendations on how to improve the design of such systems are provided in the sixth section, before concluding the work in the final section.

2 Related Works

Due to the complexity of teaching and learning audio modulation and synthesis concepts through traditional methods and using GUI, a few related studies have attempted to use TUI. A previous study proposed Audio d-touch, a TUI-based system meant for time-based musical activities including sequencing, drum editing, and collaborative composition [5]. Although the system revealed different benefits of using such TUI-based system for audio composition, its focus does not relate to teaching such aspects. Furthermore, only informal user evaluation was conducted to assess the system. Another study investigated the opportunities that TUI can potentially bring related to sound synthesis through reworking capabilities of the reactable, which is a TUI-based table-top musical instrument [6]. The usability tests conducted revealed insightful findings about how such systems could be improved. As such, related works do not focus on teaching and learning audio modulation through the application of TUIs, and thus, this study becomes relevant to undertake.

3 Prototype Design and Implementation

Due to the unavailability of existing TUI-based tools to teach audio modulation concepts, a prototype was designed and developed. The proposed prototype is described as follows:

3.1 Prototype Design

The purpose of the TUI-based tool is to assist in teaching and learning concepts of audio modulation to students at the undergraduate level of courses related to computer science. As key features of this tool, the user is welcomed with a splash screen within which concepts to be learnt can be chosen. For learning a concept, users need to progress through three stages within the integrated tool, notably an interactive tutorial, a practical exercise, and a quiz. Interactive tutorials enable concepts such as theory of sound to be taught to users using multimedia (particularly text, images, sound, and animations). Following the tutorials, practical exercises are presented to users for them to immediately apply concepts learnt. For example, after presenting the key concepts and characteristics of sound waves in the first topic, a TUI-based waveform generator is presented as the first exercise, as depicted in Fig. 1. Sound is a propagation of acoustic waves and in this exercise, students can manipulate sound waves in an interactive manner through tangible objects by modifying properties including amplitude, wavelength, and speed. Digital feedback is instantly provided as variables are manipulated by the end-user. Finally, in order to assess concepts learnt, a multiple choice-based quiz activity is present where random questions are

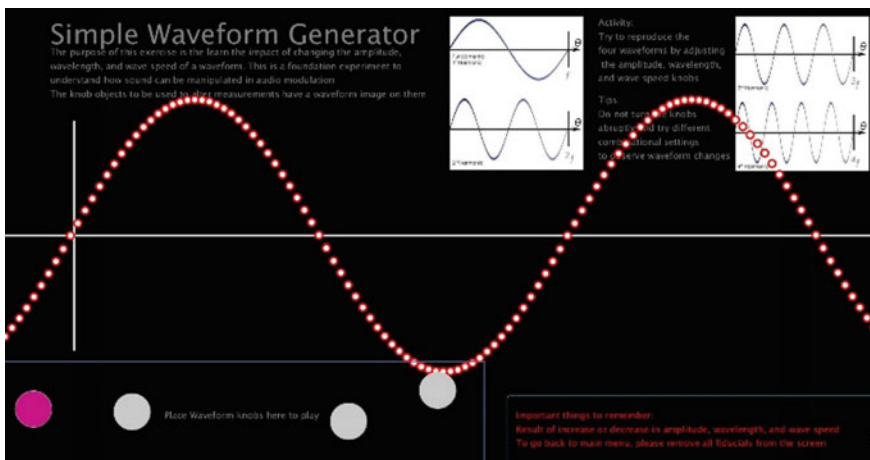


Fig. 1 Waveform generator exercise

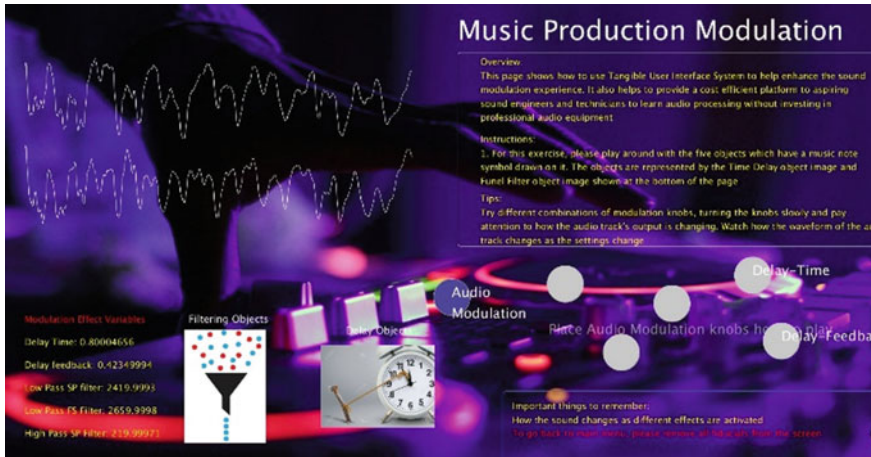


Fig. 2 Audio modulation activity

provided to the user. Self-assessment features are also available in the waveform generation exercise screen where the user is advised to try and replicate a given sample of output waveform based on four images shown onscreen.

Similar to the previous exercise, another practical activity within the integrated tool provides the opportunity to play around with a soundtrack using two important audio modulation techniques, namely delay and filters to present respective visualizations using graphs. In this exercise, the time-delay impacts the duration of the echo where feedback affects the frequency of delay repetitions. In addition, the filters consist of high pass filter to reduce high frequencies in the sound and low pass filters to reduce low frequencies. This activity is depicted in Fig. 2, where end-users can modulate and produce sound in an interactive manner.

3.2 Implementation

The proposed TUI-based tool was implemented using processing and reacTIVision. Processing is an integrated development environment that enables development of visual and electronic arts. On the other hand, reacTIVision enables rapid and robust tracking of physical objects onto which fiducial markers are stuck. In the implementation process, different libraries such as TUIO and Minim were utilized for interfacing with fiducials and for working with sound respectively.

In terms of the physical setup of the system, a table-top setup (as shown in Fig. 3) was adopted due to its popularity in TUI-based systems [7, 8], as well as its relevance for this study. This setup involves using a monitor on a table with a webcam mounted on a tripod, facing downwards to the screen in order to capture properties of fiducials attached to objects [8]. A light source is also utilized besides the screen in case

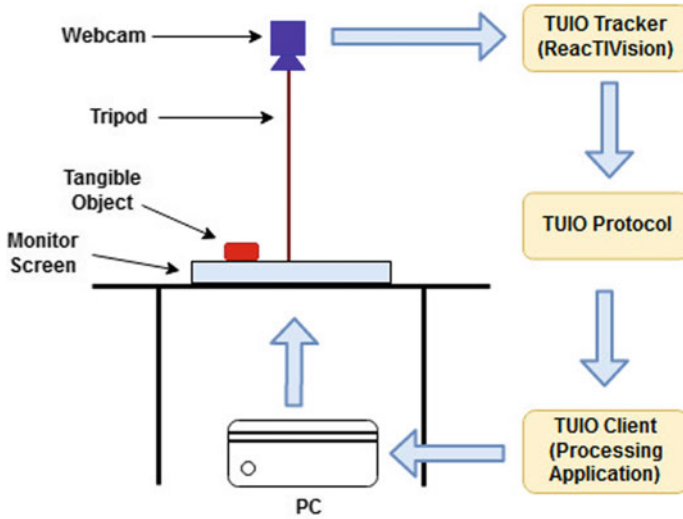


Fig. 3 Table-top setup

there is not much lighting in the room in order to ensure that the fiducials are easily detectable by the webcam.

3.3 Tangible Interaction

In order to interact with the TUI-based system, physical objects such as cubes were created, with fiducial markers attached to them. These objects are then detected by reactIVision, and using the TUIO library of processing, user actions on the objects are detected in order to trigger respective methods. Some of the objects utilized in the system is depicted in Fig. 4. For proper implementation of tangible interaction



Fig. 4 Objects used for interaction

within the TUI tool, the framework for TUI by Koleva et al. [9] was adopted. Three key types of interaction with the physical objects are possible within this study where the first one involves bringing an object from off screen to the interaction space. For example, this type of interaction is used in the system when users select specific objects to interact within the menu of the TUI tool or when responding to multiple choice questions within the quiz. Secondly, objects brought into the interaction space can be moved to different positions on screen, such as placing the timer object in the clock area as shown in Fig. 2. The third type of interaction involves rotating objects to change angle of fiducial markers, typically implemented to change levels of amplitude, wavelength, and wave speed within the interface as shown in Fig. 1.

4 Evaluation

For evaluating the TUI-based solution proposed in this study, two methods were utilized. Firstly, a TUI evaluation framework was adapted and applied due to its popularity in assessing similar solutions within previous studies [7, 8]. In addition, since similar physical setup and settings were involved, the need to assess these aspects were limited whereby focusing on evaluation of the TUI-based solution. In the adopted framework, different attributes are present to reveal insights on key aspects of such types of systems. These attributes are:

- *Learnability*

Learnability relates to the ease through which a system can be comprehended such that a user can perform fundamental tasks during first use itself [10]. Due to limited exposure with TUI-based systems for many users, evaluating the learnability of the proposed audio modulation tool becomes relevant in order to understand whether the proposed design is appropriate for first-time interaction with such systems.

- *Interaction*

Interaction relates to how a user communicates and controls a system. The introduction of an innovative mode of interaction can introduce different challenges pertaining to how the end-user communicates and controls the TUI-based system, and hence, this attribute becomes important to study.

- *Tangibility*

As defined in a previous work, tangibility refers to the ability for correctly identifying objects through the cognitive skills and sense of touch [8]. This aspect is important to study as well to understand how users are applying their cognitive skills and sense of touch when utilizing the proposed solution.

- *Enjoyment*

It was previously highlighted that enjoyment influences learning in a positive manner [8, 11]. As such, it is essential to consider this attribute as well in the evaluation process to assess how fun and engaging it is when using the proposed prototype.

- *Intention for future use*

This attribute relates to whether end-users of the system have established conscious intentions to use or not to use the system in their future practices [12]. This aspect is essential to study in order to comprehend whether users plan to make use of the solution in the future in their learning endeavours of related concepts.

Each of the above constructs can be assessed through a set of measured items, adapted from previous studies that applied similar framework [7, 8]. These measured items for each attribute are given in Table 1 as follows:

As second evaluation method, observation was used where end-users of the system were observed by the research team during use of the TUI-based system. Users were also asked to think-aloud when performing key tasks on the system in order to identify perceptions on the system, confusions, and issues faced during their activities [13]. Based on the planned evaluation methods, data collection instruments were then prepared. For heuristics evaluation from the framework described earlier, a questionnaire was prepared containing the constructs and measured items listed in

Table 1 TUI evaluation framework

Construct	Measured item
Learnability	L1-It was easy to learn about how to utilize the TUI-based system
	L2-The use of the TUI-based system made it easy to understand the basic concepts of audio modulation
	L3-Information provided within the system was clear and concise
Interaction	I1-Interacting with the system was easy
	I2-Enough information was provided to interact with the system
	I3-It was easy to interact with the system using objects
	I4-There were no signs of malfunction pertaining to interaction
Tangibility	T1-The real-life objects were proper representations of digital graphics
	T2-The selection of physical objects was instinctive
	T3-The physical objects were easily manipulated
	T4-The objects could be used to play the game naturally
Enjoyment	E1-Using the system was engaging
	E2-It was fun to use the system
Intention for future use	F1-I intend to use the system again in future
	F2-I recommend to the use of this system as a learning tool

Table 1. In the questionnaire, the measured items could be rated through a Likert-5 scale, ranging from strongly disagree (representing score of 1) to strongly agree (score of 5). Comment boxes were also present so that participants can provide positive or negative feedback related to each construct. For gathering observation data, relevant sheets were prepared containing information on participant number and observations for each feature of the system.

Once the data collection instruments were prepared, a pilot study was conducted involving three participants. This helped to finalize aspects such as the setup, questionnaire, software, and evaluation process. Then, participants were invited to participate in the study, and as target audience, first-year students of computer science-related courses were targeted to meet curriculum requirements of the proposed tool and students of these courses are taught similar concepts in their programmes. 29 students agreed to participate in the study and evaluation was conducted in the computer laboratories within Middlesex University Mauritius. As such, the minimum participant requirements of this study were met [7].

In terms of experimentation procedures, small groups of five participants were gathered to participate in the study after their classes. The session started by explaining to the participants the rationale behind the study, data-related aspects, in addition to ethical issues, among other aspects. Following the briefing, informed consent was sought from each participant and any clarifications needed were addressed prior to the experimentation. The research team also ensured that the TUI-based system was restarted and that objects involved were properly arranged next to the system. Each participant then had to utilize the system to explore the features available and learn audio modulation at the same time. Each participant had to proceed through each feature (as described in the previous section), and during this process, he/she had to think-aloud. Key observations and notes were recorded by the research team for later analysis. In addition, details on any key challenges faced were noted. After using all the features of the system, any questions pertaining to the topic or use of the system were addressed. Then, participants had to individually fill in the questionnaire, and following completion of same, the research team ensured that the submitted document was valid. The same process was repeated with all participants who agreed to be part of the study. Following the data collection process, statistical analysis was performed.

5 Results and Discussions

Statistical analysis revealed insightful information related to the application of TUI for teaching concepts of audio modulation. Results obtained are given in Table 2.

The findings for each construct are discussed as follows:

Table 2 Results

Construct	Measured item	Mean score
Learnability	L1-It was easy to learn about how to utilize the TUI-based system	3.92
	L2-The use of the TUI-based system made it easy to understand the basic concepts of audio modulation	3.94
	L3-Information provided within the system was clear and concise	4.23
	Mean construct score:	4.03
Interaction	I1-Interacting with the system was easy	3.62
	I2-Enough information was provided to interact with the system	3.67
	I3-It was easy to interact with the system using objects	3.44
	I4-There were no signs of malfunction pertaining to interaction	3.36
	Mean construct score:	3.52
Tangibility	T1-The real-life objects were proper representations of digital graphics	3.77
	T2-The selection of physical objects was instinctive	3.31
	T3-The physical objects were easily manipulated	3.82
	T4-The objects could be used to interact with the system naturally	3.75
	Mean construct score:	3.67
Enjoyment	E1-Using the system was engaging	4.23
	E2-It was fun to use the system	4.42
	Mean construct score:	4.33
Intention for future use	F1-I intend to use the system again in future	3.74
	F2-I recommend to the use of this system as a learning tool	4.03
	Mean construct score:	3.89
Overall mean score:		3.89

5.1 Learnability

Results showed that the simplicity in the transition and interaction between the different screens and features of the learning tool enabled the user to grasp the concepts related to audio modulation and sound wave processing reflected in L1. Users perceived that the learning process was incremental, and that the complementary instructions provided made it easy to learn how to use the system. However, two users perceived otherwise and felt that instructions could be made more dynamic to better assist the user based on actions being performed on the system. For L2, even though a positive mean score of 3.94 was received and that participants perceived

that the TUI-based system helped to simplify concepts of audio modulation, six participants were neutral about this statement. The group perceived that only limited information could be provided by such systems as compared to online learning. Even though the most positive score was achieved for L3, three users found it challenging to read information provided and interact with the system at the same time, especially on screens where much information is available (e.g. in Fig. 2). Overall, a mean construct score of 4.03 was achieved for the learnability of the system, thus confirming that it was easy to learn about how to use the system.

5.2 Interaction

Results showed homogenous mean scores among the four measured items for this construct, with an overall mean score of 3.52. For I1, 12 users were neutral due to challenges involving selection and use of physical objects for interaction. As compared to traditional GUI interfaces, end-users had to select specific objects on different screens for interacting with the system and this involved further time and reference to instructions. Similarly, for I2 and I3, the mean scores ranged between 3 and 4, with users mostly neutral on interaction. The lowest score was obtained for I4 and there were various instances where users' hands were obstructing the webcam's area of vision, thus hampering interaction. Similar challenges have been revealed in previous studies involving the use of such table-top setups [7, 8].

5.3 Tangibility

As explained earlier, tangibility refers to the ability for correctly identifying objects through the cognitive skills and sense of touch. Four measured items were involved to assess this construct and an overall mean score of 3.67 was obtained. Among the constructs, although a score above 3 was obtained for T1, 37.9% of participants were neutral for this measured item. The participants perceived that some objects could have been better represented. This could be attributed to the fact that many concepts were abstract (e.g. frequency, amplitude, wavelength) and representing these into physical objects is a challenging process [14]. This factor also led to a relatively low score for T2, where eight participants disagreed that the selection of physical objects was instinctive. At some point, it could be seen that participants were hesitant on where to place the objects onscreen and some participants perceived that this could be improved by using coloured targets onscreen for each object. On the other hand, from T3 and T4, it could be deduced that the objects were relatively easy to manipulate due to their physical dimensions and weight. However, since the interaction technology used was new to the participants, most took time to get accustomed to the different objects used to explore the features of the system. However, once this was figured out, interaction with the system was natural (T4).

5.4 *Enjoyment*

During evaluation, enjoyment was also assessed to understand how fun and engaging it was to use the system. Results for this construct were positive where most participants agreed or strongly agreed about E1 and E2 and a mean score of 4.33 was recorded for enjoyment. As the participants were using the prototype, it was observed that users were eager to try the different features of the system and play around with them which explains an overall score of 4.23 for user engagement. The interaction technology and the use of various objects for interaction played a huge role in enhancing the user engagement as participants were exposed to such type of system for the first time. Two participants were however neutral about this statement and further analysis showed that the participants felt that more interactive features could be included including puzzles and integrated small games to further improve engagement during the learning process. Similarly, according to the participants, it was fun to learn audio modulation, and during evaluation, participants were amazed, and also, appreciation of the interaction between the tangible objects could be noted with the expression through words such as “wow”, “interesting” and “amazing”. Overall, from the score received, it could be deduced that TUI helps to improve fun and engagement during learning.

5.5 *Intention of Use*

For this construct, mixed perceptions were received for the measured items and an average score of 3.89 was obtained. For F1, only 15 participants agreed or strongly agreed that they have the intention to use the system again in the future. For this group of participants, the new interaction and fun learning is the key reason behind intention to re-use the system in the future. The remaining who were neutral or disagreed about this statement attributed this decision to the lack of dynamic features that run every time the system is used. In other words, the group felt that they have already used the system and will only get to experience the same thing when running the system again. However, participants were more positive on F2 where most participants recommended the use of this proposed system as a learning tool (F2). This also implies that audio and sound modulation teaching through TUIs is valuable for teaching students.

5.6 *General Discussions*

Results revealed that the computed scores varied across the constructs, although a mean score above 3 was achieved for all attributes investigated as depicted in Fig. 5. Amongst, enjoyment obtained the highest mean score where applicants found the

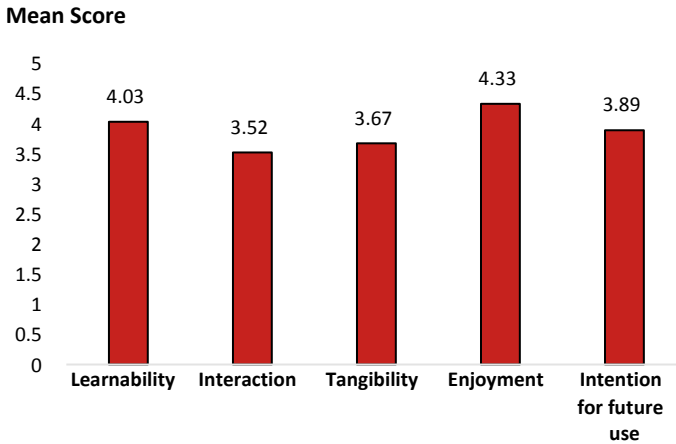


Fig. 5 Mean score distribution among attributes

application fun and engaging to use. This finding also aligns with a previous study [8] that applied similar evaluation framework. As discussed earlier, the key reason behind the relatively high score for this construct was that participants were exposed to this interaction technology for the first time and had a positive experience when using the prototype. On the other hand, interaction received the lowest score. As discussed earlier, users faced challenges to properly select and use physical objects to communicate with and control the system. Overall, a mean score of 3.89 was obtained for the different constructs, thus highlighting an inclination towards agreement for the different constructs investigated. This finding also highlights the prospects of integration of TUIs in labs for teaching and learning of concepts pertaining to audio modulation, although this will not be without challenges. Studying and evaluating these challenges were however beyond scope of this paper.

In addition to these challenges, different aspects could adversely influence findings revealed in this study. Firstly, other types of evaluation could be considered such as expert evaluation to gain insights from developers and researchers experienced in TUI systems. Furthermore, evaluation attempts could have been increased in order to get participants to use the system more than once, to collect data for repeated use. Such data would help to determine how scores vary after different attempts.

6 Recommendations

Following evaluation, a few design issues were noted that need improvements. The issues noted along with proposed recommendations are enumerated as follows. These proposed solutions could be considered by implementers of TUI-based educational systems as well as researchers for enhancing work in this area.

- Improved assistance on selection and use of objects for interaction

First-time users of the TUI-based system in this study had some issues and were hesitant at times in selecting and using physical objects for interaction. In order to address this issue, more dynamic assistance could be provided to the user based on actions being performed on the system. This could be done through the use of text or even voice instructions while highlighting the next action possible for the user.

- Extending information made available for learning

A small group perceived that limited information was provided for learning through the TUI system as compared to learning using online learning. In order to address this issue, such systems can be complemented with links to online resources related to the topic being taught.

- Solving issues with table-top setup

Within the study, the table-top setup was utilized, and users' hands were obstructing the webcam's area of vision at times, thus hampering interaction. A possible way to solve this problem would be to use alternative setups such as placing the webcam beneath an interactive transparent display to enhance the experience.

- Better representation of abstract concepts

According to a group of participants, abstract concepts could be better represented. In order to address this issue, complementary instructions could be provided in the system in order to simplify selection and use of such objects.

- Enhancing engagement through challenges

A few participants suggested that more interactive features could be included including to further improve engagement during learning and for this, challenges in the forms of puzzles and games could be integrated.

- Better planning of re-use

The current system did not cater for re-use and consequently, participants felt that when re-using only get to experience the same thing in the future. In order to address this issue, random and dynamic contents (information, quiz, puzzles) could be provided within the system.

7 Conclusions

This paper explored the application of a TUI-based system to assist in teaching and learning concepts of audio modulation. The system was implemented using processing and reactIVision using a table-top setup and aimed at teaching key

concepts including sound propagation and key audio modulation techniques. For evaluating the system, a TUI evaluation was applied to assess key constructs notably, learnability, interaction, tangibility, enjoyment, and intention for future use. 29 participants undertaking computer science-related courses were involved in evaluation to practically use the system and provide feedback on the constructs assessed. As results, an overall mean score of 3.89 was obtained for five constructs and this implies an inclination towards agreement constructs studied. Among the attributes, the highest mean score obtained was for enjoyment, whereas interaction received the lowest mean score. The overall positive results also denote the prospects of integration of TUIs in labs for teaching and learning of concepts of audio modulation. However, different limitations of the system were also identified during evaluation and recommendations have been made in this paper for developers and researchers of such types of system to consider when producing systems of this kind. As future works, the recommendations proposed in this paper could be taken on board to further enhance the prototype produced. In addition, further evaluation could be conducted to apply frameworks such as system usability scale and technology acceptance model, among others, to better study usability and acceptance of the system.

References

1. Hechmer, A., Tindale, A., Tzanetakis, G.: LogoRhythms: Introductory audio programming for computer musicians in a functional language paradigm. In: Proceedings of 36th Annual Conference on Frontiers in Education (2006)
2. Reid, G.: An introduction to Frequency Modulation,” Sound on Sound, 2020. [Online]. Available: <https://www.soundonsound.com/techniques/introduction-frequency-modulation>. Accessed 28 Mar 2020.
3. Nathoo, A., Bekaroo, G., Gangabissoon T., Santokhee, A.: Using tangible user interfaces for teaching concepts of internet of things. In: Interactive Technology and Smart Education (2020)
4. Marshall, P.: Do tangible interfaces enhance learning? In: Proceedings of the 1st International Conference on Tangible and Embedded Interaction (2007)
5. Costanza, E., Shelley, S., Robinson, J.: Introducing audio d-touch: A tangible user interface for music composition and performance. In: Proceedings of the 6th International Conference on Digital Audio Effects (DAFX-03). London (2003)
6. Potidis, S., Spyrou, T.: Spyactable: a tangible user interface modular synthesizer. In International Conference on Human-Computer Interaction. Cham (2014)
7. Waldner, M., Hauber, J., Zauner, J., Haller, M., Billinghurst, M.: Tangible tiles: design and evaluation of a tangible user interface in a collaborative tabletop setup. In: Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments (2006)
8. Gajadur, D., Bekaroo, G.: TangiNet: a tangible user interface system for teaching the properties of network cables. In: 2019 Conference on Next Generation Computing Applications (NextComp) (2019)
9. Koleva, B., Benford, S., Ng, K., Rodden, T.: A framework for tangible user interfaces. In: Physical Interaction (PI03) Workshop on Real World User Interfaces (2003)
10. Fjeld, M., Fredriksson, J., Ejdestig, M., Duca, F., Bötschi, K., Voegtli, B., Juchli, P.: Tangible user interface for chemistry education: comparative evaluation and re-design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2007)

11. Lucardie, D.: The impact of fun and enjoyment on adult's learning. In: *Procedia-Social and Behavioral Sciences* (2014)
12. Davis, F.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 319–340 (1989)
13. Couture, N., Rivière G., Reuter, P.: GeoTUI: a tangible user interface for geoscience. In: *Proceedings of the 2nd International Conference on Tangible and Embedded Interaction* (2008)
14. Skulmowski, A., Pradel, S., Kühnert, T., Brunnett, G., Rey, G.: Embodied learning using a tangible user interface: the effects of haptic perception and selective pointing on a spatial learning task. *Comput. Educ.* **92**, 64–75 (2016)

Performance Investigation of Dipole and Moxon Antennae for VHF Communication



Akshay Jain, Pranay Chavan, Pratik Maradiya, and Kiran Rathod

Abstract VHF communication links are mostly used by defense personnel and amateur radio operator or commonly known as HAM radio operators. The HAM radio operators communicate in the frequency band ranging from 144 to 146 MHz. Hence, in this paper, the dipole and Moxon antennae are designed for 145 MHz and are resonating at 144 MHz and 144.44 MHz, respectively. Both the dipole and Moxon antennae are simulated on HFSS using copper material and an element radius of 1.5 mm. The antennae are majorly investigated based on gain, directivity and return loss, and other parameters. The observed gain for dipole antenna and Moxon antenna is 2.4 dB and 6.2 dB, respectively, whereas the return loss of dipole antenna is -16.78 dB, and for Moxon antenna, the return loss is -30.26 dB. The Moxon antenna is highly directive as compared to the dipole antenna. As a result, Moxon antenna provides a reliable solution for VHF communication.

Keywords VHF communication · Dipole antenna · Moxon antenna · Performance investigation · HFSS · HAM radio operator

1 Introduction

Radio communication is one of the oldest methods for communication. Antenna plays a major role in any communication system, and the characteristic of an antenna affects the effectiveness of the system. Many including various emergency response teams,

A. Jain (✉) · P. Chavan · P. Maradiya · K. Rathod
K J Somaiya Institute of Engineering & Information Technology, University of Mumbai, Mumbai, Maharashtra, India
e-mail: Akshay.JainKJSIEIT@Somaiya.edu; akshay.hj@somaiya.edu

P. Chavan
e-mail: Pranay.ChavanKJSIEIT@Somaiya.edu

P. Maradiya
e-mail: Pratik.MaradiyaKJSIEIT@Somaiya.edu

K. Rathod
e-mail: Kiran.RathodKJSIEIT@Somaiya.edu

policemen, air traffic control rooms, navies, air forces, amateur radio operators, and even VVIPs use such antenna-based systems to communicate with each other till now. Since the VHF communication links are used majorly by defense personnel and amateur radio operator (HAM radio communication) at times of emergencies, they require such system which is not only efficient in terms of transmitting and receiving signals but also in terms of portability. This paper suggests a way to improvise the efficiency of the communication system by computing several parameters of the two antennae, namely dipole and Moxon antennae.

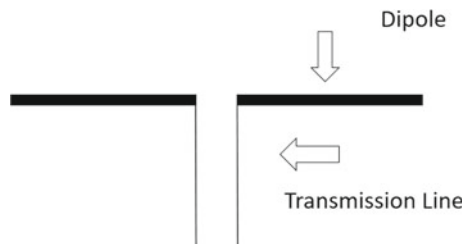
1.1 Dipole Antenna

Dipole antenna was first demonstrated by German physicist Heinrich Hertz. It is the most simple and commonly used antenna and is said to have omnidirectional radiation pattern, thus making it weakly directive antenna [1]. In the dipole antenna, two metal rods are placed parallel to each other and separated by small distance called as feed gap [2]. As in [2, 4, 5], the gain of the antenna is around 2 dB and narrow bandwidth (Table 1; Fig. 1).

Table 1 Comparison of existing work of dipole antenna with current work

Ref. No.	Center frequency (MHz)	RL (dB)	Gain (dB)	Bandwidth (MHz)	VSWR	Radius (mm)
[2]	1995	-36.10	-	290	-	0.58
[4]	689.66	-26.74	2.06	84.94	1.09	2.5
[5]	2600	-17	2.43	-	1.3	0.115
[6]	1000	-24.54	1.81	-	-	0.0025
This work	144	-16.78	2.4	10.58	-	1.5

Fig. 1 Structure of dipole antenna



1.2 Moxon Antenna

Moxon antenna was originally designed by Les Moxon, inspired by the design of Caton square antenna by Fred Caton [11]. The antenna has a couple of bent radiators, separated by a small distance called as feed gap and a bent reflector as shown in Fig. 2. The parasitic reflector beam is electromagnetically coupled to the driven element via a dielectric material (usually air). This arrangement makes the antenna capacitive in nature. Its capacitance can be calculated using the expression given below.

$$C = \epsilon_0 A/d. \tag{1}$$

where

C = Capacitance.

ϵ_0 = Dielectric constant.

A = Cross section area.

d = Distance between the conductor.

This capacitive arrangement of the antenna (load) makes it much better than inductive loading with coils, resulting in less losses [8]. The antenna has directional radiation pattern with the major lobe having an aperture of up to 100° and suppressed minor lobes at the back and at the sides, thus resulting in high front to back ratio [7]. Due to this, the antenna makes a great choice to track low earth orbiting satellites as reported in [9] (Table 2).

Hence, the above parameters prove that the antenna in this work is desirable as it exhibits more gain and has minimum return loss.

Further, the paper is formulated as follows: Sect. 2 justifies the approach for comparative analysis between antennae. Section 3 compares both antennae based on simulated results. Lastly, Sect. 4 concludes the investigation.

Fig. 2 Structure of Moxon antenna

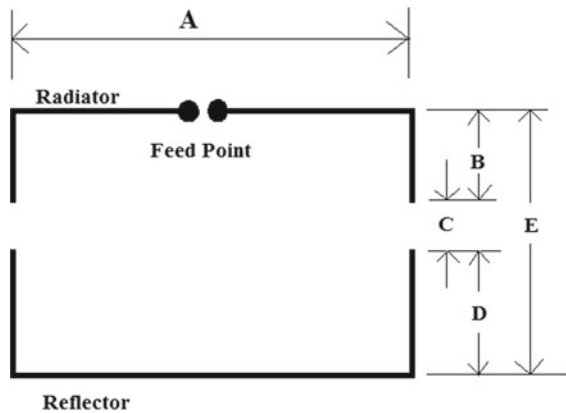


Table 2 Comparison of existing work of Moxon antenna with current work

Ref. No	Center frequency (MHz)	RL (dB)	Gain (dB)	Perimeter of antenna (m)	Antenna material	Radius (mm)
[9]	137.5	-28.48	6.09	2.148	Copper tube	2.5
This work	144.44	-30.26	6.2	1.95	Copper rods	1.5

2 Methodology

To investigate any antenna for the communication system, some sought after parameters are gain that should be preferably high to ensure stronger signals; bandwidth could be broader or narrower that is characterized based on the application; directivity which maximizes the potency of long-range communication [4] and return loss that interprets the performance of the antenna [3].

2.1 Dipole Antenna

The design methodology as in [2] reports that the length of dipole is not exactly equal to 0.5λ ; instead, it is 0.475λ due to the effect of fringing fields, whereas the radius of the element is 0.001λ . The other parameter such as feed gap is given by the length of the dipole arm divided by factor 200 (Table 3).

$$\lambda = c/f . \tag{2}$$

$$R = 0.001 \lambda . \tag{3}$$

$$L = 143 * 10^6 / f . \tag{4}$$

$$g = L/200 . \tag{5}$$

Table 3 Dimensions of dipole antenna at 145 MHz

Parameters	Calculated in (mm)	Optimized in (mm)
Wavelength	2068.9655	2068.9655
Radius	2.068	1.5
Length of dipole	986.206	980
Feed Gap	4.931	4.931

Table 4 Dimensions of Moxon antenna at 145 MHz

Parameters	Calculated n (mm)
A	703.9735
B	104.3430
C	19.9083
D	148.2820
E	272.5334

where

- λ = Wavelength,
- c = Speed of light,
- f = Design frequency,
- R = Radius of elements,
- L = Length of Dipole,
- g = Feed gap.

2.2 Moxon Antenna

The design methodology as in [7–9] suggests the use of online Moxon calculator software developed by an amateur radio operator [10]. The below shown table shows the dimensions obtained by the calculating software (Table 4).

3 Results

The antennae are simulated on HFSS and are designed for 145 MHz. The material assigned for both the antennae is copper, and the radius of elements for both antennae is 1.5 mm. The results are discussed below.

3.1 Return Loss Versus Frequency

The simulated results are carried over a frequency sweep of 100–200 MHz. Figure 3 shows the return loss vs. frequency graph for dipole antenna and Moxon antenna. The dipole and the Moxon antennae resonate on 144 and 144.44 MHz with a return loss of -16.78 dB and -30.26 dB, respectively, whereas the bandwidth of the dipole antenna is 10.58 MHz (138.38–148.96 MHz, 7.34%), and for Moxon antenna, it is 10.27 MHz (141.78–152.05 MHz, 7.11%).

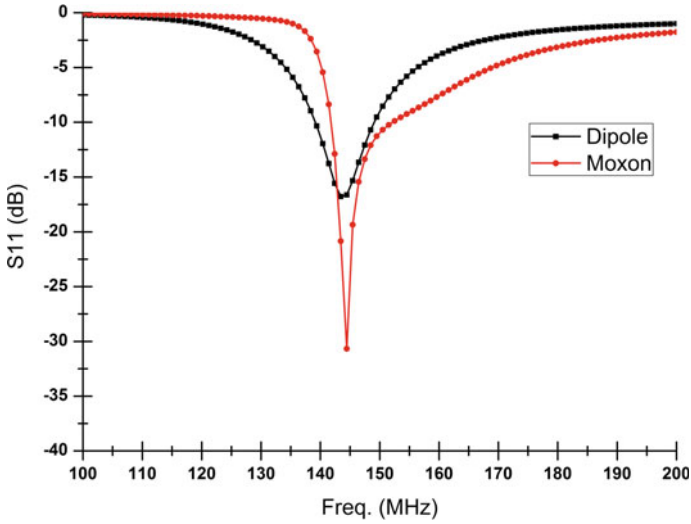


Fig. 3 Return loss versus frequency plot for dipole antenna and Moxon antenna

3.2 Radiation Pattern

The dipole antenna radiates in all directions perpendicular to the plane of elements, and the Moxon antenna has narrow beam width as compare to the beam width of the dipole antenna (Figs. 4 and 5).

Fig. 4 E-plane (Solid line) and H-plane (Dotted line) of dipole antenna

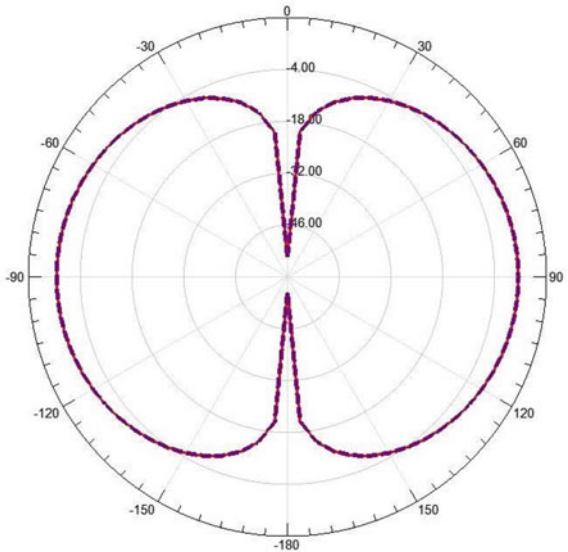
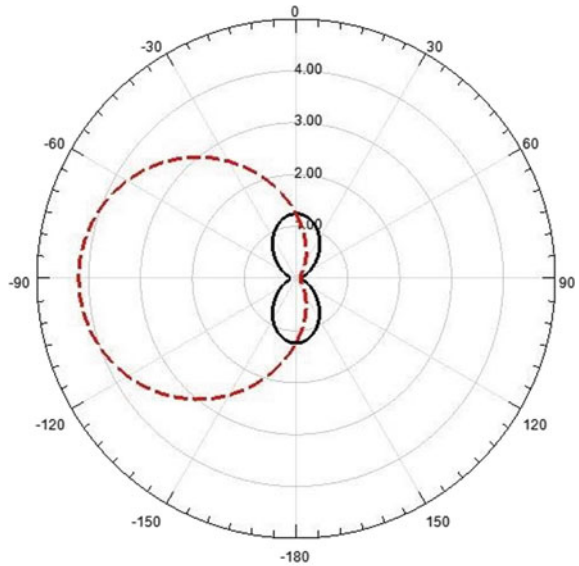


Fig. 5 E-plane (Solid line) and H-plane (Dotted line) of Moxon antenna



3.3 Gain and Directivity

The gain of the dipole antenna is 2.4 dB, and that of the Moxon antenna is 6.2 dB. The radiation pattern well clarifies Moxon as more directive.

4 Conclusion

The dipole and the Moxon antenna are designed and simulated on HFSS using copper rods as the element material and keeping the radius of elements 1.5 mm for both antennae. Since the amateur radio communication frequency ranges from 144 to 146 MHz, the minimum required bandwidth is 2–3 MHz, but the proposed antennae in this work, i.e., the dipole and the Moxon antennae exhibit bandwidth of 10.58 MHz and 10.27 MHz with percentage bandwidth of 7.34% and 7.11%, respectively, as shown in Table 5 Sect. 3. As discussed in Table 5, the return loss is -16.78 dB and -30.26 dB for the dipole and the Moxon antenna, respectively, which in ideal communication system should be as minimum as possible. The gain of the dipole is 2.4 dB, whereas for the Moxon antenna is 6.2 dB. The radiation pattern of dipole antenna is omnidirectional in nature, i.e., it will receives signals from all directions due to which two signals may interfere with each other and hence results in decreased SNR, whereas the Moxon antenna has directional radiation pattern with increased SNR. Hence, Moxon antenna provides a reliable solution for communication over VHF.

Table 5 Comparisons between dipole and Moxon antennae

Parameters	Dipole antenna	Moxon antenna
Center frequency	144 MHz	144.44 MHz
Return loss	−16.78 dB	−30.26 dB
Bandwidth	10.58 MHz,	10.27 MHz
Percentage bandwidth	7.34%	7.11%
Radiation pattern	Omnidirectional	Directional
Gain	2.4 dB	6.2 dB

Acknowledgements I would take this opportunity to thank all those people who helped me in this project work, and without them, this wouldn't have been possible. I would like to acknowledge my deepest gratitude toward Dr. Jayashree Khanapuri (HOD of Electronics and Telecommunication Engineering Department) whose invaluable guidance supported me in my project work. Also, I would like to thank Dr. Suresh Ukarande, the Principal of our college for giving me such an opportunity. I am also immensely grateful to Mr. Pankaj Jani (HAM radio operator) for their comments on technical aspects of an antenna. I would like also to thank all the staff members of Microwave engineering laboratory of our college and my friends who help me directly or indirectly for the successful completion of this project work.

References

1. Stutzman, W., Thiele, Gary.: *Antenna Theory and Design*, 3rd edn., 74–74. Wiley (2012)
2. Singh, P., Sharma, A., Uniyal, N., Kala, R.: Half wave dipole antenna for GSM applications. *Int. J. Adv. Comput. Res.* **2**(6), 354–357 (2012)
3. Return Loss and VSWR, <https://bit.ly/3734f6b>
4. Enggar Fransiska, D.W., Pratama, N.I.H., Rahmatia, A., Wulandari, P.: Design and performance investigation of dipole antenna using aluminum and iron At 644 MHz –736 MHz. *Int. Conf. Eng. Technol. Appl. Sci.* 01–05 (2017)
5. Osman, A., Alaa A., Yassin, B., Ali, H., Ahmed, S.N.: Design and simulation of high performance half wave-dipole antenna for LTE applications. In: *International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering*, 472–474 (2015)
6. Revathy, E., Ananth Kumar, T., Rajesh, R.S.: Design of highly efficient dipole antenna using HFSS. *Asian J. Appl. Sci. Technol.* **3**(1), 01–09 (2019)
7. Chougale, N.A., Magdum, A.K., Patel, S.D., Dongale, T.D., Vasmbeakar, P.N.: Design and development of UHF Moxon antenna. *Natl. J. Sci. Inform. Special Issue*, 93–95 (2012)
8. The Moxon beam, <https://www.qsl.net/dk7zb/Moxon/Moxon.htm>
9. Valdés Abreu, J.C., Sánchez, Y.S., Cimino Quiñones, L.: Moxon antenna for land stations of satellites weather of polar orbit. *Rielac*, **36**(1), 79–94 (2015)
10. MoxGen open source software, <https://tippete.net/cgi-bin/moxgen.pl>
11. Moxon Antenna Project, <https://bit.ly/2x8w3bB>

Cloud Computing Load Balancing Using Amazon Web Service Technology



Nagarjuna Hota and Binod Kumar Pattanayak

Abstract Amazon.com is a high and most important trafficked web services around the globe. It is an infrastructure-based web service which provides a wide range of products and services via Internet. AWS is a kind of platform which allows development of different applications by providing solutions for messaging, data storage with high scalability. Though AWS provides a wide range of services, there is a chance that servers are getting heavily loaded due to increasing in numbers of task being requested for the execution by the end users. Therefore, load balancing is required to balance the loads on different servers located on different regions available around the globe.

Keywords Amazon web service · Cloud computing · Load balancing · ELB · Auto scaling

1 Introduction

1.1 Cloud Computing

The term cloud computing has a variety of definitions according to different authors. But from the layman point of view, “cloud computing is a web-based technology which provides different services like storing, fetching, and execution of different data items that may be that is a .txt, .jpg, .mp3, or any video file via Internet” [6, 7, 9]. According to NIST, cloud computing can be defined as “cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be quickly provisioned and released with least management effort or service provider interaction” [4]. Cloud computing

N. Hota (✉) · B. K. Pattanayak
Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India
e-mail: nagarjunahota831@gmail.com

B. K. Pattanayak
e-mail: binodpattanayak@soa.ac.in

provides different resources in the form of services to the end users on-demand basis. It also enables businesses and users to use applications without installing them on physical machines and allows access to required resources over the Internet [6, 9, 11]. It can also be called as Internet-based computing on cloud servers in which a large number of remotely accessed servers are allow for online accessing of computer services and resources, centralized data storage by networked together. Industries or organizations can share their computing and storage resources instead of operating, building, and improving their own inbuilt infrastructures [6, 11]. Cloud computing has the following characteristics features through which the users can user cloud services as a pay-as-you-go model such as on-demand-self-services, broad network access, rapid elasticity and scalability, multi-tenacity–resource pooling, and measured services [5, 7].

In Sect. 2, we briefly describe about the AWS and its wide range of services. Section 3 is about the load balancing and ELB, and in the Sect. 4, we mentioned our paper work, result, and analysis along with the different types load balancers provided by AWS.

1.2 *Types of Cloud*

There are basically four types of cloud which are also called as cloud computing deployment models such as [5–7, 11]:

Public cloud: This type of cloud provides free access to the end users for general purposes through third-party service which provides to own their infrastructure.

Private cloud: This type of cloud provides quite same service as like the public cloud, but in private cloud, the services are managed by the third-party vendor for any organization. Here, the security types of issue can be minimized due the control over infrastructure.

Hybrid cloud: This type of cloud can be formed by combining of both private and public cloud, where the simpler issues are handled by public cloud and critical issues by private cloud.

Community cloud: This type of cloud can be used by various organizations or one organization that can share all the cloud services and resources by securing their own data not being sharable.

1.3 *Cloud Computing Service Model*

There are basically three types of cloud computing service models [5–7, 11]:

SaaS: It stands for Software as a Service. This service provides by the third-party vendor to the end user by limiting some administrative level of accessing of resources. Basically, application software is provides to the users which can manage by themselves.

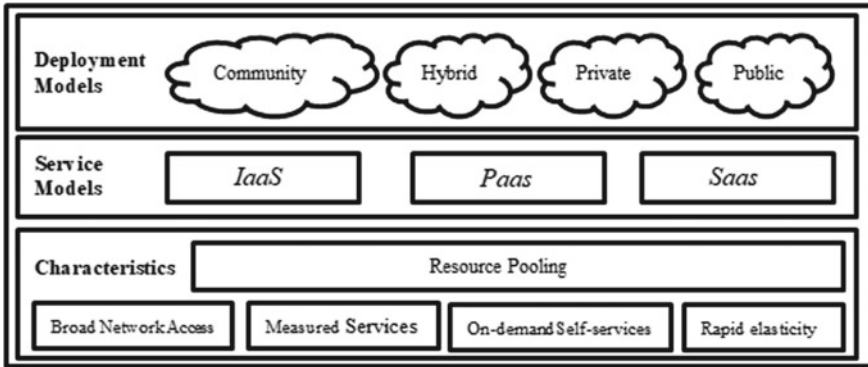


Fig. 1 Different models of cloud computing [5]

PaaS: It stands for Platform as a Service. Here, different services are provided to the clients by which they can build their own application, run, and manage instead of installing any resource on their own system. Here, the data backups are taken care by the service providing vendors. They focus on more functionality that should provide to the end user.

IaaS: It stands for Infrastructure as a Service. By this service model, the end users are capable for provisioning storage, processing, and network connectivity on their demand basis. By using IaaS, users can develop their own applications (Fig. 1).

1.4 Numerous Advantages of Cloud Computing

Some of the most important advantages offered by cloud computing are [6, 7, 10]:

Cost-efficient: It will time-consuming and cost-efficient when we are trying to develop our own application because we have to order, pay, install, and configure our required services with expensive hardware requirements also. But by using cloud service, we only have to pay for what we required instead of taking extra over head for configuration and installation. That’s why cloud computing is cost-effective.

Reliability: It is one of the most important advantages of cloud computing that it provides much more managed and reliable services to the IT infrastructure. It ensures 24*7 and 365/366 days of service providing.

Unlimited storage capacity: Cloud computing provides wide storage capacity for which we need not to be worried about running out of instances or increasing of current storage capacity. We can access more as our need and can decrease if we don’t need.

Backup and recovery: Cloud service providing vendors have enough technologies by which they can recover our data if there will any loss of data, and due to large storage capacity, user data can stored as backup for future use at any time.

Easy access to information: Once if we register and crate our own cloud, we can access all the services and resources of cloud from anywhere at any time around the globe having Internet connection.

2 What is AWS

AWS stands for Amazon Web Service which was developed to provide IT related services to various organizations in the year 2006 which is called as cloud computing now days [6, 11]. AWS is basically a platform, where the users can develop their own applications, infrastructures, or any others computations using different services of AWS as pay-as-you-go model. That means that the user needs not to give extra amount of money to the service provider instead only have to pay according to their uses [6, 7]. There is no requirement of third party and need not to go for any person to crate the cloud for our user. We just have to login into the home page of AWS through the url (<https://aws.amazon.com/>) which brings the home page directly to us [8]. First, we have to create our own instance; first, it will never take any charges from us but we have to enter our login details asking by the web page. Then, we have to enter our debit card or credit card details only for the confirmations, and after some days, it will refund to our account. After successful creation of the instance, the user is now eligible for use of all the services provided by Amazon [7]. Cloud computing has enormous impact on Amazon Web Service. The services of amazon.com represent world's largest infrastructure as a service in the today's marketplace. This one is also of the best examples of Service Oriented Architecture(SOA) [11]. It also includes HTTP, REST, and SOAP transfer protocols, the operating systems which are commercial and also open source, different application servers, and web browser-based access. Here, the virtual private servers can be prepared for virtual private clouds connected through virtual private networks for security and system administration [1, 11].

2.1 Basic Architecture of AWS

See Fig. 2.

2.2 Components of AWS

Amazon EC2: It stands for Amazon Elastic Compute cloud is a web-based interface of AWS that provides resizable compute capacity in the AWS cloud. It is particularly designed for the developers for complete control over web scaling and computes resources. The instances of EC2 can be automatically scale up or down as per the requirement. It is the central application of AWS, and we can

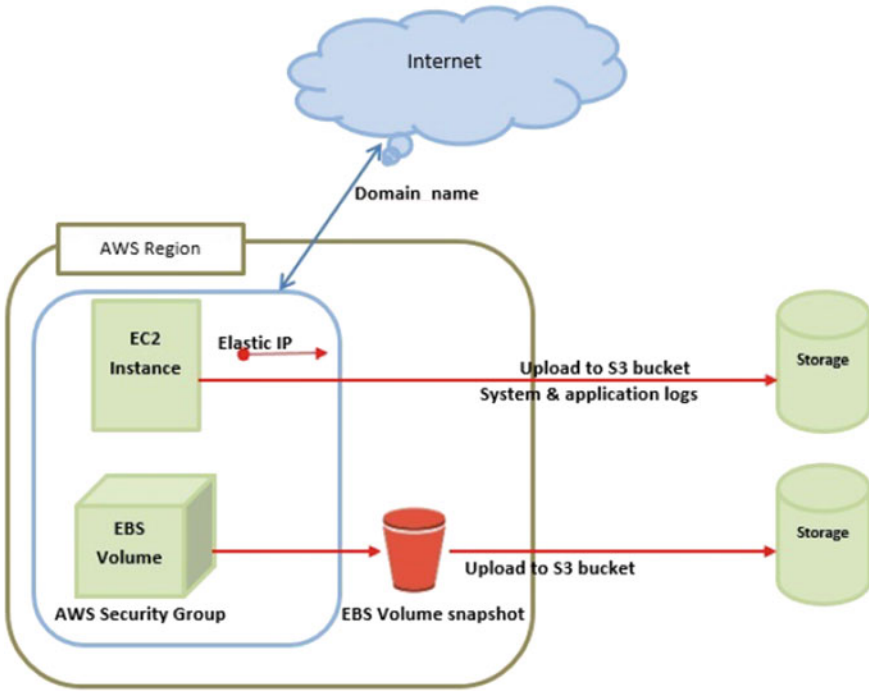


Fig. 2 Architecture of AWS [8]

access it through the url (<https://aws.amazon.com/ec2/>). The applications of EC2 are highly scalable, reliable, and fault tolerant [6, 11].

Amazon S3: It stands for Amazon Simple Storage System. It is designed as a high speed, scalable, and low-cost web service for online backup and storage of data. It allows any kind of data to upload, store, and download up to a size of 5 TB. The user can access this service through the url (<https://aws.amazon.com/s3/>) [6, 11].

Amazon EBS: It stands for Amazon Elastic Block Store which is designed to create virtual storage disks that can be used by EC2. This service can access through (<https://aws.amazon.com/ebs/>). EBS has three types of volumes that are general purpose SSD, provisioned IOPS SSD, and magnetic and are differs in terms of cost, performances, and characteristics [6, 11].

Amazon simple DB: It is a simple structured data store that supports indexing which can be used by both EC2 and S3 for query processing. It is not a full database implementation but scalable. This service can be avail by (<https://aws.amazon.com/simpledb/>) [6, 11].

2.3 Services Provided by AWS

AWS provides a wide range of services to its user. Some of the services are mentioned bellow [6]:

i. Compute	viii. Security
ii. Storage	ix. Identity
iii. Database	x. Compliance
iv. Block chain	xi. Customer engagement
v. Satellite	xii. Business application
vi. Management and governance	xiii. Desktop and app streaming
vii. Analytics	xiv. Internet of Things

3 Load Balancing

The term load balancing refers to balance the load among the nodes. When more than one server is performing task on a remote server, there is a chance that the nodes are getting overloaded, and the performance may slow down. So we have to distribute all the tasks in such an order that the loads among all the nodes will be equal. In the AWS platform, ELB that is the elastic load balancer performs as a load balancer by properly distributing the tasks among different servers for execution [1, 6, 11].

3.1 ELB

It stands for Elastic Load Balancer which automatically distributes the incoming job requests across multiple EC2 instances by achieving high fault tolerance. While jobs keep arriving for execution, it first detects the under load instances and then sends the job to that instances for processing the job in a round robin manner [6, 7, 11]. There are three basic components are available for ELB and are:

Load balancer: It monitors and handles all the requests coming through the Internet or Intranet and distributes them to the registered EC2 instances [6].

Control service: It checks the fitness of the EC2 instances. It performs automatic scaling by adding and removing instances with and from the requests [6].

SSL termination: This is optional in EC2; we can also terminate it. Its basic function is to save CPU cycle, encoding, and decoding SSL the EC2 instances attached with the ELB [6].

3.2 Auto Scaling

Auto scaling allows the user or the organization to automatically scale up or scale down the EC2 instances according to their use. It can be done through the instructions provided by the user. If any application fluctuate hourly, then daily or weekly auto scaling is particularly helpful and effective for those kinds of applications. It is enabled by amazon cloudwatch without any extra cost. CPU utilization, network traffic can also be measured by AWS cloudwatch [6, 7, 11]

4 Implementation, Result, and Analysis

Before configuring load balancer, first we have to create an EC2 instance because it is the only one web-based interface of AWS available for users. After creation of an EC2 instance, we can avail different types of services of AWS through which we can perform load balancing. After creation of EC2 instance, now we have to configure the load balancer by choosing the appropriate type according to our requirement. After configuration of load balancer, we have to register to the target container in the server through which we can perform our tasks. At last, we have to choose auto scaling for our application to scale up or scale down our instance [8, 9].

There are basically three types of load balancers available on AWS, and these are [7, 8, 11] (Fig. 3):

- i. **Application load balancer:** An application load balancer makes and support path-based routing at the application layer through HTTP and HTTPS. In application load balancer, there are different types of container instances which allow dynamic mapping through which we can perform multiple tasks in a single sever. There are different types of files are such as .txt, .mp3, .jpg, and .mp4. In application load balancer, if we are trying to do load balancing, then we have to send our file to the sever instance, and then, the server will verify which type of file we sent to the server for execution. According to the type of file, the server will

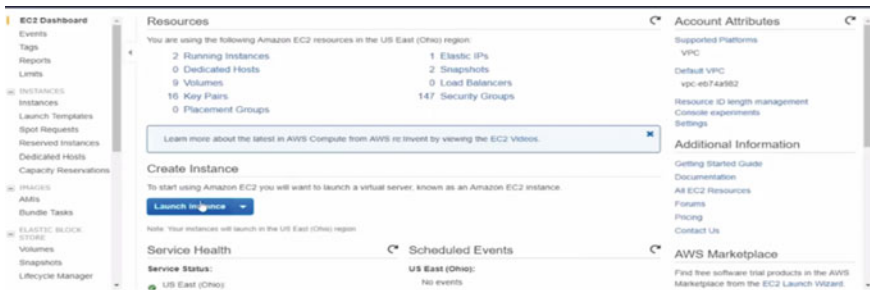


Fig. 3 Creation of EC2 instance [8]

send the file to the corresponding container and perform its execution. By doing this type of load balancing, the network traffic can be reduced, and if the corresponding container is not available, then by dynamic routing, it will choose or create another container virtually and then send for execution. By this response time and throughput will also increase. This is the main benefit for which we use this type of load balancer for our execution.

- ii. **Network load balancer:** Network load balancer makes and support routing at the transport layer through TCP/SSL. Millions of requests per second are handled here. After receiving a connection with the load balancer, it targets the target group; it uses a default rule of flow hash routing algorithm. It doesn't change the network header.
- iii. **Classic load balancer:** A classic load balancer supports routing either at the transport layer through TCP/SSL or at application layer through HTTP/HTTPS. It behaves as a static routing mapping type. In this type, the task should be waiting in the queue till the availability of the container. Once the container is available, then the task will enter and perform its execution. Here, the network traffic will be heavy due to lots of tasks waiting in the queue. Here, response time and throughput will decrease. In classic load balancer, there is a chance of fault tolerance because it detects both healthy and unhealthy instances.

By using EC2 instances, we can create virtual private networks and by ELB, we can perform load balancing. Hence, ELB is an effective load balancing technique for deployment and management of different types of applications. The load balancer maintains and improves security of the applications by providing reliable functions. From the above three types of load balancers application, load balancer is widely used one to maintain the traffic across the network for the deployment of any type of user application of web server. A snapshot of load balancing is depicted in Fig. 4.

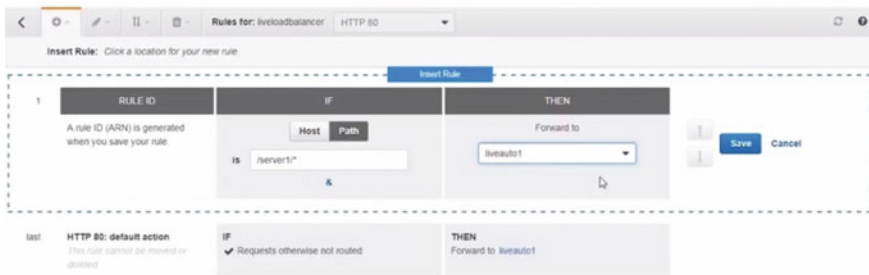


Fig. 4 Load balancing from target-1 to target-2 [8]

5 Conclusion

In this paper, we implement different types of load balancing services provided by AWS form which application load balancer is the effective one for the distribution of work load among the different availability zones or data centers. By using load balancer, the user can achieve high performance and throughput by increasing the response time of the distributed server. AWS is a complete web-based and cloud-based technology and is a best example of Service Oriented Architecture (SOA). AWS is one of the best cloud service provider among others providers available in the market nowadays. It contributes 40% of the total market shares around the globe and provides more than 100 services to their customers with high computation and storage capacity. It is never going to take any extra amount of cost. It also provides free services for a complete year to a new user if he/she doesn't going use it as commercial purpose. As it creates all the services of AWS, at a time for a new user so the down time is quite high from others.

References

1. Rodge, A.S., Chandan, P., Joy, B., Sandeep K.S.:Multicast routing with load balancing using Amazon web service.In: 2014 Annual IEEE India Conference (INDICON), pp. 1–6. IEEE (2014)
2. Zinke, J., Bettina, S.:The impact of weights on the performance of server load balancing systemsIn: 2013 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS), pp. 30–37. IEEE (2013)
3. Shah, V., Harshal, T.:A distributed dynamic and customized load balancing algorithm for virtual instances.In: 2015 5th Nirma University International Conference on Engineering (NUiCONE), pp. 1–6. IEEE (2015)
4. Mell, P., Grance, T.: The NIST definition of cloud computing recommendations of the National Institute of Standards and Technology. Nist Spec. Publ. **145**, 7 (2011)
5. Kumari, P., Kaur, P.: A survey of fault tolerance in cloud computing. J. King Saud Univ. Comput. Inf. Sci.<https://doi.org/10.1016/j.jksuci.2018.09.021>.
6. https://www.tutorialspoint.com/amazon_web_services/amazon_web_services_quick_guide.htm
7. <https://www.edureka.co/aws>
8. <https://aws.amazon.com>
9. <https://www.aws.training>
10. <https://www.guru99.com>
11. Sosinsky, B.: Cloud Computing Bible. Wiley (2011)

New Management Algorithms for Smart Electricity Network: Designing and Working Principles



Ando Ny Aina Randriantsoa, Ali Hamada Damien Fakra, Manitra Pierrot Ranjaranimaro, Mohamed Nasrouline Mohamed Rachadi, and Jean Claude Gatina

Abstract The energy needs considerably increase with the economic development of a country. The promotion of renewable energy sources was enhanced because of global awareness of fossil fuel depletion and climate changes induced by their use. Renewable energy sources are, however, intermittent by nature. Alternative solutions such as the integration of renewable energy systems into the conventional electric grids have been recommended. The integration approach uses photovoltaic, wind, and other renewable energy sources to supply sustainable energy to the built spaces during peak loads or electric backup. In this article, we propose a new approach to manage the energy flow in a smart grid. Two algorithms were developed to manage the integration of renewable sources combined with a storage system in a conventional power grid. The first algorithm aims to smooth the consumption peak and reduce the extraction from the conventional grid, while the second aims to maximize the use of renewable energy depending on the energy demand. Reliability tests of the algorithm behavior have been conducted in comparison with the HOMER software, and the results show a maximum relative error of 4.78% on the management of grid extraction. These algorithms are based on scenarios and operational parameters to optimize the

A. N. A. Randriantsoa (✉) · A. H. D. Fakra · M. P. Ranjaranimaro · M. N. M. Rachadi · J. C. Gatina

Ministry of Higher Education and Scientific Research, Higher Institute of Technology of Antananarivo (I.S.T), Iadiambola Ampasampito, Po Box 8122, Antananarivo 101, Madagascar
e-mail: ando.randriantsoa.istt@gmail.com
URL: <https://piment.univ-reunion.fr/>

A. H. D. Fakra
e-mail: alihamada.fakra@univ-reunion.fr

M. P. Ranjaranimaro
e-mail: pierrot.isttana@gmail.com

M. N. M. Rachadi
e-mail: m.nasrouline@gmail.com

J. C. Gatina
e-mail: jean-claude.gatina@univ-reunion.fr

PIMENT Laboratory, University of La Réunion, 117 rue du General Ailleret, Le Tampon - La Réunion 97430, France

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_55

integration of renewable energy in the current electricity network. Their application will reduce the negative impact of fossil energy and enhance the energy transition.

Keywords Algorithm · Electricity · Renewable energy · Management · Smart grid · Simulation

1 Introduction

Nowadays, most activities depend on energy. For instance: transport, industry, housing, agriculture, air conditioning, health, etc. However, global energy production is dominated by fossil energy sources [1]. The exhaustible nature of this energy [2], pollution issues, and the growth of global energy demands are the main factors that prompted us to adopt the strategy of sustainable development: which meets the needs of the present generation without compromising future generations. Consequently, this challenge inspired us to search for concrete solutions for energy production and management. On the other hand, the varying geographical distribution of energy resources generates geopolitical tensions and holds down the economic growth of countries which do not have enough resources [3]. The economic growth of a country depends on the growth of its energy production, consumption and management [4]. Thus, we are using renewable energy sources to address the problems of depletion and pollution of fossil energy sources. These renewable energy sources may be produced by solar radiation or wind energy, infinite energy sources [5]. Nevertheless, the renewable energy sources encounter the problem of intermittency and its dependence to external factor such as weather. For that reason, we need a storage system to contain the energy during its production phase and reuse this stored energy when needed [6].

In order to measure up to the energy needs, we use an energy-mix system that consists of coupling several energy sources, with at least one renewable energy source. This energy mix has to be managed to optimize its use, and the current electrical network becomes a smart grid. According to the European technology platform: “A smart grid is an electricity network that can intelligently integrate the actions of all users connected to it—generators, consumers, and those that do both—in order to efficiently deliver sustainable, economic and secure electricity supplies” [7]. Our work deals with the development of management algorithms for that smart grid applied to a built space environment. The distribution of electrical energy flow among connected buildings will be the main objective of this study; to smooth the energy extracting energy peaks on the electricity grid and to maximize the use of renewable energy sources by combining a consumption system, a renewable energy production system and a storage system. Two management algorithms are proposed for that purpose. We will commence by presenting the different systems involved and the management strategies adopted for our study, followed by the flowcharts of these algorithms. Then, we will illustrate the results of a comparative study between the new algorithms developed in this work and the HOMER tool which is the reference. Finally, the conclusion and perspectives of this work will be presented.

2 Designing Method of the New Algorithms

The proposed tool can manage the electric energy flow of several buildings. The buildings, presumably, have their own renewable energy production system (solar or wind energy) associated with a storage unit. They are also connected to an electricity network. This study will deal then with four energy systems to manage the electric energy flows (see Fig. 1).

- SP: Renewable energy production system is represented by the renewable energy (in kWh) generated by each building. This renewable energy can be extracted from the sun or the wind;
- SC: Consumption system, which is represented by the energy consumed by each building (in kWh). It may include lightning, electrical appliances, heating, ventilation, and air conditioning (HVAC). A load profile is then attributed to each type of building through these needs;
- STO: Storage system, which is represented by the available stored energy (in kWh) for each building. This storage system is a major part of a smart grid because it is used to store energy when the energy production is high and smooth the grid extraction during consumption peaks;
- RES: Grid withdrawal system, which is represented by the energy extracted from the current electrical grid (in kWh) by each building. Each country has its specific energy sources to feed that electrical grid so that our work will consider the extracted energy without taking into consideration the source of that energy.

The proposed algorithms interact with these four systems every hour to manage the distribution of the electric energy flow. The designing method is built under an algorithmic logic made up of possible scenario cases of an energy-mix system. Every hour, the algorithm takes into consideration the renewable energy production as well as the energy demand and then takes a decision on where the energy will be extracted

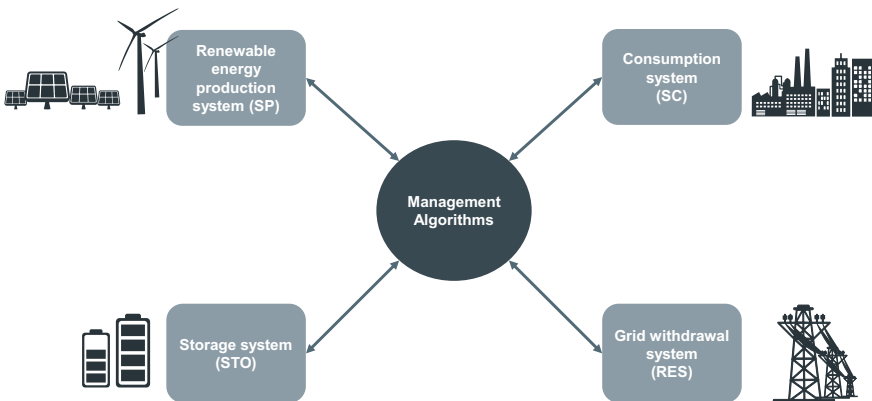


Fig. 1 The energy systems

from. The algorithm alerts if there is an excess amount of energy resulting from the energy mix. In our study, the energy data is modeled in a matrix format. Thus, the calculation process has been developed in a MATLAB environment which is widely used for matrix calculation. Thus, the energy data is described by $m \times n$ matrix:

- m : rows which represent the time step (an hour);
- n : columns which represent the energy data for each connected building.

The system will be presented in the following paragraphs.

2.1 The Consumption-Smoothing Algorithm

The first algorithm proposed in this work has been designed to minimize consumption peaks and to smooth the energy consumption curve. When the consumption is close to a defined limit, energy is extracted from the storage system to avoid excessive extraction from the grid. Priority is given to storage system charging, and this strategy saves the energy cost from the grid extraction.

The control system takes into consideration the following operating parameters:

- LH: upper limit of the storage capacity (in kWh);
- LB: lower limit of the storage capacity (in kWh);
- LR: grid withdrawal limit (in kWh);
- STO_{i-1} : storage situation for the previous time step (in kWh).

The management of the electric energy flow depends on the state of each one of the four systems and the values of the operating parameters. Then, at each hour of the day, the tool evaluates the power available at the renewable source, the power required by the load, the power that can be withdrawn from the grid and the power that can be supplied by the storage system. Depending on the values of these powers and the operating parameters of the storage systems and the grid, the tool decides how much energy to draw from the grid and to supply or take from the storage system. The algorithm detects if there is an excess energy production (EXC). The flowchart of the first algorithm is represented in Figs. 2 and 3.

According to the first algorithm, there are two main cases which determine the management of the energy flow: The first main case is when the energy demand is less than or equal to the grid withdrawal limit. Consequently, if there is no renewable energy production, the energy will be completely drawn from the grid. On the other hand, if there is renewable energy production and the storage unit has the capacity to receive the entire renewable energy production: First, the storage system will be charged, without exceeding the higher limit of storage, by allocating all of the renewable energy production to charge it. The second main case is when energy demand is higher than the grid withdrawal limit. If there is no renewable energy production but the energy from the storage system is available, the excess consumption will be covered by the storage system. If the energy from the storage system is not enough, the rest will be drawn beyond the grid extraction limit from the grid system. In the

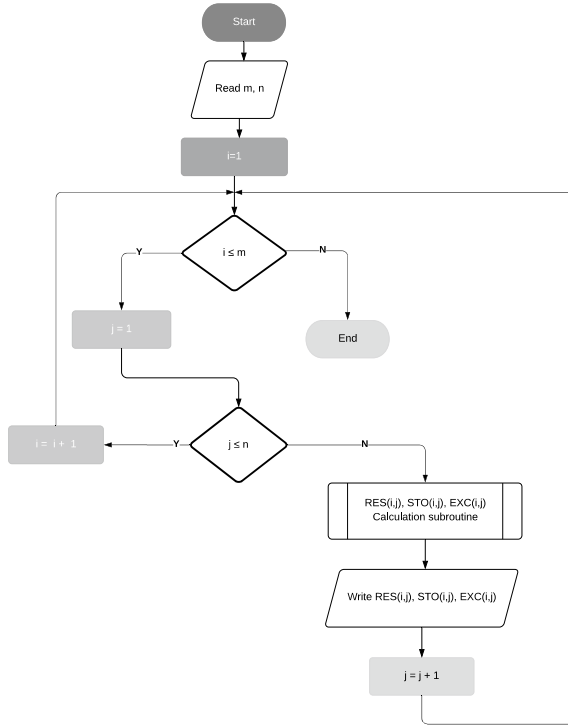


Fig. 2 Consumption-smoothing algorithm

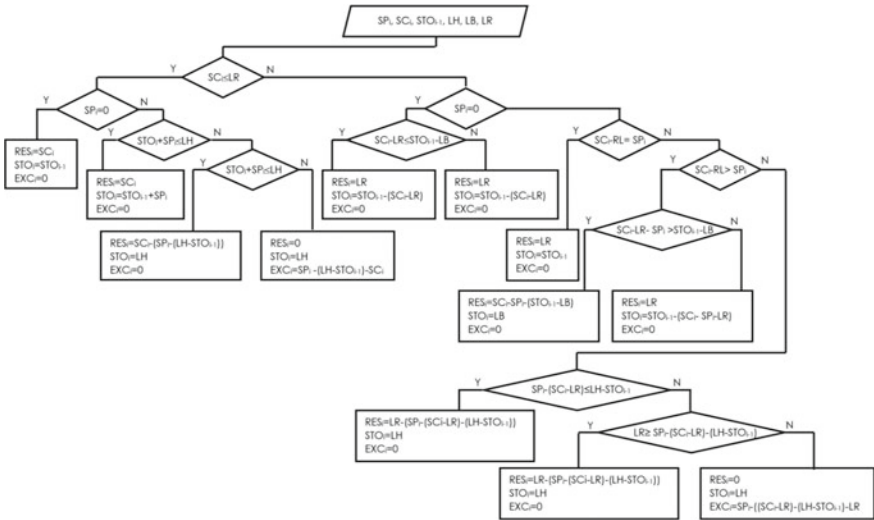


Fig. 3 Consumption-smoothing calculation subroutine

case of sufficient renewable energy production to support the excess consumption, a part of the energy demand will be ensured by the production of renewable energy. If there is no renewable energy production and the storage system is at the lower capacity limit, all the energy demand will be met by the grid.

2.2 The Renewable Energy Algorithm

The second algorithm proposed in this work has been designed to control energy flow and maximize the use of renewable energy source. For this purpose, the energy demand is met by renewable energy sources when they are available. Therefore, the storage system is recharged only when the energy demand is satisfied. The control system of the second algorithm takes the following operating parameters in consideration:

- LH: upper limit of the storage capacity (in kWh);
- LB: lower limit of the storage capacity (in kWh);
- LR: grid withdrawal limit (in kWh);
- STO_{i-1} : storage situation for the previous time step (in kWh);
- $rend_r$: electrical grid efficiency (–);
- $rend_{bat}$: storage battery efficiency (–);
- $rend_{ond}$: converter efficiency (–);
- ond_{out_max} : maximum output power of the converter (in kW);
- ond_{in_max} : maximum input power of the converter (in kW).

The second algorithm also deals with the production system and the consumption system, similar to the first algorithm. Besides, it measures the battery's charging power and discharging power to decide how the energy should be extracted from the storage system and verify if there is any excess energy production (EXC1 corresponds to the DC excess energy, while EXC2 corresponds to the AC excess energy). The algorithm behavior also depends on the operational parameters just like the first algorithm. Although the additional parameters have been included: the electrical grid performance, the storage battery capacity, the effectiveness of the converter (maximum output power and maximum input power). Figures 4, 5, 6 and 7, respectively, demonstrate the main flowchart of the second algorithm and the subroutine calculation. These flowcharts specify how the proposed algorithm displays the energy flow according to each scenario.

For the second algorithm, three main cases have been identified to regulate the electric energy flows: If there is no renewable energy production, the algorithm checks if the energy consumption is less than the grid withdrawal limit. Thus, the electricity network is adequate for the energy demand. If there is renewable energy production, and the energy produced is more than the demanded energy, the algorithm will check if the operating parameters can totally extract from this renewable energy source. Finally, if there is produced renewable energy, but the energy demand is higher, the algorithm will prioritize the use of this renewable energy and thereafter extract from

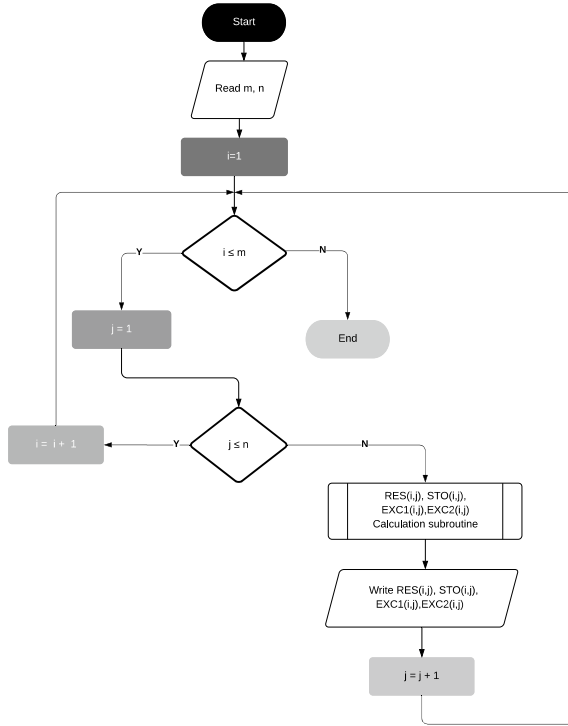


Fig. 4 Renewable energy algorithm

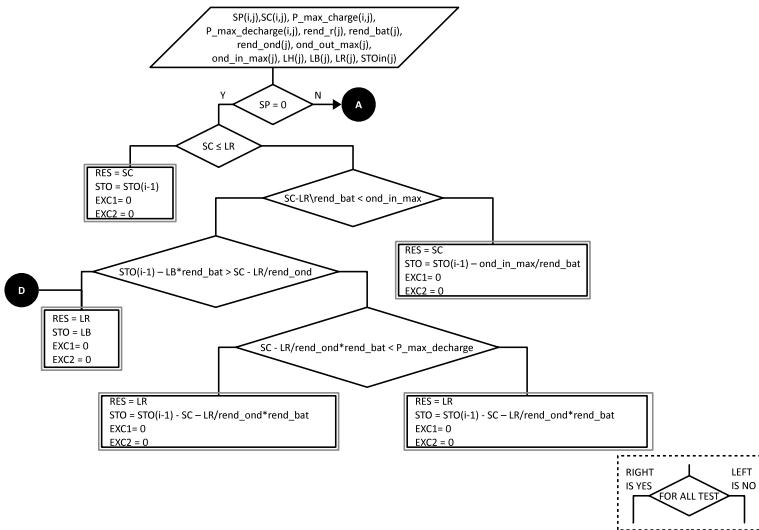


Fig. 5 Renewable energy calculation subroutine—part 1

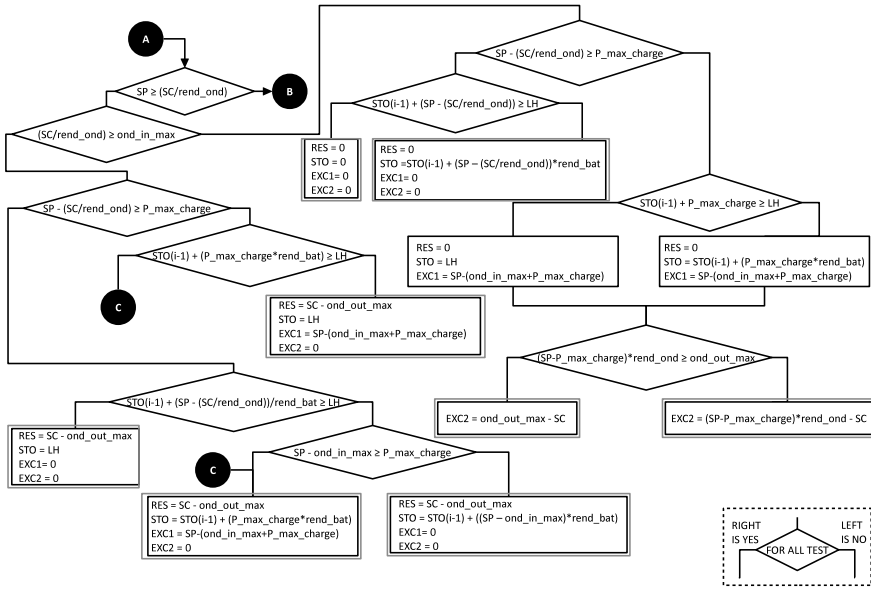


Fig. 6 Renewable energy calculation subroutine—part 2

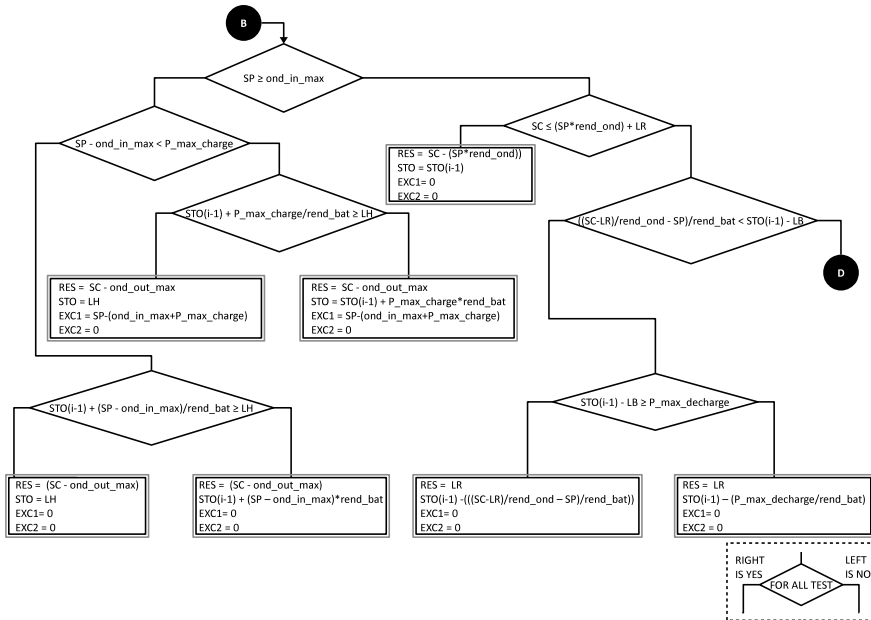


Fig. 7 Renewable energy calculation subroutine—part 3

the electric grid. In a nutshell, the second management algorithm initially extracts from the renewable energy source, then from the electric grid and eventually from the storage system. The storage system will only be charged if the energy demand is met.

3 Simulation Results of the Algorithms

Two new algorithms were designed for the management of smart grid. They have been applied to similar data to evaluate their behavior. The following paragraphs will describe how the simulation has been conducted.

3.1 Hypothesis of the Simulation

A commercial building related to its load profile has been chosen for the simulation. The production system (SP) used for the simulation consists of solar energy production through photovoltaic panels, while the storage system consists of batteries. The data related to the production system has been provided from the NASA annual database in Colorado State of the United States from January 1 at 00:00 to December 31, 2007 at 23:00 which correspond to 8760 hours of simulation.

The parameters used for the simulation are listed as following: The upper limit of storage capacity (HL) is 379.30 kWh; the lower limit of the storage capacity (LB) is 151.72 kWh; the grid withdrawal limit (LR) is 255 kWh; the electrical grid efficiency is 0.8; the storage battery efficiency 0.89; the converter efficiency is about 0.95; the maximum power output of the converter is 137.72 kW, and the maximum input power of the converter is 144.97 kW.

3.2 Behavior of the Algorithms

The algorithms have been included in a MATLAB environment for the matrix calculation process. This process represents the behavior of the two algorithms. Thus, the behavior of the two algorithms can be displayed as curves to be visualized. Figures 8 and 9, respectively, explain the behavior of the first and the second algorithms. The time range varies from 4870 h (2007-07-22 at 21:00) to 4950 hours (2007-07-26 at 05:00).

The simulation shows that the management of the storage unit is distinctive for the two algorithms. The first algorithm is prior to charging storage batteries, while the second algorithm is prior to the use of renewable energy to fulfill the energy consumption. For the renewable energy algorithm, the storage batteries are used to respond the energy consumption when that consumption exceeds the grid withdrawal

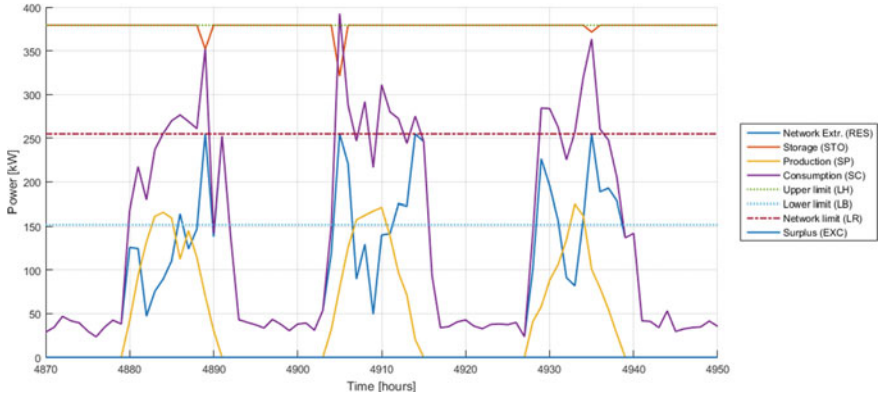


Fig. 8 Consumption-smoothing algorithm behavior

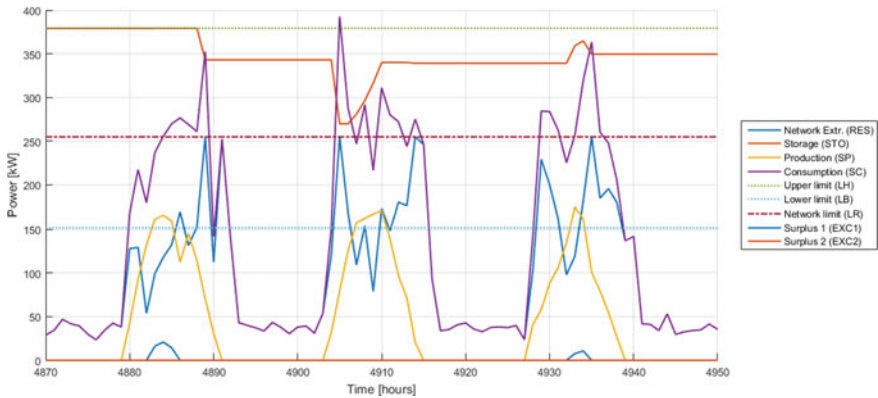


Fig. 9 Renewable energy algorithm behavior

limit. We also notice that the energy extraction from the grid is slightly low for the consumption-smoothing algorithm compared to the renewable energy algorithm without exceeding the grid withdrawal limit.

3.3 Reliability Tests of the Algorithms

The algorithms proposed in this work have been validated by comparison with another tool (HOMER software) under the same parameters to confirm their reliability. HOMER is commonly used in research and validated by the researchers' community. The HOMER software has been developed at the National Renewable Energy Laboratory (NREL, USA) [8], and it is widely used by researchers to conduct simulations on new configurations of renewable energy systems [9–12].

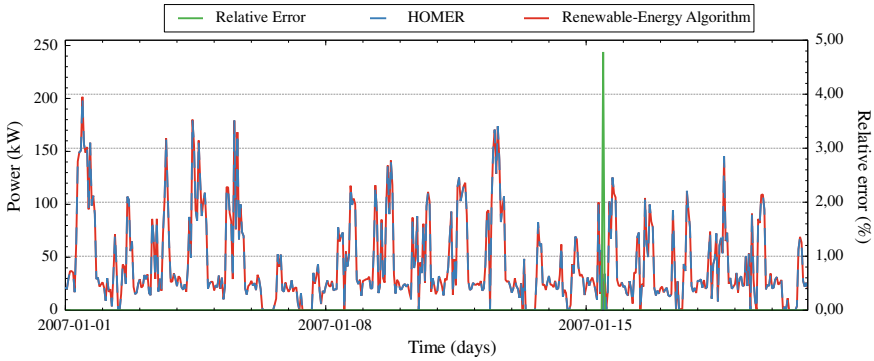


Fig. 10 Grid extraction of the proposed algorithm compared to HOMER

The comparison results (see Fig. 10) show that the renewable energy algorithm meets the comparison with HOMER. The maximum relative error observed relates to the grid extraction system. The difference in power is about 45.2 W, and then, the maximum relative error value is 4.78%. The grid withdrawal limit for the commercial building used in this simulation is about 255 kWh, and the difference in power between HOMER and the renewable energy algorithm is about 5.34W. Consequently, the power difference compared to the grid withdrawal limit is negligible ($1.77 \times 10^{-2}\%$).

4 Conclusion

In order to overcome the current problems of energy and improve the productivity of systems integrating renewable energy sources, two new algorithms for managing the flow of electrical energy are suggested. For each time step, the proposed algorithms either decide the amount of energy to be withdrawn from the grid and the storage system or used to supply the load and the storage unit. The decisions made by these algorithms are determined from data related to the production system and the energy demand from the consumption system. The first algorithm has been developed to reduce consumption peak by optimally managing the extraction of energy from the conventional electricity grid. The second algorithm aims to maximize the use of renewable energy to meet the energy demand. Thus, the choice of the algorithm to be used depends on the final objective of the users. Finally, reliability tests of the algorithm have been conducted by comparing the results obtained in the simulation of a commercial building with those given by the HOMER software. The maximum relative error obtained between the two tools is 4.78% on the power withdrawn from the grid. These values demonstrate the reliability of the proposed algorithm. In conclusion, the two algorithms have one common point: the efficient management of an energy mix integrating renewable energy. However, the objectives of each

algorithm are different: The first one is based on the financial economy of the network and the second one on the availability of renewable energy in this electricity network.

Acknowledgements This work has been developed with the collaboration of the Higher Institute of Technology of Antananarivo (I.S.T), Ministry of Higher Education and Scientific Research, Iadi-ambola Ampasampito, Po Box 8122, Antananarivo 101, Madagascar and the PIMENT Laboratory, University of la Réunion, 117 rue du General Ailleret - 97430 Le Tampon, France and has been funded by the ERASMUS+ project and “La Région Réunion”.

References

1. Martins, F., Felgueiras, C., Smitková, M.: Fossil fuel energy consumption in European countries. *Energy Proc.* **153**, 107–111 (2018)
2. Fantazzini, D., Höök, M., Angelantoni, A.: Global oil risks in the early 21st century. *Energy Policy* **39**, 7865–7873 (2011)
3. Krane, J., Medlock, K.B.: Geopolitical dimensions of US oil security. *Energy Policy* **114**, 558–565 (2018)
4. Kibria, A., Akhundjanov, S.B., Oladi, R.: Fossil fuel share in the energy mix and economic growth. *Int. Rev. Econ. Fin.* **59**, 253–264 (2019)
5. Sørensen, B.: 2—Origin of Renewable Energy Flows, pp. 39–218. *Renewable Energy* (Fifth Edition). Éd. Academic Press, Boston (2017)
6. Trainer, T.: Some problems in storing renewable energy. *Energy Policy* **110**, 386–393 (2017)
7. Dileep, G.: A survey on smart grid technologies and applications. *Renewab. Energy* **146**, 2589–2625 (2020)
8. HOMER—Hybrid Renewable and Distributed Generation System Design Software, <https://www.homerenergy.com/>. 19 Mar 2020
9. Prasetyaningsari, I., Setiawan, A., Setiawan, A.A.: Design optimization of solar powered aeration system for fish pond in Sleman Regency, Yogyakarta by HOMER Software. *Energy Proc.* **32**, 90–98 (2013)
10. Kim, I., James, J.-A., Crittenden, J.: The case study of combined cooling heat and power and photovoltaic system for building customers using HOMER software. *Electr. Power Syst. Res.* **143**, 490–502 (2017)
11. Zahboune, H., Zouggar, S., Krajacic, G., Varbanov, P.S., Elhafyani, M., Ziani, E.: Optimal hybrid renewable energy design in the autonomous system using modified electric system cascade analysis and Homer software. *Energy Convers. Manage.* **126**, 909–922 (2016)
12. Shahzad, M.K., Zahid, A., Rashid, T.U., Rehan, M.A., Ali M., Ahmad, M.: Techno-economic feasibility analysis of a solar-biomass off-grid system for the electrification of remote rural areas in Pakistan using HOMER software. *Renewab. Energy pp.* 264–273 (2017)

Analyzing Key Barriers for Adoption of Digitalization in Indian Construction Industry: A Case Study



Avirag Bajpai and Subhas Chandra Misra

Abstract In India, infrastructure and construction sectors are rapidly growing segments but the main issues faced by the industry are lesser productivity due to slow and delay in operational procedures. Due to this slowness, project trio components, i.e., cost, time, and quality, all are critically compromised. Therefore, innovative plans in construction process are required to take up further to build excellence business model by use of digitalization in construction. To take it forward as a general practice to use digitalization in construction, main barriers are essential to study first. From the extensive literature review and input taken from industry experience, professional's fourteen main barrier alternatives are identified and further ordering are established using Decision-Making Trial and Evaluation Laboratory technique (DEMATEL) method. After the analysis alternative 'improper management approach for digitalization' is having utmost association with remaining alternatives whereas; 'lack of IT policy and standard' is the alternative having minimum association with remaining alternatives. Present-case study reviews the key barriers to implement digitalization as a practice and this research work will be useful to the construction professional in their respective areas to find constraint to implement digitalization in construction processes and assess accordingly.

Keywords Digitalization in construction · Main barriers · DEMATEL

1 Introduction

Currently, we are living in a global world where countries are becoming strong to work on economic scale up and in this construction industry is becoming the back bone of this scenario. It not only generates the proper infrastructure, whereas it also

A. Bajpai (✉) · S. C. Misra
Department of Industrial Management and Engineering, Indian Institute of Technology (IIT)
Kanpur, Uttar Pradesh, Kanpur 208016, India
e-mail: avirag@iitk.ac.in

S. C. Misra
e-mail: subhasm@iitk.ac.in

helps in creating opportunities to other sectors and enables other industries to move fast forward. And this growth scenario will be there for construction sector across the globe. We can say that investment in any economy is always linked to infrastructure. The more structured infrastructure you have, it is more likely to have domestic and international investments in your country [10]. But the vast country like India has some problems in this development process which may hinder the progress. The main reason is the employment generation ability of this industry. Issues which continue to be there are:

- **Process:** Process needs to be digitized but currently there is a limited scope.
- **Digital:** Digital transformation is highly required for real-time management.
- **Qualified employees:** Gaps and unavailability of the skilled work force.
- **Worker productivity:** Currently very low, need to increase it.

These problems are everywhere in the industry. And it resulted in the performance gaps on delivery and quality of the Indian construction projects and this trend is growing very fast and now its alarming [6]. There are associated challenges which cause delay of 20% extra time and 80% excess cost for large construction projects [18] the factors which contribute more here are lack of team management skill and lacuna in innovative digital plans [27]. In this paper, we are exploring fourteen critical points which can be a barrier in adoption of innovative digitalization plan in construction. We arrived to these points after rigorously reviewing literature and several face to face interviews with industry veterans. Along with that, six industry experts had answered a survey to find the implication, relationship, and influence on each other.

This paper contains six sections. Section 1 states the introduction of construction industry followed by a brief literature review regarding key barriers for adoption of digitalization in construction in Sect. 2. Research technique DEMATEL and data analysis as in case company are presented in Sects. 3 and 4, respectively. Finally, paper closes with stating the result and ranking of key barriers in Sect. 5.

2 Literature Review

Let us start with some data—The Labor Bureau Report of 2014 said that in India, we have only 2% skilled workman. This is far lower than global standers and of developed countries. Further, only 6.8% is the number for workmen who are aged 15 years and are skilled or are yet to be skilled [4]. These figures are not encouraging and sending red signals to the sector, which required both skilled and unskilled workers.

This lack of skill has the consequence; unskilled workmen cannot carry out projects as per the new technical transformation. They have struggles with information technology skills, which in turn affect their ability to cope up with new evolving digital systems. Companies are willing to adopt new technology as this is the new competitive edge for them. But these obstacles and barriers need to be

identified before the new adoption. So from the extensive literature review, fourteen barriers alternative have been identified. Those are discussed as below in Table 1.

3 Research Method- DEMATEL Technique

DEMATEL stands for Decision-Making Trial and Evaluation Laboratory technique. It was given by Fontela E. and Gabus A. [7]. DEMATEL is basically used to determine the interrelationship among various alternatives into a cause and effect group through visual diagram. In a complex condition, it is very difficult to identify the precise decision priority in a direct crisp format. So fuzzy function [26] and gray numbers are introduced in DEMATEL method; this will avoid ambiguity in decisions. This method is used in various exploratory researches in several decision-making perspectives such as to find out success factor in hospital industry by Shieh et al. [22] and as well as to find out critical factor for total quality management by Jamali et al. [12].

The DEMATEL method involves the following steps [20].

Step 1: The fourteen barriers alternatives are established for pairwise evaluation.

Step 2: Select the experts from respective knowledge areas to evaluate the pairwise comparison for each alternatives.

Step 3: Pairwise comparison will be framed from experts input based on the following linguistic terms and further converted in to crisp numbers as shown in Table 2.

Step 4: Calculate average direct relationship matrix.

$$A = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix} \tag{1}$$

In a matrix A; $a_{ij} = \frac{1}{k} \sum_{p=1}^s y_{ij}^p$ where ‘k’ represents no of respondents in initial matrix.

Step 5: Establish normalized direct relationship (n) matrix by dividing maximum of column or row sum.

$$n = X * A \tag{2}$$

$$x = \min \left[\frac{1}{\max_{1 \leq i \leq z} \sum_{j=1}^z a_{ij}}, \frac{1}{\max_{1 \leq j \leq z} \sum_{i=1}^z a_{ij}} \right] \tag{3}$$

Step 6: Establish the total influence relationship matrix

Table 1 List of barrier alternatives, description, and references

	Barrier alternatives	Description	Refs.
C1	Absence of interoperability in various modules	Different modules such as finance, planning, store, purchase, quality, IT are not interoperable due to unavailability of common platform in existing business structures	[1, 9]
C2	Lack of system integration among various users	Unintegrated platform for various users such as supervisor, engineers, contractors, and material vendors so that proper communication is not smoothly processed among all	[5, 9]
C3	Lack of data protection and privacy	Lack of stakeholder verification in project so the data is visible to various users because of that it is important to protect data from unwanted handlers	[1, 9]
C4	High digital startup expenditure	For startup, high expenditure is required for installation of software and hardware equipment's	[8, 14]
C5	Prolong pay off period	It is considered that if the startup expenditure is high, then the return on the fund employed is very high	[5, 8, 16, 23]
C6	Lack of IT policy and standard	In organization, new innovative development is going to be collapse due to lack of policy and standards. So it is important to make standard procedures for implementation	[2–4]
C7	Lack of broadband data connectivity	Broadband data connectivity is first and most requirements to implement digitalization in construction	[2, 4]
C8	Lack of technical training from expertise	Due to lack of digital awareness training form expert, it is not possible to develop in house expert to build and maintain digital environment in organization	[9, 23]
C9	Staff technical knowledge for digital implementation	Staff IQ and technical skills plays a vital role for faster and smoother implementation digital platform	[4]
C10	Improper management approach for digitalization	Management approach is significant for any innovation because of policy and vision regarding innovation is approved and monitored by them only	[24, 25]

(continued)

Table 1 (continued)

	Barrier alternatives	Description	Refs.
C11	Users high rigidity to switch toward digitalization	User high rigidity to change from paper-based work to mobile computing and work on real-time environment	From expert
C12	Multiple reporting systems	Multiple reporting system for one entity is the another barriers to implement digitalization in construction	[21]
C13	Lack business IT infrastructure	Company IT infrastructure is also prime mover to implement digitalization such as digital resources and digital setup	From expert
C14	High in seniority and aged staff	If the maximum users are more in age and seniority, then they are more friendly to use conventional setup instead of digital setup	From expert

Table 2 Linguistic scale and respective crisp numbers

Scale (Linguistic)	Crisp numbers
(NO) No influence	0
(VL) Very low influence	1
(L) Low influence	2
(H) High influence	3
(VH) Very High Influence	4

$$\text{Total} = \lim_{l \rightarrow \infty} (n^1 + n^2 + \dots + n^l) = n(I - n)^{-1} \quad \text{when} \quad \lim_{l \rightarrow \infty} n^l = 0 \quad (4)$$

Step 7: Find out the row ($\sum D_{ij}$) and column ($\sum C_{ij}$) sum of total influence relationship matrix and calculate causal criteria as ‘prominence’ ($D + C$) values and influenced criteria as a ‘relation’ ($D - C$) values.

Step 8: Finally, draw the causal diagram as a prominence–relation map in between prominence and relation values. Importance of barriers alternative will be given by the prominence values and cause and effect category of barriers alternative will be given by relation values.

4 Case Study to Identify Main Barriers Alternatives for Digitalization in Construction Firm

This case study data has been taken from one of the large Indian construction organization, six experienced professional, having similar experience completed the survey

Table 3 Average direct relationship matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	0.00	3.33	2.00	2.16	2.16	1.50	2.66	2.33	2.33	3.50	3.16	2.83	2.66	2.50
C2	3.50	0.00	3.16	3.00	2.16	0.83	2.16	2.33	2.50	3.00	2.50	2.16	3.16	2.16
C3	2.83	3.33	0.00	2.83	2.66	2.66	2.33	2.33	2.33	2.83	1.16	2.16	2.33	1.33
C4	3.16	3.00	2.33	0.00	3.00	2.00	2.16	2.50	1.83	3.00	1.16	2.00	2.33	0.83
C5	2.33	2.50	2.50	2.83	0.00	1.83	2.00	1.66	1.16	2.83	1.50	1.00	2.00	0.50
C6	2.33	2.50	2.16	2.66	1.83	0.00	2.50	1.33	1.50	3.16	1.33	1.50	1.16	0.66
C7	3.16	2.66	2.50	2.16	2.00	1.83	0.00	1.83	1.83	2.83	1.83	1.83	2.33	1.16
C8	2.66	2.66	2.50	1.33	1.66	1.50	2.00	0.00	3.00	2.50	2.83	2.16	3.16	2.16
C9	3.00	2.83	3.00	1.33	1.50	0.83	2.00	3.16	0.00	2.83	2.50	2.33	3.00	2.16
C10	3.00	2.66	3.00	3.00	3.00	2.33	3.33	2.83	2.33	0.00	2.83	2.50	3.33	2.50
C11	2.66	2.50	2.50	2.00	2.00	1.00	2.33	2.83	3.33	2.66	0.00	3.00	2.16	3.00
C12	2.50	2.33	1.66	2.83	1.83	1.33	2.00	2.33	2.66	2.83	2.16	0.00	2.50	2.50
C13	2.83	3.00	3.50	3.50	3.00	2.16	2.33	2.50	2.66	3.33	2.16	2.16	0.00	2.00
C14	2.83	2.33	1.83	2.33	1.83	0.50	1.50	2.66	3.33	3.16	3.50	2.66	2.16	0.00

and evaluated the impact of all barriers alternatives. They evaluated all barrier alternatives on the scale of 0, 1, 2, 3, and 4, i.e., no influence, very low influence, low influence, high influence, and very high influence, respectively [15]. Further, linguistic scale has been converted in to crisp numbers as per Table 2 and average direct relationship matrix has been established with the help of Eq. (1) as represented in Table 3. The normalized direct relationship matrix has been established as per Eq. (2) and Eq. (3), as represented in Table 4 and the total influence relationship matrix has been formulated as per Eq. (4) as shown in Table 5.

The values of row ($\sum D_{ij}$) and column ($\sum C_{ij}$) sum of total influence relationship matrix are calculated and further ‘prominence’ ($D + C$) and ‘relation’ ($D - C$) values are derived as represented in Table 6. Finally, causal diagram (prominence–relation map) is plotted in between $D + C$ and $D - C$ values as shown in Fig. 1.

5 Discussion and Conclusion

For analyzing the interdependencies in all alternatives, causal diagram is drawn, and as per values of $D + C$ and $D - C$ in Table 6, alternatives can be categorized in following groups.

- (1) $D + C$ higher and $D - C$ positive—C9, C11, C13
- (2) $D + C$ higher and $D - C$ negative—C1, C2, C3, C4, C8, C10
- (3) $D + C$ Smaller and $D - C$ positive—C6, C12, C14
- (4) $D + C$ Smaller and $D - C$ negative—C5, C7

Table 4 Normalized direct relationship matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	0.00	0.08	0.05	0.05	0.05	0.03	0.06	0.06	0.06	0.0	0.08	0.07	0.06	0.06
C2	0.09	0.00	0.08	0.07	0.05	0.02	0.05	0.06	0.06	0.07	0.06	0.05	0.08	0.05
C3	0.07	0.08	0.00	0.07	0.06	0.06	0.06	0.06	0.06	0.07	0.03	0.05	0.06	0.03
C4	0.08	0.07	0.06	0.00	0.07	0.05	0.05	0.06	0.04	0.07	0.03	0.05	0.06	0.02
C5	0.06	0.06	0.06	0.07	0.00	0.04	0.05	0.04	0.03	0.07	0.03	0.02	0.05	0.01
C6	0.06	0.06	0.05	0.06	0.04	0.00	0.06	0.03	0.03	0.08	0.03	0.03	0.03	0.01
C7	0.08	0.06	0.06	0.05	0.05	0.04	0.00	0.04	0.04	0.07	0.04	0.04	0.06	0.03
C8	0.06	0.06	0.06	0.03	0.04	0.03	0.05	0.00	0.07	0.06	0.07	0.05	0.08	0.05
C9	0.07	0.07	0.07	0.03	0.03	0.02	0.05	0.08	0.00	0.07	0.06	0.06	0.07	0.05
C10	0.07	0.06	0.07	0.07	0.07	0.06	0.08	0.07	0.06	0.00	0.07	0.06	0.08	0.06
C11	0.06	0.06	0.06	0.05	0.05	0.02	0.06	0.07	0.08	0.06	0.00	0.07	0.05	0.07
C12	0.06	0.06	0.04	0.07	0.04	0.03	0.05	0.06	0.06	0.07	0.05	0.00	0.06	0.06
C13	0.07	0.07	0.09	0.09	0.07	0.05	0.06	0.06	0.06	0.08	0.05	0.05	0.00	0.05
C14	0.07	0.06	0.04	0.06	0.04	0.01	0.03	0.06	0.08	0.08	0.09	0.06	0.05	0.00

Table 5 Total influence relationship matrix

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	<u>0.30</u>	<u>0.37</u>	<u>0.32</u>	<u>0.31</u>	<u>0.29</u>	0.21	<u>0.31</u>	<u>0.31</u>	<u>0.31</u>	<u>0.39</u>	<u>0.31</u>	<u>0.30</u>	<u>0.33</u>	0.26
C2	<u>0.38</u>	<u>0.29</u>	<u>0.34</u>	<u>0.33</u>	<u>0.29</u>	0.19	<u>0.29</u>	<u>0.31</u>	<u>0.31</u>	<u>0.38</u>	<u>0.29</u>	0.28	<u>0.34</u>	0.25
C3	<u>0.35</u>	<u>0.35</u>	0.25	<u>0.31</u>	<u>0.29</u>	0.22	0.28	<u>0.29</u>	<u>0.29</u>	<u>0.36</u>	0.25	0.27	<u>0.31</u>	0.21
C4	<u>0.34</u>	<u>0.33</u>	<u>0.29</u>	0.23	0.28	0.20	0.27	0.28	0.26	<u>0.34</u>	0.24	0.25	<u>0.29</u>	0.19
C5	0.28	0.28	0.26	0.26	0.18	0.17	0.23	0.23	0.22	<u>0.30</u>	0.21	0.20	0.25	0.16
C6	0.28	0.28	0.25	0.26	0.22	0.13	0.24	0.22	0.22	<u>0.31</u>	0.21	0.21	0.23	0.16
C7	<u>0.33</u>	<u>0.31</u>	<u>0.29</u>	0.27	0.25	0.19	0.21	0.26	0.26	<u>0.33</u>	0.24	0.24	0.28	0.20
C8	<u>0.34</u>	<u>0.33</u>	<u>0.31</u>	0.27	0.26	0.19	0.27	0.23	0.30	<u>0.34</u>	<u>0.29</u>	0.27	<u>0.32</u>	0.24
C9	<u>0.35</u>	<u>0.34</u>	<u>0.32</u>	0.28	0.26	0.18	0.27	<u>0.31</u>	0.23	<u>0.36</u>	0.28	0.27	<u>0.32</u>	0.24
C10	<u>0.40</u>	<u>0.38</u>	<u>0.36</u>	<u>0.36</u>	<u>0.33</u>	0.24	<u>0.34</u>	<u>0.34</u>	<u>0.33</u>	<u>0.34</u>	<u>0.32</u>	<u>0.31</u>	<u>0.37</u>	0.27
C11	<u>0.35</u>	<u>0.34</u>	<u>0.32</u>	<u>0.30</u>	0.28	0.19	<u>0.29</u>	<u>0.31</u>	<u>0.32</u>	<u>0.36</u>	0.23	<u>0.30</u>	<u>0.31</u>	0.26
C12	<u>0.33</u>	<u>0.32</u>	0.28	<u>0.30</u>	0.26	0.18	0.26	0.28	<u>0.29</u>	<u>0.34</u>	0.26	0.21	<u>0.30</u>	0.24
C13	<u>0.38</u>	<u>0.38</u>	<u>0.36</u>	<u>0.36</u>	<u>0.32</u>	0.23	<u>0.31</u>	<u>0.32</u>	<u>0.33</u>	<u>0.40</u>	<u>0.30</u>	<u>0.30</u>	0.28	0.25
C14	<u>0.35</u>	<u>0.33</u>	<u>0.30</u>	<u>0.30</u>	0.27	0.17	0.26	<u>0.30</u>	<u>0.32</u>	<u>0.36</u>	<u>0.31</u>	0.28	<u>0.30</u>	0.19

In this study, group (1) and group (3) alternatives, above the $D + C$ axis (X axis) in causal diagram shown in Fig. (1), are categorized as influential alternatives. It means alternative, which are present in these groups, i.e. C6, C9, C11, C12, C13, and C14 are going to influence to the other remaining alternatives. While, group (2) and group (4) alternatives, i.e., C1, C2, C3, C4, C5, C7, C8, and C10 below the $D + C$ axis (X axis)

Table 6 Prominence and relation values

	Sum <i>D</i>	Sum <i>C</i>	<i>D</i> + <i>C</i>	Rank	<i>D</i> - <i>C</i>
C1	4.379	4.808	9.187	2	-0.429
C2	4.331	4.669	9.000	3	-0.338
C3	4.081	4.314	8.396	5	-0.233
C4	3.867	4.203	8.071	8	-0.336
C5	3.292	3.824	7.116	13	-0.532
C6	3.282	2.762	6.044	14	0.520
C7	3.723	3.897	7.620	11	-0.175
C8	4.015	4.060	8.075	7	-0.045
C9	4.077	4.049	8.125	6	0.028
C10	4.761	4.978	9.740	1	-0.217
C11	4.224	3.804	8.027	9	0.420
C12	3.916	3.763	7.678	10	0.153
C13	4.588	4.310	8.898	4	0.279
C14	4.091	3.186	7.276	12	0.905



Fig. 1 Causal diagram (prominence–relation map)

in causal diagram shown in Fig. (1) are categorized as impacted alternatives. The highest rank of prominence or influencing characteristic of alternative is determined by the maximum value of $D + C$. In this case, maximum $D + C$ value of 9.740 for the alternative 'improper management approach for digitalization' (C10), so it is having highest interrelationship with remaining alternatives. While 'lack of IT policy and standard' (C6) is having minimum $D + C$ value of 6.044 so it is having least interrelationship with others. The ranks of alternatives are indicated in terms of influencing characteristic as C10-C1-C2-C13-C3-C9-C8-C4-C11-C12-C7-C14-C5-C6.

Further, alternatives are ordered into cause and effect set based on the derived $D-C$ values. The cause set is having utmost positive $D-C$ values and having extreme direct causal relationship on the other alternatives. While the effect set is having negative $D-C$ values. In this case, alternative C14 (high in seniority and aged staff) is having highest causal relationship with others as a maximum $D-C$ value of 0.905. C6 (lack of IT policy and standard) and C11 (users high rigidity to switch toward digitalization) are the second and third most causal alternatives and having $D-C$ values 0.520, 0.420, respectively. In this case, it is also observed that C6 (lack of IT policy and standard) and C14 (high in seniority and aged staff) are influencing all the other barriers, but not be influenced by any other barriers because both are having lesser threshold value. (Threshold value is a mean of total relation matrix and in this case it is 0.289). The ranks of alternatives are indicated in terms of positive $D-C$ values (cause set) as C14-C6-C11-C13-C12 and in terms of negative $D-C$ values (effect set) as C9-C8-C10-C7-C3-C2-C4-C1-C5.

After observation, alternative 'improper management approach for digitalization' is having maximum interrelationship with remaining alternatives, whereas 'lack of IT policy and standard' is the alternative having minimum interrelationship with remaining alternatives. However, alternative 'high in seniority and aged staff' is the most causal alternative, whereas 'multiple reporting systems' is the least causal alternative among all. While 'staff technical knowledge for digital implementation, is most effected barriers alternative and 'prolong pay off period' is least effected alternative in all barriers. At last, this case study offers definite idea to the construction industry professionals regarding key barriers to implement digitalization in construction as a practice.

The main limitations of the case study are discussed below, which can be used in future to develop more dimension in this research:

- This case study was conducted in one of the large Indian construction organization with only selected number responses from working professional. Further, it can be used in multiple numbers of organizations with additional numbers of respondents.
- In this case study, DEMATEL method is used judiciously for exploratory analysis. In future, other MCDM techniques or other descriptive methods can be used to avoid ambiguity and vagueness in results.
- In this case study, fourteen barriers are identified but some other barriers can also be incorporated depend on the organization resources, structure, and geographical location.

References

1. Alaba, F.A., Othman, M., Hashem, I.A.T., Alotaibi, F.: Internet of things: a survey. *J. Netw. Comput. Appl.* **88**, 10–28 (2017)
2. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. *IEEE Commun. Survey Tutorials* **17**, 2347–2376 (2015)
3. Bandyopadhyay, S.: Internet of things: applications and challenges in technology and standardization. *Wirel. Pers. Commun.* **58**, 49–69 (2011)
4. Bedekar, A.: Opportunities & Challenges for IoT in India. Accessed on November 29, 2019. Online available at: <https://technology.siliconindia.com/viewpoint/exoinsights/opportunities-challenges-for-iot-in-india-nwid-3444.htm> (2017)
5. Da Xu, L., He, W., Li, S.: Internet of things in industries: a survey. *IEEE Trans. Indust. Inform.* **10**(4), 2233–2243 (2014)
6. Doloi, H., Sawhney, A., Iyer, K.C., Rentala, S.: Analysing factors affecting delays in Indian construction projects. *Int. J. Project Manage.* **30**(4), 479–489 (2012)
7. Fontela, E., Gabus, A.: DEMATEL, Innovative Methods. Rep. No. 2, ‘Structural Analysis of the World Problematique (Methods). Battelle Geneva Research Institute, Geneva. (1974)
8. Granjal, J., Monteiro, E., Silva, J.S.: Security for the internet of things: a survey of existing protocols and open research issues. *IEEE Commun. Survey Tutorials.* **17**(3), 1294–1312 (2015)
9. Hussain, M.: Internet of things: challenges and research opportunity. *CSI Trans. ICT.* **5**(1), 87–95 (2016)
10. IBEF: Infrastructure Sector in India. Accessed on November 29, 2019. Online available at: <https://www.ibef.org/industry/infrastructure-sector-india.aspx> (2019)
11. India Services: Ministry of Commerce and Industry, Overview-Infrastructure and Construction. Accessed on November 29, 2019. Online available at, <https://www.indiaservices.in/construction/> (2019)
12. Jamali, G., Ebrahimi, M., Abbaszadeh, M. A.: TQM implementation: an investigation of critical success factors. In *Education and Management Technology (ICEMT), 2010 International Conference on*, 112–116 (2010)
13. KPMG International: Global Construction Survey: Make it, or Break it. . Accessed on November 29, 2019. online available at: <https://home.kpmg/xx/en/home/insights/2017/10/global-construction-survey-make-it-or-break-it.html> (2017)
14. Lee, I., Lee, K.: The internet of things (IoT): applications, investments, and challenges for enterprises. *Bus. Horiz.* **58**(4), 431–440 (2015)
15. Lin, Y.T., Yang, Y.H., Kang, J.S., Yu, H.C.: Using DEMATEL method to explore the core competences and causal effect of the IC design service Company: an empirical case study. *Expert Syst. Appl.* **38**(5), 6262–6268 (2011)
16. Luthra, S., Garg, D., Mangla, S.K., Berwal, Y.P.S.: Analyzing challenges to internet of things (IoT) adoption and diffusion: an Indian context. *Procedia Computer Science.* **125**, 733–739 (2018)
17. Makeinindia.com: Construction Sector. Accessed on November 29, 2019. Online available at: <https://www.makeinindia.com/sector/construction> (2019)
18. McKinsey & Company: Imagining Construction’s Digital Future. Accessed on November 29, 2019 Online available at: <https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/imagining-constructions-digital-future> (2016)
19. Oesterreich, T.D., Teuteberg, F.: Understanding the implications of digitisation and automation in the context of Industry 4.0: a triangulation approach and elements of a research agenda for the construction industry. *Comput. Industry.* **83**, 121–139 (2016)
20. Singh, S., Misra, S.C.: Success determinants to product lifecycle management (PLM) performance. In: *Proceedings of the 5th International Conference on Industrial Engineering and Applications*, 386–390 (2018)
21. Saaksvuori, A., Immonen, A.: *Book—Product Lifecycle Management*. Springer, Berlin (2008)

22. Shieh, J.I., Wu, H.H., Huang, K.K.: A DEMATEL method in identifying key success factors of hospital service quality. *Knowl.-Based Syst.* **23**(3), 277–282 (2010)
23. Talavera, J.M., Tobon, L.E., Gomez, J.A., Culman, M.A., Aranda, J.M., Parra, D.T., Quiroz, L.A., Hoyos, A., Garreta, L.E.: Review of IoT applications in agro-industrial and environmental fields. *Comput. Electron. Agricul.* **142**, 283–297 (2017)
24. Wong, A., Scarbrough, H., Chau, P., Davison, R.M.: Critical failure factors in ERP implementation. In: *Pacific Asia Conference on Information Systems (PACIS)* (2005)
25. Yeo, K.T.: Critical failure factors in information systems projects. *Int. J. Project Manage.* **20**(3), 241–246 (2000)
26. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**(3), 338–353 (1965)
27. Zhang, C., Zou, Y.: Role of operational excellence in construction industry: A review. *Int. J. Eng. Appl. Sci.* **4**(5), 25–28 (2017)

Improved Peak-to-Average Power Ratio Reduction Method for OFDM/OQAM System



V. Sandeep Kumar

Abstract A low complexity peak-to-average power ratio (PAPR) reduction method is proposed for orthogonal frequency division multiplexing with offset quadrature amplitude modulation (OFDM/OQAM) systems. Firstly, the time-domain oversampling is performed, and then power normalization is adopted and then followed by clipping. Finally, the output clipped signals are filtered by a low-pass filter. In addition, effects of clipping with oversampled OFDM/OQAM signals are analyzed in terms of signal noise distortion ratio (SNDR), signal distortion ratio (SDR) and signal-to-noise ratio (SNR). Simulation results demonstrate that the proposed method can effectively reduce the peak regrowth and PAPR after filtering.

Keywords OFDM/OQAM · SNR · SDR · SNDR · PAPR · Clipping and filtering

1 Introduction

Orthogonal frequency division multiplexing with offset quadrature amplitude modulation (OFDM/OQAM) can effectively eliminate inter-symbol interference (ISI) caused by multipath propagation due to its high spectral efficiency [1]. In addition, OFDM/OQAM can be effectively combined with multiple-input multiple-output (MIMO) with image processing applications [2–6] and is adopted by many wireless communication broadband standards, such as European digital video broadcasting, 3GPP LTE system, WLAN IEEE802.11a and WiMAX broadband wireless access standards, and the OFDM/OQAM technique is widely used in future. Similar to OFDM system, OFDM/OQAM system also suffers from the problem of high peak-to-average power ratio (PAPR), which will result in two adverse effects: the higher the PAPR value, the larger is the dynamic range of the signal, the higher is the accuracy requirements of the analog-to-digital converter, and higher the cost of the device. From the perspective of power conversion, when the RF signal with a high PAPR

V. Sandeep Kumar (✉)

Department of ECE, SR Engineering College, Hasanparthy, Telangana 506371, India
e-mail: sandeep_kumar_v@srcwarangal.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_57

695

passes through a power amplifier, some distortion is introduced. For example: In-band distortion and out-of-band radiation are generated, which results in adjacent channel interference, thereby degrading bit error rate performance. So, to reduce PAPR for OFDM/OQAM system, some methods are proposed in literature. The commonly used methods are clipping [7–9], selective mapping [10–13], partial transmission sequence and precoding [14–16]. Considering the implementation complexity, the algorithm for reducing the peak-to-average power ratio should meet the following requirements: (1) compatible with existing modulation schemes; (2) high spectral efficiency; (3) low computational complexity; (4) the receiver structure must not be changed. Taking into account the above points, the conventional clipping method is most convenient, but it will introduce clipping noise. Filtering the clipped signal reduces the noise, but it increases the peak regrowth.

In this paper, a WiMAX system using OFDM/OQAM modulation is taken to study the time-domain oversampling, then clipping and filtering method. The simulation results show that the proposed method not only effectively reduce the PAPR, but also improve the bit error rate (BER) performance, which is a simple and effective method for OFDM/OQAM systems.

2 System Model

The block diagram of OFDM/OQAM system is shown in Fig. 1. Since channel coding and channel model have less impact on PAPR, for simplicity, no channel coding is used and the channel is assumed to be an ideal additive white Gaussian noise channel (AWGN). The nonlinear effects of the power amplifier introduce signal distortion, including amplitude/phase (AM/PM) and amplitude/amplitude (AM/AM) distortion. According to the Rapp model of the solid-state power amplifier [17], its AM/AM property $A(r) = Lr/[1 + (Lr/A_0)^{2p}]^{(1/2p)}$, where L is the amplifier linear gain, r is the amplitude of input signal, A_0 is amplifier output power, and p is amplifier smoothing factor, larger p , indicates the higher degree of linearization. The typical values are $p = 10$, $A_0 = 20$, AM/PM attribute $\phi(r) = 0$, which assumes no phase distortion.

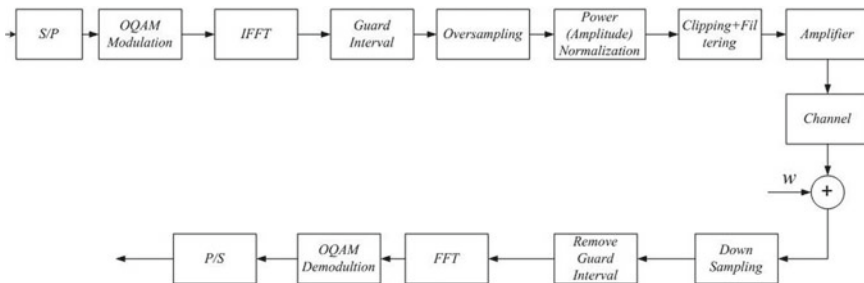


Fig. 1 Block diagram for the proposed method

3 PAPR Calculation and Clipping Process

3.1 PAPR Calculation

The real-valued symbols modulated by offset QAM are transmitted on each subcarrier, and the transmitted signal is written as

$$x(n) = \sum_{k=0}^{N-1} \sum_{m=0}^{M-1} X_m^k h(n - m\frac{N}{2}) e^{jk(\frac{2\pi n}{N} + \frac{\pi}{2})}, n = 0, 1, \dots, (L + M - 1)N \quad (1)$$

Since the adjacent time-domain signals of OFDM/OQAM data blocks overlap, the conventional definition of PAPR for OFDM systems no longer applicable to OFDM/OQAM systems. So the OFDM/OQAM signal length is divided into $(M + L)$ segments and the PAPR is calculated for each segment individually.

$$PAPR_l = \frac{\max_{lN \leq n \leq (l+1)N-1} |x(n)|^2}{E|x(n)|^2}, l = 0, 1, \dots, L + M - 1 \quad (2)$$

where $|x(n)|$ is the magnitude of $x(n)$, $E[\cdot]$ represents expectation operation. From the central limit theorem, when the number of symbol subcarriers N is large, the real part $\text{Re}(x(n))$ and the imaginary part $\text{Im}(x(n))$ of the OFDM/OQAM signal. Here, X_m^k is assumed as independent of each other and define a variance of σ_x^2 . Obviously, $x_k(n)$ values are independent of each other. Let σ_x^2 be the $s_x(n)$ variance, i.e.,

$$E\{x_k(n)\} = 0; \sigma_x^2 = E\{x_k(n)x_k^*(n)\} = \sigma_x^2 \sum_{m=0}^{M-1} h(n - m\frac{N}{2}) \quad (3)$$

Thus, the mean and variance of $x_k(n)$ are independent of k . According to the central limit theorem, if N value is large enough. $x(n)$ obeys the complex Gaussian distribution, with mean 0, variance 1, where σ_x^2 represents the variance of the real and imaginary parts of $x(n)$.

3.2 Clipping Principle

The amplitude of output signal x_n after IFFT is given by (1), and the mathematical expression for clipped signal is given by

$$y_n = \begin{cases} -Th, & |x_n| < -A \\ x_n, & -A < |x_n| \leq A, (0 < n < (L + M - 1)N) \\ Th, & |x_n| > A \end{cases} \quad (4)$$

where Th is clipping threshold, y_n is the time-domain signal after clipping. The OFDM/OQAM symbol magnitude needs to be normalized with mathematical expectation $E(x_n)$. After normalization, the clipping ratio is expressed as

$$CR = \frac{Th}{E(x_n)} = Th \tag{5}$$

$CR = \infty$ indicates signal without any clipping process.

Based on (4) and (5), (2) can be re-represented as

$$PAPR = 10 \log(CR)^2 \tag{6}$$

Assuming that the OFDM/OQAM signal x_n satisfies the complex Gaussian distribution, its magnitude obeys the Rayleigh distribution. Therefore, the power of output signal after clipping is

$$P_{out} = (1 - e^{-CR^2})P_{in} \tag{7}$$

where P_{in} is the average power of the OFDM/OQAM signal before clipping, and P_{out} is the average power after clipping. A complementary cumulative distribution function (CCDF) curve is typically used to characterize the distribution of PAPR. In this paper, the CCDF curve represents the statistical probability that PAPR is higher than any given value.

Figure 2 illustrates the PAPR comparison for different oversampling values with respect to clipping and filtering method, but low-pass filtering the clipped OQAM

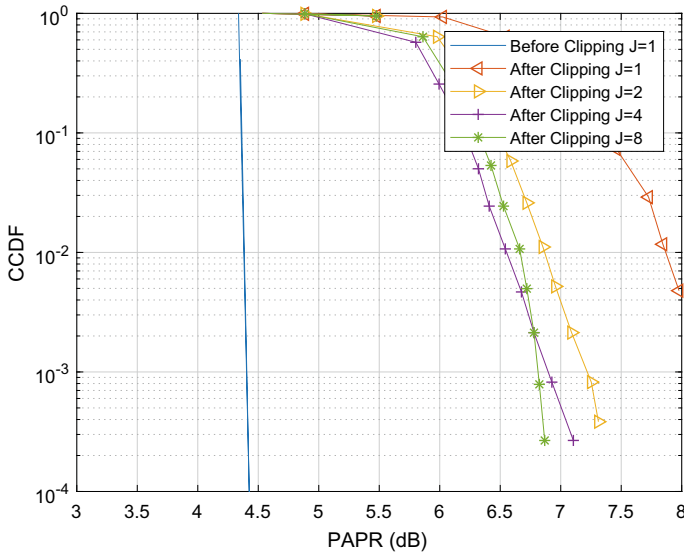


Fig. 2 CR = 1.6, CCDF under different oversampling values J

signal usually increase the PAPR (due to peak regrowth). In order to reduce the peak regrowth, filtering is done after time-domain oversampling. The principle of this method is that time-domain oversampling can increase the correlation between the OFDM/OQAM signals after clipping. So, after the signal is reconstructed, the oversampled signal can avoid peak regrowth more effectively than the Nyquist-sampled signal.

For discrete signals, the PAPR is calculated using (2). Figure 3 illustrates that after oversampling, the PAPR value is reduced by more than 1 dB than the PAPR value of the Nyquist sampling and the higher the oversampling J , the lower is the PAPR value. However, when the oversampling value is greater than 4 ($J > 4$), there is a marginal improvement in PAPR performance. In order to remove out-of-band radiation from the clipping operation, the clipped signal requires low-pass filtering. After clipping y_n is sampled, its frequency domain expression is given by

$$Y_k = Y_0, Y_1, \dots, Y_{(L+M-1)N-1}, Y_{(L+M-1)N}, \dots, Y_{(L+M-1)JN-1} \tag{8}$$

where $Y_k = \text{IFFT}((L + M - 1)JN, y_n)$, J is an oversampling multiple. It is obvious that low-pass filtering effectively removes the out-of-band radiation of the clipped signal. After low-pass filtering, the average power of the clipped OFDM/OQAM signal is defined as

$$P_{av} = \frac{1}{N} \sum_{k=0}^{N-1} E[|Y_k|^2] \tag{9}$$

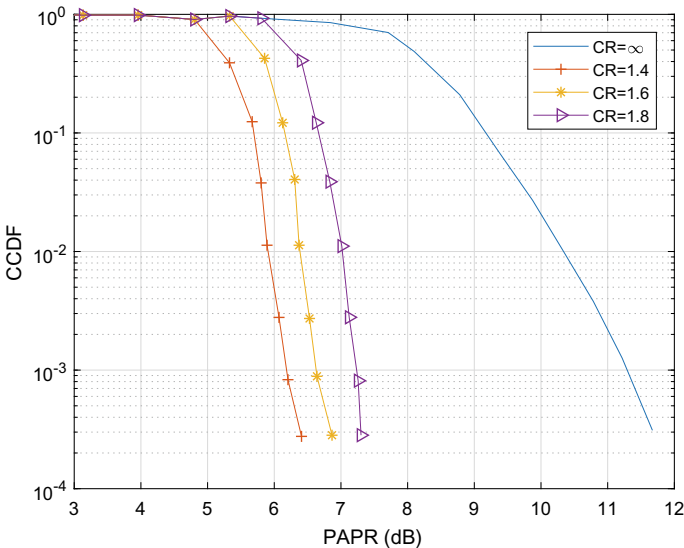


Fig. 3 CCDF under Different CR, with $J = 4$

when $P_{av} \leq P_{Out}$ the out-of-band radiation is filtered out. To facilitate the subsequent derivation, the power attenuation factor λ is defined to characterize the power ratio before and after low-pass filtering, i.e.,

$$\lambda = \frac{P_{av}}{P_{out}} = \frac{\sum_{k=0}^{N-1} E[|Y_k|^2]}{\sum_{k=0}^{MN-1} E[|Y_k|^2]}, \lambda \leq 1 \tag{10}$$

Then, the clipped and filtered signal is amplified by the power amplifier and transmitted.

4 Performance Analysis

In this section, the effect of clipping is analyzed on oversampled OFDM/OQAM signals. The analysis is of three levels; the first one is signal-to-distortion ratio (SDR); the second is signal-to-noise ratio (SNDR); the third is total signal-to-noise ratio (SNR). In the following analysis, it is assumed that the OFDM/OQAM signal is a complex gaussian stationary stochastic process. According to the Bussgang’s principle, the clipped OFDM/OQAM signal can be expressed as

$$y_n = \alpha x_n + d_n, n = 0, 1, \dots, (L + M - 1)JN - 1 \tag{11}$$

where N is the number of subcarriers, J is the oversampling multiple, d_n is the distortion introduced by clipping and d_n is un-correlated to x_n , i.e., $E[x_n d_n^*] = 0$. α is the attenuation factor introduced by the clipping and can be expressed as.

$$\alpha = \frac{E[x_n^* y_n]}{E[|x_n|^2]} = 1 - e^{-CR^2} - \frac{\sqrt{\pi CR}}{2} \text{erfc}(CR) \tag{12}$$

where $erfc$ is error function, i.e., $\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$ The received signal (11) is inverse Fourier transformed, and the signal at k th subcarrier is given by

$$y_k = \text{IFFT}(y_n) = \sum_{n=0}^{(M+L-1)N} y_n e^{-j2\pi(k/MN)n}$$

$$y_k = \alpha \sum_{n=0}^{(M+L-1)N} x_n e^{-j2\pi(k/MN)n} + \sum_{n=0}^{(M+L-1)N} d_n e^{-j2\pi(k/MN)n} \tag{13}$$

where $X_k = \sum_{n=0}^{(M+L-1)N} x_n e^{-j2\pi(k/MN)n}$ is the signal at k^{th} subcarrier and $D_k = \sum_{n=0}^{(M+L-1)N} d_n e^{-j2\pi(k/MN)n}$ is the distortion term at k^{th} subcarrier. As seen from distortion term D_k , D_k is the sum of $(L + M - 1)JN$ random variables. According to

the central limit theorem, when the number of subcarriers is large enough, D_k is approximated to Gaussian random variable.

In general, distortion terms located on different subcarriers can be considered statistically independent, i.e.,

$$E[D_k D_j] = \begin{cases} p_{in,k}, & j = k \\ 0, & j \neq k \end{cases} \quad (14)$$

where $p_{in,k} = E[|D_k|^2]$ is the average power of the distortion term.

The signal distortion ratio of the k^{th} subcarrier can be written as:

$$\text{SDR}_k = \frac{E[|\alpha X_k|^2]}{E[|D_k|^2]} = \alpha^2 \frac{p_{in,k}}{p_{d,k}} \quad (15)$$

where $P_{in,k} = E[|X_k|^2]$ is the average power of the k th subcarrier signal before clipping.

In clipped OFDM/OQAM system, the signal-to-noise ratio at the receiving end is defined as the ratio of average power of signal to the average power of white Gaussian noise, i.e.,

$$\text{SNR} = \frac{P_{av}}{P_{noise}} = \lambda \frac{P_{out}}{P_{noise}} \quad (16)$$

where λ is the power attenuation factor defined by (10), and P_{noise} is the average power of Gaussian noise. Since the signals of different subcarriers are independent of each other, and the distortion parts of different subcarriers are also independent of each other, the power attenuation factor λ can be re-represented using (15) as

$$\lambda = \frac{\alpha^2 P_{in}}{P_{out}} \left(1 + \sum_{k=0}^{N-1} \frac{p_{in,k}}{P_{in} \text{SDR}_k} \right) \quad (17)$$

If the signal powers of all subcarriers are the same, i.e., $p_{in,k} = P_{in}/N$, then (17) is given as

$$\lambda = \frac{\alpha^2 P_{in}}{P_{out}} \left(1 + \frac{1}{N} \sum_{k=0}^{N-1} \text{SDR}_k^{-1} \right) \quad (18)$$

Since the signal-to-noise ratio given by (15) does not consider the effects of signal distortion after clipping, this section will jointly consider the effect of Gaussian noise and signal distortion on the system, which is characterized by the signal-to-noise ratio. As discussed in Sect. 4.1, considering the k th subcarrier, the signal-to-noise ratio is defined as

$$\text{SNDR} = \frac{E[|\alpha X_k|^2]}{E[|D_k + W_k|^2]} = \frac{\alpha p_{in,k}}{p_{d,k} + \frac{P_{noise}}{N}} \quad (19)$$

where W is the Gaussian noise of the k^{th} subcarrier. Using (15) and (16), the inverse of equation (19) is given by

$$\text{SNDR}^{-1} = \text{SDR}_k^{-1} + \text{SNR}^{-1} \frac{\lambda P_{\text{out}}}{\alpha^2 N_{p_{\text{in},k}}} \quad (20)$$

The relation between signal noise distortion ratio (SNDR), signal distortion ratio (SDR) and signal-to-noise ratio (SNR) is given by (20). The SNR is used to calculate the theoretical capacity of OFDM/OQAM system after clipping and is an important parameter to measure system performance.

5 Simulation Results

In this section, the WiMAX system parameters are used as the simulation model to obtain the optimal clipping ratio and oversampling factor from the implementation. The simulation parameters are shown in Table 1.

The BER performance curves of WiMAX systems with different oversampling factors J and different CR ratios are shown in Figs. 3 and 4. It is observed from Fig. 4 that the oversampling can improve BER performance of the system when the clipping ratio is constant. For example, at 10^{-3} BER, the SNR of the oversampling system is improved by 7 dB when compared with Nyquist sampling. With the increase in oversampling factor, the SNR is greater than 17 dB. The simulation results in Fig. 5 show that reducing CR also increases the bit error rate, so CR cannot be reduced indefinitely. Considering the joint optimization of PAPR value and bit error rate performance, with $CR = 1.6$, the BER performance basically meets the requirements. If channel coding is adopted, the BER performance can be further improved. Choosing $CR = 1.8$ can further improve the BER performance of the system, but it comes at the cost of increase in PAPR by 0.5 dB (as shown in Fig. 3).

Table 1 Simulation parameters

Parameter	Value
Bandwidth	8 MHz
Subcarriers	256
Data subcarriers	11–100, 157–256
Subcarrier spacing	35.525kHz
Guard interval	1/4 of OFDM/OQAM symbol length
Sampling frequency	9.12MHz
Modulation	16QAM
Channel coding	None

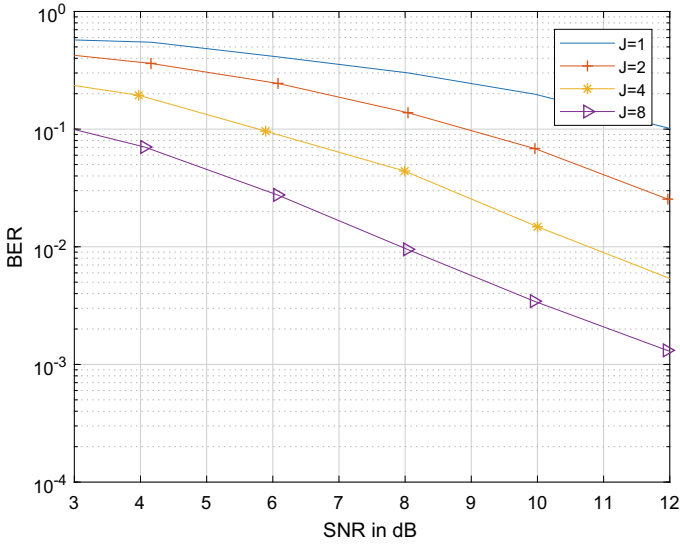


Fig. 4 BER curve with CR = 1.6, under different J values

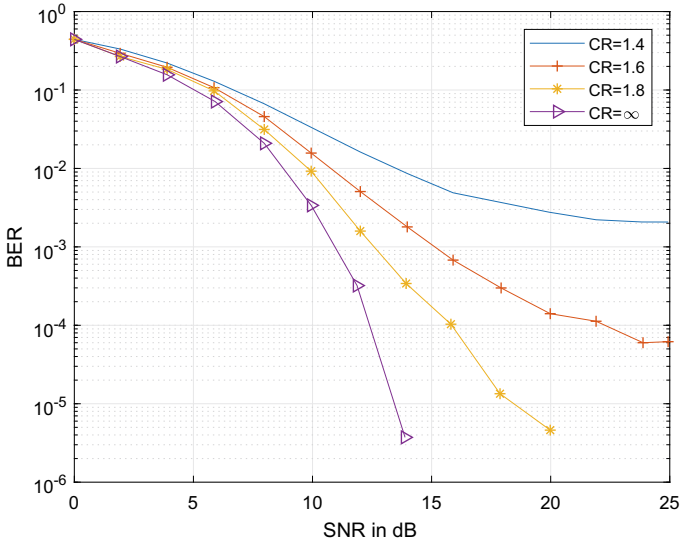


Fig. 5 BER curve with different CR, for $J = 4$

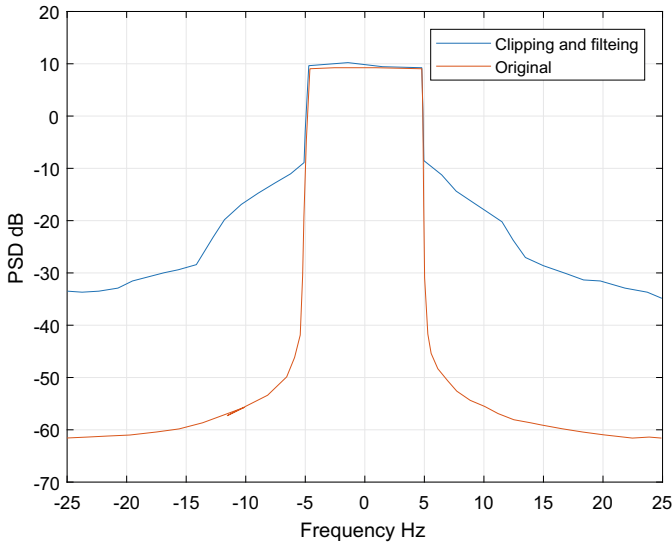


Fig. 6 PSD before and after clipping and filtering with $CR = 1.6$ and $J = 4$

The clipped OQAM signal introduces in-band and out-of-band noise, and in-band noise degrades the BER performance. This is analyzed in Sect. 4.1; out-of-band noise increases the out-of-band radiation, causing adjacent channel interference. In order to reduce out-of-band radiation, a low-pass filter is designed to filter clipped signal, so that its out-of-band radiation meets the requirements. Figure 6 shows the power spectral density of WiMAX using clipping method before and after filtering. From Fig. 6, it is observed that the in-band filtering has only a difference of 15 dB between in-band and out-of-band power spectral density. After filtering with the 85-order low-pass filter designed by the equal-wave method, the out-of-band radiation of the system is very effectively suppressed. The in-band and out-of-band power spectral density differ by more than 50 dB.

6 Conclusion

In summary, the clipping and filtering is a simple and effective method for reducing PAPR of OFDM/OQAM system. However, the clipping and filtering operation of the Nyquist-sampled OFDM/OQAM signal will give a better PAPR performance. In this paper, the method of oversampling and then clipping and filtering is used to effectively solve the problem of PAPR. This method has low implementation complexity and is suitable for any system using OFDM/OQAM transmission technique, such as IEEE 802.11a and IEEE 802.16 protocols. From the analysis and simulation, it is concluded that for WiMAX systems with OFDM/OQAM modulation, $4\times$ oversampling and clipping ratio $CR = 1.6$ is an optimal choice.

References

1. Vangelista, L., Laurenti, N.: Efficient implementations and alternative architectures for OFDM-OQAM systems. *IEEE Trans. Commun.* **49**(4), 664–675 (2001)
2. Kofidis, E., Katselis, D.: Preamble-based channel estimation in MIMO-OFDM/OQAM systems. In: 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE, New York (2011)
3. He, J., et al.: Experimental investigation of inter-core crosstalk tolerance of MIMO-OFDM/OQAM radio over multicore fiber system. *Opt. Express* **24**(12), 13418–13428 (2016)
4. Bader, F., Shaat, M.: Pilot pattern adaptation and channel estimation in MIMO WiMAX-like FBMC system. In: 2010 6th International Conference on Wireless and Mobile Communications. IEEE, New York (2010)
5. Kollem, S., Reddy, K.R.L., Rao, D.S.: Modified transform-based gamma correction for MRI tumor image denoising and segmentation by optimized Histon-based elephant herding algorithm. *Int. J. Imaging Syst. Technol.* **22429**, 1–23 (2020)
6. Kollem, S., Reddy, K.R.L., Rao, D.S.: Denoising and segmentation of MR images using fourth-order non-linear adaptive PDE and new convergent clustering. *Int. J. Imaging Syst. Technol.* **29**(3), 195–209 (2019)
7. Varghese, N., Chunkath, J., Sheeba, V.S.: Peak-to-average power ratio reduction in FBMC-OQAM system. In: 2014 Fourth International Conference on Advances in Computing and Communications. IEEE, New York (2014)
8. Sandeepkumar, V., Anuradha, S.: Adaptive clipping-based active constellation extension for PAPR reduction of OFDM/OQAM signals. *Circuits Syst. Signal Process.* **36**(7), 3034–3046 (2017)
9. Lu, S., Qu, D., He, Y.: Sliding window tone reservation technique for the peak-to-average power ratio reduction of FBMC-OQAM signals. *IEEE Wireless Commun. Lett.* **1**(4), 268–271 (2012)
10. Krishna Chaitanya Bulusu, S.S., et al.: Reduction of PAPR for FBMC-OQAM systems using dispersive SLM technique. In: 2014 11th International Symposium on Wireless Communications Systems (ISWCS). IEEE, New York (2014)
11. Skrzypczak, A., Javaudin, J.-P., Siohan, P.: Reduction of the peak-to-average power ratio for the OFDM/OQAM modulation. In: 2006 IEEE 63rd Vehicular Technology Conference, vol. 4. IEEE, New York (2006)
12. Kumar, V.S., Tarun Kumar, J.: NC-OFDM/ OQAM Based Cognitive Radio Network. LAP Lambert Academic Publishing, 978-620-0-45455-3, 22 Oct 2019
13. Chaitanya, P., Rajendra Prasad, Ch.: Performance evaluation of paper reduction. OFDM system using non linear companding transform. *Int. J. Eng. Comput. Sci.* **4**(09) (2015)
14. Ye, C., et al.: PAPR reduction of OQAM-OFDM signals using segmental PTS scheme with low complexity. *IEEE Trans. Broadcast.* **60**(1), 141–147 (2013)
15. Wang, H., et al.: Hybrid PAPR reduction scheme for FBMC/OQAM systems based on multi data block PTS and TR methods. *IEEE Access* **4**, 4761–4768 (2016)
16. Cheng, Y., et al.: Precoder and equalizer design for multi-user MIMO FBMC/OQAM with highly frequency selective channels. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, New York (2015)
17. Balti, E., Guizani, M.: Impact of non-linear high-power amplifiers on cooperative relaying systems. *IEEE Trans. Commun.* **65**(10), 4163–4175 (2017)

Implementation of a Smart Parking System for a University Campus



M. Saheer Jhugroo, M. Shahil Kataully, and Soulakshmee D. Nagowah

Abstract In recent times, the amount of vehicle users is spreading rapidly requiring additional parking spaces. A university campus can be considered as a small city and it usually undergoes the same parking problems encountered by smart cities. Searching and getting a parking space has always been a concern in university campus or even in smart cities. This paper aims to present a smart parking system based on Internet of things (IoT) to eliminate the stress, frustration and the dissatisfaction of users and to help them spot an available parking space to station his/her vehicle in a university campus. When there is no available parking space in a parking area, certain users have the tendency to roam around in search of a parking space, thus wasting time and fuel. Hence, one possible solution is to create a mobile application, which shows a real-time status of each parking space in the parking area and takes into consideration users' weekly time plan to allocate a parking space. Ultrasonic sensor and ESP8266 DevKitC microcontroller are used to implement the system. A real-time database records the status of each parking space. The system prototype is presented in the paper.

Keywords Smart city · Campus · Smart parking · Parking allocation · Algorithm

1 Introduction

The quantity of vehicles moving on the road is expanding day by day, while the quantity of free spaces is still the same. It might be problematic and costly to develop additional parking area for motor vehicles. In many smart cities, there is a big concern

M. S. Jhugroo · M. S. Kataully · S. D. Nagowah (✉)
Department of Software and Information Systems, FoICDT, University of Mauritius, Reduit,
Mauritius
e-mail: s.ghurbhurrun@uom.ac.mu

M. S. Jhugroo
e-mail: mohammad.jhugroo@uom.ac.mu

M. S. Kataully
e-mail: mohammad.kataully@uom.ac.mu

regarding parking and its availability, which leads to frustration of drivers and traffic congestion. Many endeavors are being effected in the field of Internet of things (IoT) to expand the dependability and profitability of urban foundations and several issues are being addressed by IoT like reducing congestion of traffic, restricted parking space and roadway security [1].

IoT is an innovative paradigm that helps in making smart cities, smart parking and other modern entities a reality. Thus, to solve this parking problem, a smart parking system will no doubt help the user in getting the necessary and required information about the availability status of the parking slot and most importantly about the quantity of free parking slots. Nowadays, smart parking systems are implemented with smarter technologies to be able to comfort parking users with various kinds of necessities [4, 5].

The most important aspect of IoT is to provide smart and seamless services to users. IoT makes use of a combination of technologies for sensor data acquisition, data exchange, machine learning, big data and cloud computing [2, 6]. Sensors are used to obtain data associated with a parking space and then inform the user if the parking is available or not [4]. Most of the work done are without human intervention, although people can interact with the IoT devices, for example, setting them up, giving access data or instructions. Smart parking solves parking problems by informing the drivers in advance about the availability of parking spaces at and around their intended destination [1]. There is a need for the system to find the path to an available parking spot smartly [2]. Smart parking mainly focuses on reducing the time taken to find parking lots and to avoid unnecessary travel through filled parking lots in a parking zone. Thus, the fuel consumption is reduced which, in turn, reduces carbon footprints in the atmosphere [2].

The work described in this paper addresses the modeling and development of a smart parking system for a university campus, which will help in better management of the parking spaces at the campus. The rest of the paper is structured as follows: Sect. 2 describes existing smart parking systems. Section 3 presents the proposed system and the algorithms adopted. Discussion about the system is given in Sect. 4. Finally, Sect. 5 concludes the paper and presents future works.

2 Related Work

Currently, there are several applications which provide users with information regarding free parking space and these applications presently give an interface either through a portable device or a desktop application. [1] designed a smart parking model based on IoT. The system comprises of an IoT device, which tracks and monitors whether each individual parking slot is available or not. The user can therefore check and identify the real-time accessibility of the parking slot using the provided mobile application, which is linked to the cloud and he can also book a parking slot. [8] devised an intelligent parking system making use of IoT that provides the parking slot availability in the corresponding parking field and helps the user to find

the nearest parking slot. The intelligent system is especially built to minimize time in finding a parking space and prevents futile traveling in a filled parking area. In turn, there is less release of carbon in the air and a decrease in gasoline consumption. [3] came up with an IoT parking system which keeps all its data on the cloud. This system comprises of an ultrasonic device that is employed to track and monitor whether each individual parking slot is vacant or not. Additionally, a mobile application is developed to help the owner of the car to easily find a closest parking space. The user also has the option to book a parking slot based on its availability and all the data are saved and updated in cloud database. As such, several parking problems are solved when implementing this system and thus parking users can have a better standard of living. [9] proposed an intelligent system which mainly focuses on implementing an extensive parking result to both car owners and parking area owner. Firstly, it checks the occupancy of a parking slot and then based on the dimensions of motor vehicles, it finds the parking space. The user also has the option to reserve a parking and as such, the system assigns the user a parking slot for a specific amount of time. The user can also navigate in the implemented mobile application to visualize the status and analytics of the parking space. [2] devised a system using IoT, which provides parking status information to help users in finding a proper parking slot. The user can identify several parking spaces and can select the available parking space on a web application. The parking ground in the system will only be open if there are available parking slots. The web application also helps the user in checking free slots before coming there. As a result, it saves time and reduces traffic congestion in front of the parking zone. [11] developed an algorithm to assign free parking spaces. It takes total travel time into consideration to assign a user a parking slot closer to destination. It delays the parking assignment, makes an accumulation and then performs an efficient slot assignment. A queue is used and parking requests are processed in a first-come-first-served basis. The user is assigned a parking slot, which matches best the user's destination. [12] presented a popular optimization algorithms and made a comparison. A hybrid genetic algorithm is proposed in the study. Hybridization techniques are used to solve optimization problems. The proposed system makes use of a proven crossover operation, which selects random characteristics of the parents and pass it to their children. Each parking zone is encoded as chromosomes. It expresses the index zone, which is assigned to the user. The feasibility of the generated children is checked by ensuring that each parking space capacity is not exceeded.

The described systems are compared in Table 1. They take into consideration parking identification and parking slot assignment. The systems make use of IoT devices namely sensors to capture data about the availability of the parking. A parking slot may be either occupied or unoccupied. The IoT devices allow sharing of information among the different applications through a mobile application or a web application. [11, 12] describe algorithms for parking slot assignment. [11] use time to calculate shortest distance and the system uses a queue. The system proposed by [9] uses RFID technology, which is costly and each vehicle will require a unique tag. Because the number of parking users in a campus is very high, it is very difficult to use this technology. Instead, other sensors like light-dependent resistor (LDR) or ultrasonic sensor can be used to implement the system. The different systems described

Table 1 Comparison of smart parking systems

Features	[1]	[8]	[3]	[9]	[2]	[11]	[12]
Application	Mobile	Web	Mobile	Mobile	Web	Mobile	Mobile
IoT sensors	✓	✓	✓	✓	✓	✓	✓
RFID	×	×	×	✓	×	×	×
Cloud	✓	×	✓	×	×	×	×
Parking identification	✓	✓	✓	✓	✓	✓	✓
Authenticate vehicles	✓	×	✓	✓	✓	×	×
Identify legitimate users	×	×	×	×	×	✓	✓
Check parking availability	✓	✓	✓	✓	✓	✓	✓
View parking status	✓	✓	✓	✓	✓	✓	×
Parking algorithm defined	×	×	×	×	×	✓	✓
Parking allocation	×	×	×	×	×	✓	✓

above are about parking in general. They do not take into consideration how the system will operate in the case of a campus. In the case of university campus, every student and staff are likely to have a timetable/time plan. The proposed parking system aims to use this information along with parking slots availability to efficiently allocate parking slots based on priorities. The existing systems focus mainly on finding the nearest available parking spaces and they do not cater for prioritization of users in scarce parking areas.

3 Proposed Smart Parking System

By taking into account the increased number of parking users at the University of Mauritius (UoM), a smart parking system is developed which makes use of IoT devices and a mobile application connected to a cloud system. The system first captures data and allocate parking slots to users depending on their role (staff, student, delivery guy) and based on the availability of parking slots. The user can submit data like his weekly timetable via a mobile application and can view details and availability of a parking space. Additionally, the user can visualize the parking zone and may request for parking zone status.

Figure 1 shows how the proposed system is modeled and implemented in UoM campus, whereby there are 8 parking zones in total; 2 for students and 6 for staff. In the proposed prototype, there will be 3 parking slots which are allocated to each of the faculty. The 3 parking slots belong to 1 parking zone which can be tested with several scenarios. At the end, a staff is assigned a parking slot nearest to his faculty based on a priority level. A staff having a lecture earlier and is assigned a parking space closest to his faculty compared to the one who has a lecture later during the day in case of scarcity. Each student is assigned a parking space in the student

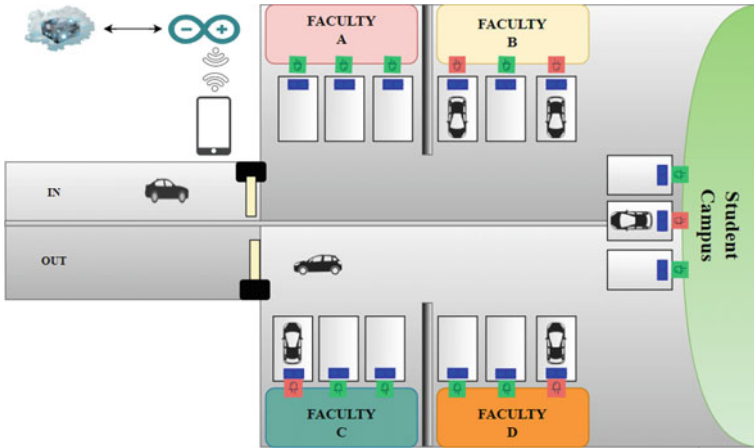


Fig. 1 Proposed smart parking system

campus parking zone on a first-come-first-served basis. There can also be staff who have reserved parking. Our prototype is modeled in a similar way just like a TreeSet model [13] ensuring that the closest parking slot will be given to the user all the time.

3.1 IoT Sensors and Data Capture

Ultrasonic sensors (HC-SR04) are positioned at individual parking slot to sense the existence of a vehicle in the parking slot as shown in Fig. 2. The chosen sensors have good distance accuracy varying from 3 cm to 3 m.

As soon as the ultrasonic sensor detects the existence of a vehicle, it automatically sends the data to the ESP8266 microcontroller. The ESP8266 microcontroller then sends the data about the status of the parking slot to the mobile application by the means of the Wi-Fi module that is already integrated in the microcontroller. Light emitting diodes (LED) are also connected to the microcontroller to visually display the parking status to user: Red light denotes that the parking slot is occupied and green light demonstrates that the parking slot is available. The system prototype is shown in Fig. 3 and the circuit diagram is shown in Fig. 4.

3.2 Data Storage

Factors that we took into consideration and that led the association of cloud and IoT are: the size of storage, the computational ability, communicational facilities and finally interoperability. In this system, Firebase cloud technology is used as it

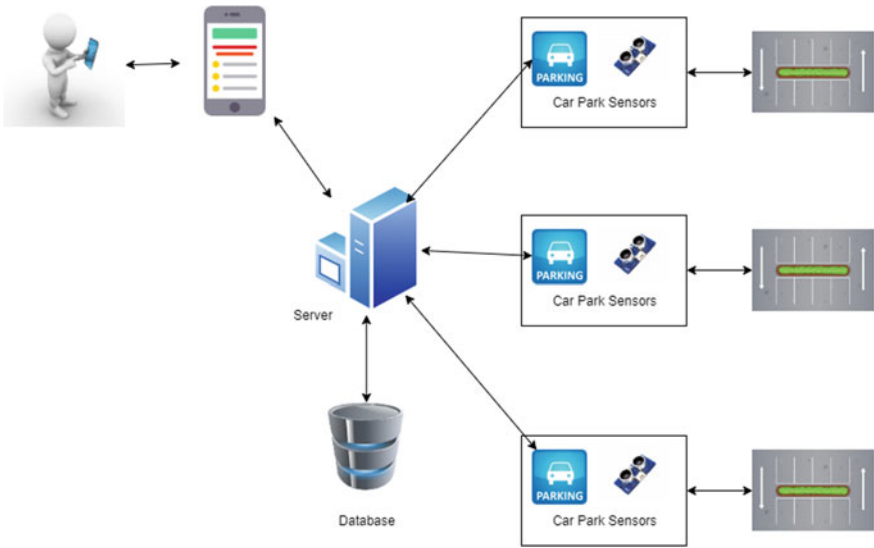


Fig. 2 Data capture



Fig. 3 System prototype

provides authentication services and storage facilities. Data captured by the sensors and user details are stored and updated in real time in the Firebase database. The database allows access from multiple devices and interfaces from mobile and Web applications. Cloud technology is suitable for use since the population of parking users is very high at University of Mauritius campus. The powerful authentication feature of Firebase is used to authenticate the users using the mobile application. Figure 5 shows the parking availability of every parking slots in Firebase. In this case, the first two parking slots are occupied while the third parking slot is available.

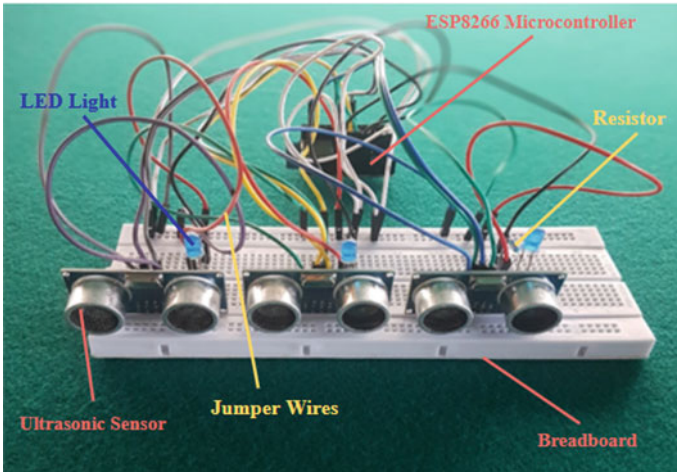


Fig. 4 Circuit diagram

Fig. 5 Parking availability database



3.3 Mobile Application

For the user interaction with the parking system, a mobile application is developed using android studio. The mobile application provides many facilities and services to the user. The user registers into the application the first time he uses it, and henceforth, he will just have to login to check out the status of the parking slot at the University of Mauritius. The user can even submit his weekly timetable as shown in Fig. 6 or he can enter the starting time and ending time himself. In the mobile application, the user is assigned the first parking slot if it is free, otherwise, it will keep on processing till it gets the second nearest parking space. In Fig. 7, the user is assigned slot 3 because slots 1 and 2 have already been assigned to another user. The occupied parking slots are colored red while the available parking slots are colored green and a user cannot book more than one parking slot. As mentioned earlier, the parking allocation is implemented using a TreeSet model [13] ensuring that the nearest parking slot is allocated to the user all the time.

Fig. 6 Timetable details

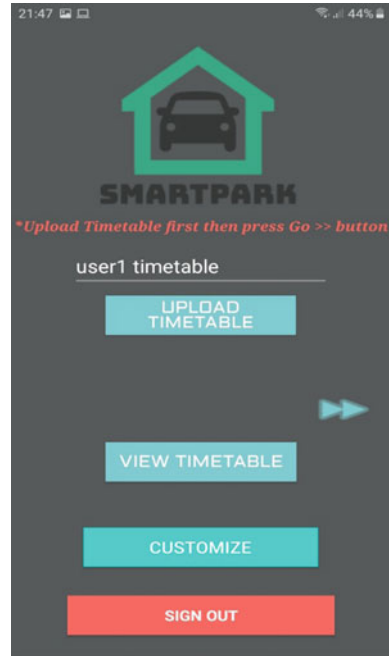


Fig. 7 Parking allocation



3.4 Parking System Algorithm

The flowchart in Fig. 8 illustrates the different steps of the smart parking system.

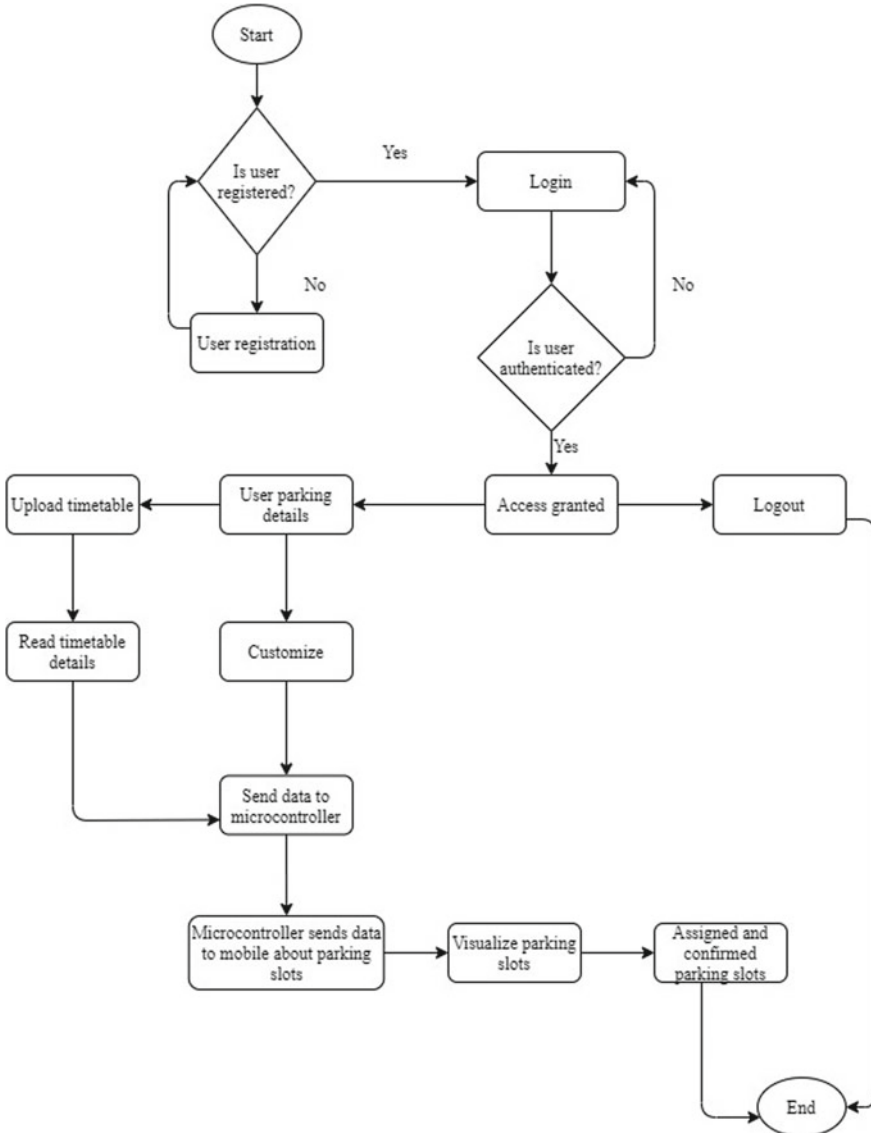


Fig. 8 Proposed smart parking system algorithm design

3.5 Prioritization of Parking Users

In the method proposed, all parking users are being considered with different levels of priorities when parking slots will be assigned. Some parking slots are reserved for high profile staff and delivery guy. University students are assigned parking slots from the two parking zones reserved for students. Students are not allowed to park into spaces reserved for staff. For staff, there is a total of six parking zones. Each of these parking zones is close to a certain faculty. Each staff must form part of only one faculty and will be assigned a parking slots from the parking zone assigned to this faculty. If it happens that the parking space belonging to his faculty is full, the staff is assigned a parking slot in another faculty, which is closest to his. Each user is assigned only one parking slot.

3.6 Optimization of Slot Assignment Time

Parking slot assignment time is optimized because the driver no longer has to drive in order to find an available parking slot. In this system, the user requests for a parking slot based on his destination and arrival time and are alerted about the available spot in advance. Moreover, the database is updated in real time.

3.7 Use of Hungarian Algorithm to Solve the Problem

The systems described in related works cater for parking slots assignment differently. In this paper, the Hungarian algorithm is used in the proposed system. It works by assigning jobs to workers and total cost must be minimized. Each worker must be assigned only one job and each job can be assigned to only one worker. Each assignment must ensure uniqueness in each row and column [10]. In our case, the workers are parking users while jobs are parking slots and the cost assigned to the jobs is the distance to the faculty. The algorithm finds minimum number of lines, that is, it finds the closest faculty to each parking users and assigns the parking slot to him. Therefore, using this method, staff are assigned a parking slot closest to his faculty.

4 Discussion

The evaluation of the present smart parking systems shows that the majority of these systems utilize sensors in order to detect the presence of a vehicle. However, they do not consider the prioritization of parking spaces for some users. Our proposed

system provides several services to the user such as visualizing the parking slot status and information via the mobile application. Additionally, our system provides the user with the nearest available parking slot in the campus closest to the faculty as described in Hungarian algorithm. Compared to the previously analyzed system, our proposed system allows the user to submit his weekly timetable based on which he can therefore book a parking space. Concerning the sensors, we have used ultrasonic sensor in order to detect the presence of a vehicle in a parking slot. LED lights are used to show the occupancy of the parking slot; red lights show that the parking slot is occupied, green lights for available parking slot and yellow lights for reserved parking slot. However, in case there is an obstacle in the parking slots, the ultrasonic sensor does not return correct values. We propose to place the ultrasonic sensor at a certain height taking into consideration the different heights of vehicles so that the sensor detects only vehicles in the parking slot.

5 Conclusion and Future Works

In this work, we have studied parking problem faced in a campus environment. The paper proposes a smart parking system that allocates the user a parking space closer to his faculty, which takes into consideration the weekly timetable of staff and students. Hungarian algorithm is used to allocate a user a parking space closer to his faculty while considering the user's weekly timetable. In the future, IR sensor could be used at the main gate to detect the presence of a vehicle, which enters and leaves the parking area. A DC servo motor could be placed at the main entrance that will help to open and close the gate and a LCD screen could be used to display the parking space information. In addition, we shall conduct simulation experiments to show how our algorithms perform on different user demands. At last, we plan to employ machine learning techniques to learn about parking patterns for different users.

References

1. Khanna, A., Anand, R.: IoT based smart parking system. In: 2016 International Conference on Internet of Things and Applications (IOTA), 266–270 (2016)
2. Rahman, S., Bhoumik, P.: IoT based smart parking system. *Int. J. Adv. Comput. Electron. Eng.* **4**(1), 11–16 (2019)
3. Deshpande, S.S., Gound R.S.: An approach for smart parking system based on cloud using IoT. In: Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering (2017)
4. Nagowah, S.D., Ben Sta, H., Gobin-Rahimbux, B.: An ontology for an IoT-enabled smart parking in a university campus. In: IEEE International Smart Cities Conference (ISC2), 474–479 (2019)
5. Rahman, M.A., Asyhari, A.T.: The emergence of internet of things (IoT): connecting anything, anywhere. *Computers* **8**(2), 40 (2019)

6. Rouse, M., Rosencrance, L., Rouse, M., Rouse, M.: What is internet of things (IoT)? Definition from WhatIs.com, IoT Agenda. [Online]. Available: <https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT>. Accessed 01 Nov 2019
7. Gupte, S., Younis, M.: Participatory-sensing-enabled efficient Parking Management in modern cities. In: 40th Conference on Local Computer Networks (LCN), 241–244 (2015)
8. SR, M.B.: Automatic smart parking system using internet of things (IOT). *Int. J. Sci. Res. Publicat.* **5**(12), 629–632 (2015)
9. Cynthia, J., Bharathi, P., Gopinath, P.: IOT based smart parking management system. *Int. J. Recent Technol. Eng. (IJRTE)* **7**(4) (2018)
10. “Hungarian algorithm,” Wikipedia, 27-Aug-2019. [Online]. Available: https://en.wikipedia.org/wiki/Hungarian_algorithm. Accessed 01 Nov 2019
11. Abidi, S., Krichen, S., Alba, E., Molina, J.: A new heuristic for solving the parking assignment problem. *Proc. Comput. Sci.* **60**, 312–321 (2015)
12. Hakeem, A., Gehani, N., Ding, X., Curtmola, R., Borcea, C.: On-The-Fly Curbside Parking Assignment. *MobiCASE* **16**, 1–10 (2016)
13. GeeksforGeeks. Treerset In Java—Geeksforgeeks. [online] Available at: <https://www.geeksforgeeks.org/treeset-in-java-with-examples/>. Accessed 11 March 2020 (2020)

High-Resolution Wind Speed Mapping for the Island of Mauritius Using Mesoscale Modelling



Tyagaraja S. M. Cunden and Michel R. Lollchund

Abstract In the context of developing wind energy for the generation of electricity in many countries, proper knowledge and understanding of the availability of wind resources over these countries are of great importance. This paper demonstrates how the Weather Research and Forecasting (WRF) mesoscale model can be used to simulate the wind flow patterns over a complex terrain such as the island of Mauritius at a horizontal spatial grid resolution of $1 \text{ km} \times 1 \text{ km}$ in view of generating reliable wind speed data. The simulation period spans from January 2015 to December 2017. The WRF results are validated against ground truth data for various climatic conditions such as calm days and the occurrences of two extreme events, namely a cyclone and an anticyclone. Statistical metrics such as Bias, RMSE and Pearson correlation are used for comparing both the datasets. Excellent agreement with r -values above 0.6 is obtained for all the cases considered. Finally, high-resolution ($1 \times 1 \text{ km}$) mean monthly, seasonal, and yearly wind maps are generated for the island of Mauritius.

Keywords Wind speed mapping · Mesoscale modelling · Complex topography · Weather research forecast

1 Introduction

The use of wind energy for the generation of electricity is considered as one of the promising avenues for the promotion of low carbon technologies in many Small Island Developing States (SIDS). In this context, a proper knowledge and understanding of the availability of wind resources over these states is of paramount importance. According to Seshaiyah and Indhumathy [1], the knowledge of wind energy distribution at a particular site is an important factor for the design of wind farms.

T. S. M. Cunden (✉)

Department of Electromechanical Engineering and Automation, Université Des Mascareignes, Rose Hill, Mauritius

e-mail: tcunden@udm.ac.mu

M. R. Lollchund

Department of Physics, Faculty of Science, University of Mauritius, Réduit, Mauritius

Hence, studies of the wind resources are conducted by setting up a campaign of wind measurements over the whole study area to collect data over a long period of time (normally decades). The collected data are then used to perform statistical analyses to produce estimates of the wind resource. Such a study was conducted in 1986 by the American firm, Batelle Pacific Northwest laboratory for the island of Mauritius [2]. However, due to a restricted number of measuring stations (11 stations over an area of 1800 Km²), the results of the study were too coarse to produce high-resolution wind resource maps for Mauritius. Unfortunately, conducting a wind resource assessment using more measuring stations can prove to be very costly and time consuming. In a recent study, Dhunny et al. [3] have attempted to study the wind flow patterns over the island using a microscale model. However, the model used has certain limitations. Firstly, it is a steady-state model; therefore, it cannot generate time series data which is important in understanding the variable and turbulent nature of wind, and secondly, the model does not consider the thermal stratification of the planetary boundary layer which limits its capabilities in predicting the wind speed at different heights.

In this paper, we propose the use of a Numerical Weather Prediction (NWP) model, the Weather Research and Forecasting (WRF) modelling system to develop high-resolution wind maps for the island of Mauritius. This model has been used in many researches related to wind studies. Some examples are highlighted in Cunden et al. [4]. We aim at showing that the WRF model can effectively simulate the local weather over Mauritius for various climatic conditions such as calm days and occurrence of two extreme events, namely a cyclone and an anticyclone. The main objective is to use the model to develop high-resolution wind resource maps for Mauritius that can be used by decision-makers and wind energy planners. The authors believe that the methodology presented can also be used as a guideline for wind resource analysis in other countries, especially SIDS.

The paper is structured as follows: Sect. 2 describes the study region followed by the setting up and validation of the WRF model in Sect. 3. Section 4 details the results of the simulation and the mapping of wind resources over the study region. The main conclusions of the study are presented in Sect. 5.

2 The Region of Study

The study region is the island of Mauritius and its surrounding islets (see Fig. 1). It is an island of volcanic origin situated within the tropical belt at latitude 20°17' South and longitude 57°50' East. The main land of Mauritius spans about 61 km from North to South and about 45 km from East to West and covers an area of approximately 1865 km². The topography of the island is composed of a central plateau at more than 400 m above sea level, which is surrounded by coastal plains in the northern part of the island and a complex terrain in the south and south-western parts of the island. The central plateau is surrounded by mountain ranges which seem to be the remains of a caldera that was at the origin of the volcanic activity which led to the creation of the island. As shown in Fig. 1, the main land of Mauritius is surrounded

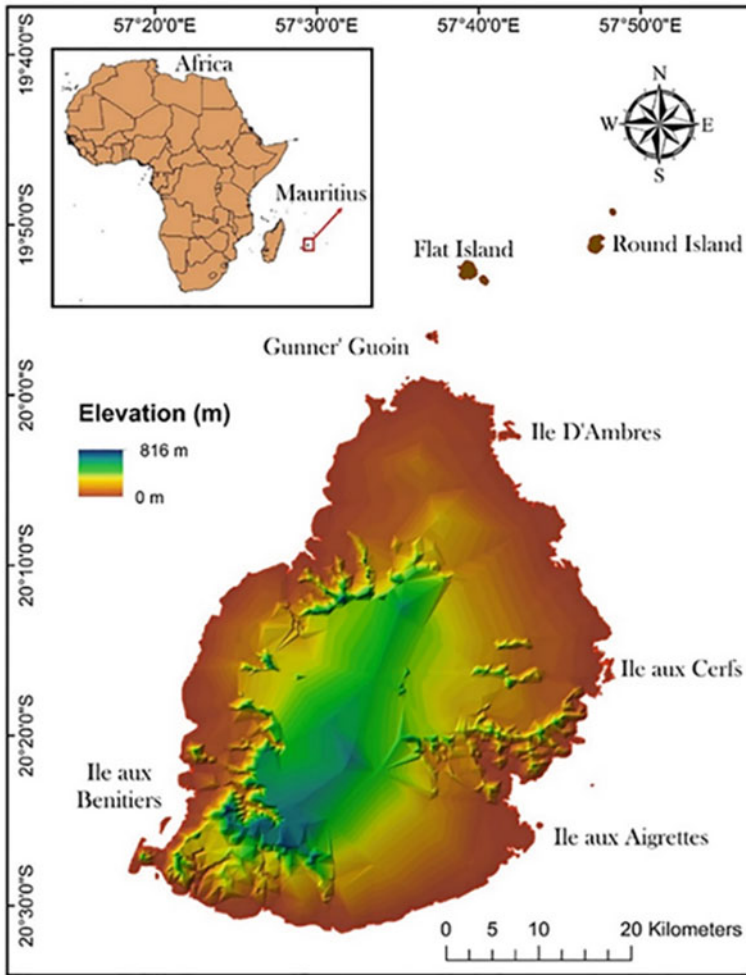


Fig. 1 Topographic map of Mauritius and surrounding islets (Illustration by the authors)

by a few islets which are unpopulated. The areas of the islets are in the range of 0.25–2.6 km².

The climate of Mauritius is characterized by two seasons: summer (in the period November–April) and winter (in the period May–October). As the island is at the edge of the tropic of Capricorn, it is subject to the South–East Trade Winds all year round. These trade winds have higher speeds in winter. However, westerlies are also recorded during the period June–October [5]. Studies of the upper air winds indicate that the trade winds at 900 hPa are fairly strong on the average throughout the year with maximum speeds in the three months August–October [5]. The depth of the trade winds varies seasonally and is dependent on the synoptic time scale. However,

in the upper air at 700 hPa, during winter season, there are the westerlies, the same situation arises in summer whence the Inter Tropical Convergence Zone (ITCZ) lies to the north of Mauritius.

The dominant synoptic system, which causes trade winds to flow over the region, originates from the St Helena region in the Atlantic Ocean, transiting off the south-eastern coast of the African continent into the Indian Ocean. These anticyclones establish over the southern Indian Ocean as very large, persistent, zonally elongated high-pressure zones (known as the Mascarene highs). Fine weather is usually associated with such systems with widespread stratocumulus. Anticyclones have a very stable layer of cold interface with the low troposphere from the potentially warm (with low humidity) air aloft. The latter is the result of persistent downward motion. The stable layer in the low troposphere of the anticyclone is called the subsidence layer [6].

3 Data and Methodology

This section describes the setting up of the domains for the WRF model as well as the configuration and the initial conditions used. The validation of the model is also discussed.

3.1 *Brief Description of the WRF Model*

WRF is an open-source numerical weather prediction (NWP) and atmospheric simulation system designed for multiple applications [7]. It solves the physical and chemical laws which govern the different processes in the atmosphere (see Fig. 2). The physical parameterization of the WRF model is a very crucial component of the modelling and influences the dynamics of the state of the system being modelled. In a previous work [4], the authors have conducted a sensitivity analysis for choosing the best combinations of the physics options for the region of Mauritius. These results are used for parameterising the WRF model in the present study.

3.2 *Model Configuration and Initialization*

The first step to configure WRF for a region is to define the domains sizes for the WRF pre-processing system (WPS). In the present work, the downscaling is done in four steps, starting with a coarse domain, d01 (33×33 grid points) with a grid resolution of 27 by 27 km. Then, using a downscaling grid ratio of 3:1, the grid resolution is refined to 9 by 9 km in domain d02 (37×37 grid points), 3 by 3 km in domain d03 (52×52 grid points) and ultimately 1 by 1 km in domain d04 (61×91

grid points). The domains are represented in Fig. 3. It is to be noted that the nested domains (d02, d03 and d04) are two-way in nature.

The WRF model also consists of a vertical grid following a pressure levels coordinate system. 41 pressure levels are used to define the vertical grid and it is ensured that 12 levels fall within the first km of the vertical grid. The first level is situated at 14 m above ground level (agl). The WRF model also requires meteorological datasets as inputs for the initial boundary conditions. The NCEP FNL dataset, ds083.2, is used in this study. For further details on the WRF model and the dataset, the interested reader is referred to the work of Cunden et al. [4].

As suggested by Hahman et al. [8], simulations for consecutive eleven-day periods were run in series for the whole period starting on January 1, 2015–December 31, 2017. The last day of an eleven-day period and the first day of the next consecutive eleven-day period are overlapping and is kept as spin up period for the model. Hence, the data simulated for the first day (24 h) of each eleven-day run is not considered in the study. Runs are started (cold start) at 00:00 UTC at the beginning of every eleven-day period and are integrated for eleven days. The model is configured to generate an output data file every one hour in the NetCDF format. For the purpose of this study, only data files for the innermost domain d04 are used.

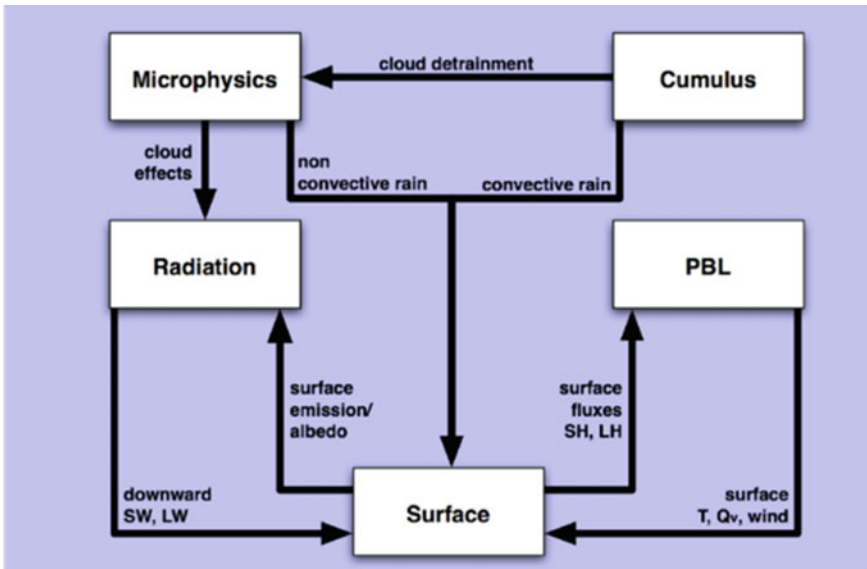


Fig. 2 Physics processes in WRF (Adapted from Dudhia 2010)

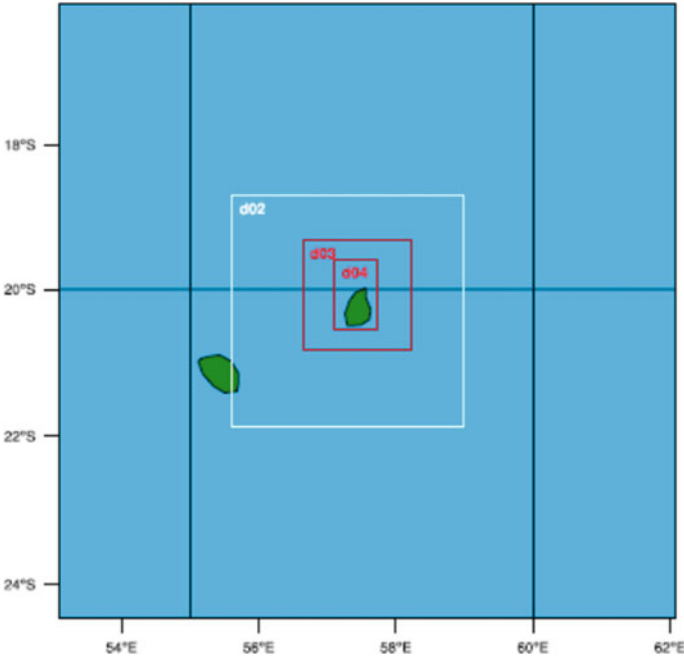


Fig. 3 Definition of the simulation domains in WPS

3.2.1 Model Validation

The simulated data are extracted for the location of Plaine des Roches (20.13° S, 57.73° E) where a commercial wind farm of capacity 9.5 MW has been installed in 2016. This region is chosen because of the availability of reliable wind data. The simulated data for the period of January 2015 to December 2017 are compared with the available measured data and statistical analyses, following the work of Cheng et al. [9], are conducted for validation purposes. The statistical metrics used for the analyses, namely the bias, the root mean square error (RMSE), and the Pearson’s product-moment correlation coefficient, r , are, respectively, defined in Eqs. (1)–(3) [4].

$$\text{Bias} = \frac{1}{N} \sum_{i=1}^N (M_i - O_i) \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (M_i - O_i)^2} \tag{2}$$

$$r = \frac{\sum_{i=1}^N (M_i - \bar{M})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^N (M_i - \bar{M})^2} \sqrt{\sum_{i=1}^N (O_i - \bar{O})^2}} \quad (3)$$

where M_i is the simulated data, O_i is the measured data, \bar{M} and \bar{O} are the average of the simulated and measured data, respectively, and N is the number of data points available for comparison. The results obtained in this validation processed are presented in the next section.

4 Results and Discussions

The island of Mauritius is subjected to various climatic conditions throughout the year. These include cyclones and anticyclones. It is, therefore, important to ensure that the model behaves well in the different climatic conditions.

4.1 Validation of the Model for Different Climatic Conditions

As mentioned in the previous section, simulated results are compared with ground truth data to validate the applicability of the WRF model for the various climatic conditions that prevail over Mauritius. Initially, simulations were run for the period February 2016–September 2017. Monthly mean data were the extracted for Plaines des Roches (20.13° S, 57.73° E, 40 m altitude) and the statistical metrics applied to compare the datasets. Table 1 presents the results obtained for each month during the period of study. A positive bias is obtained throughout, indicating that the model tends to overestimate the wind data. It is also observed that r -values are in the range of 0.6–0.9, which is an indication of strong correlation between the modelled and observed data. The model also shows a good accuracy given that all the values of the RMSE do not exceed 11%.

The WRF model is now applied to study wind flow in clear days and during some particular extreme climatic events which occurred over the island. The sections below discuss about these events and present the results obtained from WRF.

4.1.1 Mascarene Highs (October 2009)

It is not unusual to have a strong anticyclone influencing the local weather during the month of October, even though October is the period when there is a transition towards summer from winter. One such episode occurred between October 21–25, 2009, and is shown in Fig. 4.

Table 1 Comparison of WRF simulated data and measured data at Plaine des Roches at 40 m

Month	Mean wind speed (m/s)	Bias (m/s)	RMSE (m/s)	r-value
Feb-16	6.79	1.69	0.75	0.68
Mar-16	6.5	1.28	0.5	0.81
Apr-16	7.71	3.78	0.82	0.81
May-16	7.49	2.75	0.84	0.58
Jun-16	8.89	1.74	0.91	0.85
Jul-16	9.47	2.80	0.53	0.57
Aug-16	9.08	2.26	0.24	0.75
Sep-16	4.71	0.73	0.49	0.65
Oct-16	4.73	2.25	0.52	0.72
Nov-16	4.91	3.14	0.39	0.54
Dec-16	5.26	2.01	0.27	0.75
Jan-17	6.13	3.05	0.45	0.61
Feb-17	5.10	2.41	0.57	0.70
Mar-17	5.94	2.96	0.23	0.64
Apr-17	6.2	2.61	0.67	0.78
May-17	6.28	1.43	0.64	0.89
Jun-17	4.43	1.94	0.25	0.66
Jul-17	5.58	2.40	0.61	0.71
Aug-17	7.47	1.37	0.78	0.79
Sep-17	4.31	0.63	0.49	0.88

Figures 5, 6 and 7 show the WRF model results for Mascarene Highs of October 2009 for mean wind speed at three different heights (10, 60 and 100 m.a.g.l respectively). It can be noticed that gusts of the order of 20 m/s were produced in some regions especially at heights 60 m and above.

4.1.2 Tropical Cyclone (Giovanna, February 2012)

On February 13, 2012, killer Tropical Cyclone (TC) Giovanna, Fig. 8a, located some 465 km east of Antananarivo, caused extensive damage in Madagascar. Formed in the south west Indian ocean region, this storm reached Hurricane Category 4 on February 13, only to significantly lose its strength and quickly dissipate on February 20. The complete track of TC Giovanna is shown in Fig. 8b. The Event Time Line details for this tropical cyclone, with emphasis on Madagascar, can be found in [10]. The US Navy's Joint Typhoon Warning Center (JTWC) reported that the cyclone had maximum sustained winds of 230 km/h and gusts up to 280 km/h.

Figure 9 shows the WRF simulation for Tropical Cyclone Giovanna for February 2012 for mean wind speed at 60 m.a.g.l. These agree with the corresponding satellite

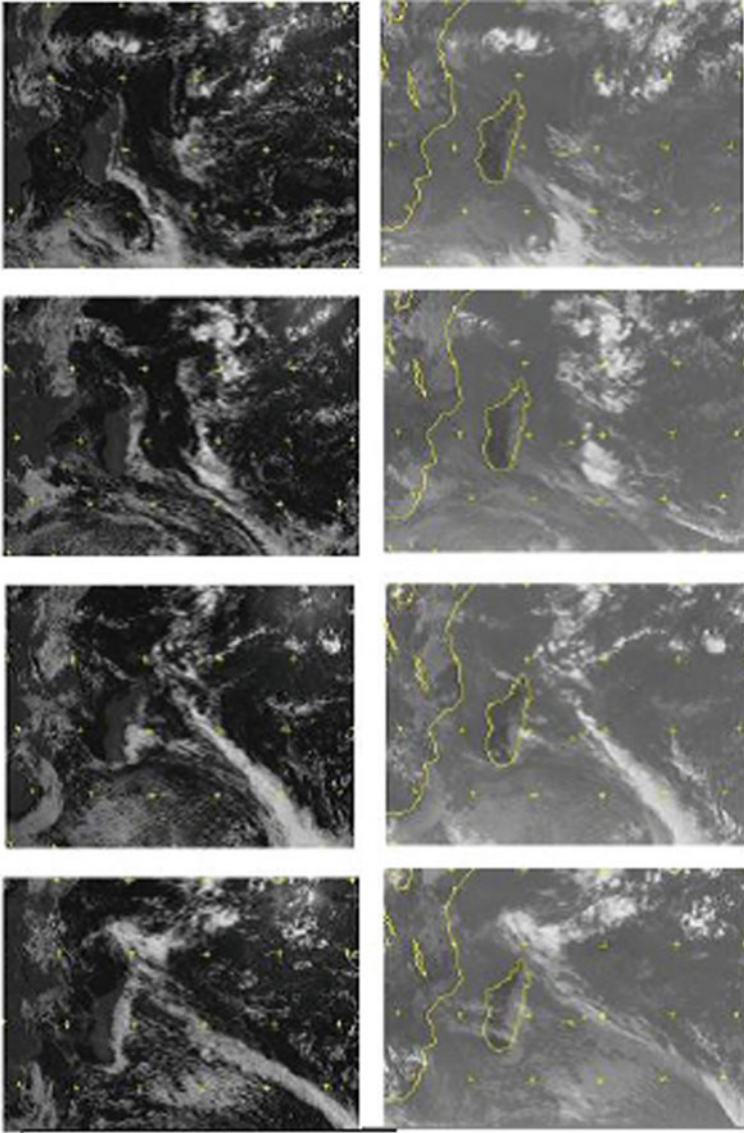


Fig. 4 Visible and infra-red satellite pictures for period 21st to October 21–24, 2009 (from top to bottom)

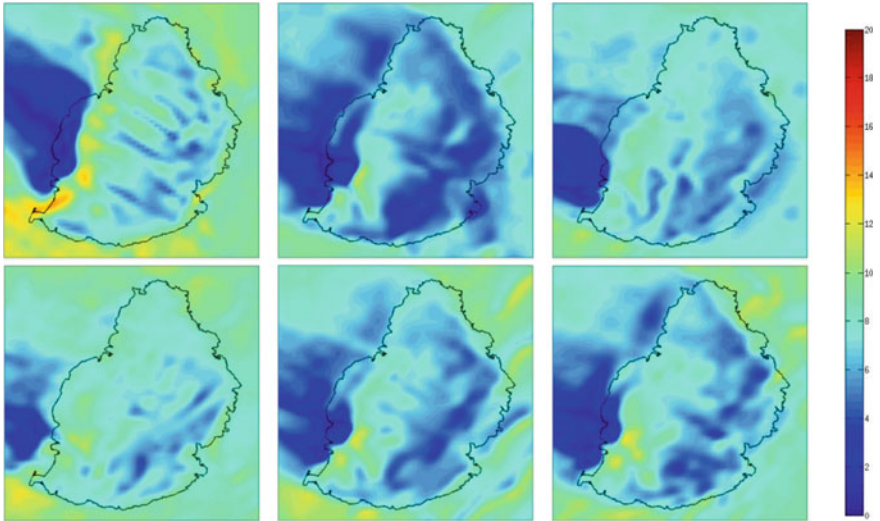


Fig. 5 WRF simulations for Mascarene Highs of October 2009: Mean wind speed (in m/s) at height 10 m.a.g.l. Top—From Left Right: 21st at Midnight; 21st at 6.00 AM; 21st at 12.00 PM. Bottom—From Left to Right: 21st at 6.00 PM; 22nd at Midnight; 22nd at 6.00 AM

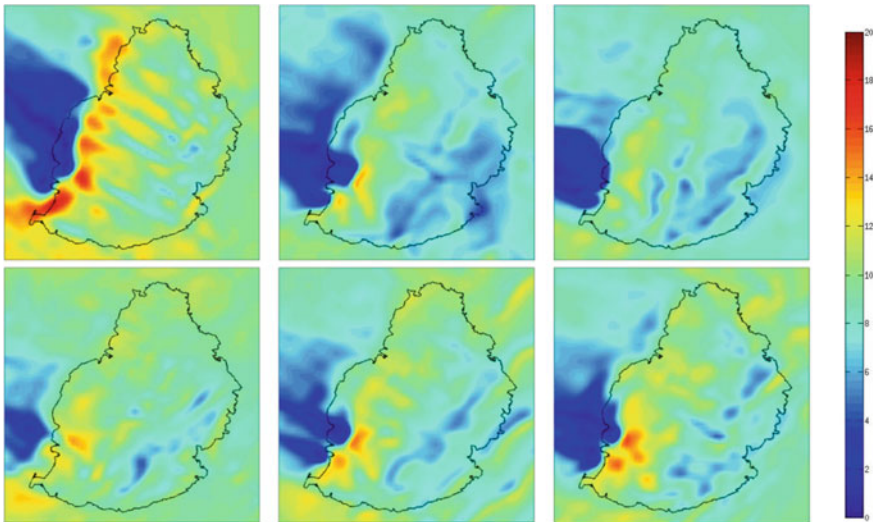


Fig. 6 WRF models for Mascarene Highs of October 2009: Mean wind speed (in m/s) at height 60 m.a.g.l. Top—From Left Right: 21st at Midnight; 21st at 6.00 AM; 21st at 12.00 PM. Bottom—From Left to Right: 21st at 6.00 PM; 22nd at Midnight; 22nd at 6.00 AM

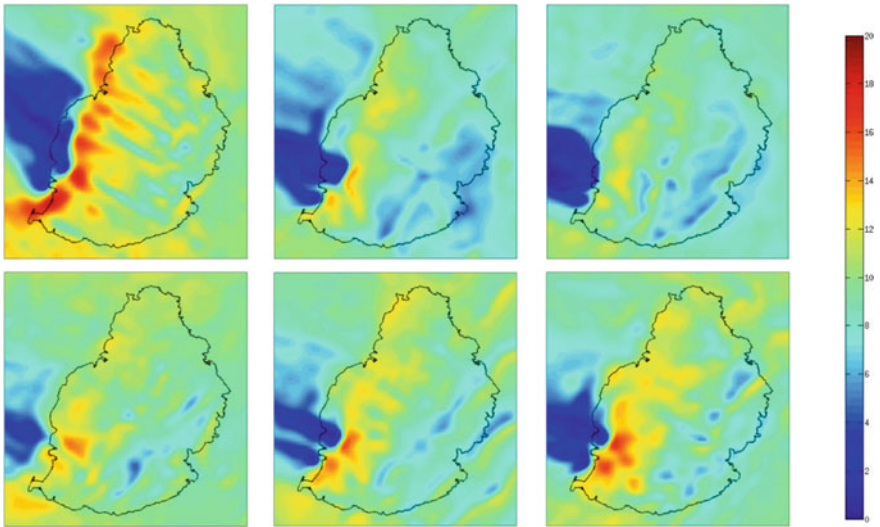


Fig. 7 WRF models for Mascarene Highs of October 2009: Mean wind speed (in m/s) at height 100 m.a.g.l. Top—From Left Right: 21st at Midnight; 21st at 6.00 AM; 21st at 12.00 PM. Bottom—From Left to Right: 21st at 6.00 PM; 22nd at Midnight; 22nd at 6.00 AM

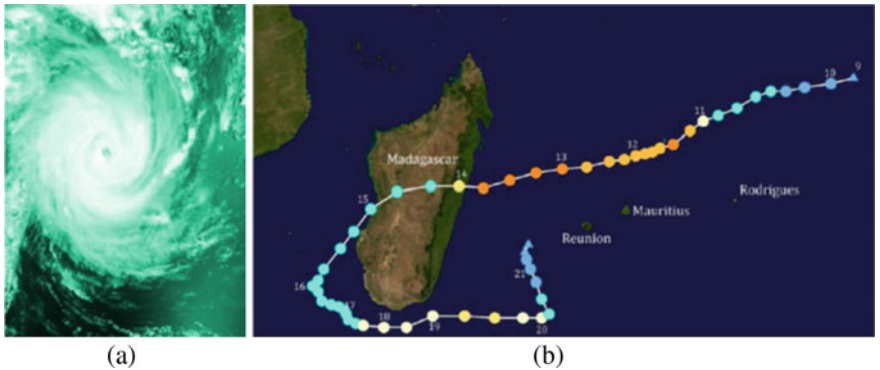


Fig. 8 **a** Satellite Picture of Tropical Cyclone (TC) Giovanna on February 13, 2012. **b** Trajectory of TC Giovanna. Dates shown are 00 UTC. *Source* https://en.wikipedia.org/wiki/Cyclone_Giovanna#/media/File:Giovanna_2012_track.png

images which can be consulted on Web site of Zoom Earth [11]. It can be noticed that the TC Giovanna produced winds of the order of 22 m/s although the center of the cyclone was more than 200 km from the shores of Mauritius. There are historical cyclones such as, Carol, 1960 with gust attaining the speed of 71 m/s, Gervaise, 1975 with gusts of 78 m/s, Claudette, 1979 with gusts of 61 m/s and Dina, 2002 passing 50 km north of Mauritius and yet producing gusts of the order of 63 m/s (MMS,

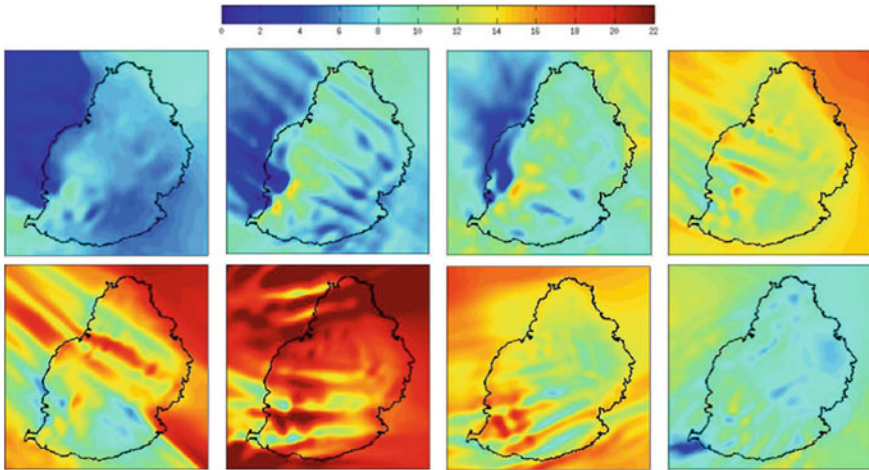


Fig. 9 WRF models for TC Giovanna (February, 2012): Mean wind speed (in m/s) at height 60 m.a.g.l. Top—From Left Right: 10th at Midnight; 10th at 6.00 AM; 10th at 12.00 PM; 10th at 6:00 PM Bottom—From Left to Right: 11th at Midnight; 11th at 6.00 AM; 11th at 12.00 PM; 11th at 6:00 PM

2019), just to name a few. This is a very important factor to be considered in the design of wind turbines for the island of Mauritius.

Three locations were selected based on availability of ground truth data: Plaisance airport (20.4° S, 57.7° E), in the south of the Mauritius; Plaine des Roches wind farm (20.13° S, 57.73° E), situated in the north east of Mauritius; and Round Island, an islet having an area of 1.69 km² located 22.5 km from the northern coast of Mauritius (see Fig. 1). The data were compared at different heights depending on what height measured data were collected at these locations. This allowed the model to be verified at different heights also. The *r*-values are presented in Table 2. It is observed that the *r*-value is in the range of 0.6 to 0.9 which indicates strong correlation.

Table 2 *r*-values between WRF simulations and measured data for the extreme events

	Plaisance at 10 m	Round Island at 10 m	Plaine des Roches at 40 m
Clear days	0.62	0.61	0.64
Anti-cyclone	0.74	0.81	0.91
Cyclone	0.71	0.92	0.92

4.2 Mapping of High-Resolution Wind Resource Maps

WRF simulations were finally performed for the whole three-year period of 2015, 2016, and 2017 to generate high-resolution wind resource maps for Mauritius in terms of mean monthly, seasonal, and yearly wind speed at the height of 60 m.a.g.l which is the hub height of most large-scale commercial wind turbines. The results are displayed in Figs. 11, 12 and 13, respectively.

It can be observed that throughout the year, the region which has the highest mean wind speed is located in the south-west of Mauritius. This region has the highest elevations (see Fig. 1 which are the cause of acceleration of the wind. The mean wind speed in this region varies between 9 and 13 m/s. It is also worth noting that the western coast of the island is always having the lowest wind speed (in the range of 2–7.5 m/s). This can be explained by the following: Mauritius being in the South West Indian Ocean region, is under the influence of the South–East Trade Winds throughout the year. As the wind blows over the island (which stands as an obstacle in the middle of the ocean), a wake region is created in the western part of the island and is the cause for low wind speeds in that region. The month of October has the highest mean wind speed while the lowest mean wind speed is in the month of February. The mean seasonal wind maps Fig. 12 show that mean wind speed is generally higher in winter than in summer. The mean wind speed in winter varies from 5 to 13 m/s whereas in summer, it varies from 4 to 10 m/s. The mean yearly wind map Fig. 13 also indicates similar trends with highest wind speeds in the south-west region.

5 Conclusions

The application of the WRF model for high-resolution wind speed modelling over the island of Mauritius has been presented. The model has been applied to simulate the wind speeds over Mauritius for calm days and during the occurrences of two extreme events, namely an anticyclone, Mascarene High in 2009 and Tropical Cyclone Giovanna in 2012. Data from the model has been validated with ground truth measurements using statistical metrics such as Bias, RMSE and Pearson correlation. Excellent agreement is obtained with r -values above 0.6 for all cases considered. It has been noticed that in both the anticyclonic and cyclonic events, gusts of the order of 20 m/s were observed at 60 m.a.g.l and above, and this must be considered in the design of wind turbines for the island of Mauritius. The high-resolution mean monthly, seasonal, and yearly wind maps for Mauritius Island were obtained from the WRF data, and from the maps, it was possible to identify the areas having high potentials for wind farming. These high-resolution maps will be used to produce the first mesoscale wind atlas for the island of Mauritius which will be useful for decision makers and wind energy planners. The methodology presented in this paper can be used in other countries, especially SIDS, to study multi-level wind flow patterns.

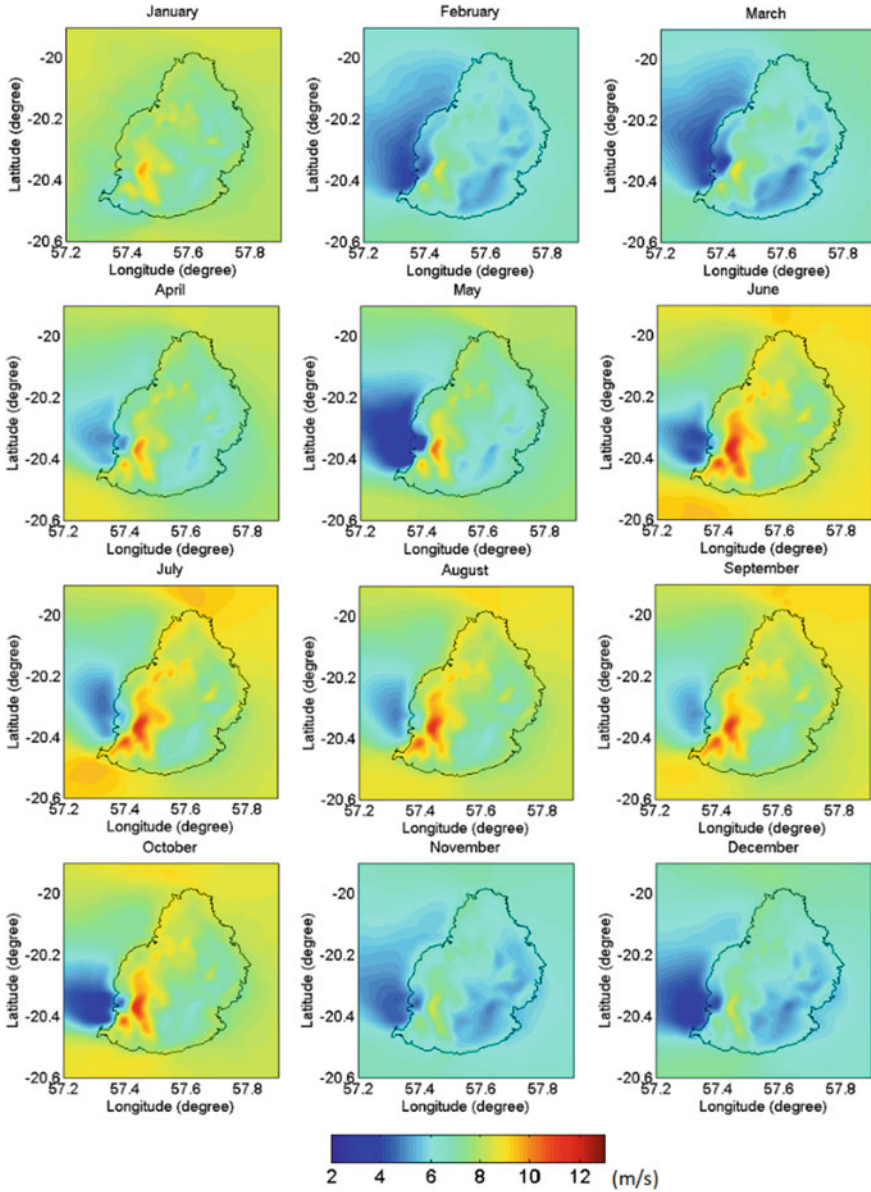


Fig. 11 Mean monthly wind maps (in m/s) at 60 m.a.g.l. (Jan 2015 to Dec 2017) (North is upwards)

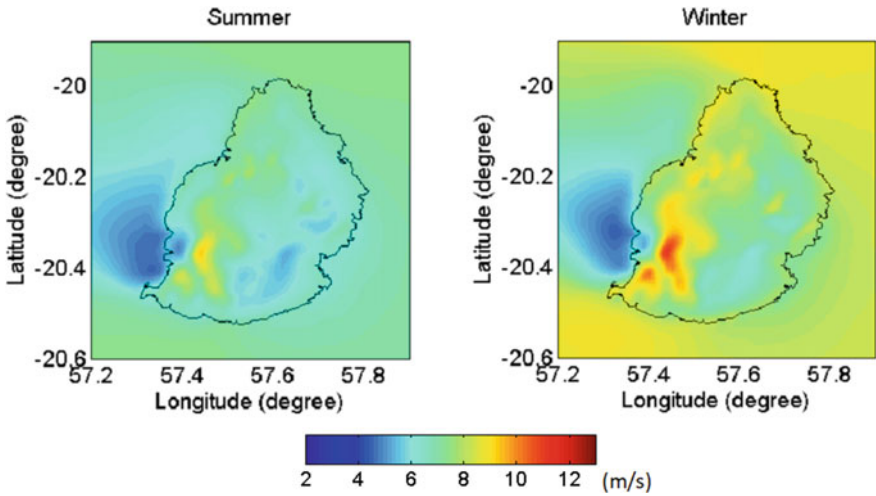
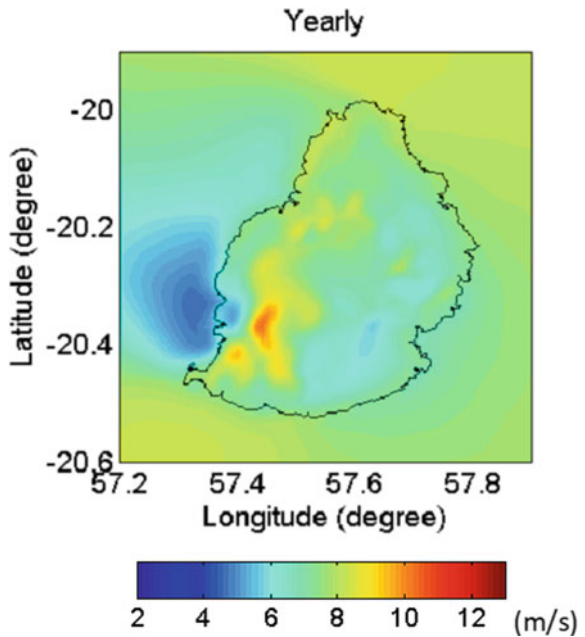


Fig. 12 Mean seasonal wind maps for the period of Jan 2015 to Dec 2017 at 60 m.a.g.l. (North is upwards)

Fig. 13 Mean yearly wind map (in m/s) at 60 m.a.g.l. (North is upwards)



Acknowledgements The authors wish to thank the Mauritius Meteorological Services (MMS) and Eole Wind farm for providing ground truth data. Special thanks are dedicated to Professor SDDV Rughooputh and Dr Z Dhunny for their fruitful discussions throughout this research.

References

1. Seshaiyah, C.V., Indhumathy, D.: A bimodal Weibull distribution-capacity factor for different heights at sulur. *Wind Struct* **28**(1), 63–70 (2019)
2. Elliot, D.L., Aspliden, C.I., Barnard, J.C., Severtsen, R.H.: *Wind Energy Resource for Mauritius*. Battelle Pacific Northwest Laboratories, Richland, Washington, USA (1986)
3. Dhunny, A.Z., Lollchund, M.R., Rughooputh, S.D.D.V.: Wind energy evaluation for a highly complex terrain using computational fluid dynamics (CFD). *Renew. Energy* **101**, 1–9 (2016)
4. Cunden, T.M., Dhunny, A.Z., Lollchund, M.R., Rughooputh, S.D.D.V.: Sensitivity analysis of WRF model for wind modelling over a complex topography under extreme weather conditions. In: *Proceedings of 5th International Symposium on Environment-Friendly Energies and Applications (EFEA)*, pp. 1–6. IEEE (2018)
5. Padya, B.M.: *Weather and climate of Mauritius*, Geography of Mauritius Series Mahatma Gandhi Institute (1989)
6. McIlveen, R.: *Fundamentals of Weather and Climate*. Psychology Press (1992)
7. Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Duda, M., Huang, X., Wang, W., Powers, J.: A description of the advanced research WRF version 3, NCAR technical note, mesoscale and microscale meteorology division. National Centre for Atmospheric Research, Boulder, Colorado, USA (2008)
8. Hahmann, A.N., Lennard, C., Badger, J., Vincent, C.L., Kelly, M., Volker, P.J., Argent, B., Refslund, J.: Mesoscale modelling for the wind atlas of South Africa (WASA) project. *DTU Wind Energy* **0050**, 80 (2014)
9. Cheng, X.L., Li, J., Hu, F., Xu, J., Zhu, R.: Refined numerical simulation in wind resource assessment. *Wind Struct.* **20**(1), 59–74 (2015)
10. Joint Research Centre (JRC). Tropical Cyclone GIOVANNA, Madagascar. P. Probst, G. Franchello, A. Annunziato, T. De Groeve, L. Vernaccini, A. Hirner, and I. Andredakis, Reference Report by the Joint Research Centre of the European Commission (2012)
11. Interactive, N.: Cyclone Giovanna 2012—Zoom Earth. [online] Zoom Earth. Available at: <https://zoom.earth/storms/giovanna-2012/>. Accessed 16 May 2020. (2018)
12. MMS. Mauritius Meteorological Services [Online] Available at: <https://metservice.intnet.mu>. Accessed on 8 Feb 2020 (2019)

Analysis of Wind Energy Resources for the Island of Mauritius Using Concepts of Thermodynamics



Tyagaraja S. M. Cunden, Naafeera B. R. Abdel Hassan,
and Michel R. Lollchund

Abstract Exergy is the term which relates to the ‘quality’ of the work potential of a system. It is usually split into kinetic exergy, physical exergy, chemical exergy and potential exergy. Hence, giving a more realistic background on the performance of the system. For wind energy systems, kinetic exergy and physical exergy are the only relevant quantities. However, kinetic exergy is more related to the wind turbine system, while physical exergy results from the environmental factors around the system. In this paper, the physical exergetic analysis of the wind energy resources over the island of Mauritius is conducted. It is calculated by using the meteorological variables such as temperature, humidity ratio, pressure and wind speed at typical turbine hub height (60 magl), obtained from running the Weather Research and Forecasting (WRF) model. A high-resolution physical wind exergy map is generated which can help energy planners and decision makers to identify potential hotspot regions for the installation of wind turbine systems that will operate at maximum efficiency.

Keywords Exergy analysis · Wind energy systems · Weather research and forecasting

1 Introduction

The use of wind energy as a promising alternative for the replacement of fossil fuel has been gaining a lot of interest worldwide in the recent years. As a matter of fact, the worldwide overall installed capacity reached 600 GW at the end of 2018, representing around 6% of the global electricity demand [1]. In the Republic of Mauritius, the installation of a 9.35 MW wind farm (Eole wind farm) in December 2016, consisting

T. S. M. Cunden

Department of Electromechanical Engineering and Automation, Université Des Mascareignes,
Rose Hill, Mauritius

N. B. R. Abdel Hassan · M. R. Lollchund (✉)

Department of Physics, Faculty of Science, University of Mauritius, Réduit, Mauritius
e-mail: r.lollchund@uom.ac.mu

of 11 turbines in the northern part of the island at Plaine des Roches, has given a boost to the deployment of wind power for utility scale. Hence, the government is looking forward to invest on similar projects in order to reach the target of using renewable energy to 35% by 2025.

The distribution of electricity in Mauritius is controlled by the Central Electricity Board (CEB). 40% of the electricity is produced by the CEB itself and the remaining 60% by Independent Power Producers (IPP's). It is worth noting that in 2017 about 79% of the electricity production was made from fossil fuels, and the remaining 21% was made from renewables (constituted of Bagasse—14.7%, Hydro—2.8%, Landfill gas, Wind and Solar—3%) [2]. Several studies conducted since the 1980's have shown that wind energy constitutes a very promising avenue for the development of renewable energy in Mauritius [3–5].

The harnessing of energy from the wind is mainly considered to be dependent on the following factors: the wind speed, the power extracting capability of the wind turbine and the ability of the wind turbine to deal with wind turbulences [6]. The power of the wind is proportional to the cube of the wind speed. For this reason, most wind energy resource assessment studies focus mainly on one meteorological variable, the wind speed [3, 7]. Moreover, in wind engineering, the air is generally assumed to be dry, neglecting the humidity variations. However, many studies [8–10] have shown that the performances of wind energy systems are influenced by other factors such as losses due to irreversibilities in each conversion process of the system. Hence, in order to optimise the harnessing of wind energy, it is important to fully understand the thermodynamics behind the wind energy system and consider the use of thermodynamic analyses (exergy) for the evaluation of the maximum power output of the system.

Exergy analysis provides a meaningful tool for evaluating and comparing several processes judiciously. It is based on the second law of thermodynamics. In addition to the quantity of energy, exergy also considers the quality of energy in a reversible process with respect to a reference state. Hence, an exergy analysis can be helpful for calculating the maximum available work in a system, as well as, for locating and sizing the inefficiencies (irreversibilities) of the system.

In the literature, Hepbasli and Alsuhaibani [9] have presented a systematic review of the application of exergy analyses for the assessment of wind turbine systems. They advocate that exergy analyses can also have widespread usage in achieving sustainable development goals. They have investigated the economic aspect of wind energy systems design and operation from an exergetic point of view (exergoeconomic analysis) and suggest that 'such studies may provide key information for people working in the area for better design, analysis, performance improvement and operation of the wind energy systems'. Exergy efficiency maps were introduced in 2006 along with energy efficiency maps and 'provide a common basis for regional assessments and interpretations'.

Koroneos et al. [8] have studied the impacts of the wind speed and irreversibilities in the mechanical parts of the turbine, namely the rotor and gearbox and in the electrical components, namely the generator, on the efficiency of wind turbine systems. According to them, most commercial wind turbine generators only achieve

an efficiency of around 40% when operating at high wind speeds, which is quite far from the maximum efficiency of 60% as predicted by the Betz law. This is attributed mainly to the exergy losses due to friction in the moving parts (rotor shaft and bearings), the heat extracted by the cooling fluids used in the generator and the power electronics components (e.g. the thyristors). The exergy losses are mainly in the form of heat.

Sahin et al. [11] were the first to derive equations related to the exergy analysis of wind turbine systems. They demonstrated that the wind speed is affected by air temperature and pressure. They go on to describe the effect of wind chill and suggest that enthalpy and entropy components must be taken into consideration in a thermodynamic analysis. The Bernoulli's equation requires that the wind pressure be considered in the entropy calculation of the wind. According to the authors, these components have never been considered in calculating turbines efficiencies. They provide a new efficiency formula for wind energy systems. The same authors attempted to develop spatio-temporal exergy maps and reported differences between energy and exergy efficiencies of the order of 20–24% at low wind speeds and 10–15% at high wind speeds.

Pope et al. [10] have conducted energy and exergy studies on systems constituted of both Horizontal Axis Wind Turbines (HAWT) and Vertical Axis Wind Turbines (VAWT) while taking into consideration different commonly used designs and operating parameters of HAWT's and VAWT's. Each system was analysed with respect to both the first and second laws of thermodynamics. The authors reported a difference in the range of 50–53% between energy and exergy efficiencies of HAWT with airfoil design and 44–55% for VAWT systems. They concluded that the exergy analysis provides insight beyond the first law analysis and hence can be considered as a useful design tool for wind energy systems.

The contribution of this paper is use the concepts of thermodynamics (exergy) to analyse the wind energy resources over the island of Mauritius with the aim of producing high-resolution wind physical exergy maps that will be useful for wind energy planners and stakeholders.

2 Methodology

2.1 Study Region

The area of study is the tropical island of Mauritius Fig. 1. It is situated at latitude 20°17' South and longitude 57°50' East in the South-West Indian Ocean (SWIO) basin. The island of Mauritius is of volcanic origin and is composed of a very complex topography with an area of around 1900 km². A central plateau at altitude 500 m above sea level forms the highest part of the island and is surrounded by the coastal plains. The main island has a span of around 60 km from North to South and 45 km

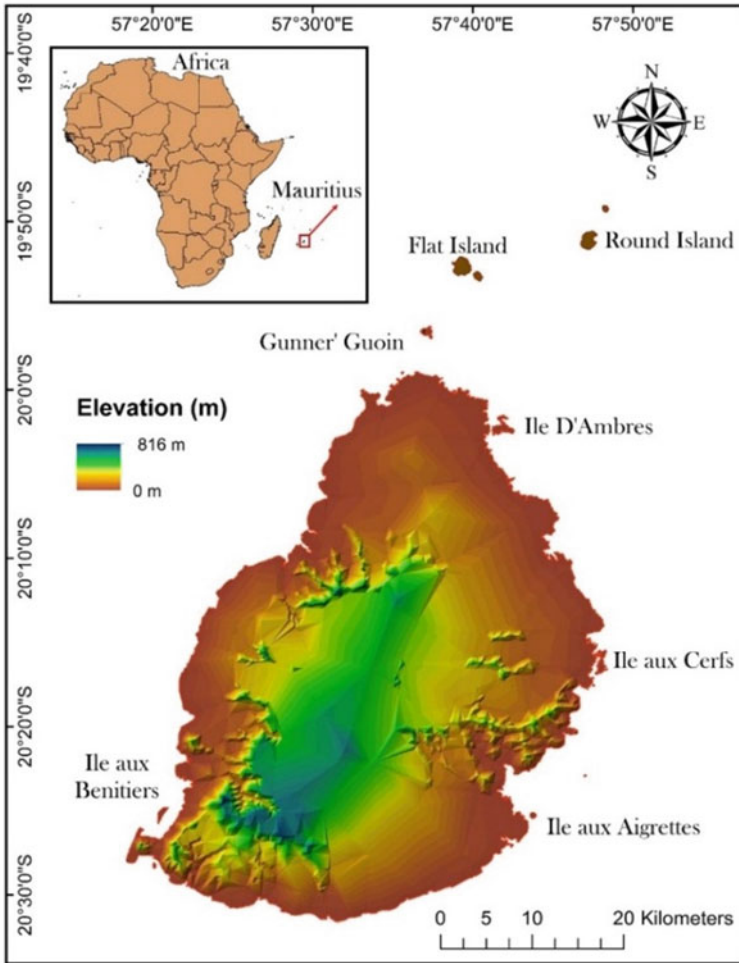


Fig. 1 Main island of Mauritius and surrounding islets (illustration by authors)

from East to West. The population of Mauritius is around 1.265 million as at 2017. The main island is surrounded by a few non-populated islets (see Fig. 1).

As Mauritius is situated in the southern hemisphere and is at the edge of the tropic of Capricorn, it is continuously under the influence of the South East trade winds. According to the local meteorological data, an average wind speed of 4.1 m/s prevails over the island. The trade winds are at stronger during the winter season and are at their peak values during the months of August and September [12].

2.2 Wind Mapping

This section is based on the previous study done by Cunden et al. [13], whereby sensitivity tests are carried out to find the best physics options for setting up the WRF model as well as the study of Cunden [3], whereby the high-resolution wind mapping is done for the study region. The WRF simulation is run for the period 01 January 2015 to 31 December 2017 on a grid resolution of 1 km by 1 km. The WRF model generates the following meteorological variables on an hourly basis: horizontal wind speeds, wind directions, air temperature, relative humidity among others, in a netCDF format. The time series data at each grid point are extracted using a MATLAB code, and the wind maps are generated. The monthly wind maps are represented in Fig. 2.

2.3 Efficiency of Wind Turbines

The evaluation of the capacities of wind energy systems involves the thermodynamic analysis accompanying the energy and exergy conversions within the systems and the computation of energy and exergy efficiencies. The balancing of a quantity in a given process can be generalised by the following equation:

$$\text{Input} + \text{Generation} - \text{Output} - \text{Consumption} = \text{Accumulation} \quad (1)$$

Equation (1) shows that the amount of a quantity accumulated in a process is the net difference between the total amount of the quantity entering and generated during the process and the sum of the consumption of the quantity within the process and the output of the process.

The stream tube representing the control volume of air moving through the rotor of a wind turbine is represented in Fig. 3. The wind speeds at the inlet and the outlet of the stream tube are denoted as V_1 and V_2 , respectively. The mean wind speed at the rotor is denoted as V_{avg} . The mass flow of air through the rotor is given by Eq. (2).

$$\dot{m} = \rho A V_1 \quad (2)$$

where ρ is the air density; A is the rotor swept area; and V_1 is the wind speed at the entrance of the stream tube. Applying the momentum theory, the thrust force on the turbine blades is equal to the rate of change of momentum of the air and can be calculated as follows:

$$P = \dot{m}(V_1 - V_2)V_{avg} \quad (3)$$

Equation (4) gives the power, P extracted from the wind by the wind turbine's rotor blades.

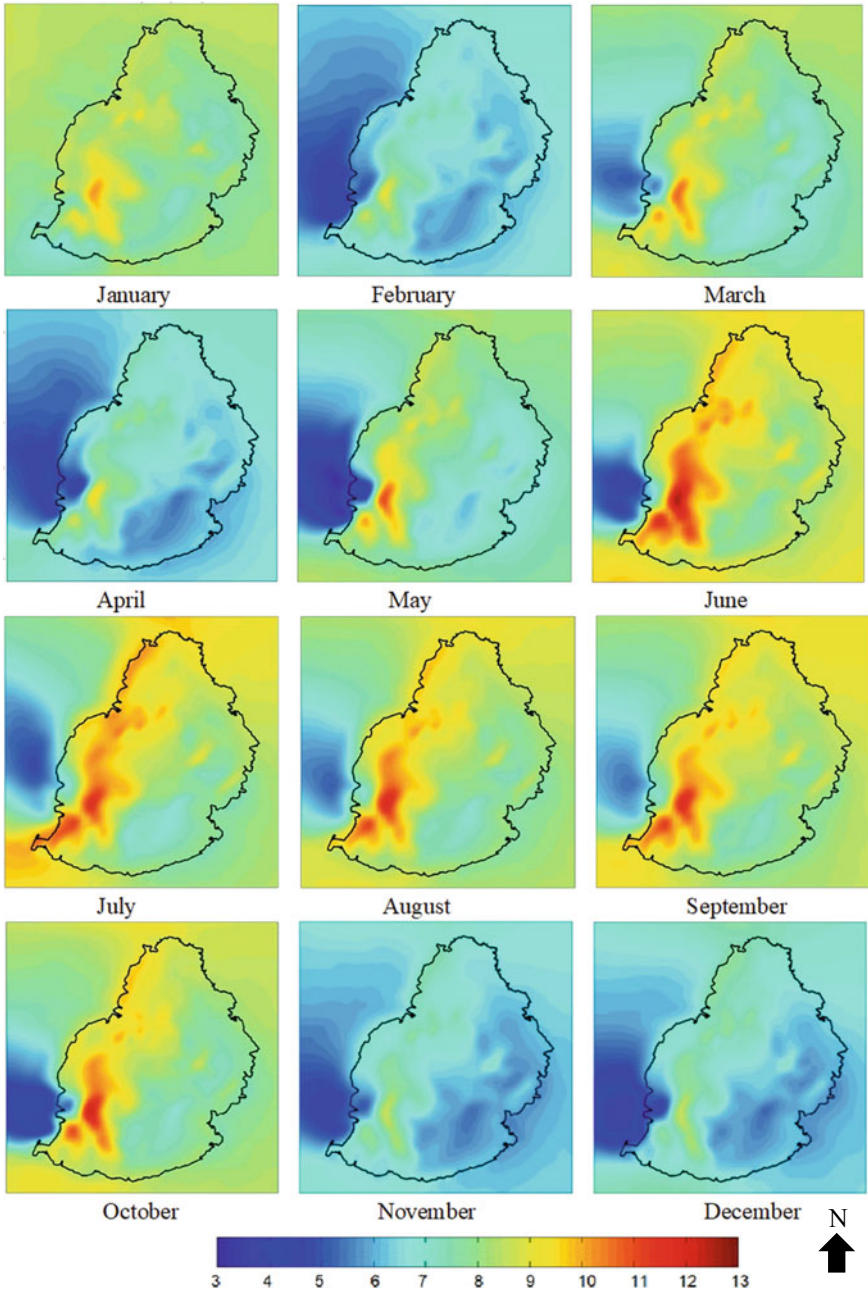
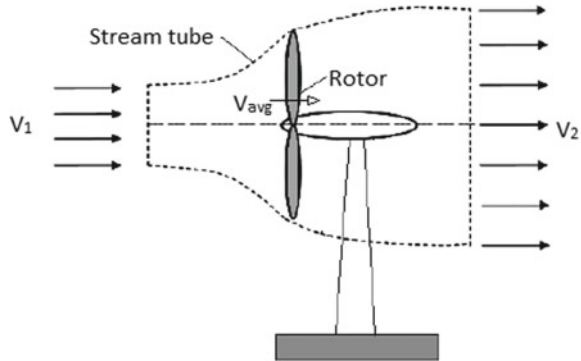


Fig. 2 Monthly mean wind speed maps (in m/s) at 100 m.a.g.l. [3]

Fig. 3 Stream tube around a wind turbine



$$P = \frac{1}{2} C_p \rho A V_1^3 \tag{4}$$

where C_p is the power coefficient; ρ is the air density; and A is the rotor swept area. The maximum value of C_p is 0.593 and is known as the Betz criterion. The air density is a function of meteorological variables such as the humidity ratio ω , the atmospheric pressure p and the temperature T . It is expressed as follows:

$$\rho = \frac{1 + \omega}{R_a + \omega R_v} \frac{p}{T} \tag{5}$$

where R_a is the gas constant, and R_v is the water vapour constant.

By replacing ρ in Eq. (4), the power coefficient can be written in terms of the meteorological variables as follows:

$$C_p = \frac{2(R_a + \omega R_v)}{1 + \omega} \frac{TP}{\rho A V_1^3} \tag{6}$$

The theoretical power extracted by the wind turbine’s rotor blades can thus be computed from Eqs. (4) and (6). However, the real power output, P_{out} of the wind turbine from the wind is lesser than that obtained from Eq. (4) due to losses occurring in the transformation of energy within the wind energy conversion system and limits imposed on power generation by the control system of the wind turbine. Figure 4 represents the typical power curve of a wind turbine. It can be seen that the wind turbine starts generating power at the cut-in wind speed, V_{cut-in} . In the region B, the power generated increases exponentially with increasing wind speed, and in region C, the rated power is generated as from the rated wind speed, V_{rated} . The turbine stops generating power at the cut-out wind speed, $V_{cut-out}$. High wind speeds, beyond $V_{cut-out}$, can cause serious physical damage to the wind turbine system. Hence, below V_{cut-in} and beyond $V_{cut-out}$ no power is generated.

The energy efficiency of a wind turbine can be defined as the ratio of the useful energy output rates with products to the total energy input rates and is represented by Eq. (7) and can be computed from Eq. (8).

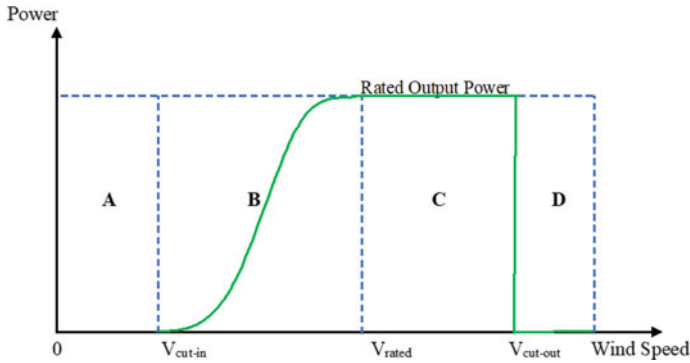


Fig. 4 A typical power curve of wind turbines

$$\eta = \frac{\text{Useful energy output rates with products}}{\text{Total energy input rates}} = \frac{\sum \dot{E}_{\text{out,useful}}}{\sum \dot{E}_{\text{in,total}}} \quad (7)$$

$$\eta = \frac{2(R_a + \omega R_v)}{1 + \omega} \frac{TP_{\text{out}}}{\rho AV_1^3} \quad (8)$$

where P_{out} is the output power defined by the power curve of the wind turbine.

An exergy analysis of the wind energy system takes into consideration the irreversibilities of the system and provides a better understanding of the maximum possible obtainable work from the system. The exergy content of matter can be considered to comprise of four components, namely: kinetic exergy (Ex_{ki}), physical exergy (Ex_{ph}), chemical exergy (Ex_{ch}) and potential exergy (Ex_p) and is represented by the following equation:

$$Ex_{\text{total}} = Ex_{ki} + Ex_{ph} + Ex_{ch} + Ex_p \quad (9)$$

In the case of a wind turbine system, chemical and potential components of exergy can be neglected since there is no chemical reaction within the system, and there are practically no elevation differences for the rotor [14]. Hence, in the exergy analysis of a wind turbine system, only the kinetic exergy and the physical exergy are considered. Hence, the exergy for a wind energy system can be written as follows:

$$Ex = Ex_{ki} + Ex_{ph} \quad (10)$$

The computation of the physical exergy can be done using Eq. (11). It can be seen that the physical exergy considers enthalpy and entropy changes within the process and are represented by the terms on the right-hand side of Eq. (11).

$$Ex_{ph} = c_p(T_2 - T_1) + T_0 \left(c_p \ln \left(\frac{T_2}{T_1} \right) - R \ln \left(\frac{P_2}{P_1} \right) - \frac{c_p(T_0 - T_{\text{avg}})}{T_0} \right) \quad (11)$$

where c_p is the specific heat of air; T_0 is the reference temperature; T_1 ; and T_2 are the temperatures at Sects. 1 and 2 of the stream tube, respectively, and T_{avg} is the temperature at the rotor section of the stream tube depicted in Fig. 3. P_1 , and P_2 are the pressures at Sects. 1 and 2 of the stream tube, respectively. R is a combined gas and water vapour constant. However, Eq. (11) requires values of temperatures and pressures at the Sects. 1 and 2 of the stream tube which are not practical to measure in a real application. It is more convenient to work with the temperature, T and pressure, p at hub height. Ozgener [15] proposes the following equation for the computation of physical exergy based on temperature and pressure at hub height and standard reference temperature, T_0 and pressure, p_0 .

$$\begin{aligned}
 Ex_{ph} = & (C_{p,a} + \omega C_{p,v})(T - T_0) - T_0 \left[(C_{p,a} + \omega C_{p,v}) \ln \left(\frac{T}{T_0} \right) \right. \\
 & \left. - (R_a + \omega R_v) \ln \left(\frac{p}{p_0} \right) \right] + T_0 \left[(R_a + \omega R_v) \ln \left(\frac{1 + 1.6078\omega_0}{1 + 1.6078\omega} \right) \right. \\
 & \left. + 1.6078\omega R_a \ln \left(\frac{\omega}{\omega_0} \right) \right] \tag{12}
 \end{aligned}$$

The kinetic exergy Ex_{ki} is given by Eq. (13) [14].

$$Ex_{ki} = \frac{1 + \omega}{2(R_a + \omega R_v)} \frac{\rho A V_1^3}{T} \tag{13}$$

The exergy efficiency is calculated as follows:

$$\psi = \frac{P_{out}}{Ex_{total}} \tag{14}$$

where

$$\begin{aligned}
 Ex_{total} = & \frac{1 + \omega}{2(R_a + \omega R_v)} \frac{\rho A V_1^3}{T} + (c_{p,a} + \omega c_{p,v})(T - T_0) \\
 & - T_0 \left[(c_{p,a} + \omega c_{p,v}) \ln \left(\frac{T}{T_0} \right) - (R_a + \omega R_v) \ln \left(\frac{p}{p_0} \right) \right] \\
 & + T_0 \left[(R_a + \omega R_v) \ln \left(\frac{1 + 1.6078\omega_0}{1 + 1.6078\omega} \right) + 1.6078\omega R_a \ln \left(\frac{\omega}{\omega_0} \right) \right]
 \end{aligned}$$

2.4 Validation of the Model

Equations (8) and (14) have been applied to calculate the energy and exergy efficiencies, respectively, for a given wind turbine system. In an attempt to verify the

Table 1 Energy and exergy efficiencies computed by our model and compared with published data by (Hu et al. [16])

δ	u (m/s)	P ($\times 10^4$)	T (K)	ω ($\times 10^{-3}$) kg/kg	η		ψ	
		(Pa)			(Hu et al. [16])	(Our model)	(Hu et al. [16])	(Our model)
1	7.37	9.6	272	4	0.42	0.41	0.3	0.31
2	4.89	9.7	260	1.3	0.42	0.4	0.22	0.21
3	5.01	9.6	261	6.21	0.42	0.4	0.33	0.33
4	9.4	9.4	273	4.8	0.46	0.45	0.38	0.37

validity of our model, the values of the efficiencies obtained are compared with the corresponding values obtained by Hu et al. [16] for a similar system and are tabulated in table 1. It is observed that the values of the efficiencies computed by our model tally well with the values of Hu et al. [16]. On the basis of the excellent agreements obtained, it is considered that our model is fit for studying wind turbine systems. This is elaborated in the next section.

3 Results and Discussion

3.1 Mapping of the Physical Exergy for Wind Over Mauritius

Using the model developed in Sect. 2, data are generated for calculating the monthly mean physical exergy over the study region. The physical exergy is computed using Eq. (12). The monthly physical exergy maps for Mauritius are generated using the mapping methodology described by Cunden [3] and are presented in Fig. 5.

It is observed that the maximum exergy is obtained in the month of July while the minimum exergy is in the month of January. January is peak summer and July peak winter in Mauritius. However, a lot of concordance is observed between the monthly mean wind speed maps in Fig. 2 and the monthly exergy maps in Fig. 5. The region towards the south-west of the island where maximum mean wind speed is observed; throughout the year is also the region, where maximum physical exergy is observed. The western coast always presents a low physical exergy. Figures 6 and 7 represent the seasonal and yearly physical energy geographical distribution, respectively, for Mauritius. It can be observed that the physical exergy in winter is generally higher than in summer. This can be attributed to the lower relative humidity of the air and the lower temperatures in the winter season (see Figs. 8 and 9).

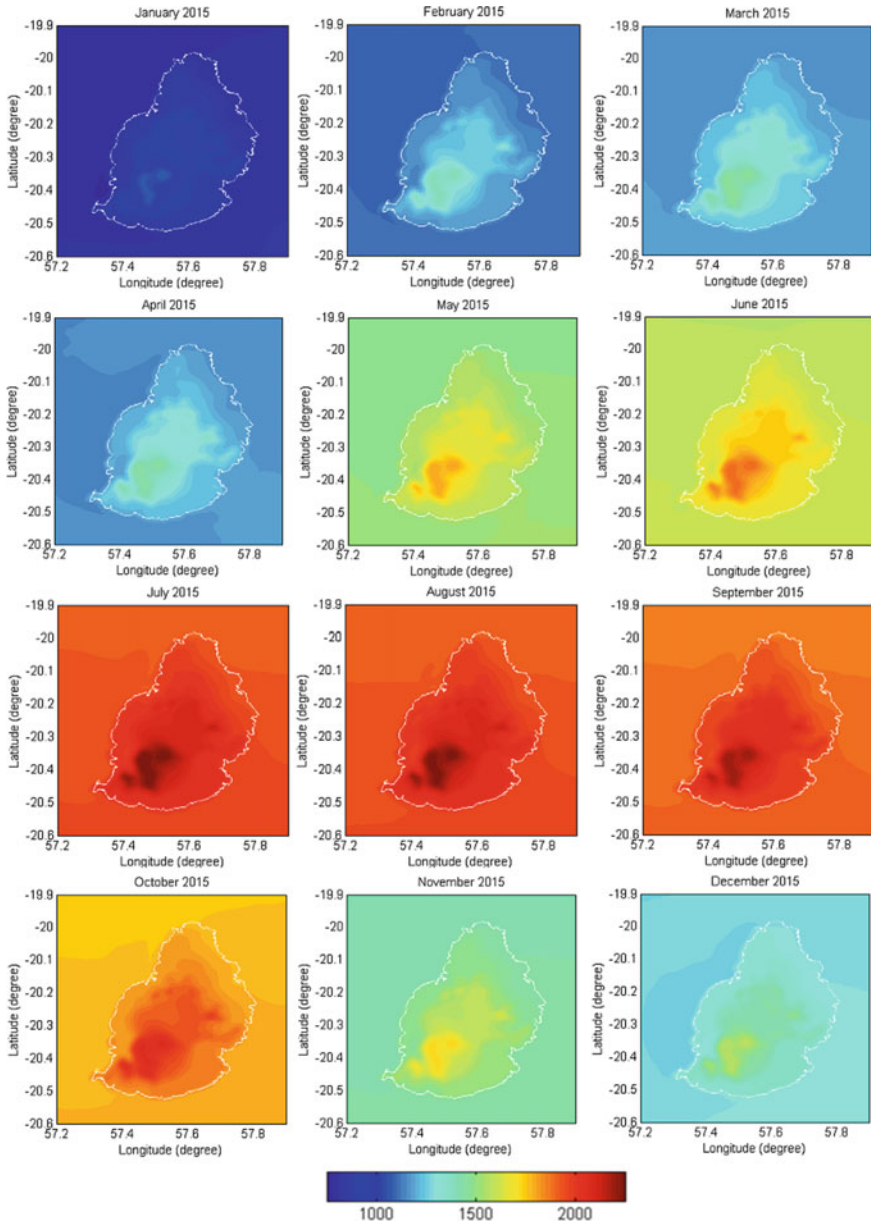


Fig. 5 Mapping of monthly physical energy (W) for Mauritius

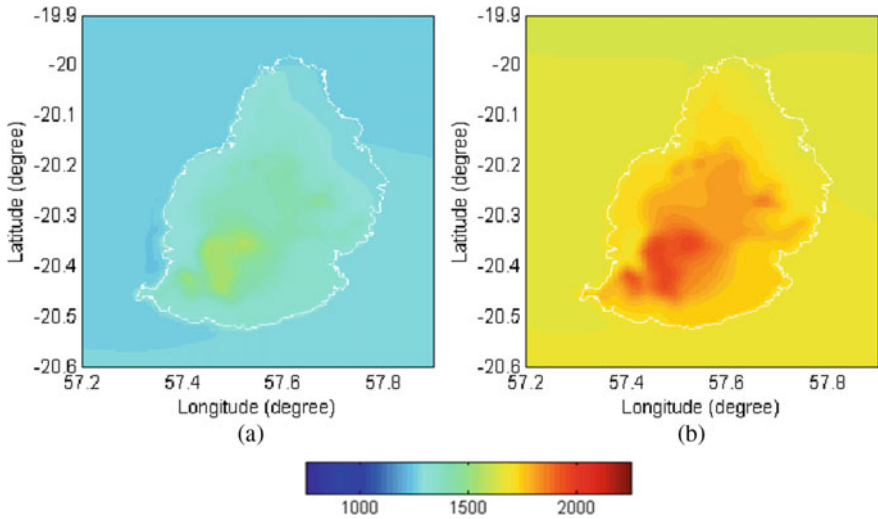
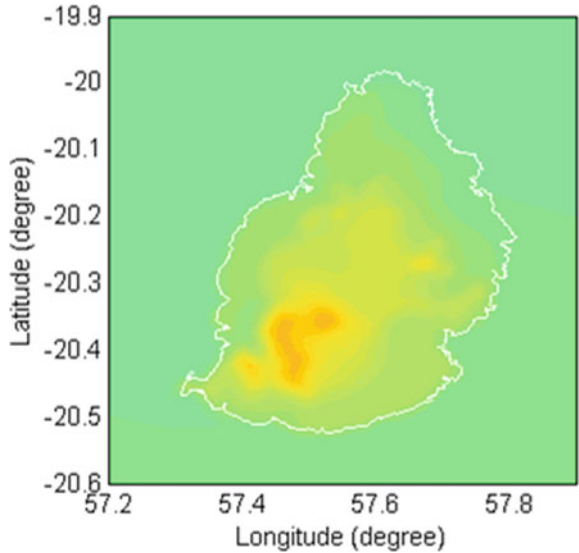


Fig. 6 Mapping of seasonal physical energy (W): **a** Summer, **b** Winter

Fig. 7 Mapping of yearly physical energy (W) for Mauritius



4 Conclusions

This paper has presented the methodology for conducting an exergetic analysis of wind energy systems. The equations for calculating the energy and exergy efficiencies of a wind turbine have been presented. Using data obtained by running the WRF model over the island of Mauritius, the meteorological data relative to wind speed,

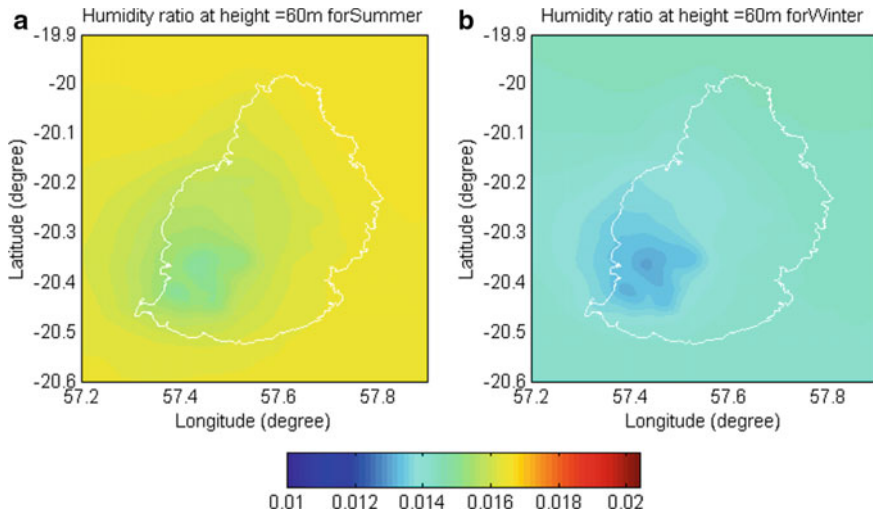


Fig. 8 Mapping of seasonal humidity ratio (kg/kg): a Summer, b Winter

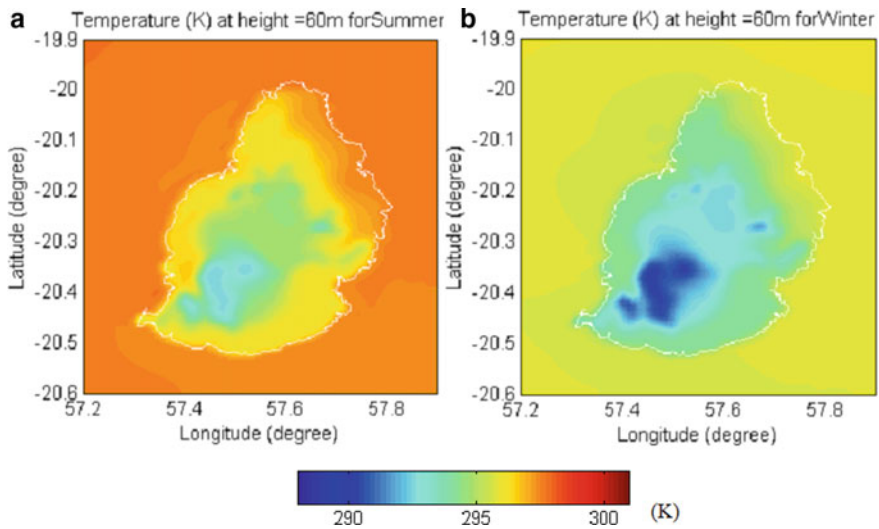


Fig. 9 Mapping of seasonal temperature (K): a Summer, b Winter

pressure, temperature and humidity ratio were extracted and were used to generate a high-resolution maps of monthly, seasonal and yearly physical exergy for the island of Mauritius. The physical exergy is found to be generally higher in winter when the temperature and humidity are lower. Further, studies are directed towards analysing the energy and exergy efficiencies of commercial wind turbines at potential locations of wind farms over the island.

References

1. WWEA: World Wind Energy Association <https://www.wwindea.org/new-record-in-worldwide-wind-installations/>
2. Central Statistics Office: Economic and Social Indicators on Energy and Water Statistics report 2017. Mauritius. [online] Available at: https://statsmauritius.govmu.org/English/Publications/Documents/2018/EI1386/Energy_Yr17.pdf. Accessed 13 Feb 2020
3. Cunden, T.S.M.: Mapping of Wind Resources for Mauritius with Multilevel Constraints Under Normal and Extreme Weather Conditions. Ph.D. thesis, Université des Mascareignes, Mauritius (2020)
4. Dhunny, A.Z., Lollchund, M.R., Rughooputh, S.D.D.V.: Wind energy evaluation for a highly complex terrain using computational fluid dynamics (CFD). *Renew. Energy* **101**, 1–9 (2016)
5. Elliot, D.L., Aspliden, C.I., Barnard, J.C., Severtsen, R.H.: Wind Energy Resource for Mauritius. Battelle Pacific Northwest Laboratories, Richland, Washington, USA (1986)
6. Draxl, C., Hahmann, A.N., Peña, A., Giebel, G.: Evaluating winds and vertical wind shear from weather research and forecasting model forecasts using seven planetary boundary layer schemes. *Wind Energy* **17**(1), 39–55 (2014)
7. Hahmann, A.N., Lennard, C., Badger, J., Vincent, C.L., Kelly, M., Volker, P.J., Argent, B., Refslund, J.: Mesoscale modelling for the wind atlas of South Africa (WASA) project. *DTU Wind Energy* **0050**, 80 (2014)
8. Koroneos, C., Spachos, T., Moussiopoulos, N.: Exergy analysis of renewable energy sources. *Renew. Energy* **28**(2), 295e310 (2003)
9. Hepbasli, A., Alsuhaibani, Z.: Exergetic and exergoeconomic aspects of wind energy systems in achieving sustainable development. *Renew. Sustain. Energy Rev.* **15**(6), 2810–2825 (2011)
10. Pope, K., Dincer, I., Naterer, G.: Energy and exergy efficiency comparison of horizontal and vertical axis wind turbines. *Renew. Energy* **35**, 2102–2113 (2010). <https://doi.org/10.1016/j.renene.2010.02.013>
11. Sahin, A.Z., Aksakal, A.: Wind power energy potential at the North-eastern Region of Saudi Arabia. *Renew. Energy* **14**(1–4), 435–440 (2006)
12. Padya, B.M.: Weather and Climate of Mauritius, Geography of Mauritius Series. Mahatma Gandhi Institute (1989)
13. Cunden, T.S.M., Dhunny, A.Z., Lollchund, M.R., Rughooputh, S.D.D.V.: Sensitivity analysis of WRF model for wind modelling over a complex topography under extreme weather conditions. In: 2018 5th International Symposium on Environment-Friendly Energies and Applications (EFEA), pp. 1–6. IEEE (2018)
14. Redha, A.M., Dincer, I., Gadalla, M.: Thermodynamic performance assessment of wind energy systems: An application. *Fuel Energy Abstr.* **36**, 4002–4010 (2011). <https://doi.org/10.1016/j.energy.2011.05.001>
15. Ozgener, O., Ozgener, L.: Exergy and reliability analysis of wind turbine systems: a case study. *Renew. Sustain. Energy Rev.* **11**, 1811–1826 (2007)
16. Hu, W., Zhenyu, L., Jianrong, T.: Thermodynamic Analysis of Wind Energy Systems (2019). <https://doi.org/10.5772/intechopen.85067>

Design and FPGA Implementation of an Efficient 8×8 Multiplier Using the Principle of Vedic Mathematics



Smitha Bhat Kaje and Jagadish Nayak

Abstract This paper is regarding the design and analysis of an efficient 8×8 Vedic multiplier by the principle of Vedic mathematics [1]. Here, a unique Vedic multiplier architecture is adapted, which is not based on the regular method of multiplication (addition, shifting). The design is as per the "Urdhwa-Tiryakbhyam Sutra" of Vedic mathematics. Two numbers (binary of 8-bit each) are multiplied using the methodology of this principle/sutra. "Urdhwa-Tiryakbhyam" means "vertically and crosswise", wherein the partial products are computed at once, thus lessening the delay and hence making the multiplication faster. This 8 by 8 bit multiplier is coded in Verilog HDL and tested on a DE10-lite FPGA kit.

Keywords Vedic maths · Principles · Urdhwa-tiryakbhyam sutra · Verilog HDL · DE10-lite FPGA

1 Introduction

Multiplication is an operation much needed in various applications. A multiplier is a prominent block in the processing systems. Multipliers are vital components for computation purpose in several high-performance systems such as microprocessors. Faster the multiplier, the better is the system's performance. Usually, the multiplier is the part of the system that takes most computational time. There is a requirement of faster multiplier due to the domination of multiplication in the execution time of most of the applications. Hence optimizing in terms of delay and speed of the multiplier is an important requirement in designing.

DE10-lite has an efficient platform for hardware design, which is setup on the Altera MAX 10 FPGA. The inputs are entered through the DIP switches, and the LEDs and the seven segment display are used to show the outputs. The results of the

S. B. Kaje (✉)

M E Microelectronics, BITS Pilani, Dubai, UAE

e-mail: kajesmita@gmail.com

J. Nayak

HOD, EEE Department, BITS Pilani, Dubai, UAE

Vedic multiplier [1] and a regular Booth's multiplier [4] are tabulated and compared for performance in terms of factors like the logic elements used and propagation delay. This architecture can be further developed to multiply complex numbers with high speed and reduced delay. This project's intention is to increase the recognition of Vedic mathematics in the field of engineering and to get considered in the design of components in commercial products of everyday use.

2 Literature Survey

In the literature, there have been various algorithms proposed to perform the multiplication. All of them have their own advantages and provide trade-offs with respect to speed, delay, circuits complication, area, and power. The literature survey of the references used in this paper is as follows.

Swami [1] researched for eight years on Vedas and has written a book on Vedic mathematics. He bifurcated the mathematics into 16 simple sutras (formulae) and 16 sub-sutras which are the backbones of Vedic mathematics. The greatest advantage of Vedic mathematics is that it converts the complex calculations of traditional mathematics into simple ones, thereby optimizing processing time, area, and delay. This book gives an insight on all these principles in a detailed manner.

In John et al. [2] described how to reduce the delay and consumed power in multipliers, by the carry select addition technique. Nine full adders and a special 4-bit adder are used to design a 4×4 Vedic multiplier, resulting in reduced delay. Subsequently, an 8-bit multiplier is formed using four 4-bit multipliers with three-carry select adders.

In this paper, Gadakh et al. [3] described about the design of up to 32-bit Vedic multiplier that reduces power, delay, etc. Xilinx 14.7 software is used for simulation. The results are compared for various Vedic multipliers of different architectures.

In Sunitha et al. [4] used four 4X4 multipliers, designed using Urdhwa-Tiryakbhyam sutra for generating partial products. Optimization in terms of power and delay is achieved by using high speed 4:2 compressor architecture. Three modified ripple carry adders are used for accomplishing the addition of the partial products.

In Ingle et al. [5] compared Vedic multipliers with conventional multipliers like Booth's multiplier. Multipliers are coded in Verilog, and simulation is on XILINX software 14.3. Further, the performance metrics of multipliers such as area and delay are determined and compared.

In Dhole et al. [6] describe about the design of three Vedic multipliers using Vedic mathematics that help in reducing overall power and delay. Xilinx 14.7 software is used for the simulation. The report consists of the delay comparison too.

Here, there is a mention about the sutras [7].

From [8], [9] and [10] references on Vedic multiplication are taken.

This book provides a broad coverage of Verilog HDL and has practical designs of Verilog [11].

The DE10-lite system CD contains the documentation and materials such as the user manual, reference designs, and datasheets of the device. User manual provides the complete reference for the FPGA [12].

3 Vedic Mathematics

Vedic mathematics [1] is a primitive method of mathematics of India, which was developed 2500 years ago. "Vedic" comes from "Veda" meaning knowledge. Vedic mathematics has arithmetic rules that allow faster implementation. The sixteen principles called as sutras are the main basis of Vedic mathematics [7]. These are very attractive, and they give effective algorithms which can be adapted in several areas of engineering.

3.1 *Urdhwa-Tiryakbhyam*

The multiplier is based on an algorithm "Urdhwa-Tiryakbhyam" of the ancient Indian Vedic mathematics. "Urdhwa" and "Tiryakbhyam" words are derived from Sanskrit. This principle is called the "crowning gem of Vedic mathematics" [1], because it is the ultimate technique; we can multiply any number into any number within a single line and in less time.

The following example is based on a generic formula that can be used for all cases of multiplication involving large numbers.

Example 1: 88×97 .

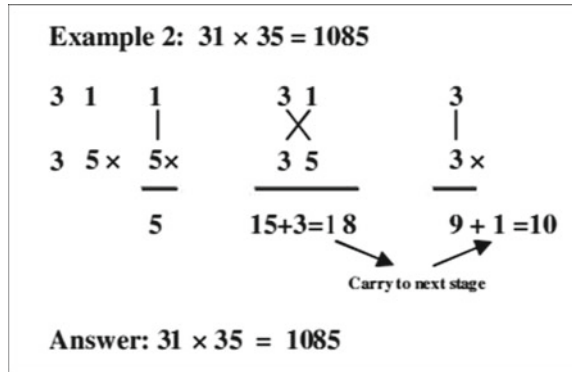
1. Subtract $100-88 \rightarrow 12$ and $100-97 \rightarrow 3$

2. Subtract crosswise to get equal results: $88 - 3 = 97 - 12 = 85$, the first figure of the answer is 85.

3. Multiply vertically: 12×3 to Get 36, the last figure of the answer. Hence, $88 \times 97 = 8536$ (Fig. 1).

Some conventional methods used for multiplication operation, such as Booth's method are not very efficient in speed and number of elements. In the conventional method, partial products are added only after every partial product is obtained. In Vedic principle, as shown in the figure, partial products are calculated vertically, and once all the elements of a column are obtained, respective partial products are added simultaneously. Hence, speed advantage is obtained over the conventional method. This method can be extended to higher bits like 16, 32, 64 multipliers, and their performance can be compared. The hierarchical structure and the parallel addition of the partial product terms play a prominent role in increasing the speed of Vedic multiplier. This design hence saves resources in FPGA for all applications.

Fig. 1 Example of Vedic multiplication [8]



4 Implementation of Urdhwa-Tiryabhyam (Algorithm)

Based on the “Urdhwa-Tiryakbhyam” Sutra, an 8×8 bit multiplier is designed as follows.

The basic module is a 2-bit multiplier as per Vedic principle.
 Considering a 2×2 bit multiplication using Vedic methodology,
 Take $a = 01$ and $b = 11$ (Fig. 2).

- (1) AND operation of $a[0]$ with $b[0]$
- (2) AND operation of $a[0]$ with $b[1]$ —(Cross)
- (3) AND operation of $a[1]$ with $b[0]$ —(Cross)
- (4) AND operation of $b[1]$ with $a[0]$ —(Cross)

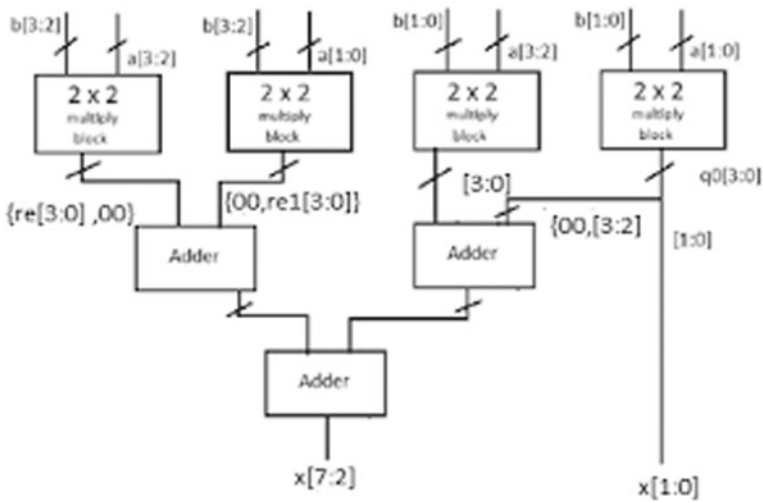


Fig. 2 Block diagram of a 2×2 Vedic multiplier [12]

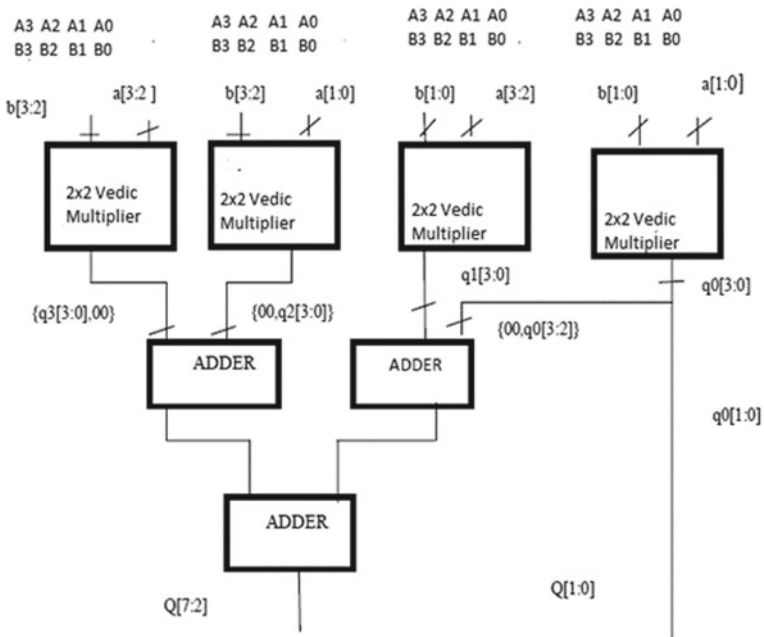


Fig. 3 Block diagram of a 4×4 Vedic multiplier [9]

(5) ADD the above side by side

Similarly, the next block is a 4×4 bit multiplier.

Example: $a = 1100$ and $b = 1001$ (Fig. 3).

- (1) 2×2 multiplier for $a[3:2]$ and $b[3:2]$ then concatenate 00; total bits are now 6 bits
- (2) 2×2 multiplier for $a[1:0]$ and $b[3:2]$ then concatenate 00; total bits are now 6 bits
- (3) Add the results of step 1 and step 2
- (4) 2×2 multiplication of $a[3:2]$ and $b[1:0]$
- (5) 2×2 multiplication of $a[1:0]$ and $b[1:0]$
- (6) Let the result of 6th step be q , then concatenate 00 before $q[3:2]$
- (7) Now, adding the results of steps 6 and 5
- (8) Then, adding the results of steps 7 and 3
- (9) $x[7:2]$ is the result.
- (10) and $x[1:0]$ is same as $q[1:0]$
- (11) Concatenating the above two, we get the result $x[7:0]$

8-bit Vedic multiplier consists of both 4 and 2 bit modules.

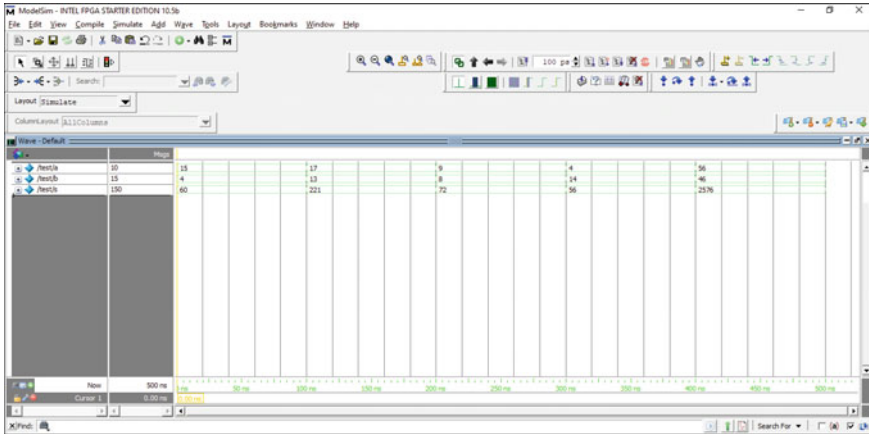


Fig. 4 Testbench simulation results of the 8×8-bit Vedic multiplier

5 Software and Hardware Implementation and Results

The code of the 8×8 bit Vedic multiplier is written in Verilog HDL. Modelsim-Altera 10.5b is used in conjunction with Intel Quartus Prime (18.1) tool for simulation and logic synthesis. Modelsim (Mentor Graphics) is used for simulating Verilog HDL (Hardware Description Languages). The testbench simulation results of an 8-bit multiplier are as in Fig. 4.

To verify the results, a 4-bit multiplier is implemented and tested on a DE10-lite FPGA board. The inputs are two 4-bit numbers. Input A (SW3 SW2 SW1 SW0) is displayed on the HEX0 location of the seven segment display. Input B (SW7 SW6 SW5 SW4) is displayed on the HEX2 location of the seven segment display. The result is displayed on the HEX4 and HEX5 locations. The inputs and outputs are displayed as in Fig. 5.

6 Comparison with Booth’s Multiplier

For comparing the performance of the Vedic multiplier, a multiplier based on Booth’s algorithm Fig. 6 is simulated and verified for various parameters.

The testbench simulation results of the Booth’s multiplier [4] is as in Fig. 7.

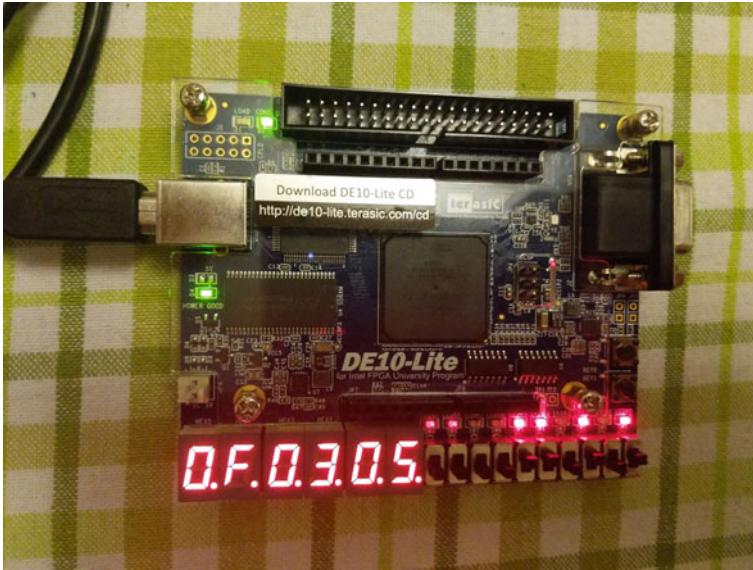
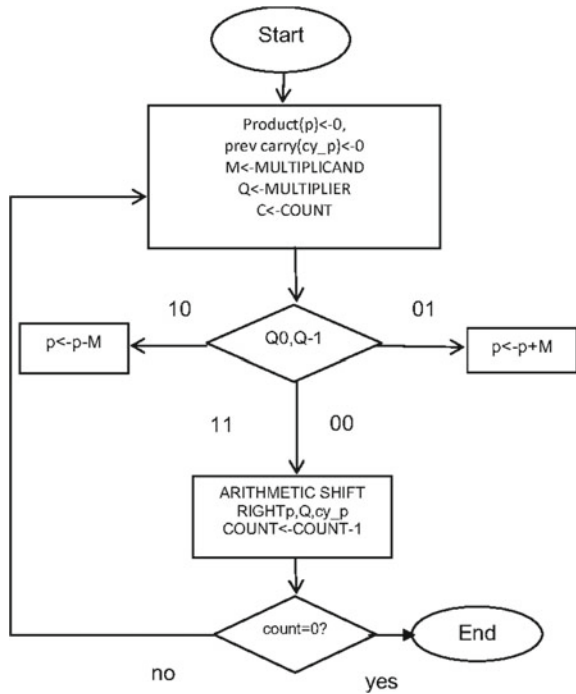


Fig. 5 Inputs and output on a DE10-lite board

Fig. 6 Booth's algorithm [4]



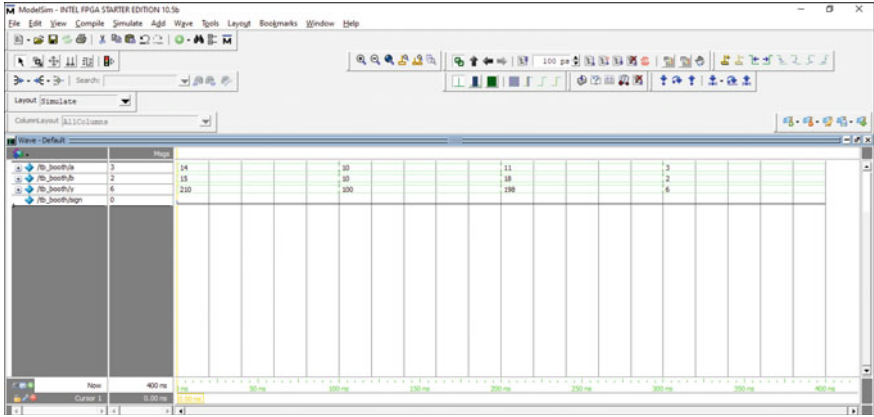


Fig. 7 Testbench results of the Booth's multiplier [4]

6.1 Logic Elements

The number of LE's used in the Vedic multiplier is **161**, whereas the LE's used in the Booth multiplier is **184**.

The "Analysis and Synthesis" Report of the 8×8 Vedic multiplier gives the flow summary as in Fig. 8.

The "Analysis and Synthesis" Report of the Booth multiplier gives the flow summary as in Fig. 9.

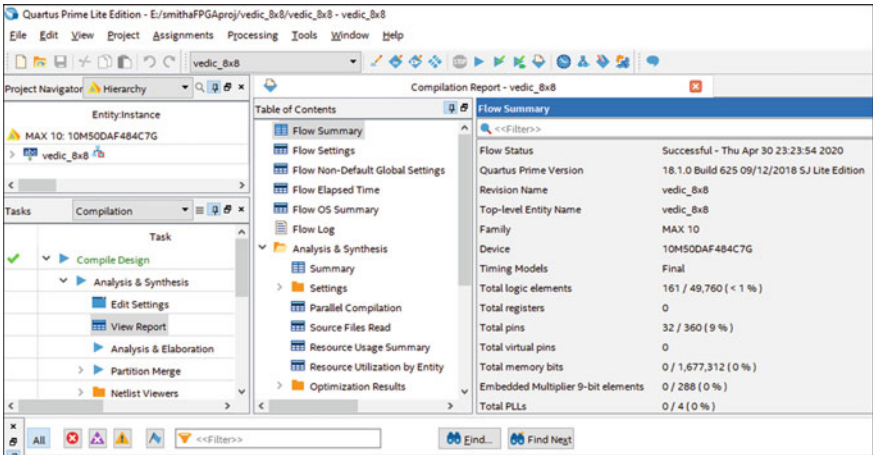


Fig. 8 Flow summary of the 8×8 Vedic multiplier

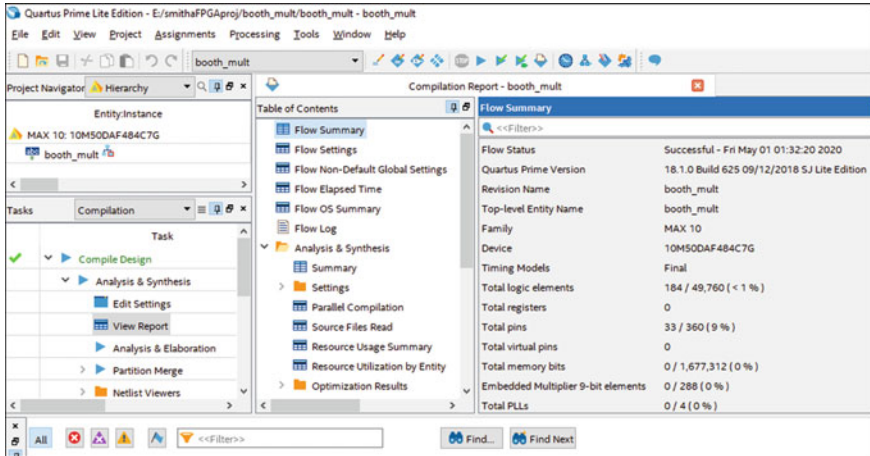


Fig. 9 Flow summary of the booth multiplier

6.2 Propagation Delay

The timing analyzer provides the minimum propagation delay reports of the 8×8 Vedic multiplier and the Booth’s multiplier as below. The values are lesser in the Vedic multiplier.

Propagation delay—is the longest delay between the edges of a signal propagating from an input port to an output port. It is in nanoseconds (Figs. 10 and 11).

- RR—rising edge to rising edge longest delay.
- RF—rising edge to falling edge longest delay.
- FR—falling edge to rising edge longest delay.
- FF—falling edge to falling edge longest delay.

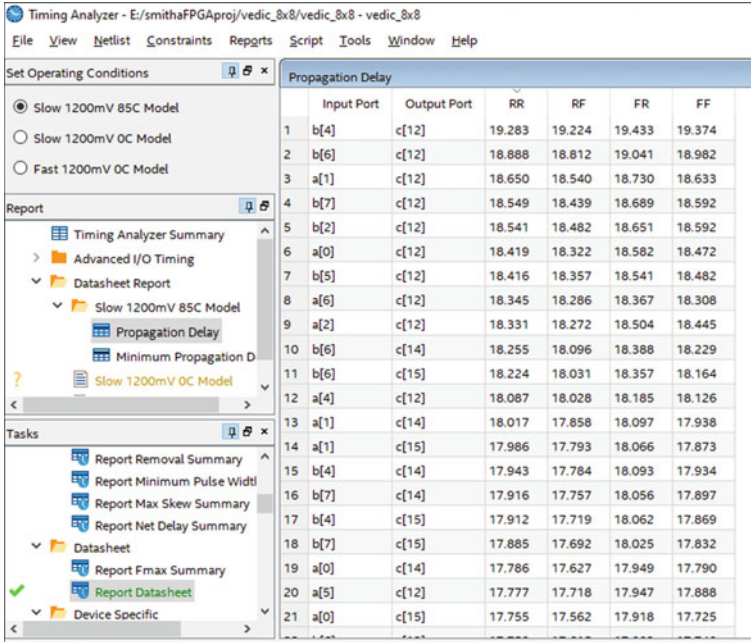
6.3 Structure

The RTL viewer diagram of both the multipliers clearly indicates that Vedic multiplier is more structured. Figures 12 and 13 show the various RTL viewer diagrams of the Vedic multiplier and the Booth’s multiplier.

6.4 Power Analyzer Summary

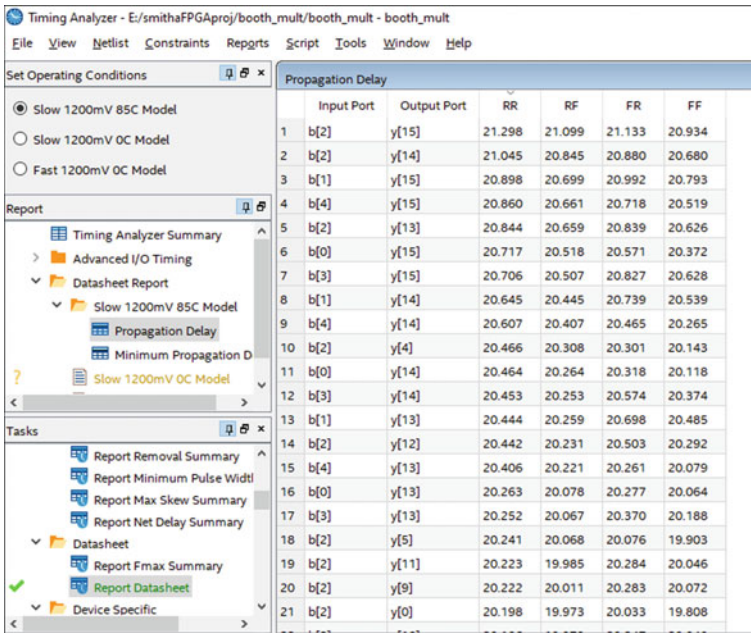
Both multipliers show similar power dissipation (Fig. 14).

The below Table 1 comprises of the comparison of the various performance parameters of both the multipliers:



	Input Port	Output Port	RR	RF	FR	FF
1	b[4]	c[12]	19.283	19.224	19.433	19.374
2	b[6]	c[12]	18.888	18.812	19.041	18.982
3	a[1]	c[12]	18.650	18.540	18.730	18.633
4	b[7]	c[12]	18.549	18.439	18.689	18.592
5	b[2]	c[12]	18.541	18.482	18.651	18.592
6	a[0]	c[12]	18.419	18.322	18.582	18.472
7	b[5]	c[12]	18.416	18.357	18.541	18.482
8	a[6]	c[12]	18.345	18.286	18.367	18.308
9	a[2]	c[12]	18.331	18.272	18.504	18.445
10	b[6]	c[14]	18.255	18.096	18.388	18.229
11	b[6]	c[15]	18.224	18.031	18.357	18.164
12	a[4]	c[12]	18.087	18.028	18.185	18.126
13	a[1]	c[14]	18.017	17.858	18.097	17.938
14	a[1]	c[15]	17.986	17.793	18.066	17.873
15	b[4]	c[14]	17.943	17.784	18.093	17.934
16	b[7]	c[14]	17.916	17.757	18.056	17.897
17	b[4]	c[15]	17.912	17.719	18.062	17.869
18	b[7]	c[15]	17.885	17.692	18.025	17.832
19	a[0]	c[14]	17.786	17.627	17.949	17.790
20	a[5]	c[12]	17.777	17.718	17.947	17.888
21	a[0]	c[15]	17.755	17.562	17.918	17.725

Fig. 10 Propagation delay report of the Vedic multiplier



	Input Port	Output Port	RR	RF	FR	FF
1	b[2]	y[15]	21.298	21.099	21.133	20.934
2	b[2]	y[14]	21.045	20.845	20.880	20.680
3	b[1]	y[15]	20.898	20.699	20.992	20.793
4	b[4]	y[15]	20.860	20.661	20.718	20.519
5	b[2]	y[13]	20.844	20.659	20.839	20.626
6	b[0]	y[15]	20.717	20.518	20.571	20.372
7	b[3]	y[15]	20.706	20.507	20.627	20.628
8	b[1]	y[14]	20.645	20.445	20.739	20.539
9	b[4]	y[14]	20.607	20.407	20.465	20.265
10	b[2]	y[4]	20.466	20.308	20.301	20.143
11	b[0]	y[14]	20.464	20.264	20.318	20.118
12	b[3]	y[14]	20.453	20.253	20.574	20.374
13	b[1]	y[13]	20.444	20.259	20.698	20.485
14	b[2]	y[12]	20.442	20.231	20.503	20.292
15	b[4]	y[13]	20.406	20.221	20.261	20.079
16	b[0]	y[13]	20.263	20.078	20.277	20.064
17	b[3]	y[13]	20.252	20.067	20.370	20.188
18	b[2]	y[5]	20.241	20.068	20.076	19.903
19	b[2]	y[11]	20.223	19.985	20.284	20.046
20	b[2]	y[9]	20.222	20.011	20.283	20.072
21	b[2]	y[0]	20.198	19.973	20.033	19.808

Fig. 11 Propagation delay report of the Booth's multiplier

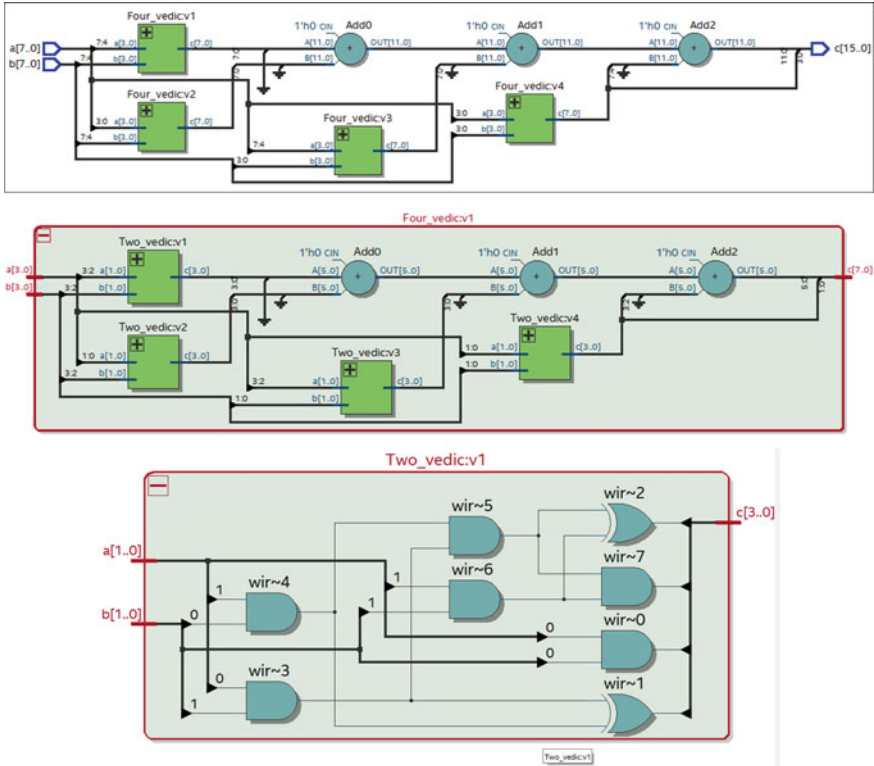


Fig. 12 RTL structure of the Vedic multiplier

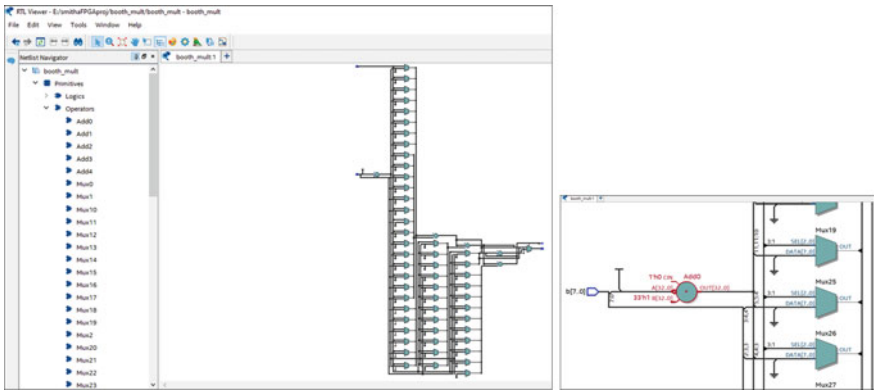


Fig. 13 RTL structure of the Booth's multiplier

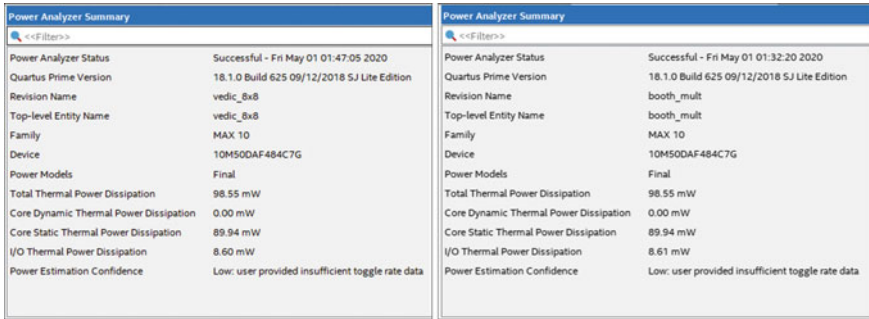


Fig. 14 Power analyzer summary diagrams of Vedic and Booth’s multipliers

Table 1 Performance comparison table of the Vedic and Booth’s multipliers

Sr. No.	Parameters	Vedic multiplier	Booth’s multiplier
1	Number of LE’s	161	184
2	Worst propagation delay	19.433 ns	21.298 ns
3	Thermal power dissipation	98.55 mW	98.55 mW
4	RTL structure	Structured	Not structured

7 Conclusion

An 8×8-bit multiplier is successfully simulated and implemented using Vedic techniques. It is compared with an 8×8-bit Booth’s multiplier. Based on the simulation, FPGA outputs and performance analysis arrived at a conclusion that the vedic multiplier is more structured, uses lesser logical elements, has lesser propagation delays and hence faster. The performance would be even more evident and more beneficial in Vedic multipliers of further higher bits like 16×16, 32×32, and so on.

References

1. Maharaja, B.K.T.: Vedic Mathematics, Motilal Banarsidass Publishers (1965)
2. Johny, N.: Area and Delay efficient Multiplier using Vedic Mathematics. New Horizon College of Engineering, Bangalore, Karnataka, India (2016)
3. Gadakh, S.N., Khade, A.A.: FPGA implementation of high speed Vedic multiplier. Int. Conf. Workshop Electron. Telecommun. Eng. ICWET Mumbai 184–187 (2016). <https://doi.org/10.1049/cp.2016.1144>

4. Sunitha, G.S., Rakesh, H.M.: Performance Comparison of Conventional Multiplier with Vedic Multiplier Using ISE Simulator. Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India (2018)
5. Ingle, A., Oza, Dr.: FPGA Implementation of Novel High Speed Multiplier. BVDU's college of Engineering, Pune, Maharashtra, India (2015)
6. Dhole, S., Shembalkar, S., Yadav, T.: Design and FPGA Implementation of 4×4 Vedic Multiplier using Different Architectures. RCOEM, Nagpur, India (2017)
7. Y. Bansal, C.M., Kaur, P.: High speed vedic multiplier designs-a review. In: 2014 Recent Advances in Engineering and Computational Sciences (RAECS), Chandigarh (2014)
8. Sriraman, L., Prabakar, T.N.: Design and Implementation of Two variable Multiplier using KCM and Vedic Mathematics. Oxford Engineering College, Trichy, Tamilnadu, India (2012)
9. Dheeksha S.S., Kiran Kumar V.G.: VLSI Implementation of Arithmetic Operation, Sahyadri College of Engineering and Management, India (2016)
10. Magesh, V., Selvaoviya, S., Soundarya, B., Pooja, S.V.: FPGA implementation of pipelined FIR filter using vedic multiplier. Int. J Appl. Eng. Res. **14**(6), ISSN 0973-4562 (2019)
11. Samir, P.: Verilog HDL—A guide to Digital Design and Synthesis, 2nd edn. Dorling Kindersley and Pearson Education, Inc (2008)
12. www.intel.com DE10-Lite System CD and DE-10 Lite Manual.

Fault Detection and Isolation in a Leaky Water Distribution Network Using Fuzzy Logic Control Based on Residual Pressure Analysis



Lekhramsingh Latchoomun and Tsiatsipy Durand Brunel

Abstract In this research paper, we try to address the problem of leakage in an experimental water distribution network using the method of fault detection and isolation. The algorithm is based on the residual generation and analysis of nodal pressures for the faulty network compared to the ideal one. A fuzzy controller then determines the location of the critical leaking node based on a Mamdani set of inference rules for isolation using flow control valves. Results show that the residual pressure analysis for FDI is very effective when network is embedded with two leaks, whereby isolation of the faulty pipes is successfully carried out to prevent wastage of water.

Keywords Fault detection and isolation (FDI) · Fuzzy logic controller · Water distribution network (WDN) · Flow control valve · Step test and system identification toolbox

1 Introduction

State of the art FDI [1] techniques using principal component analysis, observer-based techniques, parameter estimation, sensitivity analysis or pattern recognition are not suitable for an old leaky network because there is no information on the state of the network when it was new or faultless, and there is a very limited number of sensors in the network as is the case here. In the light of the above problem, a new method based on pressure step testing is proposed. The idea is to get the pressure response without leak or after repairs and use it eventually for comparison in the event of suspected leakage or pipe bursts. Unlike other knowledge or model-based FDI schemes [2] where dynamic data are analyzed with a reference model before taking

L. Latchoomun (✉) · T. D. Brunel
Université Des Mascareignes, Ave. De la Concorde, Roches Brunes, Rose Hill, Mauritius
e-mail: nlatchoomun@udm.ac.mu

T. D. Brunel
e-mail: tsiatsipydurandbrunel@yahoo.fr

decisions, here, the test is done online at a specific time, preferably during minimum night flow (MNF) [3], where pressure is highest, and disruption of water supply is not really a problem. In this method, a step test is performed on the network and the pressure profile obtained through one or more sensors normally placed at the start of distribution, and at a few other sparse nodes are compared with that of the initial good state. The criticality of the situation is assessed based on the evaluation of the residuals generated, and the appropriate decisions are taken especially in case of high leakage or pipe bursts. The complexity of analysis of the residual matrix increases with the number of nodes and measuring points. For simplicity, it is assumed that there are no faulty sensors during the test. In a real case situation, this test can be performed based on the strategic location of a few sensors, where the frequency of repairs is highest in the network. On the other hand, adaptive FDI is not very suitable for a distribution network because of the delay inherent in the distribution of water. Overall, isolation not only helps to prevent excessive leakage of water but also helps to improve the energy efficiency of the pumping and distribution system [4]. The next section outlines the methodology and the materials used in this research work, whereas Sect. 3 gives a comprehensive description of residual pressure analysis for membership functions of the fuzzy logic approach. Section 4 is dedicated to the fuzzy logic controller design on Simulink, and the simulation results are analyzed in the following section. Finally, conclusions are drawn based on the results obtained, and recommendations are made accordingly in the last part.

2 Methodology

Before supplying water in a leaky distribution network which may be old and sometimes not worth repairing [5], it is essential to detect critical leaking points so that they can be isolated in order not to penalize other customers by a subsequent pressure drop. In this connexion, a step test can be performed prior to distribution, whereby the flow valve is opened fully and closed after a predetermined time depending on distance of the investigation nodes from the valve and then analyzing the pressure transients at different points of interest. In the experimental setup shown in Fig. 1, pumping and storage are carried out first, and then, distribution is done in two loops which are embedded with leaks at points P2 and P4. All the nodes are fitted with pressure sensors, and demand valves are partially opened. Solenoid valves EV1, EV2, EV3 and EV4 are included for the isolation of sections of the pipeline when leak becomes critical.

First, a simulation without leak of the pressure response to a step input of V5 or V4 is recorded through all the pressure sensors PS1 through PS7 installed at the nodes. Both flow and pressure measurements are available for analysis at the concerned point but since pressure is more sensitive to leaks, they are preferred to flow. Then, using the identification toolbox of MATLAB, a model of the leak free state is built.

The exercise is repeated for all combinations of leak possibilities in pipes P2 and P4, whether low or high, and the pressure characteristics are compared to the ideal

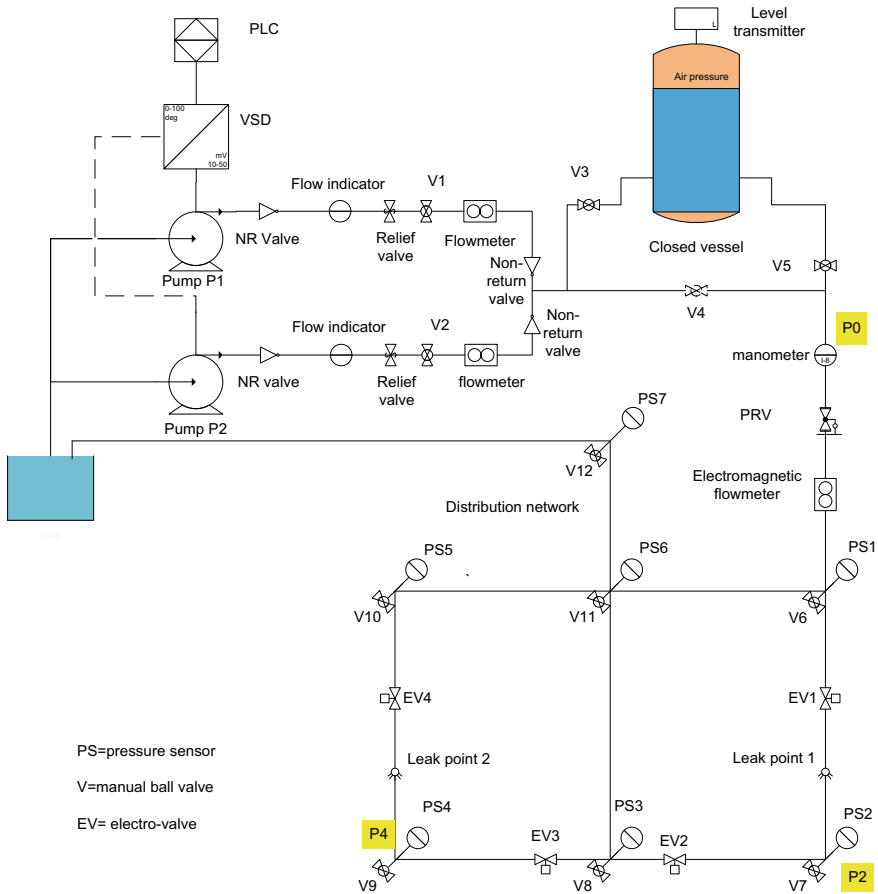


Fig. 1 Experimental setup [6]

model for residue generation [6]. The concept of low leak is based on the fact that less than 10% of the total volume of water produced is lost, and high leak assumes that it is greater than or equal to 10%. In a real situation, these leaks can be simulated through fire flows installed in subsidiary trunks. As shown in Fig. 2, pressure residues will be generated characterizing the behaviour of all possible faults at different nodes. Analysis of this matrix gives valuable information about parameters like the time constant, the time delay, the amplitude and the decay time for each of the fault combinations in terms of nodal pressure residues. Since there is no clear demarcation between the low and high leakage, there can be overlapping of residue characteristics. Therefore, a fuzzy logic control approach is used for this analysis. Based on the inference rules of the Mamdani method [7], a decision of the fuzzy logic controller is then taken before implementing any isolation, especially when the situation becomes critical. The system will normally react automatically either by adjusting the pumping

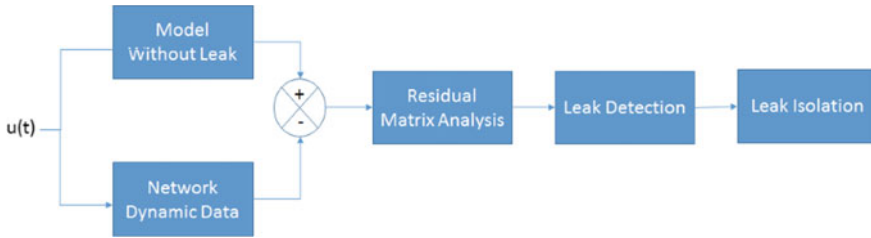


Fig. 2 Block diagram of scheme

regime in the network or isolating the damaged pipe without disrupting water supply for repair in case of high leakage using the solenoid valves is also envisaged.

The first parameter that will determine the level of leakage in the network is the magnitude of the residue generated after comparison with the leak free model. This amplitude is also affected by the distance of the leak from the measuring point. The second parameter is the decay time after removing the step input. Finally, the rise time of the response also provides some information about the nature of the leak. A close analysis and comparison allows a good decision to be taken promptly by the controller.

2.1 Transient Analysis of Pressure

The test has been conducted for the scenario of pumping with a VSD, whereby pressure at the entrance of the network is adjusted according to consumption. Simulation of the step input is carried out by simply opening abruptly the valve at the entrance of the network either V4 or V5 with the flow modulating valve of demand set at its minimum opening of 30%. The response of the system without leak with a step input for pressure approximates a first order as shown in Fig. 3. Three parameters are available for analysis with the pressure profile as mentioned above. Since the experimental network is small with only eight pipes, pressure residual analysis will be the easiest method. For a complex real network, flow rate or a combination of both will be more appropriate when doing the step testing.

2.2 System Identification

First, we obtain the leak free model of the system using the system identification toolbox of MATLAB. In order to obtain its transfer function, the experimental data is imported with the step input, preprocessed or filtered, validated, and then an estimated model is built using the polynomial approach as shown in Fig. 4.

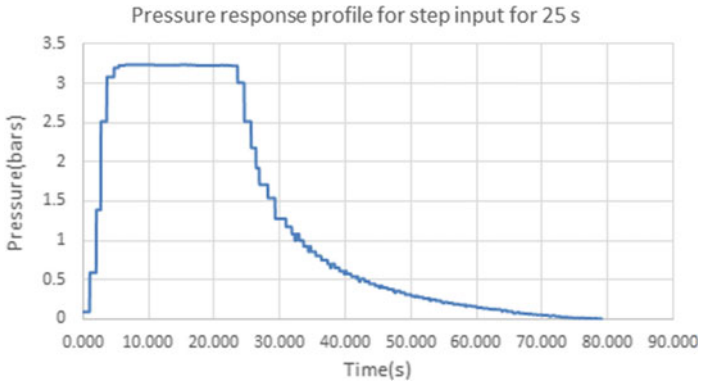


Fig. 3 Pressure response to step input

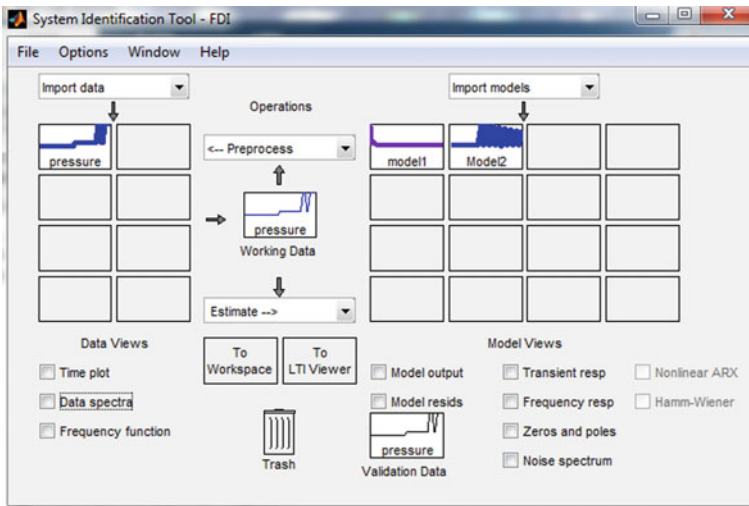


Fig. 4 Identification of system

In the example of Fig. 5, the best fit is 94.63% for model 2 compared to 93.76% for model 1 with satisfactory autocorrelation and cross-correlation given by Fig. 6. The model is then stored in memory for comparison with faulty pressure measurements.

3 Residual Generation and Analysis

For the experimental hydraulic simulation, a small leakage is typically of the order of 3–5% of overall consumption, whereas a high leakage is above 10%. Objectively,

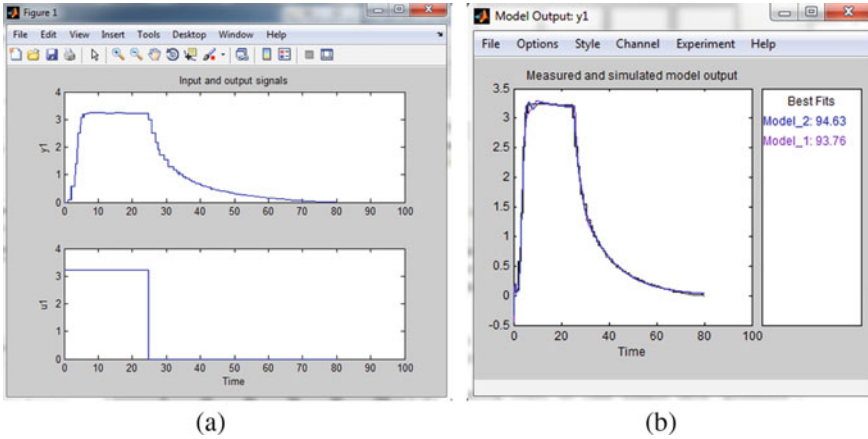
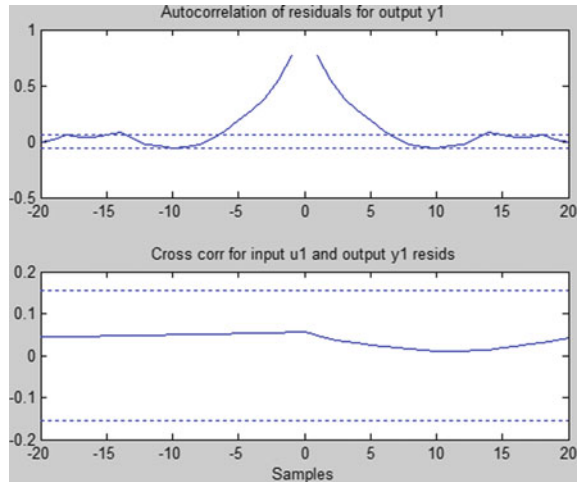


Fig. 5 Modelling the response of pressure nodes

Fig. 6 Auto and cross-correlations of model



even 3% can be considered a high leakage depending on the daily consumption according to the international water association [5]. Leak detection methods for small leakage such as background are not very accurate and precise in nature. For this showcase, we shall stick to the same percentages for high and low leakages using manual valves in order to distinguish between critical and non-critical situations. For example, there can be a single high leak at P4 (case 5) or a high leak at P2 and a low leak at P4 (case 7) as depicted in Table 1. Overall eight scenarios are possible as described below.

The first step of the scheme is to generate the residuals for each of the pressure profiles associated with the eight above cases. Step input to the system with and without leaks is actually simulated by opening and closing of the main valve at

Table 1 Possible scenarios of leakage at P2 and P4

Case	1	2	3	4	5	6	7	8
P2	LL	X	LL	HL	X	HL	HL	LL
P4	X	LL	LL	X	HL	HL	LL	HL

LL Low Leak; HL High Leak; X No leak

the entrance after a period of 25 s. The graphs of Fig. 7 compare the step pressure response profiles of the leaky system with that of the leak free at measuring points.

First, it can be observed that the amplitude of the pressure profile is highly dependent on the extent of leakage. For example, it ranges from 1.76 to 2.23 bars for both high leaks. The second parameter that can be used for identification is the decay time of pressure when the valve is closed. For the same case, it can vary between 6.6 and 8.4 s. Finally, the third differentiating factor that can be used for identification is the rise time. For the particular case, it varies between 8 and 8.4 s as can be observed in the last row of Table 2.

Checking the consistency of analytical redundancy is normally achieved by comparing measured signals with their estimates in model-based FDI. The difference usually known as the signal residual is given by:

$$r_i = y_i - \hat{y}_i,$$

where y_i denotes the i th measured system output, and \hat{y}_i denotes i th estimated system output.

The graph of Fig. 8 shows the residual R1 generated for the first case of Table 1, i.e. a low leak at P2 only. The amplitude especially at steady state and sign of these residuals (R1 to R8) are stored in a database. They will be used as a benchmark for comparison with online residuals that are produced when the step test is carried out without any prior knowledge of the state of the network.

In order to distinguish between the leaks, it is important to evaluate the residuals' characteristics using a reliable algorithm which can be easily implemented and does not require much resources computationally. With regard to the amplitude, the decay time or rise time of the pressure profiles and the associated residuals of the different leak cases, a series of simulations were carried out for each kind of leak to determine the mean (μ) and its standard deviation ($\mu-\sigma$), σ being the variance, by different % opening of the manual leak valve within the limit of high and low leakage. Table 3 summarizes the results obtained.

With regard to the mean values and their deviation for the pressure signals or their residuals, it has been observed that there is an overlapping of states of the parameters. One of the simplest methods that can be applied to this detection problem along with the human intuition is the fuzzy logic approach. A fuzzy inference set of rules can be defined using the data contained in Table 3. The Simulink implementation of the fuzzy system is performed using the inference rules governing each of the eight output possibilities with the three input parameters of residual magnitude, decay

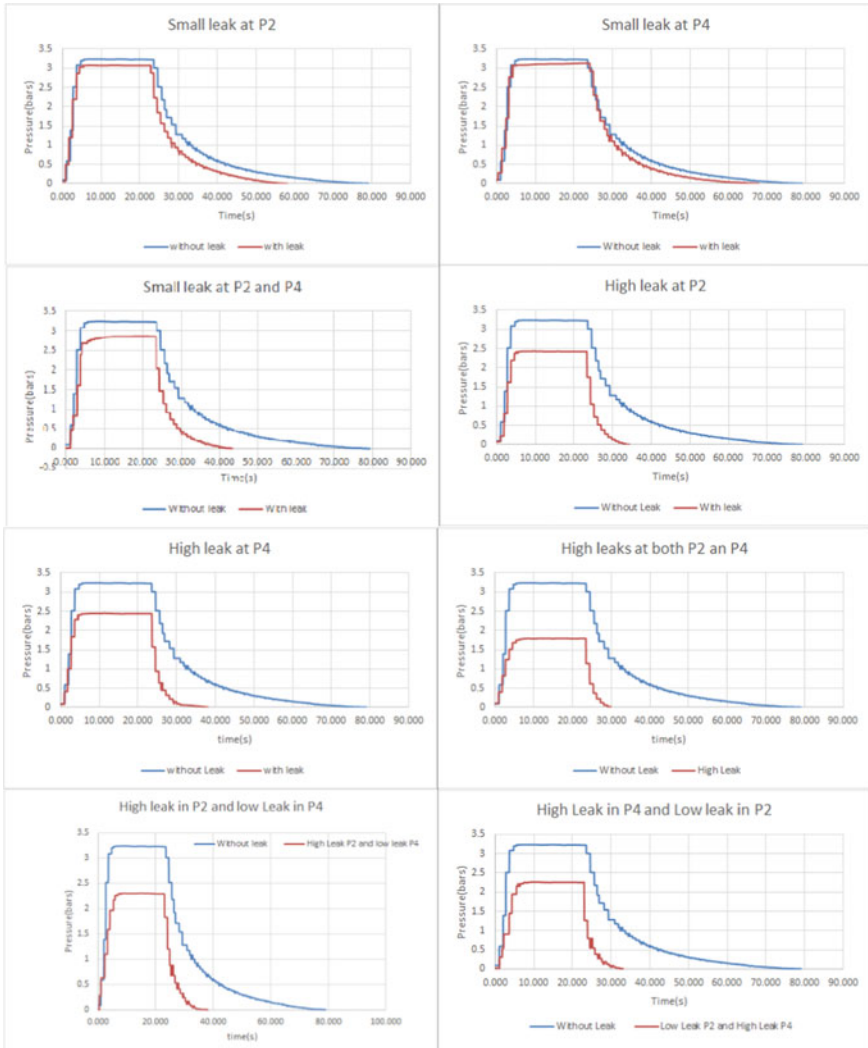


Fig. 7 Leaky and leak free step responses

time and rise time of pressure profiles. The Mamdani-type [7] fuzzy system has been adopted for residual analysis, and the procedures are outlined below:

1. Design the membership function for each input fault using the statistical data gathered in Table 3. It is assumed that all faults of interest are known (100%), and for each input parameter (amplitude, decay time and rise time), there are three membership functions states, namely low, medium and high as shown in Fig. 9. The Gaussian law is applied with the mean value and the deviation found in Table 3. For each fault, there are two output membership functions for the

Table 2 Range of variation the pressure transient characteristics

Category of leakage	Step Ampl. (bars)		Rise time (s)		Decay time (s)	
	Min	Max	Min	Max	Min	Max
Single low	2.98	3.23	5.2	6.2	34.8	55.4
Both low	2.79	3.07	6.2	6.3	20	34.8
Single high	2.39	2.47	6.3	6.5	14	20
Both high and low	2.21	2.42	6.5	8	10	14
Both high	1.76	2.23	8	8.4	6.6	10

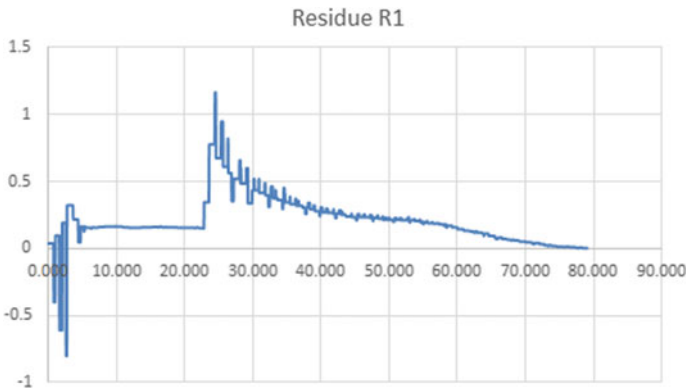


Fig. 8 Residue behaviour of first pressure node

Table 3 Mean and standard deviation of input parameters

Leak/residue	Magnitude		Decay time		Rise time	
	μ_R	$(\mu_R - \sigma_R)$	μ_d	$(\mu_d - \sigma_d)$	μ_r	$(\mu_r - \sigma_r)$
High	1.44	0.05	6.53	0.42	5.51	0.365
Medium	0.92	0.06	12.4	1.06	6.26	0.37
Low	0.82	0.03	13.4	0.8	8.13	0.21

pipes P2 and P4, either true or false, (Fig. 10) because we are interested only in high leakage cases in either P2 or P4 or both.

2. In the presence of each leak, the behaviour of the residual magnitude with respect to the nominal value is determined.
3. The inference rules are derived in terms of the membership functions for each of the outputs as shown below. There is a set of nine rules which govern the outputs. If more leaks are present, then the set of rules is likely to increase but the algorithm would be more complex to figure out. In such a case, further truncation of the system with valves may be necessary before applying the method. For the

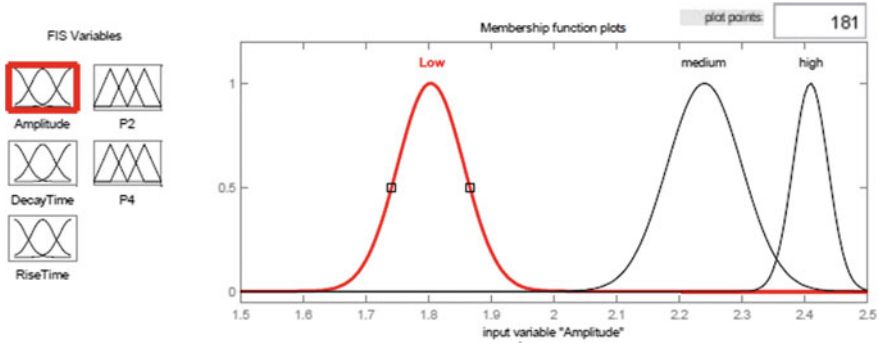


Fig. 9 Membership input functions

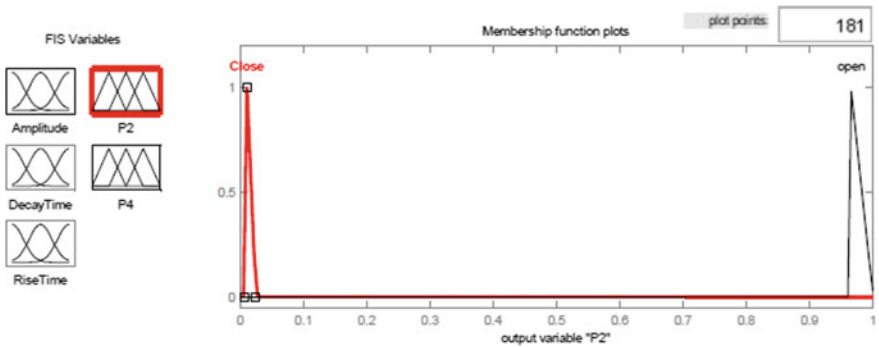


Fig. 10 Output variables

purpose of isolation here, only high leaks are taken into consideration. This summarized in the rule base of Fig. 11. Weights are assigned to the rules based on their relevance and validity.

1. If (Amplitude is Medium) and (DecayTime is Low) and (RiseTime is Low) then (P2 is open)(P4 is Close) (1)
2. If (Amplitude is Medium) and (DecayTime is Low) and (RiseTime is Medium) then (P2 is Close)(P4 is open) (0.5)
3. If (Amplitude is Medium) and (DecayTime is Low) and (RiseTime is high) then (P2 is Close)(P4 is open) (1)
4. If (Amplitude is Medium) and (DecayTime is Medium) and (RiseTime is Medium) then (P2 is Close)(P4 is Close) (0.5)
5. If (Amplitude is High) and (DecayTime is Low) and (RiseTime is high) then (P2 is Close)(P4 is Close) (1)
6. If (Amplitude is High) and (DecayTime is Medium) and (RiseTime is Low) then (P2 is open)(P4 is Close) (1)
7. If (Amplitude is High) and (DecayTime is Medium) and (RiseTime is high) then (P2 is Close)(P4 is open) (1)
8. If (Amplitude is High) and (DecayTime is Medium) and (RiseTime is high) then (P2 is Close)(P4 is Close) (0.5)
9. If (Amplitude is High) and (DecayTime is Low) and (RiseTime is Low) then (P2 is Close)(P4 is Close) (1)

Fig. 11 Inference rules

4 Fuzzy Logic Controller Design

Isolation is here applied to extreme cases of heavy leaks in P2 and P4. For small leaks, the pressure or flow rate can be adjusted using the VSD when demand is low. The fuzzy system’s output has been defined as either true or false but based on the membership functions’ behaviour, we may fall in between especially if the input states are medium. In such a situation, one can use intuition to take a decision of whether to isolate or not. Here, we simply set a rule, whereby a nominal value of 0.5 is taken as reference. If the concerned output is more than 0.5, then we go for isolation. Otherwise, no decision is taken. This algorithm is embedded into a fuzzy controller that is used to implement the LDI [8] scheme online under Simulink in Fig. 12 below. The signal builder inputs the step like pulse to the leak free model generated by the system identification toolbox. Simultaneously, the entrance valve on system is automatically opened. The pulse lasts for only 25 s, and the valve is then closed.

Residuals are dynamically generated by comparison of the leaky response from the OPCRead block and the leak free model. In the MATLAB function block, the amplitude of the residual, the decay time and the rise time of the response are extracted and fed to the fuzzy controller, one for each output. Next, defuzzification of the controller takes place to obtain crisp values of the output based on the fuzzy set of rules constructed from available database. The output can take either a true or a false state. If the threshold in the switch linked with controller output is greater than 0.5, then a value 1 is written to the PLC tags of the concerned solenoid valves to isolate the leaky pipe.

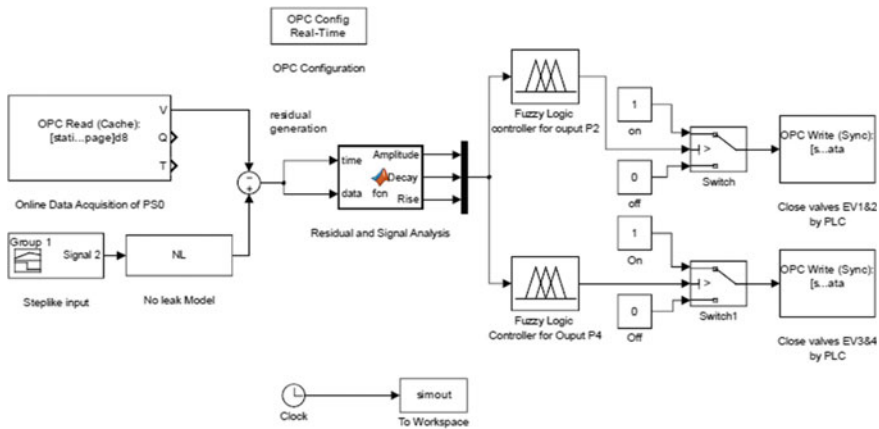


Fig. 12 Simulink implementation of fuzzy logic control

5 Results of Simulation

Having taken a decision after the step test, its implementation is carried out by the PLC. An example of heavy leak is simulated in the network at pipe P2. The corresponding residual signal and the rule viewer output are shown in Figs. 13 and 14, respectively. As can be seen, the amplitude of the generated residual at steady state is 0.92, the decay time of the signal (not the residue) is 8.47 s and the rise time is 8.34 s. From Fig. 13, one can see that the amplitude of the response is close to the peak of the high distribution state, the decay time of 8.47 s lies within the characteristics of medium leaks from Table 3, and finally, a rise time of 8.34 s confirms this high leak from the same table. The combined effect of the magnitudes of the three parameters from the inference rules of Fig. 11 leads to P2 taking value of 0 meaning ‘close’, whereas P4 = 1 (open). Consequently, this entails the isolation of pipe P2.

It can be observed from Fig. 14 that the system reacts after 17 s to close down valves EV1 and EV2 such that there is an increase in the network pressure from 2.25 to 3.2 bars and hence isolating P2. Other scenarios of heavy leak simulation as per Table 1 have produced the graphs of Fig. 15.

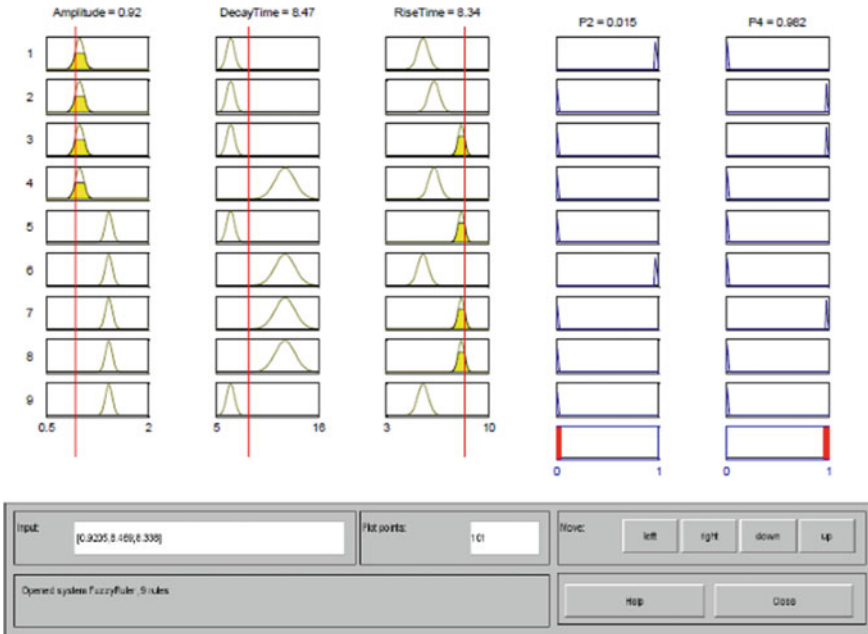


Fig. 13 Simulation of fuzzy system

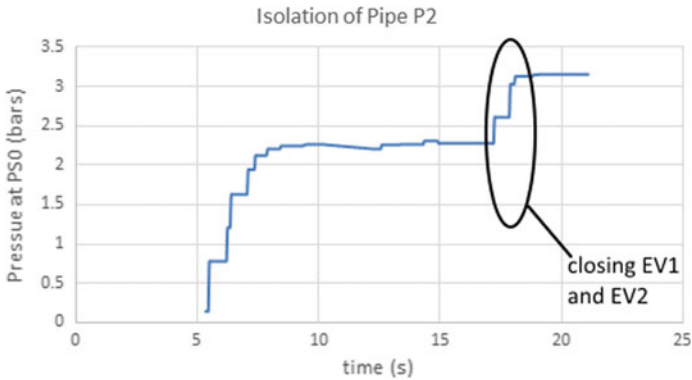


Fig. 14 Leak isolation of faulty pipe P2

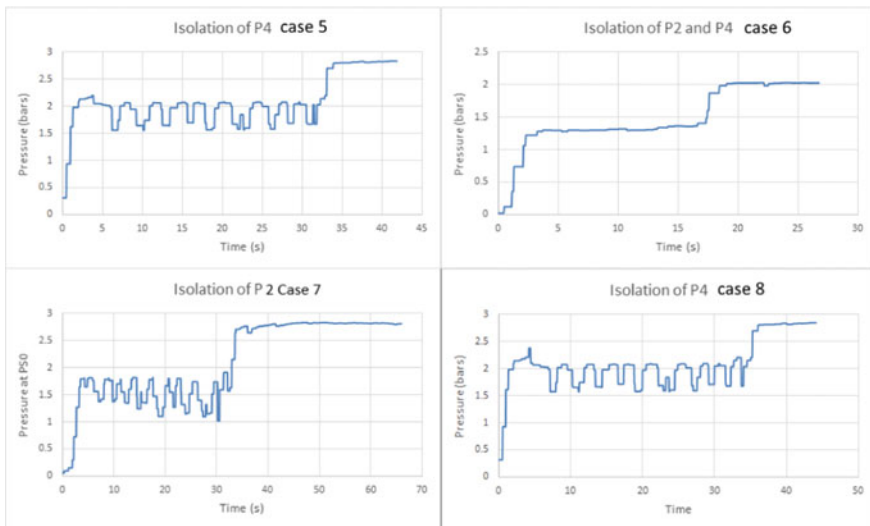


Fig. 15 Leak isolation of other cases

6 Conclusion and Recommendations

A simple leak detection and isolation scheme based on the fuzzy logic approach have been adopted in this section. The method is inspired from the conventional step testing which is still practiced today for leak detection. This innovative approach is more precise as it first determines the residual generated by the leak free system and the faulty system and then extracts the residuals' magnitude, the decay time and the rise time of the dynamic response signal which is then fed to a fuzzy controller that takes a decision whether to isolate the faulty pipe (s) or not. This decision is executed

by the PLC which is driven online by an OPC server. The results obtained through the LDI scheme are very promising but it needs to be tested on a real network.

References

1. Massoumnia, M. A., Verghese, G., Willsky, A.: Failure detection and identification. *IEEE Trans. Automat. Contr.* **34**(3), 316–321 (1989)
2. Bedjaoui, N., Weyer, E.: Algorithms for leak detection, estimation, isolation and localization in open water channels. *Control Eng. Pract.* **19**(6), 564–573 (2011). <https://www.sciencedirect.com/science/article/abs/pii/S0967066110001498>
3. Araujo, L., Ramos, H., Coelho, S.: Pressure control for leakage minimisation in water distribution systems management. *J. Water Res. Manage.* **20**, 133–149 (2006)
4. Latchoomun L., Ah King R.T.F., Busawon K. K., Ginoux J. M.: Harmonic Oscillator Tank: a new method for leakage and energy reduction in a water distribution system with pressure driven demand. *Energy* (2020). <https://doi.org/10.1016/j.energy.2020.117657>
5. International Water Association: <https://iwa-network.org/>. Accessed 20th May 2020
6. Latchoomun L.: Pressure control for leakage reduction in the water distribution network of Mauritius, Ph.D. thesis, chapter 8, 196–215 (2019)
7. Pourjavad, E., Shahin, A.: The application of mamdani fuzzy inference system in evaluating green supply chain management performance. *Int. J. Fuzzy Syst.* **20**, 901–912 (2018). <https://doi.org/10.1007/s40815-017-0378-y>
8. Boaz, L., Kaijage, S., Sinde, R.: An overview of pipeline leak detection and location systems. In: *Proceedings of the 2nd Pan African International Conference on Science, Computing and Telecommunications (PACT 2014)* 133–137 (2014)

Robo-Friend: Can a Social Robot Empathize with Your Feelings Effectively?



Eshtiak Ahmed, Ashrafal Islam, Atiqul Islam Chowdhury,
Mohammad Masudur Rahman, Shahnaj Chowdhury, and Md Imran Hosen

Abstract Social robots are becoming more popular everyday because of their resemblance with human behavior and interaction styles. They should be treated more like companions rather than just a fancy source of entertainment. Recent studies have shown great promise for robots to act as teachers, companions, caregivers and so on. In this study, primarily, a feasibility analysis is done to find out the way how a social robot can be used as a companion where it can sense the emotions of the users, empathize with their feelings and provide feedback with a view to changing their mood. A context study was conducted to make design implications where user experience goals such as inspiration, sense of control, relaxation, accomplishment and confidence were considered in the implementation. Furthermore, a prototype was designed based on a social robot, Pepper which then was interacted with potential users, either in groups or individually. The results support the fact that, a well-implemented social robot can effectively empathize with human users' emotions.

E. Ahmed (✉) · S. Chowdhury · M. I. Hosen
Tampere University, Tampere, Finland
e-mail: eshtiak.ahmed@tuni.fi

S. Chowdhury
e-mail: shahnaj.chowdhury@tuni.fi

M. I. Hosen
e-mail: md.hosen@tuni.fi

A. Islam
University of Louisiana at Lafayette, Louisiana, USA
e-mail: ashrafal.islam1@louisiana.edu

E. Ahmed · A. Islam
Daffodil International University, Dhaka, Bangladesh

A. I. Chowdhury
United International University, Dhaka, Bangladesh
e-mail: achowdhury201036@mscse.uju.ac.bd

M. M. Rahman
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
e-mail: masudurism@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*,
Advances in Intelligent Systems and Computing 1299,
https://doi.org/10.1007/978-981-33-4299-6_63

Keywords Human–robot interaction · User experience · Experience goals · Experience driven design

1 Introduction

Robots are becoming more and more common everyday. They are being adopted into almost all the aspects of human life ranging from day to day life to even critical medical technology. Robots have enough potential to be useful if used in proper context. There are so many situations in human life where robots can still be introduced especially in social situations.

Social or conversational companions are one of the important needs for humans to have sound mental health. They always try to express their feelings to others whenever they are sad, angry or feeling negative emotions, giving them a sense of empathy from the opposite and provide them with the sense of belonging. However, the society nowadays is becoming more and more monogamous [1], and people are starting to rely on technology as more of a companion than products. In all these circumstances, a robot could work as a viable companion to human beings [2]. They can be employed as companions that empathize with the user and provide feedback based on the emotions of the user. As an example, if the user appears to be sad, the robot can provide some feedback to make the user feel better.

There has been some previous work [3–5] which emphasizes robots being teachers or companions. But there are a lot of contexts where more work can be done. In this project, the primary target is to evaluate a social robot's functionality in having and performing social behavioral skills. We also aim to evaluate if a social robot can act as a friend who can react and empathize with people's feelings and provide consolation or feedback at the same time.

This study initially focuses on the feasibility of a robot being able to work as a companion. User studies are conducted to know and analyze the perception of the users in the scenario where robots can empathize with them and provide meaningful feedback depending on their mental conditions. Finally, a prototype evaluation has been done to support the facts.

This paper is organized as follows. Section 2 will discuss related works regarding this area. The design approach and context study will be discussed in Sects. 3 and 4. The explanation of the prototype design and its evaluation will be elaborately described in Sects. 5 and 6, respectively. The rest of the section is about the discussion and conclusion of this article.

2 Related Work

In recent times, there has been a good number of studies involving robots in the daily lives of people. Especially, there have been many studies on robots acting as teachers of a second language. In [3], several aspects of using a social robot as a language

learning companion were assessed which included the potential application as well as limitations. This study mainly focused on improving the vocabulary and grammar as well as speaking and reading skills. This paper reported 33 different studies which focused on robots' effect on children's learning motivation, and the ways a robot should behave to maximize learning outcomes. In [4], some features of design were proposed keeping a social robot in mind which would be child-friendly, for assisting in learning a second language. Both studies discuss the issues of motivation factors as well as social behaviors of learners.

Robot-assisted language learning (RALL) has been focused in some of the studies which include the types of robots that can be employed [3]. Robots have also been employed as behavior teachers for children with autism spectrum disorder having delayed gestural development [6].

There have also been some studies which have tried to find out the role and influence of empathy in human-robot interaction. Robots seem much more friendly and approachable if they empathize with people [7]. In this study, a robot was used during a chess match which was showing different types of emotions to a specific player and was neutral for the other. Thirty-one participants were recruited for this study, and all of them found the robot to be emphatically friendly. These studies show that robots can evoke emotions into human minds and can be used as companions and provide feedback based on the users' situations. These kinds of applications in robots can be developed using AI and intelligent algorithms [8].

3 Design Approach

Whenever a new product is designed, it needs to follow certain characteristics in order to be appealing to the end user. User experience models usually provide a better understanding of how people perceive and value objects as well as providing common operationalization and easy measurement of key elements. In this project, the components of user experience (CUE) model [9] has been used to evaluate user experience values. This model takes into account many aspects of a system such as interaction characteristics, emotions of users, both instrumental and non-instrumental qualities, all in all the overall perception of the system's quality. It takes interaction characteristics into account to evaluate user experience components. There are specific components of user experience by which the behavior of any application or system is defined. A recent study [10] has shown the use of this specific model where they employed this model to understand the players' appraisal of their newly designed game.

The experience-driven design approach has been followed here while designing the prototype. This starts with a "why" question which helps define the reason why the user will be using the product. Then comes a question which tries to identify the needs and emotions involved with the product and what the users actually can do with the product. Then, there is the "how" question which tries to answer how the users interact with the product. Why goals are also called the be-goals which

define the experience of the product to the user. Defining experience goals rather than functionalities can help connect with the user more which could result in better response from the user [11]. The idea of creating an experience-driven design is to define which experiences we want the users to have while using a particular design. While this is opposite of features driven designs, it is more effective in the long run as it is the experience that makes or breaks a system for a user.

4 Context Study and Conceptualization

To understand the users and the expectations better, context studies are very important. For this project, the context study was conducted in order to understand how potential users perceive the idea of having a social robot as a companion which can also empathize with their mental conditions. We plan to conduct studies with users and report in this section on later phases.

To get a better understanding of what a potential user would want, 16 university students, age ranging from 21 to 28, were interviewed. Seven of them had previous experience of interacting with a robot, while the other nine did not have any or had seen people interacting with robots at best. Most of them reported positive experiences with robots, while two of them reported some awkward and frustrating experiences.

The interview questions were designed to ask users about how they would like a social robot as a companion. Responses were analyzed on a topic basis, such as “what a robot should be able to do as a companion”, “how a robot should behave in a certain situation”. This provided an idea about the design implications of the prototype. This also helped to understand what a user can expect from a companion robot in the context of providing feedback for different mental states. The participants were also asked about what type of response they expect from the robot in mental conditions like sad, angry, stressed, bored and depressed.

When asked about what a companion robot should look like, most of them answered that it should look like a human and mimic human behavior. Two participants were neutral about their appearance but did prefer some physical appearance. In the user study, we asked participants what a robot should do when they are sad, angry, stressed, bored or depressed. Following are the suggestions that were collected from the results of the questionnaire.

- **Sad:** Touching or patting, singing a song, dance, offer to play fun games, show nice pictures, being funny in appearance, have a conversation.
- **Angry:** Music, show negative effects of anger, say motivational quotes about anger management, listen to the yelling and confession of the user.
- **Stressed:** Relaxing music, dance, tell jokes, make food
- **Bored:** Tell a joke, dance, play music, read a book
- **Depressed:** Motivational words, tell stories, show achievements of user.

5 Design Explained

This section will discuss our experience goals and prototype design in the following subsections. For this research, we had analyzed some activities of potential users based on some different phases. That analysis is described in the Sect. 5.1. After analyzing these activities, then how we designed our prototype, is described in the Sect. 5.2.

5.1 Experience Goals

In the initial phase, there were some pre-defined experience goals to create initial design implications which included relaxation, inspiration, empathy, confidence, assurance, sense of satisfaction, control, accomplishment and challenges. Later, while designing the experience journey map, in Fig. 1, the experience goals became more streamlined and closer to context. The experience journey map is the hypothetical portray of a potential user's journey of using a system. Here, the journey is divided into three phases, before using the robot as a companion, during the interaction and after using it. Three activities of potential users were quantified to understand their experience during the three phases. These activities are "doing", "thinking" and "feeling". These three activities combined with the phases were then concluded into specific experience goals for each phase.

From the experience journey map, initially defined the experience goals as satisfaction, inspiration, sense of control, relaxation, accomplishment and confidence were defined. In the later phases of prototype development, the experience goals became even more defined and clear. The final experience goals were happiness, inspiration, relaxation, calmness and accomplishment.

Users' perception on both instrumental qualities such controllability and effectiveness can be portrayed as some of the experience goals like sense of control and confidence. Users' emotional reactions are also taken into account. The experience goals are discussed in the following section along with their connection with the components of the CUE model.

- **Inspiration:** A user may get inspiration by the positivity of the conversation and stories shared by robot. The conversations with the robot can be designed in such a way that it will exhibit positivity. This can be related to components like visual aesthetics and psychological reactions.
- **Happiness:** Happiness can be created in the mind of the users by listening to their favorite songs, jokes, seeing nice pictures or even by interacting successfully with the robot. This experience goal can be related to controllability and psychological reactions. Users are supposed to get a positive experience.
- **Relaxation:** A sense of relaxation can be invoked through music, dance and motivational stories. Also, the robot should talk in a voice that could be relaxing for the user. This can be related to effectiveness and haptic quality.



Fig. 1 Experiences journey map

- **Accomplishment:** After the interaction, a user generally greets the robot and leaves the premises. During this time if he/she had feelings like satisfied, happy, calm and inspired. Being able to overcome negative mental situations by using the robot can create a feeling of accomplishment in the mind of the user. This can be connected with controllability.
- **Calmness:** Calmness can be achieved by the user through relaxing music, seeing nice pictures or even by listening to a joke.

5.2 Prototype Design

When deciding which robot to use to create a prototype for, users’ perception of companions was taken into account. Majority of the users so far from the context study mentioned that they would like a human-like figure for a companion robot. Also, the robot will need to have the ability to have a proper conversation with them. Keeping all that in mind, we decided to design a prototype on the Pepper [12] robot which is a semi-humanoid robot manufactured by SoftBank Robotics.

To design the prototype, four mental conditions of the human mind were chosen. When a user approaches the robot, it provides four options to choose from to proceed.

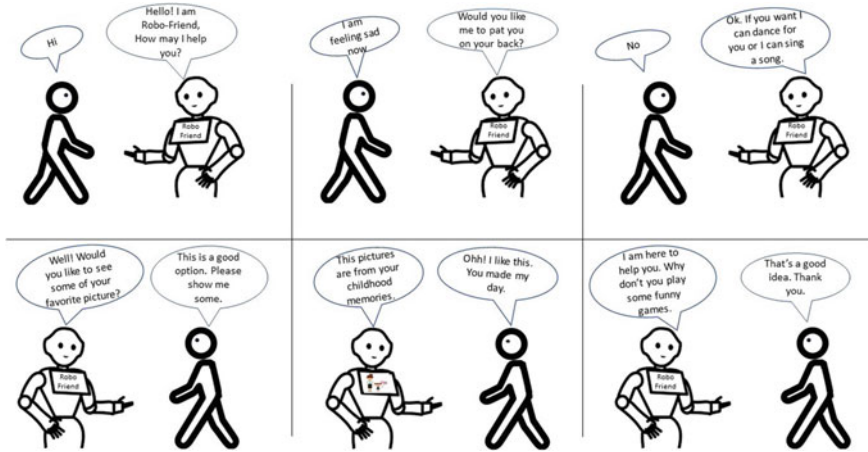


Fig. 2 Storyboard: when the user is sad

The options are “sad”, “angry” “stressed” and “bored”. Based on the choice of the user, the robot will offer to tell jokes, play music, show nice pictures, etc. A storyboard is designed to explain a scenario on how robots will deal when a user is feeling sad, in Fig. 2.

The use context of this application can be either at home or at a workplace. In this study, it has been considered more suitable for use in an organizational context where different people can come and express their feelings with a view to improve their mood. The robot could be kept in a separate room, and people can enter and use the application.

6 Prototype Evaluation

The prototype was evaluated in two phases, first one was at prototype demonstration at the university with groups of participants, and the second one was conducted with individual participants.

6.1 Evaluation Methods

In the first phase of evaluation with groups, the designed prototype was presented to the potential users before doing the prototype evaluation to make them understand the functionalities and how to use and interact with it, and also, the purpose of the evaluation was explained. After the presentation and demonstration, participants were asked to interact with the robot based on their own preferred choice of func-

tionalities. During the participants' interaction with robots, behavioral observations were conducted which were followed by semi-structured interviews. In the second phase, a mixed method [13] and meCUE [14] model for evaluation.

To get both the qualitative and quantitative data, mixed method research approach [13] was employed which consists of observation and semi-structured interview option to complement each other. Observation was done with a structured observation sheet to identify the participants appearance, approaching style to robot, selection of prototype functionalities, way of interaction, the emotional reaction with ten other observation criteria and three open-ended observation questions. Semi-structured interview was conducted with a set of 11 questions to collect the data where questions were grouped into background data, experience and emotion evoked during the interaction with robot and the last group participants own thoughts, expectation and suggestion.

Along with the mixed method research approach, meCUE [14] model was also used to identify the dimensions of usefulness, usability, visual aesthetics, positive/negative emotion which is related to our defined experiences goals.

6.2 Participants

In demonstration day evaluation, the participants were the students of the university coming in groups. A total of five groups (2–4 members) of adult male and female participants took part in it. In second phase evaluation, participants were recruited randomly from the corridor of university who were passing by. Among 13 individuals, three participants are female, and other ten were male, and the age range was 19–34. All the participants were student and from different nationalities, such as France, Vietnam and Bangladesh.

6.3 Results

From the results of the observation of both phases, it became clear that the participants were curious to interact with the robot and also they were energetic while interacting. The participants tried different functionality and features of the prototype. Most of the participants enjoyed the joke and song feature of the prototype, and they specifically mentioned that the prototype and robot interaction was cool and enjoyable. They also rated the features of song, joke, pictures and motivational quotes. Table 1 shows the features and how many participants used each feature during the evaluation. In Fig. 3, the ratings given for each of the features from 6 of the users are shown. There were only six users who explored all the features, thus this presentation.

From the interview, it was noticed that the participants' experience with the robot was good, enjoyable and they found the robot as friendly and informative. But it has some balancing and movement problem as per two participants. The feeling during

Table 1 Features and their usage

Features	Participants
Sad	13
Bored	8
Stressed	6
Angry	6

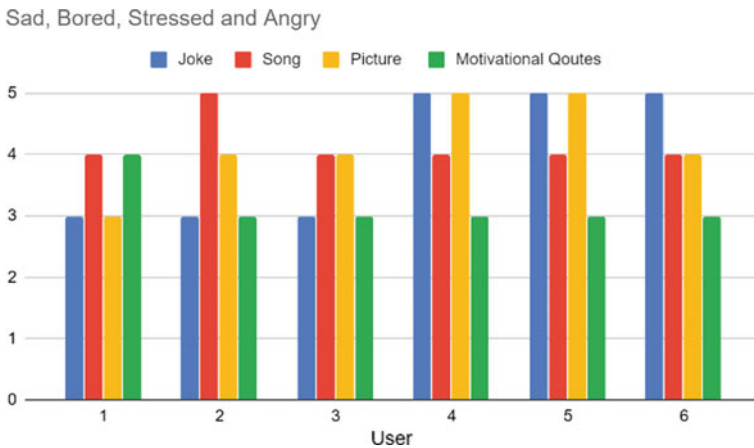


Fig. 3 Rating of different features from the range of 1–5

the interaction was varied in different participants. And they also found the robot friendly, entertaining and it did not do any scary or rude behavior to them. Figure 4 shows the participants’ impression on the prototype and its behavior.

The participants are preferred to interact with the robot through touch and speech modality. All the participants showed their interest to use the robot as companion for empathy except one and the encouragement behind using it is the humor level, intelligence, easy to share feeling to it then any human and its friendliness attitudes. Figure shows the feedback ratings from the user.

However, they think there are some challenges also in this context like identifying the emotion properly, voice detection, more human-like behavior and level of artificial intelligence. Comment from one of the participants (P2, age:19, male) “I would like to have eyes interaction and the robot can realize the distinct people talking to it , so it does not repeat something it told to me, example the joke” represents the level of intelligence user expects from this type of service.

From the results of meCUE questionnaire, it revealed that, the participants found the system as useful for the context, and the system was also appealing to them due to the overall aesthetics. It was able to evoke positive emotions to the participants while they were using it. Figure 5 shows the meCUE model results for usefulness, usability and aesthetics.

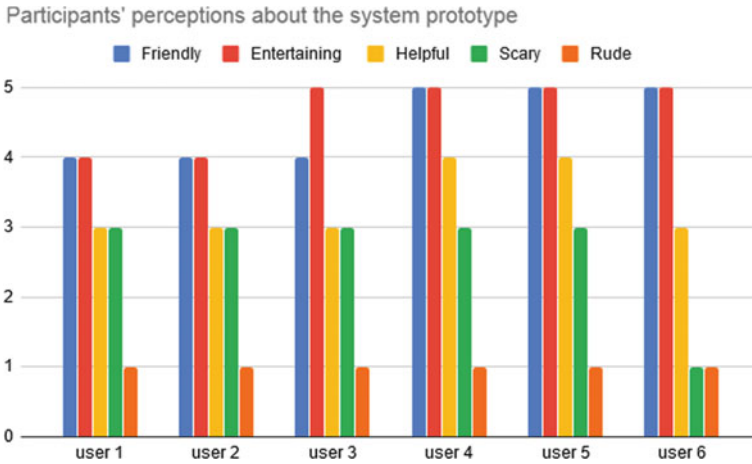
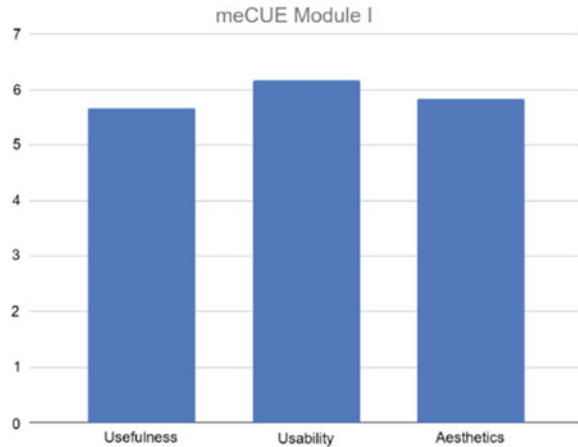


Fig. 4 Results from the interview about the feedback of robot behavior

Fig. 5 meCUE model results for module I (usefulness, usability and aesthetics)



7 Discussion

This study was focused on understanding the potential usability of a robot working as a human companion. The initial user studies point towards a growing interest in robot-based application and some specific requirements. In the beginning, a lot of experience goals were defined to go through with the experience-driven design process. However, as the time progressed, some of them seemed very high level and superficial. It was also very unclear how these experience goals can be evoked and measured at the same time. Further discussions and primary prototyping decisions helped narrow down the experience goals to some realistic ones before starting the prototyping phase.

When prototyping started, some more experience-driven aspects came to light. It was realized that some of the experience goals were still superficial and difficult to evoke. As a result, slight changes were made to them, keeping the main idea intact. While trying to measure if the experience goals were actually evoked or not, some of them were evident to be evoked, while others were difficult to decide.

The results from the prototype evaluation portray the usefulness of the system, at least according to the limited number participants. Results show that participants found the robot as funny, entertaining and helpful. One of the participants felt shy while interacting with the robot. Overall the results were very encouraging for any further investigation.

8 Conclusion and Future Work

This study focused on understanding people's perception towards using a social robot as an emotional companion. From the results, it can be said that potential users have a very positive feeling about the idea and they also find it useful. This study has been completed with a limited number of users, and also, all the participants were university students. This makes a future work to be very much possible with more users from more diverse groups based on age and profession. Also, only four states of human emotions were considered here, and in future, more specific states can be explored. In this study, the movements and gestures from the robot end were not explored which could be something very interesting to investigate.

References

1. Conley, T.D., Ziegler, A., Moors, A.C., Matsick, J.L., Valentine, B.: A critical examination of popular assumptions about the benefits and outcomes of monogamous relationships. *Personality Soc. Psychol. Rev.* **17**(2), 124–141 (2013)
2. Dautenhahn, K., Woods, S., Kaouri, C., Walters, M.L., Kheng Lee Koay, Werry, I.: What is a robot companion - friend, assistant or butler? In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1192–1197 (2005)
3. van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., Leseman, P.: Social robots for language learning: A review. *Rev. Educ. Res.* **89**(2), 259–295 (2019)
4. Vogt, P., de Haas, M., de Jong, C., Baxter, P., Krahmer, E.: Child-robot interactions for second language tutoring to preschool children. *Front. Human Neurosci.* **11**, 73 (2017)
5. Han, J.: *Emerging Technologies Robot Assisted Language Learning* (2012)
6. So, W.C., Wong, M., Lam, W., Lam, C., Fok, D.: Using a social robot to teach gestural recognition and production in children with autism spectrum disorders. *Disability Rehab. Assistive Technol.* **13**, (2018)
7. Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., Paiva, A.: The influence of empathy in human-robot relations. *Int. J. Hum.-Comput. Stud.* **71** (2012)
8. Yang, J., Wang, R., Guan, X., Hassan, M.M., Almogren, A., Alsanad, A.: AI-enabled emotion-aware robot: The fusion of smart clothing, edge clouds and robotics. *Future Gener. Comput. Syst.* pp. 701–709 (2020)

9. Mahlke, S., Thuring, M.: Studying antecedent of emotional experiences in interactive contexts, pp. 915–918 (2007)
10. König, A., Wegener, J., Pelz, A., Gripenkoven, J.: Serious games: A playful approach to reduce usage barriers of innovative public transport systems (2017)
11. Hassenzahl, M.: User Experience and Experience Design (2011)
12. Robotics, S.: Pepper the humanoid and programmable robot. <https://www.softbankrobotics.com/emea/en/pepper>. Accessed 09 June 2020
13. Johnson, R., Onwuegbuzie, A., Turner, L.: Toward a definition of mixed methods research. *J. Mixed Methods Res.* **1**, 112–133 (2007)
14. Minge, M., Thuring, M.: The Mecue questionnaire (2.0): Meeting five basic requirements for lean and standardized UX assessment (2018)

Management and Banking Applications

Using Stakeholder Expectations and Perceptions to Guide the Brand Refresh of a Tropical Airline



Vimi Neeroo Lockmun-Bissessur, Swaleha Peeroo, and David Savy

Abstract Literature on successful approaches for executing brand evolution exercises in airlines, along with effective livery designs, is scant. This paper aims to contribute to the existing literature around successful airline branding and livery design, by reporting on multiple stakeholder engagement in the brand evolution of a tropical airline, Air Seychelles. Research gleaned from different stakeholder expectations and perceptions, at different stages, is elaborated upon. A mixed method approach was used, building on both quantitative and qualitative feedback. Initial focus groups were held to analyse the situation regarding the airline's brand perception. Further focus groups were then executed to refine the Air Seychelles brand essence. Customer surveys were done on board Air Seychelles flights across different sectors, to situate customers' viewpoints on the different brand touchpoints. From analysis of data, it was evident that Air Seychelles needed a 'brand evolution' rather than a 'brand revolution', thus guiding proposals for the refreshed brand for subsequent implementation.

Keywords Branding · Brand evolution · Airline livery redesign · Mixed methods

1 Introduction

Being a very competitive industry, airlines rely heavily on their brand to appeal to their customers. A brand is what the service stands for in people's minds and includes both visual and non-visual aspects [1]. Brands differentiate a company or a product from that of their competitors. Branding is the process of executing and managing the things that make people feel a particular way towards a given brand. Airlines use liveries to promote their brands. Liveries are powerful reflections of airline brands, as *'the colours and design motifs that are applied to the world's commercial aircraft*

V. N. Lockmun-Bissessur (✉) · S. Peeroo
Université Des Mascareignes, Udm Lab, Roches-Brunes, Mauritius
e-mail: vlbissessur@udm.ac.mu

D. Savy
Seychelles Civil Aviation Authority, Mahé, Seychelles

fleet are one of the most visible and familiar expressions of an airline's brand and corporate identity' [2].

This paper therefore sets out to use Air Seychelles as a successful case study in brand refresh and livery design. Despite being a small tropical airline, Air Seychelles has been internationally recognized for its remarkable livery.

Air Seychelles started up with domestic operations in 1978, with the launch of international services as from 1983. The logo of the airline originated in 1985 and was inspired from the winning entry of a local competition, with further rework by Saatchi & Saatchi. In view of the planned renewal of the international and domestic fleet, a brand refresh project was launched for Air Seychelles. This consisted of a review of the tag line, of the logo and a transition to a modern, refreshed livery, whilst harmonizing all collaterals and brand touch points. It was thus important to situate customers' views on the different branding aspects of Air Seychelles.

Research was hence conducted at various stages to gather data. This paper aims to determine the perceptions of stakeholders on the different brand elements of the airline and to address the following research questions:

- (a) What defines the Air Seychelles brand essence?
- (b) What are the perceptions and expectations of stakeholders of the airline's brand elements at the different brand touchpoints?
- (c) How can the airline livery be made more representative of being the national carrier of the Seychelles?

The paper is structured as follows: first, a summary of the meaning of brands and branding is introduced, followed by an overview of associated concepts such as branding elements, brand touchpoints, brand essence and airline livery. Second, the methodology used in the study is elaborated upon. The findings are subsequently reported and discussed. The paper finally concludes by elaborating on this study's implications on successful airline branding, along with its limitations and future avenues for research.

2 Literature Review

2.1 Brands

Existing literature contains a multitude of definitions of brands. It is nonetheless interesting to note that the origins of the word 'brand' go back to Old Norse *brandr*, signifying *burn*. This was because man then stamped ownership of their cattle by burning, hence allowing the differentiation of cattle ownership from one farmer to another [3].

Today's definitions of brands commonly centre around the concept of identification. Definitions range from brands being 'identification markers' that allow market differentiation of one brand to another [4] to the power to influence the markets [5]. A

brand is a concept that lives in the minds of people. They are a promise that tacitly link the customer to a given product or service. Brands evoke images, emotions, words or mental associations in people's minds [6]. Whilst some definitions of brands focus only on visual aspects, there are others which go to deeper levels beyond the visuals, symbolizing the brand essence [1]. Brands that go beyond plain visuals, incorporate psychological, emotive and intangible elements and appeal to the senses are more robust and less likely to suffer from competitive erosion [7]. Emotions being impregnated with psychological, cultural and social traits, it is therefore vital to understand the sociocultural environment in which a given brand is to be marketed [8]. Rosenbaum et al. go further by defining brands as cultural artefacts and signifiers with meaning, which affect our emotions and thoughts [8].

2.2 Branding and Branding Elements

Brands and branding are two different concepts. Branding relates to the activities that create differentiated meaning of a given brand in customers' minds. Branding is thus defined as a set of deliberate, purposeful efforts which aim to establish particular *identifications, representations and affective entanglements* with a chosen brand in a given cultural environment [4]. Branding is also defined as the composition of branding elements such as the symbolism, the name, the logo, the tagline, the colours, images and other features of the brand that either individually or otherwise constitute its identity [9].

Successful branding cannot be undertaken as a cosmetic exercise that focuses only on visual branding elements, which, nonetheless, can be used as vehicles for expressing cultural understanding [10]. Successful branding entails a deep understanding of what an organization is about and using this knowledge to build competitive advantage by incorporating this into the brand elements and the brand touchpoints [3].

2.3 Brand Touchpoints

Brand touchpoints are not solely limited to customer contact points, where the organization interacts with the customer. Drawing on Baxendales's [11] definition, brand touchpoints are instances where the customer comes into direct or indirect contact with a given brand and these interactions create an impression with customers and stakeholders [8]. In the airline industry, the customer journey involves multiple touchpoints whether it is pre-flight, during the travel journey or post-flight.

2.4 Brand Essence

The brand essence remains the same irrespective of the context and is a unique intangible attribute that demarcates one's brand from others. It is defined as a brief statement that encapsulates the most important aspects of a given brand as a whole, reducing what the brand is all about to its plain essentials in a single word, phrase or sentence [12].

2.5 Branding in the Airline Industry

Airlines face intense competition and endeavour to differentiate themselves so as to increase customer-perceived values through their branding efforts. It is therefore vital for an airline to give the public a strong and lasting impression about who they are [13]. Flights have lost the magic that they used to hold in the olden days, and with flight experiences becoming increasingly similar, branding is of significant importance in this industry.

2.6 Livery as an Integral Part of Branding in the Airline Industry

At a fundamental level, airline liveries serve the basic function of preventing aircraft body corrosion. The colours and design elements displayed on the world's commercial aircraft are, admittedly, one of the most noticeable and familiar expressions of a given airline's brand and corporate identity [2]. Through their vivid combinations of colours, design characteristics, visual motifs and typefaces, liveries are also used to allow consumers to help differentiate an airline's aircraft from those of the competitors and remain an eye-catching aspect of the service touchpoints that passengers come across in their flight journey [14].

There are however more emotive dimensions to airline liveries. Indeed, valuable lessons regarding the emotive dimensions in airline livery can be learnt from the British Airways rebranding exercise carried out in 1997. Despite colossal sums amounting to £60 million spent on the rebranding of the airline, where it decided to break away from its long-time traditional livery based on the British national flag, the exercise caused massive uproar, drew considerable criticism from the public and caused media outcry. It was an immense failure. British Airways subsequently had to revert to a more recognizable British corporate image to its aircraft fleet. As a case study of airline branding, the BA story vividly exemplifies the considerable tension which sits at the heart of the airline industry's agenda nowadays [15].

Previous research reports on the positive effects between airline livery and anticipated service and purchase intention [16], and this view is reinforced by another

study which found that other sensory experiences have considerable influence on travellers [14, 17].

Through the visual content analysis of airline liveries, Budd identifies different factors that affect airline liveries, of which an airline's geographic origin. She underlines the fact that 'particular cultural meanings' are created through a mixed use of particular typefaces, colours, designs or visual *motifs*/elements [2].

However, literature on successful methodologies for implementing successful livery design is not commonplace, and there is little guiding theory of how to design livery for airlines [14]. This paper aims therefore to add to the existing literature around successful airline branding and livery design.

3 Methodology

In order to determine perceptions of stakeholders, a mixed method approach was adopted. The mixed method approach builds on both qualitative and quantitative methods. Different studies were done at different stages of the brand refresh exercise in order to provide a wealth of information to answer the research questions, thus guiding the subsequent steps for a successful brand refresh exercise. Whilst researching on place branding, Kavaratzis (2012) used participatory branding, where stakeholders are offered a chance to participate in the branding exercise, in contrast to current practice where stakeholders are only paid lip service [18].

3.1 Focus Groups

In the initial stages of the brand refresh exercise, focus groups were held with different stakeholders, of which different levels of staff include board members, top, senior, middle management, personnel from different departments, customers from the varying market segments and key airline branding partners.

Focus groups are good techniques to capture participants' perceptions, feelings and suggestions about topics, services or products or other issues [19]. The focus groups were guided through the different topics of discussion through moderators using a structured interview protocol so as to obtain input from all participants.

The first two focus groups involved carrying out a diagnosis to analyse the current situation regarding the Air Seychelles brand perception. This included a strengths, weaknesses, opportunities and threats (SWOT) analysis. Data from ongoing passenger surveys of Air Seychelles was also used to enable the SWOT exercise. Participants' perceptions of the current brand were also gathered, as well as the key considerations in developing the new brand identity. The outcome of this initial phase was the emergence of an 'only' statement, which would underpin the flow of the brand review exercise. Thereafter, further focus groups were held to

further refine the ‘only’ statement of Air Seychelles and evaluate concrete ways in which the right branding signals could be translated onto the brand touchpoints.

The focus group discussions were recorded and transcribed. Constant comparison method of analysis was used to code and categorize data to identify similarities and differences in the collected data [20]. The researchers were constantly involved in a multiple-stage iterative process based on open, axial and selective coding in the analysis of data, to ensure accuracy and consistency in the interpretation of data [20].

3.2 Surveys

Once data was gleaned from the focus groups, preliminary work on the brand refresh proposals for both domestic and international service provisions of Air Seychelles was initiated. A survey instrument was then developed and administered to different customer segments from multiple destinations (of which some were frequent flyers) in order to further examine their perceptions at the different brand touchpoints.

Data was collected through a seven-part questionnaire that included: viewpoints on the Air Seychelles slogan, the Air Seychelles logo and livery, the airline interior, the uniforms, the service. A seven-point Likert scale was used to measure satisfaction ratings. Demographics of the respondents were also collated, including age group, gender, nationality, country of residence, profession, class travelled, travel frequency, route and purpose of trip. A miscellaneous section was also included to gather additional feedback on issues like Air Seychelles brand associations of customers, reasons for selecting Air Seychelles over other airlines, reasons for visiting the website and suggestions for nomenclature on the different classes on board.

The data was analysed on Snap Survey software—which was used both for design and analysis purposes. Descriptive statistics were used to summarize and organize the data. This feedback was then used to further orient the brand development. The proposals for the refreshed brand were finally reviewed through two final focus groups, whereupon the final decision was taken regarding the refreshed Air Seychelles brand.

4 Findings

This section details the findings from the focus groups and the surveys carried out. The findings are categorized as follows: (1) initial diagnosis (general brand and livery perceptions) with the emergence of the Air Seychelles ‘only’ statement, (2) dissection of the Air Seychelles ‘only’ statement and definition of the brand essence, (3) customers’ feedback on Air Seychelles brand touchpoints, (4) livery refresh options.

4.1 Initial Diagnosis

A diagnosis of the situation at Air Seychelles revealed the SWOT matrix, as displayed in Table 1. This analysis revealed the potential to further strengthen the brand by capitalizing on the strengths and opportunities identified, namely with regard to the unique selling point that Air Seychelles possessed by being the only national carrier that could embody the local touch across the different brand touchpoints.

Participants to the initial focus groups found that the current brand seemed dated and understated, and needed to be freshened up, modernized and made ‘more expressive’. They expressed their views that the national airline needed to take better advantage of its point of difference—the essence of its ‘creole spirit’—which conveyed sense-awakening messages of colourfulness, natural hybrid beauty, tropical ‘chic’, multi-ethnicity, multi-linguicism, relaxation, joie de vivre, exoticism, informality, friendliness, cleanliness and environmental responsibility. Participants indicated that Air Seychelles should capitalize on the fact that it was uniquely positioned to create the sense that the vacation had begun the moment passengers step on the plane.

When prompted about the livery, there was a general sense that there was a level of recognition of the identity due to the striking red and green livery, which was incidentally also viewed as a nod to the former national flag. The two birds on the tailfin, which are a pair of fairy terns that pair for life and fly in formation, had long been associated with Air Seychelles and were considered an indissociable part of the Air Seychelles brand.

Table 1 SWOT analysis for Air Seychelles

Strengths	Weaknesses	Opportunities	Threats
Genuine Seychelles Creole experience	Lack of super-premium offering	Only national carrier	Increasing competition
Direct flights	Limited frequent flyer incentive	Local touch through local staff	High costs of operation in the industry
Fun, vacation experience with sense of escape	Lack of general travel agent recommendations	Support for the Seychelles economy	Vulnerability due to small size
Recommendations from specialist travel agents	Code share limitations	Name recognition associated with the islands	
Knowledgeable staff who know the Seychelles	Product and service variability		
Small size tends to agility and focus with responsiveness	Lack of sophistication in Web experience		
	Lack of control with package deals		

4.2 Dissection of the Air Seychelles ‘Only’ Statement and the Brand Essence

Through the focus groups, the ‘only’ statement that emerged regarding the definition of Air Seychelles was that *‘In an era of increased competition and premium options of air travel to the Seychelles, Air Seychelles is the only airline that creates a relaxed, uniquely Seychellois experience for travellers who are looking to disconnect from their busy lives’*. Participants therefore confirmed that the existing tagline ‘Flying the Creole Spirit’ was pertinent to the airline and encapsulated the Air Seychelles brand essence, provided it was appropriately translated onto the brand touchpoints.

The subsequent focus groups evolved a more refined definition of the Creole Spirit, which was summed up as *‘The Creole Spirit is the ‘experience of Seychelles in a nutshell’. Evoking the sights, sounds and fragrances of Seychelles, it captures the spontaneous joie-de-vivre, passion and natural warmth of the Seychellois people shining through their attitude to life and the way they interact with others. It faithfully echoes an authentic, exotic, island-style way of living rooted in the multi-ethnic harmony and family values of a soulful way of living close to the rhythms of Nature. The Creole Spirit embodies a distinct approach to, and celebration of, life that is unique to the Seychelles Islands’*.

During the focus group discussions which drilled into particular aspects of the tagline, statements emerged that collectively conveyed customer perceptions regarding the feelings, sights, sounds and scents associated with it.

Statements that surfaced regarding the ‘feelings’ associated with ‘Flying the Creole Spirit’ were: *‘It is about providing a happy, friendly, charming service with a specific joie-de-vivre’*, *‘It is about the engaging and accessible Seychellois people, which makes others feel valued’*, *‘It is about the sensual side of the creoles with their innate passion for life’*, *‘It is about the fun-loving, easy going aspect of the Seychellois, with a carefree, relaxing, laid-back and free-flowing, free spirit. UNPLUGGED for short’*, *‘A person with a true Creole spirit goes out of the way to put a smile on people’s face. He/She would have home grown values with a natural flair for welcoming people within the family’*.

Participants felt that ‘sights’ associated with the tagline are *‘evocative of island motifs, with clear turquoise seas. It reminds of old Creole scenes, of market scenes’*, *‘It evokes tropical scenes of luxuriance amongst palm leaves’*, *‘It is reminiscent of a vibrant, authentic, happy, easy going and jovial person with the typical qualities of Creole joie de vivre and a sense of humour’*, *‘It would typically be an approachable person with a strong personality, typically a ‘richly-mixed’ person, with a wide smile and shining eyes’*, *‘The food would have rich local flavours like the Seychellois grilled fish, or octopus curry, with its chilli condiment’*, *‘Typical snacks would be fried cassava, fried breadfruit or fresh slices of coconut’*.

Members associated ‘sounds’ reminiscent of the ‘Creole Spirit’ with the *‘sounds of dreamy guitars, birds, waves, surfing, sea, cockerels, laughter or kan-kan in the distance’*, or *‘the sound of musical, melodious language’*. Members unanimously concurred that the sounds are *‘suggestive of the melody of island life, rooted in natural*

rhythms of Seychelles, immediately identifiable, of the authentic sounds rooted in culture of the Seychelles' and that *'It evokes a melodious, rhythmic traditional music which can be the Sega, the moutia or the kanmtole'*.

Phrases that emerged to define typical scents evocative of the 'Creole Spirit' were *'It is those very subtle, faint scents that trigger the memory. These are tropical natural and fresh scents that can linger on board. It could be a hint of perfume'*, *'It could be the scent of a mix of spices, of cinnamon, citronella, curry leaves, vanilla, of patchouli, frangipani, casuarinas or cassant'*, *'It is evocative of hot spicy food with chilli, coconut curry and of barbecued, grilled fish'*, *'It is the marine smell of the sea, of the reef'* or *'It is the woody smell of the forest, of Morne Seychellois'*.

These were invaluable, collectively validated statements that would then be taken on board to be translated onto the touchpoints.

4.3 Customers' Feedback on the Air Seychelles Brand Touchpoints

Eight hundred and sixteen responses were obtained through the brand surveys conducted on board Air Seychelles flights across different destinations, indicating a response rate of 40%.

The respondents were of many different nationalities of which the French, Italian, British, Seychellois, German, Mauritian, Swiss, South African, US citizens and the Spanish. Regarding socio-demographic characteristics of respondents, 40% were male compared to 39% females, with the remaining 20% not having answered which gender they belonged to.

Interestingly, 64% of the respondents were not aware of the tagline 'Flying the Creole Spirit'. Of the 31% who knew of the Air Seychelles slogan (the remaining 5% not having replied), 53% liked it, 27% were indifferent to it and 9% expressed the view that they did not like it. Forty-five percentage of the respondents who knew the slogan properly understood its meaning, meaning that 55% did not have a grasp of its significance, showing that the airline had considerable scope to improve on its communication.

Feedback on the existing logo and livery of Air Seychelles clearly demonstrated the existing brand value. Forty-six percentage of respondents liked the exterior livery across a wide spread of nationalities. An impressive 70% of the respondents were confident they would readily recognize the Air Seychelles aircraft anywhere, with 20% not sure if they would, with 5% stating they would not (the remaining 5% not having replied).

Fifty-seven percent of customers stated that the birds were significant in the logo, with 50% of the firm opinion that the birds should remain in the logo and livery. Thirty-four percent of the respondents felt that the birds signified freedom, with 24% associating with it nature, which they felt was an indirect tribute to Seychelles'

laudable efforts in nature's conservation. This converged with the focus group discussions, where it was evident that the birds, which were actually fairy terns, had tacitly become associated with the Air Seychelles brand and should stay as a primary visual metaphor. Forty percent of the respondents felt the significance of red, green and white in the logo, with 45% of them associating them as a nod to the Seychelles national flag. Significance for the colour red was very diverse, with a few associating it with the sun (19 respondents) or flowers (12 respondents) or the soil of the Seychelles (10 respondents). Thirty percent of those who felt that the colours were significant associated the green with nature, flora and fauna and the environment and the blue colour with the seas of the Seychelles.

It was clear from customers' feedback that aqua colours were predominantly associated with the Seychelles, with 64% associating the Seychelles with green, 55% with turquoise and 54% with blue.

Feedback regarding interior of the aircraft revealed that a majority of respondents (67%) preferred a subdued and calming interior as opposed to an energizing and vivid one (22%). The bulk of the respondents (69%) were flying to the Seychelles for holidaying purposes, and all agreed with the statement that 'flying Air Seychelles would make them feel that their holiday would begin the moment they step onto the plane'. The aspects of the interior that customers valued the most were the seats with its colour, configuration and spacing (17%), the local crew (14%) and the interior aqua colours (9%).

Four main aspects of the inflight service that would make the on-board experience typically Seychellois were (1) the inflight entertainment with more films from or on the Seychelles (21%), (2) the food and beverages on board with more Seychellois food and drinks (19%), (3) the amenities on board with a typical Seychellois touch (8%) and (4) the duty-free goods and gifts with Seychellois products and mementoes (3%).

Fifty-six percent of customers thought the cabin crew uniforms were appropriate to the tropical airline, though there were 5% who felt that it was old fashioned and needed a modern touch, and another 5% felt it needed to be more connected with the Air Seychelles corporate identity. The overall rating for the cabin staff grooming appearance was rated 5 on a scale of 1–7, where 7 signifies very satisfied. Ground staff uniforms were less appreciated, with 47% of respondents liking it. Those who did not like it reiterated the need for ground uniforms to be modernized and freshened up with lively colours more in tune with the islands of the Seychelles.

4.4 Livery Refresh Options Based on Findings

Drawing upon the findings from the focus groups and the branding survey, it was clear that there were value and recognition of the existing Air Seychelles identity. A brand evolution was therefore to be explored, as opposed to a brand revolution. Whilst it was evident that change was required and desired by the different stakeholders, it was also deemed essential that this be carefully carried out so as to maintain enough

familiarity to promote brand recall. The livery options for the brand evolution exercise were thus refined with the incorporation of additional aqua colours and a stylized rendition of typically Seychelles elements on the livery, as epitomized through the redefined brand essence. Three main refreshed options were produced for the livery.

5 Conclusion

This study examines the perceptions of stakeholders at different stages of a small tropical airline's brand refresh exercise, in a synchronized participatory branding process [18]. It investigates a concrete methodology that integrates multiple voices at different junctures in time, with stakeholder involvement being a key aspect of this study. The aims of the study were to define the Air Seychelles brand essence, gauge the expectations and perceptions of stakeholders of the airline's brand elements at the different touchpoints and identify ways in which the livery could be made more representative of the national carrier of the Seychelles islands.

The findings revealed the untapped potential in the existing brand touchpoints and demonstrated that there was significant scope for Air Seychelles to exploit the fact that it was the sole national carrier. Whilst it was evident that there were value and recognition in the current brand, stakeholders expressed the need for a touch of modernization across its corporate identity, along with a more soulful rendering of the 'Creole Spirit' across the Air Seychelles brand touchpoints. This study further delayers the underlying connotations of the brand essence, thereby enabling more meaningful and concrete translations across the touchpoints. Stakeholders' and customers' feedback generated a wealth of information that was fed in the brand refresh exercise and the livery redesign. The findings of the study steered the exercise towards a 'brand evolution' effort rather than a 'brand revolution'.

Branding initiatives, which fail to engage, or even alienate local stakeholders, are almost always programmed to fail [18]. This study illustrates the extent to which participatory branding, in which stakeholders are actively engaged in the redesign of a brand, in a phased approach, can fuel a healthy brand refresh exercise within an airline environment.

The limitations of the research pertain to the fact that the studies undertaken were cross-sectional studies carried out at a given point in time. Future research could extend the research question to different industries and could investigate, by means of case studies, the different impacts of stakeholder participation on different industry branding. Stakeholder perceptions could be gleaned through different means and not restricted to focus groups and surveys. Other aspects of branding touchpoints could be taken into account such as the social media and website encounters or the impact of infrastructure at multiple points of contact.

This study contributes to the literature on branding by pursuing a multi-stakeholder participatory approach for a branding exercise within the airline industry. The above study and discussions are set forth in order to illustrate the potential impact of an inclusive, participatory stakeholder approach in the branding exercise of a small

tropical airline. It is the aim of this article to spark interest in such an approach, so that it may become entrenched in airline rebrand exercises, thereby enhancing the chances of success.

References

1. Dinnie, K.: *Nation Branding Concepts Issues Practice*. Routledge, London and New York (2016)
2. Budd, L.C.S.: The influence of business models and carrier nationality on airline liveries: an analysis of 637 airlines. *J. Air Transp. Manag.* **23**, 63–68 (2012). <https://doi.org/10.1016/j.jairtraman.2012.01.017>
3. Clifton, R., Simmons, J.: *Brands and Branding*. Profile Books Ltd., London (2003)
4. Bajde, D.: Branding an industry? *J. Brand Manag.* **26**, 497–504 (2019). <https://doi.org/10.1057/s41262-019-00152-y>
5. Kapferer, J.-N.: *The New Strategic Brand Management: Advanced Insights and Strategic Thinking* (Livre numérique Google) (2012)
6. Adamson, A.P.: *Brand Simple: How the Best Brands Keep it Simple and Succeed* (2007)
7. Lynch, J., de Chernatony, L.: The power of emotion: Brand communication in business-to-business markets. *J. Brand Manag.* **11**, 403–419 (2004). <https://doi.org/10.1057/palgrave.bm.2540185>
8. Rosenbaum-Elliott, Richard; Percy, L.: *Strategic Brand Management*. Oxford University Press (2018)
9. Rossidis, I., Belias, D., Varsanis, K., Papailias, S., Tsiotas, D., Vasiliadis, L., Sdrolas, L.: Tourism and Destination Branding: The Case of Greek Islands. 93–100 (2019). https://doi.org/10.1007/978-3-030-12453-3_11
10. Kladou, S., Kavaratzis, M., Rigopoulou, I., Salonika, E.: The role of brand elements in destination branding. *J. Destin. Mark. Manag.* **6**, 426–435 (2017). <https://doi.org/10.1016/j.jdmm.2016.06.011>
11. Baxendale, S., Macdonald, E.K., Wilson, H.N.: The Impact of Different Touchpoints on Brand Consideration. *J. Retail.* **91**, 235–253 (2015). <https://doi.org/10.1016/j.jretai.2014.12.008>
12. Barnham, C.: The structure and dynamics of the brand. *Int. J. Mark. Res.* **51**, 593–610 (2009). <https://doi.org/10.2501/S1470785309200840>
13. Lin, Y.H., Ryan, C.: From mission statement to airline branding. *J. Air Transp. Manag.* **53**, 150–160 (2016). <https://doi.org/10.1016/j.jairtraman.2016.02.013>
14. Lee, J., Yi, J., Kang, D., Chu, W.: The effect of travel purpose and self-image congruency on preference toward airline livery design and perceived service quality. *Asia Pacific J. Tour. Res.* **23**, 532–548 (2018). <https://doi.org/10.1080/10941665.2018.1483956>
15. Thurlow, C., Aiello, G.: National pride, global capital: A social semiotic analysis of transnational visual branding in the airline industry. *Vis. Commun.* **6**, 305–344 (2007). <https://doi.org/10.1177/1470357207081002>
16. Wang, S.W., Ngamsiriudom, W.: Celebrity livery featured aircraft, the Moneki Neko (fortune cat) of airlines. *J. Air Transp. Manag.* (2015). <https://doi.org/10.1016/j.jairtraman.2014.09.005>
17. Kim, H.C., Chua, B.L., Lee, S., Boo, H.C., Han, H.: understanding airline travelers' perceptions of well-being: the role of cognition, emotion, and sensory experiences in airline lounges. *J. Travel Tour. Mark.* (2016). <https://doi.org/10.1080/10548408.2015.1094003>
18. Kalandides, A., Kavaratzis, M., Boisen, M., Kavaratzis, M.: From “necessary evil” to necessity: stakeholders' involvement in place branding. *J. Place Manag. Dev.* **5**, 7–19 (2012). <https://doi.org/10.1108/17538331211209013>
19. Jones, C., Newsome, J., Levin, K., Wilmot, A., McNulty, J.A., Kline, T.: The qualitative report friends or strangers? A feasibility study of an innovative focus group methodology recommended APA citation. *Qual. Rep.* **23**, 98–112 (2018)

20. Daymon, C., Holloway, I.: Qualitative research methods in public relations and marketing communications: Second edition. (2010). <https://doi.org/10.4324/9780203846544>

An Analysis of Communication Strategies of Fast Food Outlets on Social Media in Mauritius



Swaleha Peeroo and A. Mooznah Auleear Owodally

Abstract Fast food outlets in Mauritius are leveraging social media as a marketing tool. However, studies on the adoption of social media by companies in Mauritius are scarce. This paper aims to examine the types of content strategies used by businesses in Mauritius to engage customers. Content analysis was used to analyze the posts on the Facebook pages of McDonald's, KFC and Pizza Hut. This study shows that KFC and McDonald's use engaging contents, while Pizza Hut posts use informative contents. By observing the interactions on the Facebook pages, we found that Pizza Hut embraces an egocentric communication strategy, while KFC and McDonald's use a conversational communication strategy. However, all three fast food outlets adopt a secretive communication strategy. This study adds to the body of knowledge by identifying the content strategies and social media communication strategies of fast food outlets on their local Facebook pages.

Keywords Social media · Facebook · Fast food · Content analysis

1 Introduction

Social media have transformed the marketing landscape and are already considered as an additional marketing tool together with the traditional tools of the marketing mix [1]. Companies have readily incorporated social media platforms within their integrated communication strategies [2] to build and nurture relationships with targeted customers [3]. More than 15 million brands use social media aiming to reach over 1 billion customers [4]. Accordingly, companies are allocating more funds for social media in their budgets seeking to harness the opportunities afforded by social media. In the past ten years, social media spending has increased by 250% [5]. Following

S. Peeroo (✉)

Université Des Mascareignes, Udm Lab, Beau Bassin-Rose Hill, Mauritius

e-mail: speeroo@udm.ac.mu

A. Mooznah Auleear Owodally

University of Mauritius, Reduit, Mauritius

the international trend, companies in Mauritius are increasingly using social media to connect with customers.

Although the adoption of social media by businesses has increased rapidly, academic research lags behind industry practice. There is still limited research on the value that companies gain by using social media [6]. There is scant research on how businesses in Mauritius are harnessing social media as a marketing tool. It is important for managers to understand how to use social media to reach the desired marketing outputs.

To address these gaps in the literature, this paper sets out to examine the types of social media content strategies used by businesses in Mauritius to engage with customers. We have chosen the fast food industry as most of the international fast food outlets, which are present locally, and are using social media platforms to engage with customers in Mauritius. The wide adoption of social media by fast food outlets in Mauritius and the rest of the world may be explained by the fact that competition is intense in the fast food industry owing to low switching costs [7]. The intense competition in this industry is compelling businesses to use innovative tools to reach and interact with customers [8]. The main target market for fast food chains is young people, and since young people are heavy users of social media, fast food outlets are using social media as a marketing tool to engage with them [7].

This study will analyze the conversations on the Facebook pages of fast food outlets to determine how these companies are using social media to engage customers. The research questions are (1) what are the types of contents that fast food outlets send to their customers? and (2) what communication strategies are used by fast food outlets?

The paper is structured as follows. First, an overview of the concept of social media is introduced, followed by a review on social networking sites and online brand communities. Second, the methodology that was adopted is presented. The findings are then reported and discussed. Finally, the paper concludes by elaborating on the managerial implications, the limitations of the study and directions for future research.

2 Literature Review

The popularity and ubiquity of social media have transformed the way people interact with each other [5]. Social media usage continues to grow, and social networking sites like Twitter, Facebook, Instagram and LinkedIn appeal to hundreds of millions of users [9]. These social networking sites are online communities where users interact and socialize with each other [10]. As a result of the large number of people using social media, businesses have joined social media platforms seeking to communicate and interact with existing and potential customers [5].

2.1 Social Media

Social media have become an essential element of the marketing communications mix of companies and brands [11]. Organizations are using social media as a communication channel [12] of strategic importance [13] to promote their products and services [14], interact with consumers [15] and engage them with marketing content [16]. Social media have witnessed greater flexibility and visibility in branded content. Additionally, the way customers interact with businesses has changed radically with the advent of social media [17].

Customers have been empowered by social media which enable them to shift from being passive recipients of marketing content to being active co-creators of brand messages [18]. Social media have transformed the marketing realm as they are interactive channels which allow triadic conversations between the company and its customers and among customers [19]. The interactive capabilities of social media platforms have improved the customer experience and have influenced several behavioral aspects such as acquiring information and purchase behavior [18]. Various studies have found that social media platforms provide the right setting for user-generated content [20], co-creation of innovative products [21, 22] and brand/product referrals [23]. When customers like, share, comment and post reviews on corporate social media platforms, there is either co-creation or co-destruction of value [24].

2.2 Facebook

The social media platform that has most captured the attention of marketers is the social networking site, Facebook which heralds 2 billion active users monthly [7]. Responding to the significant presence of customers on Facebook, companies have established their Facebook brand pages where fans have the opportunity to interact with the brand using the 'like', 'share' and 'comment' options [25]. These Facebook brand pages are online brand communities as the fans share a common interest [26] and these brand pages are the most common form of social media marketing [27]. These brand pages are created either by companies or by fans who want to show their appreciation of the brand [28].

2.3 Online Brand Communities

On their corporate Facebook pages, businesses communicate to their fans to inform them about their products or sales promotion activities, to publicize sponsored events, to gather feedback, to convey informational announcements and to entertain their fans by posting fun messages [29]. Fans join an online brand community to view and contribute to its content [21].

Users become fans of a Facebook brand page because they use the products, seek discounts and promotions, wish others to know that they like the brand, want to obtain information about the brand before other customers and/or want to access exclusive content [30].

Online community members gain four main benefits while being active on social networking sites: social benefits through interaction with other online community members, informational benefits by obtaining information on the brand, hedonic benefit while having fun and enjoying themselves on the social networking site and economic benefits by seeking promotional deals [31]. Businesses also benefit from the contents posted on corporate Facebook pages which may result in higher levels of affective commitment, customer satisfaction and positive word of mouth [32].

2.4 Content Strategy

Content on corporate Facebook pages has a critical role in increasing the level of interaction and engagement of fans with branded posts. Studies have shown that customers engage more with brands when the posts are remunerative, informational and interactive [33, 34]. Companies may choose from 6 complementary social media communication strategies for their content: egocentric, conversational, selective, openness, secretive and supportive [35]. An egocentric strategy is when firms use one-way communication, while a conversational strategy is when firms encourage dialogue on the Facebook page. A selective strategy is adopted when businesses filter only positive comments, as opposed to an openness strategy where all posts are visible to the community. A secretive strategy is used when firms resolve conflicts in a private forum and a supportive strategy is used to help users through the consumer decision-making process.

Fast food outlets in Egypt have used Facebook to target young consumers [8], and they post branded content on their corporate Facebook pages, where members of the online community interact with the content and the other members [36]. Fast food restaurants in Egypt post contents to inform their fans about the sandwiches and meals, and they also use Facebook pages to respond to queries of customers [8].

3 Methodology

This paper aims to examine the types of social media content strategies used by businesses in Mauritius to engage customers. We have chosen to examine the use of Facebook by fast food outlets in Mauritius as fast food consumption and social media usage are popular among young consumers. The fast food industry is so popular among young consumers that they have been tagged ‘the fast food generation’ [8]. The high penetration rate among young consumers on Facebook has pushed fast food outlets worldwide to create their Facebook page [8]. According to the Internet World

Stats (2018), Facebook is the social network platform with the highest number of users in Mauritius. We have selected McDonald's, KFC and Pizza Hut as these are international fast food chains which have all set up local Facebook pages.

Qualitative content analysis was used to analyze the contents posted on the Facebook pages of these three outlets. It is a method in which the content of text data is interpreted subjectively by systematically coding data and classifying them in order to identify themes or patterns [37]. Qualitative content analysis consists of examining words intensely to cluster large amounts of texts into categories expressing similar meanings, instead of counting words [37].

Data was collected from the official Facebook pages of McDonald's, KFC and Pizza Hut for 3 consecutive months ranging from 1 October to 31 December. The corporate posts as well as comments of customers were saved for analysis. The data was coded based on 4 main categories of posts: informative content, entertaining content, incentive content and engaging content to determine the types of messages that fast food outlets post on Facebook. To examine the social media communication strategies, the data was categorized according to the communication strategies identified by Floreddu and Cabiddu: egocentric, conversational, selective, openness, secretive and supportive [35].

To generate insights from the data, the constant comparative method was used to identify similarities and differences [38]. Open coding was initially applied to the data. After thorough scrutinizing and comparing of data, axial coding, which consists of linking the identified categories, was then performed to attempt to identify any patterns or trends in the phenomenon under study.

4 Findings and Discussion

This paper aims to examine the content strategies of international fast food outlets operating in Mauritius having set up local corporate Facebook pages. Table 1 shows the date of creation of the official Facebook page of each outlet and the number of fans. Pizza Hut was the first fast food outlet to create its local Facebook page, and it is KFC which has attracted the highest number of fans boasting 176,700 likes on its Facebook page. None of the fast food outlets post contents on Facebook on a daily basis. KFC is the most active on Facebook with an average of one post every 3 days, while Pizza Hut is the least active averaging one post every five days.

Table 1 Facebook pages of fast food outlets in Mauritius

Fast food outlet	Facebook page created	Number of fans
McDonald's	June 2012	105,147
KFC	May 2011	176,700
Pizza Hut	July 2010	127,284

4.1 Content Strategies of Fast Food Outlets on Facebook

Fast food outlets use Facebook pages to communicate to and interact with customers. This study shows that all the three fast food outlets use various content strategies to engage customers. Fast food outlets in Mauritius use the following content strategies: engaging, entertaining, incentive and informative. These findings are consistent with previous research carried out on the fast food industry in Egypt [8].

Informative content strategy on the fast food Facebook pages involved posting contents that provided information on the menu and meals (prices of meals and new meals), on the fast food chain or outlets such as opening of new branches and information about corporate social responsibility. An example is provided below, where KFC informed its customers about the updated version of its KFC app.

Good news, it's Friday! Even better news, the KFC Mauritius app has been updated and you can now get access to the iOS version, too. You can pay for your meals, locate your nearest store and win fidelity points! Place your orders in a short time and pick up via our drive thru, in store or enjoy your KFC meal on the spot. Download it now. On Android: <https://bit.ly/2xmsxJP> On iOS: <https://apple.co/2xYbldb>.

Entertaining content strategy is used when the content aims to make the consumer smile and is humorous. Entertaining contents included posts with funny pictures of consumers eating in restaurants, posts greeting customers for some occasions such as New Year. The two Facebook posts below illustrate how fast food outlets interact with their online community to develop positive feelings toward the brand. McDonald's, for example, posted the following message on the occasion of teacher's day, while KFC dedicated a day to its female customers.

A Teacher presents the past, reveals the present, and creates the future. #HappyTeachersDay #WeLoveTeachers #McDonaldsMauritius.

Girls, today's YOUR day and it is a great excuse to pamper yourselves with some KFC;) #GirlsDay #SorryNotSorry.

Incentive content strategy includes posts that provide an incentive or a reward for consumers to act in a desired way such as getting discounts, obtaining gifts for sharing or clicking on a particular content. Both posts below illustrate how fast food outlets induce customers to interact with their page by offering them a reward.

Hello Pizza Hut lovers, get a chance to win a voucher of Rs250 by catching a Pan Pizza. How to play? 1. Simply send us a screenshot with the pizza in the dotted circle under the comments section on our page! 2. Tag a friend who you think deserves to win too on the screenshot.

Today is #FastFoodDay! Fun fact: The KFC slogan "finger lickin' good" was, however, a little lost in translation when they opened in China, as it translated to "eat your fingers off". Impress us with a poem, a fun fact, or anything cooler than cool, and you could win a voucher worth Rs 500! We're waiting.

Engaging content strategy is used when the Facebook post asks consumers to execute an action and encourages the consumers to engage with the content. Engagement content on the fast food pages included posts asking consumers to rate their

favorite sandwiches and meals and posts asking them to name a branch. For example, when Pizza Hut launched a new drink it encouraged customers to like and share the post to generate awareness and create a buzz. Similarly, McDonald's posted a contest where customers had to count the number of burgers in the picture and write their answers in the comments section.

Welcome this summer with our new refreshing MOJITO MOCKTAIL! Like or share if you want to try !! #pizzahutmauritiu #mocktails #mojito.

Ready to play with us? How many burgers can you see in the picture? Leave your answer in the comments;) #McDoTrivia#McDonaldsMauritiu.

This study shows that the content strategies of the 3 fast food outlets differ. Though all the fast food outlets in the study used several types of contents, we found that KFC and McDonald's had a similar approach to content strategies. KFC and McDonald's posted more engaging contents, i.e., corporate posts which explicitly asked customers to carry out a specific action and encouraged customers to engage with the company. However, Pizza Hut used mostly informative contents to pass on information. This research has also shown that the fast food outlets also used their Facebook pages to recruit staff, thus not only using Facebook as a communication channel but also as a recruitment tool as illustrated below:

Looking for job during your summer holidays? We are looking for team members to work in our restaurants as part & full timers as: 1. Waiters 2. Cashiers 3. Kitchen helpers 4. Delivery Riders To apply email us only at hr@pizzahut.mu.

4.2 Social Media Communication Strategies of Fast Food Outlets on Facebook

While communicating to customers, fast food outlets adopt various complementary social media communication strategies. These social media communication strategies describe how fast food outlets communicate to their customers based on the conversations occurring on the online brand community [35]. Out of the six complementary communication strategies identified by Floreddu and Cabiddu (2016), fast food outlets in Mauritius are using four of them: egocentric, conversational, selective and secretive.

None of the fast food outlets used the openness social media communication strategy as customers complain about their posts being either deleted or ignored. Whenever customers posted complaints, the fast food outlets redirected them to their customer service department or asked them to use the private message option, thus refusing to resolve the conflict in an open forum. There is little transparency of business–customer conversations on the Facebook pages.

Fast food outlets do not use the supportive social media communication strategy which involves providing support to the customer during the consumer decision-making process. A supportive strategy is best suited for brands where customers are

highly involved in the decision-making process, while fast food is a low involvement service.

An egocentric social media communication strategy is used by Pizza Hut which uses its Facebook page mainly to share information, but does not engage in conversation with fans and customers. With an egocentric strategy, businesses do not seek to build relationships with customers; their ultimate aim is to increase the brand's visibility through social media [35].

McDonald's and KFC both implemented the conversational social media communication strategy since they engage in dialogues with customers with the aim to build strong relationships [15]. However, we observed that both fast food outlets do not respond to every single comment shared by customers. This study shows that when customers respond by commenting on a brand post, both fast food outlets respond to the comments of customers for the first few hours and then ignore the comments posted by customers at a later stage.

McDonald's and KFC also use the selective social media strategy as they seem to respond only to positive comments and ignore negative comments. This is consistent with prior research which reported that firms using the selective strategy aim to build a relationship only with fans who support the brand and avoid managing conflicting opinions [35].

All three fast food outlets apply a secretive social media communication strategy as they choose to respond to customer complaints through private channels such as private messages or emails. They prefer to manage conflicts outside Facebook.

5 Conclusion

This paper sets out to examine the types of content strategies used by businesses in Mauritius to engage customers. The qualitative content analysis of the Facebook page of McDonald's, KFC and Pizza Hut has shown that KFC and McDonald's post more engaging contents while Pizza Hut uses mostly informative contents. By observing the interactions on the Facebook pages, it was found that Pizza Hut embraces an egocentric communication strategy, while KFC and McDonald's use a conversational communication strategy. However, all three fast food outlets adopt a secretive communication strategy.

This study adds to the body of knowledge of marketing and social media in the fast food industry. Even though the chosen outlets are well-known international brands, they have set up local social media platforms to adapt to the culture. This study contributes to the literature by identifying the content strategies and social media communication strategies of fast food outlets on their local Facebook pages.

The findings in this study provide managers with some useful guidelines to engage customers on social media platforms. Brand pages on social media platforms are an important marketing channel as they offer excellent opportunities to firms to engage customers. These brand pages on Facebook provide unique opportunities for brands and fast food chains to glean information from their fans about their

preferences and dislikes. This feedback can be used to improve their products and services or develop new products. Managers should use their Facebook page as a customer service platform as customers have taken to Facebook pages to post their complaints. Instead of redirecting customers outside Facebook, managers should adopt an openness strategy to solve any issues customers are facing. When resolving conflicts, firms have the opportunity to retain customers and research has shown that when conflicts are resolved satisfactorily, customers become loyal. The conflict resolution when done on Facebook will be visible to all the fans of the brand page, and this will help to generate positive word of mouth and enhance corporate reputation.

Owing to its exploratory nature, this research has some limitations, which point to avenues for further research. The study is based on the fast food industry and in only one location; hence, generalizing the findings of this study to other contexts is not recommended. It is important to broaden our research to other industries where the level of customer involvement in decision making is higher. Also, another limitation of the study is that the research was done on one social media platform. Other social media platforms such as Instagram, which is gaining wide popularity, could be studied to gauge its usefulness as a marketing tool to engage customers.

References

1. Pletikosa Cvijikj, I., Michahelles, F.: Online engagement factors on Facebook brand pages. *Soc. Netw. Anal. Min.* **3**, 843–861 (2013). <https://doi.org/10.1007/s13278-013-0098-8>
2. Šerić, M.: Relationships between social Web, IMC and overall brand equity: An empirical examination from the cross-cultural perspective. *Eur. J. Mark.* **51**, 646–667 (2017). <https://doi.org/10.1108/EJM-08-2015-0613>
3. De Vries, L., Gensler, S., Leeftang, P.S.H.: Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. *J. Interact. Mark.* **26**, 83–91 (2012). <https://doi.org/10.1016/j.intmar.2012.01.003>
4. Stieglitz, S., Mirbabaie, M., Ross, B., Neuberger, C.: Social media analytics – Challenges in topic discovery, data collection, and data preparation. *Int. J. Inf. Manage.* **39**, (2018). <https://doi.org/10.1016/j.ijinfomgt.2017.12.002>
5. Shanahan, T., Tran, T.P., Taylor, E.C.: Getting to know you: social media personalization as a means of enhancing brand loyalty and perceived quality. *J. Retail. Consum. Serv.* **47**, 57–65 (2019). <https://doi.org/10.1016/j.jretconser.2018.10.007>
6. Estrella-Ramón, A., García-de-Frutos, N., Ortega-Egea, J.M., Segovia-López, C.: How does marketers' and users' content on corporate Facebook fan pages influence brand equity? *Electron. Commer. Res. Appl.* **36**, 100867 (2019). <https://doi.org/10.1016/j.elerap.2019.100867>
7. Rasheed Gaber, H., Elsamadicy, A.M., Wright, L.T.: Why do consumers use Facebook brand pages? A case study of a leading fast-food brand fan page in Egypt. *J. Glob. Sch. Mark. Sci.* **29**, 293–310 (2019). <https://doi.org/10.1080/21639159.2019.1622434>
8. Gaber, H., Wright, L.: Fast-food advertising in social media. A case study on Facebook in Egypt. *J. Bus. Retail* **9**, 52–63 (2014). Doi: [https://doi.org/10.1016/S0891-4222\(02\)00092-6](https://doi.org/10.1016/S0891-4222(02)00092-6)
9. Tuten, T.L., Solomon, M.R.: Social Media Marketing Strategy. In: *Social Media Marketing* (2014).

10. Dennis, C., Morgan, A., Wright, L.T., Jayawardhena, C.: The influences of social e-shopping in enhancing young women's online shopping behaviour. *J. Cust. Behav.* (2010). <https://doi.org/10.1362/147539210x511353>
11. Triantafyllidou, A., Yannas, P., Lappas, G.: Facebook Content Strategies: A Case Study of a Subsidiary Company in Greece. 191–198 (2019). https://doi.org/10.1007/978-3-030-12453-3_22.
12. Ashley, C., Tuten, T.: Creative strategies in social media marketing: an exploratory study of branded social content and consumer engagement. *Psychol. Mark.* 32, (2015). <https://doi.org/10.1002/mar.20761>.
13. Brubaker, P.J., Wilson, C.: Let's give them something to talk about: Global brands' use of visual content to drive engagement and build relationships. *Public Relat. Rev.* (2018). <https://doi.org/10.1016/j.pubrev.2018.04.010>
14. Cervellon, M.C., Galipienzo, D.: Facebook Pages Content, Does it Really Matter? Consumers' Responses to Luxury Hotel Posts with Emotional and Informational Content. *J. Travel Tour. Mark.* (2015). <https://doi.org/10.1080/10548408.2014.904260>
15. Choi, Y.G., Ok, C.M., Hyun, S.S.: Relationships between brand experiences, personality traits, prestige, relationship quality, and loyalty: An empirical analysis of coffeehouse brands. *Int. J. Contemp. Hosp. Manag.* (2017). <https://doi.org/10.1108/IJCHM-11-2014-0601>
16. Coelho, R.L.F., De Oliveira, D.S., De Almeida, M.I.S.: Does social media matter for post typology? Impact of post content on Facebook and Instagram metrics. *Online Inf. Rev.* (2016). <https://doi.org/10.1108/OIR-06-2015-0176>
17. Dolan, R., Conduit, J., Fahy, J., Goodman, S.: Social media: communication strategies, engagement and future research directions. *Int. J. Wine Bus. Res.* 29, 2–19 (2017). <https://doi.org/10.1108/IJWBR-04-2016-0013>
18. Mangold, W.G., Faulds, D.J.: Social media: The new hybrid element of the promotion mix. *Bus. Horiz.* 52, 357–365 (2009). <https://doi.org/10.1016/j.bushor.2009.03.002>
19. Peeroo, S., Samy, M., Jones, B.: Trialogue on Facebook pages of grocery stores: Customer engagement or customer enagement? *J. Mark. Commun.* 7266, 1–23 (2018). <https://doi.org/10.1080/13527266.2018.1482559>
20. Vivek, S.D., Beatty, S.E., Morgan, R.M.: Customer Engagement: Exploring Customer Relationships Beyond Purchase. *J. Mark. Theory Pract.* 20, 122–146 (2012). <https://doi.org/10.2753/MTP1069-6679200201>
21. De Vries, N.J., Carlson, J.: Examining the Drivers and Brand Performance Implications of Customer Engagement with Brands in the Social Media Environment (2014). <https://doi.org/10.1057/bm.2014.18>
22. Hoyer, W.D., Chandy, R., Dorotic, M., Krafft, M., Singh, S.S.: Consumer cocreation in new product development. *J. Serv. Res.* 13, 283–296 (2010). <https://doi.org/10.1177/1094670510375604>
23. Chu, S.C., Kim, Y.: Determinants of consumer engagement in electronic Word-Of-Mouth (eWOM) in social networking sites. *Int. J. Advert.* 30, 47–75 (2011). <https://doi.org/10.2501/IJA-30-1-047-075>
24. Peeroo, S., Samy, M., Jones, B.: Facebook: a blessing or a curse for grocery stores? *Int. J. Retail Distrib. Manag.* (2017). <https://doi.org/10.1108/IJRDM-12-2016-0234>
25. Gummerus, J., Liljander, V., Weman, E., Pihlström, M.: Customer engagement in a Facebook brand community. *Manag. Res. Rev.* 35, 857–877 (2012). <https://doi.org/10.1108/01409171211256578>
26. Pöyry, E., Parvinen, P., Malmivaara, T.: Can we get from liking to buying? Behavioral differences in hedonic and utilitarian Facebook usage. *Electron. Commer. Res. Appl.* 12, 224–235 (2013). <https://doi.org/10.1016/j.elerap.2013.01.003>
27. Luarn, P., Lin, Y.F., Chiu, Y.P.: Influence of Facebook brand-page posts on online engagement. *Online Inf. Rev.* (2015). <https://doi.org/10.1108/OIR-01-2015-0029>
28. Zaglia, M.E.: Brand communities embedded in social networks. *J. Bus. Res.* (2013). <https://doi.org/10.1016/j.jbusres.2012.07.015>

29. Dekay, S.H.: How large companies react to negative Facebook comments. *Corp. Commun.* (2012). <https://doi.org/10.1108/13563281211253539>
30. Pereira, H.G., de Fátima Salgueiro, M., Mateus, I.: Say yes to Facebook and get your customers involved! Relationships in a world of social networks. *Bus. Horiz.* (2014). <https://doi.org/10.1016/j.bushor.2014.07.001>
31. Park, H., Kim, Y.K.: The role of social network websites in the consumer-brand relationship. *J. Retail. Consum. Serv.* (2014). <https://doi.org/10.1016/j.jretconser.2014.03.011>
32. Royo-Vela, M., Casamassima, P.: The influence of belonging to virtual brand communities on consumers' affective commitment, satisfaction and word-of-mouth advertising: The ZARA case. *Online Inf. Rev.* (2011). <https://doi.org/10.1108/14684521111161918>
33. Jeon, H., Ahn, H.J., Yu, G.J.: What makes people react to the posts on the brand pages of mobile social network games? *Online Inf. Rev.* (2016). <https://doi.org/10.1108/OIR-07-2015-0236>
34. Su, N., Reynolds, D., Sun, B.: How to make your Facebook posts attractive: A case study of a leading budget hotel brand fan page. *Int. J. Contemp. Hosp. Manag.* (2015). <https://doi.org/10.1108/IJCHM-06-2014-0302>
35. Floreddu, P.B., Cabiddu, F.: Social media communication strategies. *J. Serv. Mark.* (2016). <https://doi.org/10.1108/JSM-01-2015-0036>
36. Saad, H.E., Badran, N.A.: How Successful is Fast Food Social Media Marketing? *International vs. Local Chains* (2016)
37. Hsieh, H.F., Shannon, S.E.: Three approaches to qualitative content analysis. *Qual. Health Res.* **15**, 1277–1288 (2005). <https://doi.org/10.1177/1049732305276687>
38. Miles, M.B., Huberman, A.M.: *Qualitative Data Analysis : Handout. A Sourceb. New Methods.* California; SAGE Publ. Inc. (1984).

Factors Affecting Task Allocation and Coordination in Distributed Agile Software Development



Chitra Nundlall and Soulakshmee D. Nagowah

Abstract The advantages in terms of low-cost resources are driving software industry to shift towards global markets in recent years. The distributed agile software development (DASD) companies aim at achieving high productivity by using resources around the world. Together with benefits, DASD presents some issues and challenges, which require further studies and research. Adopting an appropriate task allocation process is important for the smooth running of the DASD projects. Assigning tasks to remote teams requires a number of factors to be taken into consideration. Therefore, there is the need to identify factors that influence task allocation and coordination in agile distributed environment. In this study, a survey is conducted with agile professionals to assess the importance of the identified factors in the literature. Snowball sampling technique has been chosen due to small population size in Mauritius. The target population of the survey consists basically of agile project managers, agile team leaders, agile team members, agile researcher and other decision-makers involved in the process. The results show that some factors are valued differently from the literature.

Keywords Task allocation and coordination · Agile global teams · Factors

1 Introduction

Agile global software development is becoming an inexorable trend these days. The distributed projects solicit for more collaboration between remote teams. In agile collocated team, task allocation can be performed more easily as the availability, skills and cultures are known [5]. Deciding and knowing which team member is idle and which one is stuck with a task for long is difficult to distinguish when teams are

C. Nundlall · S. D. Nagowah (✉)

Department of Software and Information Systems, University of Mauritius, Moka, Mauritius
e-mail: s.ghurbhurrun@uom.ac.mu

C. Nundlall

e-mail: shwetanundlall@gmail.com

from different locations. Lack of communication and control of activities can be problematic when managing a team. There exist different types of dependencies between the sites. With distributed software development, barriers towards communication exist such as geographical separation, cultural and language differences [11]. After conducting a systematic literature review on agile distributed software development, it has been found that task allocation and coordination is one of the major challenges faced by distributed project managers. There are a number of factors that must be considered during the decision-making process [17]. Dependencies between tasks, people, resources, requirements, expertise and others bring the need for an effective communication and coordination strategy for a good flow of information.

The gap in the literature between factors that have been identified and those that have been used in the related works motivated us to consult the industry to identify the real needs when allocating and coordinating tasks. A survey has been conducted to gather feedback from agile professionals on the factors identified in order to determine their importance. The rest of the paper is structured as follows: Sect. 2 describes the factors and dependencies influencing task allocation and coordination in DASD. Section 3 highlights the research aims and methodology adopted in the study. Section 4 describes three survey results and discussions, limitations of the survey and recommendations for future work. Section 5 concludes on the paper. Finally, Sect. 6 presents details on ethical compliance of the survey.

2 Distributed Agile Task Allocation and Coordination

Task allocation is a project management activity that requires enough attention for the smooth running of DASD projects. Related decision, if not taken appropriately, may directly lead to communication and coordination issues, project delays and additional project costs. Literature highlights a number of factors that influence task allocation process but organizations often consider only a few resulting in project failure [2, 6]. Depending on the nature of the tasks, some teams are allocated tasks that require proximity to customers while others can be done remotely [18]. Storing as much information as possible about past projects and their influencing factors becomes important in this case. Documentation becomes necessary for a good flow of information between teams. Table 1 describes the different factors obtained while studying related papers.

Only considering the different factors identified is not enough for the success of DASD. Managing coordination dependencies is also important. In offshoring companies, issues related to coordination within projects are due to lack of communication and collaboration between remote teams. Sitting together, discussing things and resolving issues are not possible when teams are geographically distributed. Factors like time differences due to different time zones can cause latency during the development when a task is depending on an information from another team to proceed. Issues that could be resolved instantly take longer time. Table 2 describes

Table 1 Factors affecting task allocation and coordination

Factors	Details
Expertise	Expertise is considered as a major factor in task allocation and coordination. Having experts in the team helps to avoid the project progress from slowing down by providing training to unexperienced team members [2, 6, 7, 9].
Technical ability	Technical skills such as knowledge and ability on methods, programming languages and tools are important not only for team members but for project managers and team leaders as well [19]. Having technical ability helps them to teach and guide the team technically and to understand the level of difficulty being faced by the team [2, 7, 9].
Team members' knowledge and skills	Having enough knowledge and skills helps the team members to coordinate within a team and remain aligned during decision process. Shared knowledge of team helps to mitigate the negative effect of geographic distance on coordination [2, 5, 7, 9].
Personnel availability	Personnel availability in task allocation and coordination process refers to the availability of the team members during the decision-making process as well as free to perform tasks [2, 6, 7].
Project manager maturity	It includes experience in the field and maturity of the project manager in the profession [19]. As stated by [15], <i>“if there are limits to what we can know about our organization, there are limits to what we can achieve in a predetermined and planned way”</i> . Therefore, it is important for the project manager to have a certain level of maturity [7, 14, 16].
Team maturity	A team is mature when it has a certain age experience in the domain [1]. A mature team, having more experience, speeds up the completion of assigned tasks. Lack of trained team members in a team can be challenging [7, 12].
Task site specificity	Task site specificities involve application and platform experience possessed by the team members. Lack of application or platform experience impacts the task allocation decision process since the team members will take more time to complete the task [2, 5-7].
Labour cost	Distributed software development takes advantage of the low cost of professionals and other resources [5]. Labour costs play a vital role in task allocation decision-making process [3, 13, 21].
Workload at site	Even though the expertise is present at the site but if the personnel has a busy schedule due to other commitments, tasks with strict deadlines cannot be assigned to the person/ site [3, 5, 21].

(continued)

Table 1 (continued)

Factors	Details
Working time	It is difficult to communicate and coordinate with teams working at different hours and time zones [5]. However, it can also be advantageous in the case where tasks like programming can be performed and when testers from different countries come to work, they have the product to test. They do not have to wait for developers [2, 7, 9].
Cultural differences	Distributed teams have different cultures and working habits [3, 5, 13, 21].
Site locations	Site locations play an important role in task allocation when there is the need for proximity to customer for feedback [2, 6, 7].
Team willingness	Team willingness represents the motivation and interest of the team to complete a task [9, 16].
Communication	A good communication infrastructure (speed, availability) is needed when team is communicating remotely [19].
Coordination	Coordination is an important factor in distributed software development. Agile approaches lead to effective coordination. The chemistry between team members must be very good to enable coordination in a team. [20] defined coordination as a comprehensive understanding of the project and what is going on and when, what other team members are doing to fit in with other team members work.
Task size	The size of the task helps to determine the amount of time a resource will complete the task [2, 5, 6, 9].
Proximity to customer requirement	One of the agile principles is customer focus. The authors in [10] highlighted in their research that in order to follow this principle, agile teams need to work closely with the customer to gather requirements and customer feedback throughout the project [2].
Required resources	Required resources include hardware, software and human requirements. The availability of resources is as important as cost [2, 5].
Task deadline	Task deadline needs to be considered in order to allocate task [2, 9].
Effort	Effort has to be considered while assigning task. Experienced personnel requires less efforts, while less experienced ones require more [9].
Product architecture	Product architecture includes modules view, components and connectors that need to be considered for task allocation [2, 5, 7].

(continued)

Table 1 (continued)

Factors	Details
Product	The nature of the product is important. If it is a new product, the personnel will take more time to complete the implementation. If the infrastructure is already present for that product, lesser time will be needed [7, 12].
Transparency	Transparency in the team while allocating task is important to prevent team conflicts. Transparency at task level is important when a team member is stuck with a particular task [2, 4].
Prioritized delivery	Prioritized delivery is given higher priority compared to other factors when allocating tasks [2, 9].
Enough documentation	In distributed software development, documentation is important since there are communication issues. Missing information can be found in product deliverables such as documentation [2, 13].
Customer collaboration	Customer collaboration constitutes customer knowledge and involvement. At the start of a project, collection of information on customers is needed to understand the past behaviours of the client including their everyday ongoing processes. In order to respond to the client’s needs, the client should collaborate with the developer and build a customer-developer relationship [10].
Language fluency	In distributed software development, teams are located in different geographical locations or countries. The team members must be fluent in the foreign languages commonly used in the office [1].

Table 2 Dependencies affecting task allocation and coordination

Dependency	Details
Tasks	The architecture of the product illustrates the coupling between tasks for task allocation [2, 6–8].
People	People dependency englobes team members’ presence to work simultaneously [6, 7].
Requirements	Absence of a requirement can impact the work progress [2, 7].
Resources	Resources dependency englobes people, location or anything that the project requires for a work to be performed, for example, if a team member is not allocated a pc or a headset to communicate remotely, this will impact communication [6–8].
Sites	Site dependencies consist of cultural and temporal differences and working time differences and public holidays [2, 6, 7].
Expertise	If knowledge is not shared between experts, there will be a dependency on the expertise in the project [7].
Activity	Activity dependency is related to activities that depend on other works to be completed, for example, testing cannot start before coding [7].
Technical	Teams communicate when technical dependencies are present [7, 8].

the different dependencies between factors related to task allocation and coordination in DASD.

3 Research Aims and Methodology

This section describes the methodology adopted for the study. A survey was carried out with agile professionals in the software industry and agile researchers from different geographical locations. The aim of targeting agile professionals in the software industry from different countries was to get their feedback based on their everyday experience in the domain. Agile researchers were also targeted to get the latest updates on this area as there are ongoing studies being carried out. A self-administered questionnaire was designed and distributed to the respondents via mail, websites and mobile applications (WhatsApp). The survey was conducted to both local companies and offshore companies, which are well-known in providing consulting information technology services. The primary focus of this survey is to determine whether there is a gap between existing researches and real-life scenario. The survey was conducted to determine the list of factors that have an impact on the task allocation and coordination process in agile distributed software development.

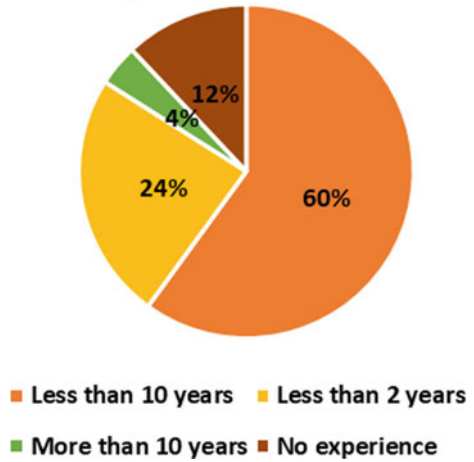
A sample size of 25 respondents was considered. Snowball sampling was used to gather responses. Snowball sampling also known as chain referral sampling is a sampling technique where participants provide referrals to recruit samples [22]. In this technique, a few known participants are asked to fill the survey, and they are requested to look for other participants in their proximity to participate in the survey. The reason behind choosing snowball sampling was because the agile distributed software development population was unknown and the agile professionals were hardly reachable. In order to increase the sample size, participants having agile profiles were obtained using the search engine of LinkedIn. It is difficult to reach project managers and team leaders in large agile companies, so the team members were asked to identify other potential participants and to motivate them to come forward. The survey does not only gather feedback on the factors provided but also contain open-ended questions where the participants are asked for other factors and dependencies if not present in the list. They are asked for the reason behind choosing for the three best factors and dependencies affecting task allocation and coordination.

4 Survey Results and Discussion

The results of the survey are presented in this section. The study used agile professionals having role of team leader, team member, project manager, researcher in the field and others (software engineer). Most of the participants were those working in agile global companies found in Mauritius. Figure 1 depicts the number of years of experience of the participants. 60% of the respondents had more than 2–10 years

Fig. 1 Years of experience

Years of experience in an agile environment



of experience. Among the 12% of the participants who had no experience, there are agile researchers and other software developers who wanted to give their opinion on the different factors that might affect task allocation and coordination in DASD. The rest have less than 10 years' experience. As depicted in Fig. 2, 72% of the respondents had experience in an agile global team, and the rest either worked in agile team from co-located locations or merely had no experience. Figure 3 depicts the role played by the participants in agile environment. Figure 3 shows that the majority of

Fig. 2 Experience in DASD

Have you worked in an agile global team?

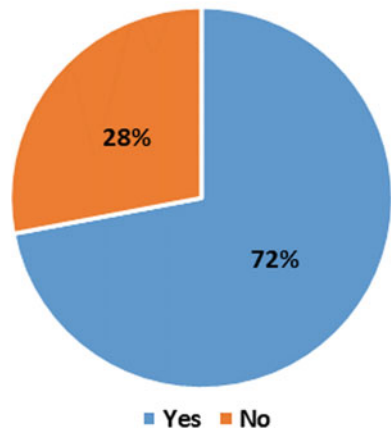


Fig. 3 Roles



the respondents were team members. The 15% who are agile project managers and 18% who are agile team leader put a certain weight on the accuracy of the results. The rest are agile researchers and non-agile participants.

4.1 Ranking of Factors

Practitioners were requested to rank the factors from Table 1 based on their importance in the task allocation and coordination process in agile distributed software development. Figure 4 depicts the value scores for the factors ranked by the participants in descending order. The value scores were obtained from the relative value of each factor ranked by the practitioners. Weights were assigned: 6 being most important factor and 1 being least important. The value scores are calculated by summing the weights given by each participants for each factor. In open-ended questions, the practitioners mentioned that global emergency crisis in that particular country might impact the product and other factors like poker planning, bug tracking and continuous integration/continuous delivery, continuous feedback, project size, task breakdown, appreciation and delivery mechanism need to be taken into consideration.

The survey response data show that communication, coordination, effort, task deadline, transparency, prioritized delivery, project manager maturity and team willingness are highly ranked by the participants. These are the factors that determine the outcome and influence the project lifecycle and delivery. They bring stability, productivity and quality delivery to the development of the product. Communication and coordination were highly valued as they are the basic requirements for proximity to customer requirements. Communication is important to be able to be

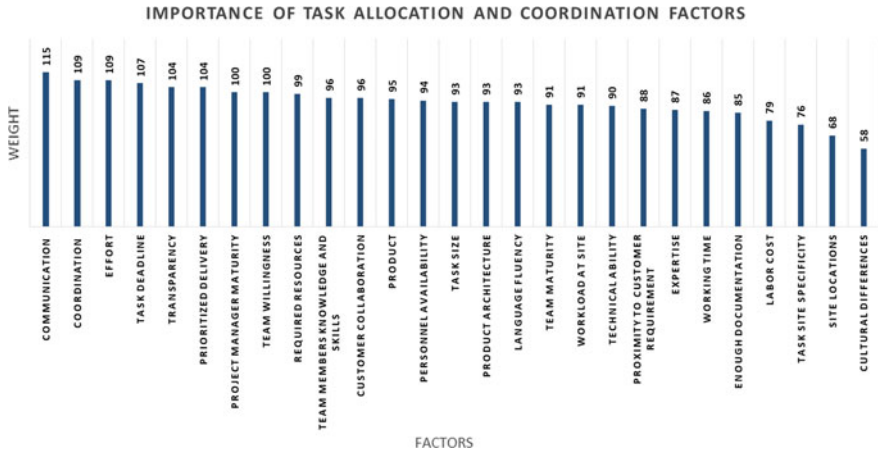


Fig. 4 Score per factor

on the same page as everyone, that is, everyone is aware what is happening in the project. Communication ensures that right product is being developed correctly. It is important to know the task deadline to know where to concentrate the effort.

Transparency was ranked important as smooth handing-over between new and existing team members is important. Having proper sign off by clients for each release is a way of promoting transparency between clients and team. Each user story will have valid and solid acceptance criteria. As a result, each sprint will have a proper scope, and client/development team will be on the same level. Transparency gives both the client and consultants confidence that things are being done in the right way. Some practitioners mentioned that proximity to customer requirement is important as it defeats the whole purpose if customer requirement is not met. A participant mentioned that technical abilities/ knowledge/skills/communication are needed to maintain coordination and deliver a product as per the client’s requirements. In agile environment, team members’ knowledge and skill level are crucial to task allocation in order to avoid rework and customer frustrations. Prioritized delivery is essential to avoid missing the deadlines. The resource and team effort are important for delivering quality products. A greater task estimation and cost can be calculated from those variables. Task deadline is a crucial element as it guides in task prioritization. Resources can be mobilized and channelized accordingly once clarity is achieved. Labour cost, task site specificity, site locations and cultural differences appear to have less importance. The respondents put less emphasis on working time, expertise, workload at site and documentation. Despite being important, these factors are less valued. This might be the cause of failures in the global agile companies.

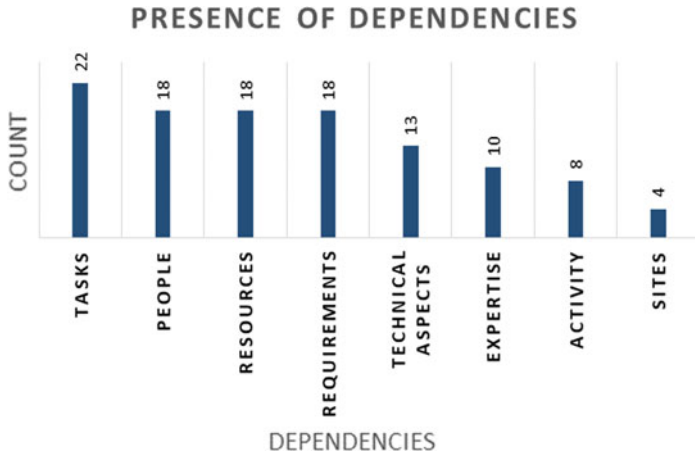


Fig. 5 Score per dependency

4.2 Existing Dependencies

The practitioners were given a list of dependencies between the factors listed in Table 2 and were asked to choose those present in the environment that need to be taken into consideration. Figure 5 depicts the total number of responses obtained per dependencies in descending order. Other than these dependencies, a practitioner mentioned about business process dependency. Others mentioned that daily feedback is required from team members and clients. Agile development is costly without the right team. Practitioners gave more value to tasks, people, resources and requirements dependency for task allocation scenario. The dependencies generate the needs for communication and coordination as well.

4.3 Limitations of the Survey

The survey presented a few limitations. The response rate for the survey was low. It was not easy to gather responses from the target group as they had limited availability. In open-ended questions, the practitioners were asked for other factors not present in the literature. Few responses were obtained, and the ranking of the factors was not mentioned. Hence, it is difficult to distinguish the level of importance of the proposed factors. Additionally, due to the anonymity of the survey responses, it was difficult to contact the practitioners for more clarifications.

4.4 Recommendations and Future Works

In this study, it has been found that task allocation is a complex decision-making process that requires multiple factors and dependencies to be taken into consideration. The survey results show that despite having multiple factors that influence task allocation, only a few of them are given attention. Factors like effort, transparency, project manager maturity, and product architecture are given precedence over expertise, working time, site locations, labour cost and cultural differences despite being highlighted by literature as influential factors during site selection. Despite working with teams from global locations, sites dependencies are least valued. The results obtained from the survey call for a correlation between the different factors and dependencies. The gap between the literature and the software industry practice motivates us to mount an ontology based on the literature and results obtained in the survey.

5 Conclusion

Many projects fail due to lack of attention to factors and dependencies between factors that influence task allocation and coordination in DASD. In this study, we first extracted the factors and dependencies influencing task allocation and coordination in agile distributed software development in the area from the literature. We then validated the extracted results across a survey with 25 agile professionals. Results of the survey highlight the importance of the factors for task allocation and coordination. Multiple factors were given high weightage, and this signifies the importance of the factors. The results also show that dependencies between technical aspects, expertise, activity and sites are ignored in the team. Difference in site locations is not incorporated in the task allocation process. There is therefore the need to define a taxonomy considering various factors that influence task allocation and coordination in DASD.

6 Ethics Compliance

Consent was obtained from all participants who participated in the survey. They were informed that the data provided would be kept anonymous and no personal nor health-related information would be recorded, prior to conducting the survey.

References

1. Almeida, L.H., Albuquerque, A.B.: A multi-criteria model for planning and fine-tuning distributed scrum projects. In: 2011 IEEE Sixth International Conference on Global Software Engineering IEEE. August, pp. 75–83 (2011)
2. Aslam, W., Ijaz, F.: A Quantitative framework for task allocation in distributed agile software development. IEEE Access. **6**, 15380–15390 (2018)
3. Banijamali, A., Dawadi, R., Ahmad, M.O., Similä, J., Oivo, M., Liukkunen, K.: An empirical study on the impact of Scrumban on geographically distributed software development. In: 2016 4th International Conference on Model-Driven Engineering and Software Development MODELWARD IEEE. February, pp. 567–577 (2016)
4. Collins, E., Macedo, G., Maia, N. and Dias-Neto A.: An industrial experience on the application of distributed testing in an agile software development environment. In: 2012 IEEE Seventh International Conference on Global Software Engineering, pp. 190–194 (2012)
5. Espinosa, J.A., Slaughter, S.A., Kraut, R.E., Herbsleb, J.D.: Team knowledge and coordination in geographically distributed software development. *J. Manag. Inf. Syst.* **24**(1), 135–169 (2007)
6. Hashmi, A., Hafeez, Y., Jamal, M., Ali, S., Iqbal, N.: Role of Situational Agile Distributed Model to Support Modern Software Development Teams. Vol. 38. July, pp. 655–666 (2019)
7. Ijaz, F., Aslam, W.: Identification of dependencies in task allocation during distributed agile software development. *Sindh Univ. Res. J. Sci. Ser.* **51**, 31–36 (2019)
8. Intiaz, S., Ikram, N.: Dynamics of task allocation in global software development. *J. Softw. Evolut. Process.* **29**, e1832 (2017)
9. Lin, J.: Context-aware task allocation for distributed agile team. In: 2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE). pp. 758–761 (2013)
10. Lohan, G., Lang, M., Conboy, K.: Having a customer focus in agile software development, pp. 441–453. In: *Information systems development*. Springer, New York (2011)
11. Marques, A.B., Carvalho, J.R., Rodrigues, R., Conte, T., Prikładnicki, R., Marczak, S.: An ontology for task allocation to teams in distributed software development. In 2013 IEEE 8th International Conference on Global Software Engineering. IEEE. August, pp. 21–30 (2013)
12. Moe, N.B., Šmite, D., Šablīs, A., Börjesson, A.L. and Andréasson, P.: Networking in a large-scale distributed agile project. In: *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, p. 12 (2014)
13. Moe, N.B., Cruzes, D., Dybå, T., Mikkelsen, E.: Continuous software testing in a globally distributed project. In: 2015 IEEE 10th International Conference on Global Software Engineering. IEEE. July, pp. 130–134 (2015)
14. Nordio, M., Estler, H.C., Meyer, B., Aguirre, N., Prikładnicki, R., Di Nitto, E., Savidis, A.: An experiment on teaching coordination in a globally distributed software engineering class. In: 2014 IEEE 27th Conference on Software Engineering Education and Training (CSEE&T). IEEE. April, pp. 109–118 (2014)
15. San Cristóbal, J.R., Carral, L., Diaz, E., Fraguera, J.A., Iglesias, G.: Complexity and project management: a general overview. *Complexity* (2018)
16. Simão Filho, M., Pinheiro, P.R., Albuquerque, A.B.: Task allocation approaches in distributed agile software development: a quasi-systematic review. In: *Software Engineering in Intelligent Systems*. Springer, Cham, pp. 243–252 (2015)
17. Simão Filho, M., Pinheiro, P.R. and Albuquerque, A.B.: Task assignment to distributed teams based on a qualitative multi-criteria approach. In: 2017 12th Iberian Conference on Information Systems and Technologies (CISTI). IEEE. June, pp. 1–6 (2017)
18. Simão Filho, M., Pinheiro, P.R., Albuquerque, A.B., Rodrigues, J.J.: Task allocation in distributed software development: a systematic literature review. *Complexity* (2018)
19. Simão Filho, M., Pinheiro, P.R., Albuquerque, A.B., Simão, R.P., Azevedo, R.S., Nunes, L.C.: A multicriteria approach to support task allocation in projects of distributed software development. *Complexity* (2019)
20. Strode, D.E., Hope, B.G., Huff, S.L., Link, S.: Coordination effectiveness in an agile software development context. In: *PACIS*. July, p. 183 (2011)

21. Szőke, Á.: Optimized feature distribution in distributed agile environments. In: International Conference on Product Focused Software Process Improvement. Springer, Berlin, Heidelberg, June, pp. 62–76 (2010)
22. Naderifar, M., Goli, H., Ghaljaie, F.; Snowball sampling: a purposeful method of sampling in qualitative research. *Strid. Dev. Med. Educ.* **14**(3) (2017)

Optimizing Recruitment Process Within Businesses: Predicting Interview Attendance Using C4.5 Algorithm



Shivianee Sandhip Laldjee, Chiamaka Ann Marie Ajufo,
and Girish Bekaroo

Abstract An essential component of the job recruitment process involves conducting interviews, during which decision is made on the appropriateness of a candidate for a particular job. An important challenge faced by interviewers during the recruitment process relates to determining whether or not selected candidates will attend the job interview. Candidates failing to attend interviews because of various reasons lead to time wasting and as such, effective prediction is needed by human resource managers about whether or not a candidate will attend an interview. In other words, if the human resource department knows exactly whether or not the job seeker is attending the interview, a more efficient interview invitation strategy could be worked upon. However, even though different machine learning techniques are available for prediction, limited work has been undertaken to address this recruitment process related issue. Taking cognizance of this problem, this paper investigates, analyzes, and predicts candidate attendance during interviews in order to optimize recruitment process within businesses. For this, C4.5 algorithm was applied on an open dataset related to job interview attendance. Following generation of a classification tree, results showed some insightful patterns on when candidates do not attend job interviews.

Keywords Interview attendance prediction · C4.5 algorithm · Classification · Human resource management · Data mining

1 Introduction

During recent years, companies have increasingly adopted e-recruitment, which enables a wider, faster, and cost-effective recruitment process [1]. Indeed, matching the right profile to the right job description has never been easier with the various tools offered by recruitment platforms [2]. Interviews form an important part of the

S. Sandhip Laldjee (✉) · C. A. M. Ajufo · G. Bekaroo
School of Science and Technology, Middlesex University Mauritius, Coastal Road, Unicity,
Flic-En-Flac, Mauritius
e-mail: shivianeesandhip@gmail.com

recruitment process and enable decision-making on whether or not the candidate should be considered for the job. The complexity of interview process depends on various factors such as company size and job description where the process can take more than one round while being time consuming [3]. A previous study showed that in 2015, the selection process took more time in the UK as compared to 2011, increasing from 13 days to 23 days [4]. As such, reducing the recruitment time can provide advantages to both the company and the applicants, where a lengthy recruitment process can frustrate potential applicants. They might then start to look for other opportunities and the company misses out on a good element [5]. This process is worsened when candidates fail to attend interviews because of various reasons, thus wasting further time of companies and interviewers [6]. A way to reduce time and efforts wasted would be to predict if the candidate will attend an interview or not, in order to allow the job seeker to decide whether to call the candidate for the interview or not. Although techniques such as machine learning could be used for this prediction, limited work has been conducted in this direction.

As related work, a previous study presented a content-based recommender system to propose jobs to users on social networks including Facebook and LinkedIn [7]. Although machine learning techniques were used in this work, its focus was different and involved predicting user interests for jobs through the use of basic similarity measures. Another study attempted to increase efficiency of the recruitment process by utilizing a machine learning prototype system so as to limit interviewing of job-seekers solely to the top candidates [1]. The proposed system in the same study intended to aid recruiters by automating the process of evaluating the candidates' profiles in order to determine the most suitable ones for the job position. Although the system showed to be effective, its direct focus was not on predicting attendance of candidates during interviews. With limited study conducted to apply machine learning in the recruitment process, this paper investigates, analyzes and predicts candidate attendance during interviews in order to optimize recruitment process within businesses. This paper is intended to be beneficial to human resource (HR) management units within businesses toward helping to improve candidate selection and to better manage time and resources during interviews.

This paper is organized as follows: In the first section, the introduction to the research problem is given followed by describing the techniques and algorithms used for prediction in Sect. 2. In the third section, the methodology used to achieve the purpose of the paper is detailed and Sect. 4 presents the findings of the study. Finally, the conclusions and avenues for future work are presented in Sect. 5.

2 Theoretical Background

In scientific research, decision trees are important structures that help to effectively lay out problems for analysis and decision-making. Fundamentally, a decision tree is an analysis diagram that uses a tree-like graph design which aids in making decisions by projecting possible outcomes [8]. A popular algorithm used to produce

decision trees is the C4.5 algorithm. This algorithm was proposed by Quinlan in 1993 to extend the ID3 algorithm, after different limitations were highlighted [9]. The C4.5 algorithm was chosen as part of this study due to its high accuracy in decision-making as compared to similar ones [8, 9]. Also, the resulting classification rules generated by this algorithm is human readable and easy to understand thus simplifying interpretation [11]. The algorithm works by taking a dataset as input and outputs a tree similar to an orientation diagram where each node (leaf) represents the decision (class) of data which verifies all tests from the root to leaf [10]. Due to its benefits, this algorithm has been used in a multitude of studies related to machine learning and prediction. For instance, the C4.5 algorithm has been used to predict targeted destination port numbers during network attacks [12], analyze the selection of exemplary teachers [13], identify promising students by predicting student performance [14], and predict breast cancer survivability [15], among others. In the field of human resource management, this algorithm has been used to predict human talents [11]. However, among the related works, no studies have been undertaken to predict attendance of candidates during interviews, thus making the gap addressed in this study relevant to investigate.

3 Methodology

To achieve the purpose of this paper and to predict attendance of candidates during interviews, the C4.5 algorithm was used. For this study, a dataset [16] related to interview attendance within companies was used, that dataset contains 23 columns of data collected between September 2014 and January 2017. It contains information related to the job position, industry, and location, among others. Following acquisition of the dataset, the next stage involved preprocessing the dataset in order to prepare it for analysis. This started by filtering the different columns available in the dataset, some of them were removed due to their irrelevance to the scope of this paper. Having an unoptimized list of columns would increase the processing time while also generating a complex decision tree. In other words, the selection of attributes is important in order to optimize the results and obtain meaningful information so as to make strategic decisions related to the effectiveness of recruitment process. After this filtering, the attributes given in Table 1 were left for processing.

For analysis, the Waikato Environment for Knowledge Analysis or Weka was utilized in this study. Weka is a suite of machine learning software written in Java and is commonly used for data mining [17]. In Weka, a Java implementation of the C4.5 algorithm is present and is called the J48 algorithm. For preparing the dataset for Weka, the file had to be converted to the Attribute-Relation File Format (ARFF) format. When opening the dataset in Weka, some attributes which had the same values but were written differently, were recognized as two different values by Weka. It occurs mainly because Weka is case sensitive and takes whitespaces into account while reading the values. For instance, no, NO, No and no (with an extra space at the end), are seen as four different values. To correct this, a standard value was chosen

Table 1 Description of attributes

Attribute/factor	Choice of answers
Expected attendance	Yes, no, NA, uncertain
Observed attendance	Yes, no
Have you obtained the necessary permission to start at the required time?	Yes, no, NA, not yet, yet to confirm
Can I call you 3 h before the interview?	Yes, no, NA
Can I have an alternative number?	Yes, no, NA
Did you print out your updated resume and understood the job description?	Yes, no, NA, not yet

by replacing all irregularities. Some changes had to be made for different attributes to ensure consistency and reliability of data being used. Once the irregularities were fixed, the C4.5 algorithm was trained on Weka using the 10-fold cross validation as it gives a more accurate estimation [18, 19] and the algorithm also works better with a larger training set [20]. This also means 90% of the data was be used as a training set and the process will be repeated ten times, and each time the training and test set will be changed. The classification tree was then generated on Weka and interpreted.

During the experiment, the key challenges faced were during the preprocessing stage. Firstly, there were compatibility issues between the original dataset file and Weka, where the dataset was not readable and required a great amount of modifications. Also, values of various fields in the dataset had to be standardized, while also ensuring that null values were not present. This process led increased the duration of the preprocessing stage due to the large number of rows present within the dataset. Another challenge was to find the right combination of attributes so as to obtain the optimum prediction for interview attendance and this involved trial and error sessions.

4 Results and Discussions

After running the C4.5 algorithm on the 1233 rows, 67.15% of the records were correctly classified while the remaining could not be classified by Weka. This high percentage for the correctly classified instances also means that the values are accurate enough to perform the prediction. The statistical summary following the execution of C4.5 algorithm on Weka is given in Fig. 1. The incorrectly classified instances were particularly due to records containing uncertain values for the different attributes given in Table 1.

Further analysis was then made pertaining to the occurrences of each attribute to ensure that each attribute only consisted of the choice of answers given in Table 1. A summary is provided in Table 2 where NA means ‘not applicable’ u means ‘uncertain’ YTC means ‘yet to confirm’ and NY means ‘not yet’. From Table 2, it could be


```

=== Summary ===
Correctly Classified Instances      828          67.1533 %
Incorrectly Classified Instances    405          32.8467 %
Kappa statistic                    0.2267
Mean absolute error                 0.4351
Root mean squared error             0.4672
Relative absolute error             93.8559 %
Root relative squared error         97.0381 %
Total Number of Instances          1233
    
```

Fig. 1 Classification summary by the J48 classifier

Table 2 Number of occurrence of different attributes

Attribute/factor	Yes	No	NA	U	YTC	NY
Expected attendance	885	93	5	250		
Observed attendance	783	450				
Have you obtained the necessary permission to start at the required time?	921	80	209		4	19
Can I call you 3 h before the interview?	955	11	267			
Can I have an alternative number?	937	29	267			
Did you print out your updated resume and understood the job description?	942	17	268			6

seen that from the 885 candidates who were expected to attend, only 783 attended, representing 88.5% of the expected candidates. In total, out of the 1223 candidates, it could be seen that only 63.5% of candidates attended interviews. This also implies that an important number of candidates, notably 36.5% are either uncertain or plan not to attend the interviews. As such, for these cases, if the human resource department knows exactly whether or not the job seeker is attending the interview, a more efficient interview invitation strategy could be worked upon. Furthermore, in Table 2, it could be seen that there are various instances where candidates respond No to different queries such as ‘Did you print out your updated resume and understood the job description?’.

After running the algorithm, the classification tree was generated in Weka as depicted in Fig. 2 to predict attendance of candidates during interviews. Visualization and analysis of the generated tree revealed different insights. Among the findings, if the applicant has a printout of the updated resume, has read the job description and does not want a follow-up call, but has an alternative number, he/she is going to attend to interview in 100% of the cases. Also, it was found that if a participant is not certain to attend an interview, does not have a printout of an updated resume, has not read the job description, and does not want a follow up call, there is a 100% chance that the candidate is not attending the interview. The same insight was found if in

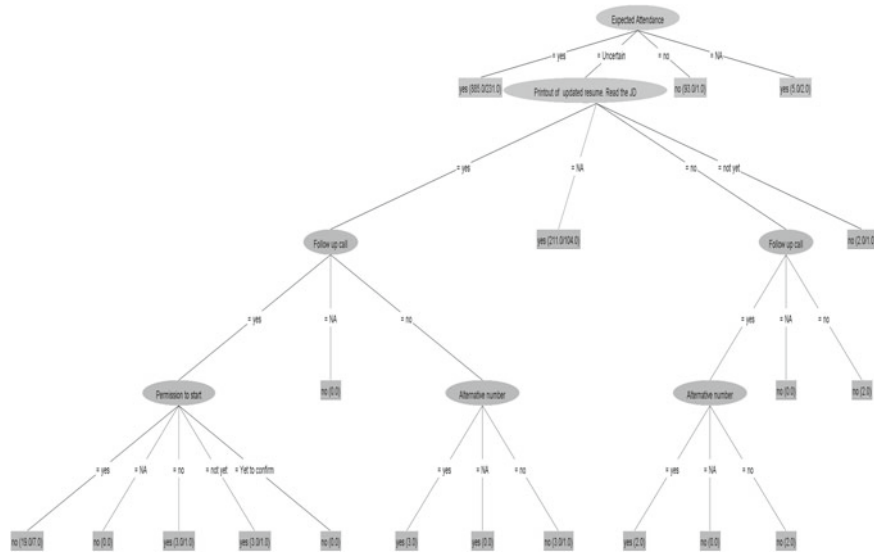


Fig. 2 Classification tree generated on Weka

another case, the job seeker wants the follow up call but does not have an alternative number. On the contrary, if the candidate wants a follow up call, and has an alternative number, there is 100% attendance chance.

Furthermore, another interesting attendance probability obtained was that if the applicant answers that he/she does not have an alternative number, he/she is not likely to attend the interview in 66.6% of the cases, as 2 instances out of 3 have been classified as a no in observed attendance for this branch of the tree. This rate is high indeed and could be taken into account by the recruiter to make certain decisions. Moreover, if an applicant has taken a printout of his/her updated resume, wants a follow up call but did not get permission to start yet, he/she is still 66.6% likely to attend. In addition to the same criteria, it was found that in 63.2% of the cases, applicants do not attend the interview even if the applicant has obtained the necessary permission to start at the required time. This result is ironical since logically an applicant is more likely not to attend an interview if he/she does not have the permission to start at the mentioned date. Another interesting pattern found is that the observed attendance rate of applicants based on those who did not answer (NA) to '3 h prior follow up call' given in Table 1. As deduced from the tree in Fig. 2, applicants who do not answer have an attendance rate of 40.1%, which is low and less than half of the total number of applicants who did not answer. As such, these candidates are more likely not to attend the interview.

Besides, the study is undermined by the limitations of the C4.5 algorithm as its run-time complexity relates to the depth of the tree and cannot be greater than the number of attributes [12]. As such, the depth of the tree is related to the tree size and thus to the number of examples [8]. Another limitation relates to the dataset

which has only 1233, which is limited in terms of details provided. For instance, details on industry (e.g., information technology, tourism or manufacturing), country in which the job was advertised or other demographic details of participants (e.g., gender or age) were unavailable, which could be analyzed to better investigate how demographic or geographic details influence interview attendance.

5 Conclusion and Future Works

This paper investigated, analyzed, and predicted candidate attendance during interviews in order to optimize recruitment process within businesses. For this, the C4.5 algorithm was applied to a selected open dataset containing interview related information for 1233 candidates. This algorithm was trained on Weka by utilizing 90% to generate decision tree and the rest was used to evaluate the tree generated. Analysis of the dataset showed that only 63.5% of candidates attended interviews, thus highlighting the need to predict attendance during interviews so as to save time and costs during recruitment. The classification tree generated showed some insightful patterns on failure to attend interviews. For instance, it was found that if a candidate mentions to be uncertain to attend an interview, does not have a printout of an updated resume, has not read the job description, and does not want a follow up call, there is a 100% chance that the candidate is not attending the interview. Likewise, participants who do not answer the follow-up call only have 40.1% chance of attending. As such, insights revealed in this paper could help the HR department decide whether or not the job seeker is attending the interview and implement a more efficient interview invitation strategy.

As future works, percentage of rows for the training and evaluation sets could be varied toward generating different classification trees in order to analyze any variance against the one presented in this study. Based on the same dataset used in this study, the influence of the removed attributes on the classification tree produced could be investigated. In addition, further datasets could be investigated in order to further confirm results produced and to also analyze how demographic or geographic details influence interview attendance.

References

1. Faliagka, E., Ramantas, K., Tsakalidis, A., Tzimas, G.: Application of machine learning algorithms to an online recruitment system. In: In Proceedings of the International Conference on Internet and Web Applications and Services (2012)
2. Crous, H.: E-recruitment versus Traditional Recruitment Methods (2016) [Online]. Available <https://www.linkedin.com/pulse/e-recruitment-vs-traditional-recruitment-methods-henneri-crous>. Accessed 3 Apr 2020
3. Huegeli, J., Tschirgi, H.: An investigation of the relationship of time to recruitment interview decision making. In: Academy of Management Proceedings, Briarcliff Manor (1975)

4. Dishman, L.: Why The Hiring Process Takes Longer Than Ever. Fast Company [Online] (2015). Available <https://www.fastcompany.com/3048421/why-the-hiring-process-takes-longer-than-ever>. Accessed 4 Apr 2020
5. Meilak, L.: The importance of a fast recruitment process. Flexi Personnel [Online] (2017). Available <https://www.flexipersonnel.com.au/blog/professional-recruitment/the-importance-of-a-fast-recruitment-process>. Accessed 10 Feb 2018
6. Ball, J.: The 6 Most Common Excuses for Missing an Interview. Coburg Banks [Online] (2017). Available <https://www.coburgbanks.co.uk/blog/candidate-tips/excuses-for-missing-interviews/>. Accessed 13 Feb 2020
7. Diaby, M., Viennet, E., Launay, T.: Toward the next generation of recruitment tools: an online social network-based job recommender system. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2013)
8. Bhargava, N., Sharma, G., Bhargava, R., Mathuria, M.: Decision tree analysis on J48 algorithm for data mining. Proc. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**(6) (2013)
9. Hssina, B., Merbouha, A., Ezzikouri, H., Erritali, M.: A comparative study of decision tree ID3 and C4.5. Int. J. Adv. Comput. Sci. Appl. **4**(2) (2014)
10. Amin, R., Sibaroni, Y.: In Information and Communication Technology (ICoICT), Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region). In: 2015 3rd International Conference on Information and Communication Technology (ICoICT) (2015)
11. Jantan, H., Hamdan, A., Othman, Z.: Human talent prediction in HRM using C4.5 classification algorithm. Int. J. Comput. Sci. Eng. **2**(8), 2526–2534 (2010)
12. Gangabissoon, T., Nathoo, A., Ramhith, R., Gopee, B., Bekaroo, C.: Improving effectiveness of honeypots: predicting targeted destination port numbers during attacks using J48 algorithm. In: International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (2018)
13. Siahaan, H., Mawengkang, H., Efendi, S., Wanto, A., Windarto, A.: Application of classification method C4. 5 on selection of exemplary teachers. J. Phys. Conf. Ser. **1**, 012005; **1235** (2019)
14. Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., Honrao, V.: Predicting students' performance using ID3 and C4.5 classification algorithms. arXiv preprint (2013)
15. Bellaachia, A., Guven, E.: Predicting breast cancer survivability using data mining techniques. Age **58**(13), 10–110 (2006)
16. Desai, R., Leo, M., Nakra, R., Prima and Trupthi: The interview attendance problem—predict which candidates will attend the interview. Kaggle (2017) [Online]. Available <https://www.kaggle.com/vishnusraghavan/the-interview-attendance-problem>. Accessed 15 Feb 2020
17. Holmes, G., Donkin, A., Witten, I.: Weka: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems (1994)
18. Quinlan, J.: C4.5: Programs for Machine Learning. Elsevier (2014)
19. Brownlee, J.: How to choose the right test options when evaluating machine learning algorithms, machine learning mastery (2014) [Online]. Available <https://machinelearningmastery.com/how-to-choose-the-right-test-options-when-evaluating-machine-learning-algorithms/>. Accessed 5 Mar 2020
20. Salzberg, S.: C4. 5: programs for machine learning by j. ross quinlan. Mach. Learn. **16**(3), 235–240 (1994)

Training Engineers as Drivers of e-Learning in a University



Nirmal Kumar Betchoo

Abstract e-learning has become unexpectedly popular in the educational sector in Mauritius. In this context, the Université des Mascareignes (UdM) has partnered with Université de Caen in developing an e-learning strategy where it trains academics to become training engineers with the perspective of implementing effective e-learning platforms. This paper considers the need to have a structured form of virtual learning that will be dependent on factors like the relevance of e-learning, the importance of training engineers in e-learning, the approaches needed to develop effective e-learning and the ways in which e-learning could be sustained. A survey was conducted among the 17 participants registered as the training engineers course. Results explained that training engineers were firstly willing to be certified as training engineers which would allow them properly develop e-learning. Potential training engineers considered learning to be provocative which would demand involvement in developing dedicated platforms for effective interaction between learner and trainer. Trainer motivation and student capacity mattered mostly in supporting the sustainability of e-learning.

Keywords e-learning · Training engineers · Education · University · Learning sustainability

1 Introduction

e-learning is an alternative to learning today in most societies including Mauritius. This form of learning blends with traditional face-to-face learning which is by far the most common method of learning in Mauritius. Face-to-face learning ensures direct contact between the educator and the student. Web 2.0 or alternately e-learning has recently been heralded as the opportunity cost to face-to-face learning. This situation has been exacerbated by the COVID-19 pandemic that affected the world since its discovery in December 2019 in the Wuhan Province in China. It claimed nearly a

N. K. Betchoo (✉)

Université Des Mascareignes, Beau Plan Campus, Pamplémousses, Mauritius

e-mail: nbetchoo@udm.ac.mu

hundred thousand lives internationally and forced more than 200 nations and 3 billion people to stay in confinement for quite long [1]. Incidentally, schools were closed in most parts of the world and the confinement directly impacted classroom attendance because all classes were closed for a couple of weeks. Concerning the Mauritian Government, the Ministry of Education, Tertiary Education and Scientific Research had no other alternative than having recourse to e-learning. This strategy had to be developed and implemented on a contingency basis through the development of learning platforms like Zoom, Moodle, Google classroom including classes directly run from television broadcast. The support form provides students time to engage in predictable and structured learning opportunities [2]. Although the initiative was an excellent one, the main problem was that there was no adequate preparation regarding the development of e-learning and the teaching activities, though commendable, did not prove to be totally effective because the training and the implementation were not planned. There was a need to see how training could be well developed. The same practice took place at the Université des Mascareignes (UdM) where a series of e-learning platforms and methodologies were adopted to immediately address the problem. Despite the fact that the learning activities took place during the national confinement, it was imperative that e-learning had to be structured to ensure that this becomes an alternative learning mode and also addresses certain immediate requirements of students.

1.1 Literature Review

There have been earlier approaches to provide alternative forms of learning in the country. The ex-Mauritius College of the Air provided distance learning courses in the form of correspondence courses to Mauritian students who could not attend a formal face-to-face course. At the same time, distance learning was promoted internationally by popular institutions like the Rapid Results College (RRC) and Wolsey Hall Oxford that provided 'O' and 'A' level courses along with professional courses like the chartered professional examinations in accountancy, marketing and management [3]. The University of London and the University of South Africa offered distance learning degree programmes to Mauritian students. These examples are just illustrations of alternative learning earlier provided in Mauritius and that helped hundreds of Mauritian students learn indirectly from educational support providers.

The profession of the training engineer is linked to that of education, and he must, *inter alia*: be able to implement and encourage training conceptualization approaches that have a solid theoretical and methodological background; know how to mobilize theories for application or as an aid to the understanding of social problems addressed in formation (risky behaviours and their modification, engaging communication, questions of changes in behaviour, etc.); be able, in a case of lack of knowledge, to produce new ones by prompting new studies; be able to design and implement concrete training proposals and understand the consequences of their modification.

In other areas close to training, the use of the term engineering is dated to the mid-1980s and helps formalize ‘skills’ from which a certain ‘employability’ and the idea of ‘reflexivity as competence’ arise [4]. The stakes are all the more pervasive in the field of consultancy, noting that ‘the ability to perform as an activity is the basis of all interactive services: psychotherapy, counselling, teaching, trade, etc., are all activities of implementation or even staging of oneself’ [5].

Companies invest in their development thanks to the skills of the training engineer: it is up to him to meet this challenge. His knowledge of information and communication technologies and their application in open and remote training are assets of the training engineer. With the utmost discretion, the engineer of training benefits from having the confidence of employees. Even more than other human resources functions, the training engineer requires a flawless ethic: to be a creative and organized educator. Having a good legal culture as well as a good competence in budgeting and project management are the other assets of the training engineer [6].

1.2 The Framework for Learner Support

There are different contributions from researchers on the term ‘accompaniment’ or learner support in the e-learning process. This term developed in the late 1990s probably due to the development of learning technologies including the Internet and online training. Learner support is viewed as governing, helping, training, teaching techniques or imbued with values of commitment. This includes all the basic ideas about training. Therefore, training is not only tutoring or teaching, but also accompanying and mentoring the learner through values such as perfecting knowledge, developing knowledge, developing skills related to a discipline [7].

Learner support is achieved through tutoring, counselling, coaching, sponsorship, mentoring, counselling, consulting, companionship and sponsorship. Tutoring means teaching, counselling involves directing, coaching is confined to training, sponsorship and mentoring rely on coaching at the time of teaching, counselling, consulting, companionship and sponsorship being broader approaches to primary objectives such as teaching and tutoring [8]. Therefore, this perspective prioritises the social, emotional relationship in teaching including the basic elements that are the essential work of the trainer or tutor.

The trainer is also considered a companion by his ability to follow the student’s progress during his journey. This is analogical to sharing bread in the form of ‘*cum panis*’ which basically means sharing knowledge, activity and mutual commitment between him and the student [9]. As a pilgrim, the student or learner must be accompanied in his learning up to a fixed period such as the duration of a course that is defined under hours of training.

Training could be ‘provocative’. This is initially manifested by distance as soon as one introduces, to transmit knowledge, a ‘pedagogical material’ that is more technological, between the teacher and the student, even if by chance this distance is *de facto* very small [10].

There is a difference between traditional teaching-centred training that promotes knowledge but other permutations such as learning-centred and the participatory approach that result in learning, according to them, better in the form of motivation, activities, production, etc., which are generally superior to what is taught [11]. In short, hybridisation is not only a technological contribution but, with this online support, frees students' skills and makes them become motivated, enthusiastic, creative learners, among others [12].

Far from conforming to a role, a standard pattern of behaviour, each trainer has his own identity and sensitivity and makes it possible for him to focus on a particular part of the range of fields and tasks on which he can intervene, in a game of adaptation and affirmation that guarantees the presence for which he is being asked. By putting their activity online, trainers obliged to have to redefine its actions, their range and limitations, and to think of them articulated to a particular strategy and other actors with whom they must agree. The evolution is profound, in terms of identity representations on the teaching/learning relationship, and, in a reflexive approach, the obligation to have to represent one's own action [13].

1.3 Research Questions

Research questions emerge from the literature review giving rise firstly to the basic importance of bringing e-learning in the university educational system to form a hybrid mode of learning. At the same time, there are questions regarding its long-term viability. The key questions are as follows:

Why is e-learning an imperative at the Udm?

What is the relevance of training engineers in e-learning?

What approaches should training engineers consider in developing e-learning platforms?

How does the training engineer sustain e-learning at the Udm?

1.4 Research Methodology

The research methodology was based on information channelled among lecturers of the Udm enrolled for the programme 'Online course for Training Engineers' offered from January to June 2020 by the Université de Caen (France) under the Formation à Distance (FOAD) programme. Seventeen lecturers were enrolled for the course, and they formed small groups of three to four to prepare for the course. All of them were included for the research survey. There were questions asked to staff within the group with respect to the training engineers. The information was gathered from the direct interventions that staff had on the 'Blue Button' platform where they directly intervened with their supervisors and trainers from Université Caen. There was a mix of formal and informal intervention. The researcher selected the salient information

during the interventions that were recorded for internal use from all the participants. Discussions and forums formed the basis of data collection. A substantive amount of primary data also came from WhatsApp that was purposely developed for the training programme. These interventions formed the basis of assessing primary data for the research undertaken.

1.5 Research Findings

The research findings evaluated four key questions that were asked and focused on the following: the importance of e-learning at the UdM, the relevance of training engineers in e-learning, an evaluation of the approaches that training engineers face in developing e-learning platforms including sustain e-learning at the UdM. The findings are presented in the following sub-section.

1.5.1 The Relevance of e-learning at the UdM

The first question concerned the relevance of e-learning at the UdM. The university has traditionally provided face-to-face learning both on a full and part-time basis. E-learning has been promoted by the INCRE (Institute de Création de Ressources en ligne) which was set as an alternative to teaching and learning. All the respondent's views were considered on this question. The results are displayed below.

Higher scores went towards the choice of being certified as a trainer or training engineer. This was a new concept at the UdM, and it was being offered to lecturers from all areas which including the ICT, engineering and business management areas. Gaining the certification meant that such lecturers would be capable of working independently and developing online training for students.

The other option chosen was obtaining a diploma. The course was positioned as a Master's level programme initially which drove a substantive number of applicants. The fact that a diploma would be granted by a reputable French public university stimulated learners to opt for this programme.

Another viable point was the possibility of gaining credit for promotion as a result of following the course. Since lecturers have been recently promoted, this alternative scored lesser but remained a viable option.

Scores were quite low regarding the real intention behind being trained as a 'training engineer'. This came from an initial evaluation where the opportunity of gaining an alternative learning method was fairly weak. This could improve during the training process.

The expected use and application of e-learning obtained a lower score because there were no initial premises to develop e-learning. At the present time, e-learning had been developed just as a means of replacing face-to-face teaching that was impossible with the COVID-19 pandemic. Today, digital learning goes mainstream to upend conventional educational practices [14] (Fig. 1 and Table 1).

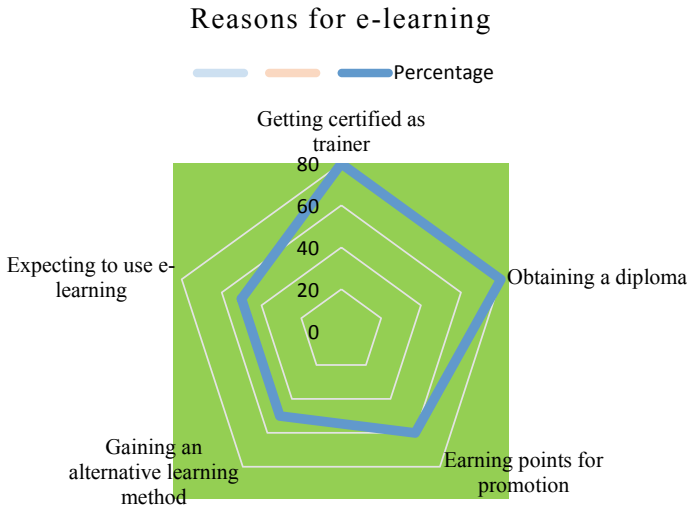


Fig. 1 Diagrammatic representation of the reasons for e-learning by UdM staff

Table 1 Choices for undertaking the training engineers’ course by UdM academics

Choice	Percentage (%)	Indexed value	Significance
Getting certified as trainer	80	80	Very high
Obtaining a diploma	80	80	Very high
Earning points for promotion	60	60	High
Gaining an alternative learning method	50	50	Moderate
Expecting to use e-learning	50	50	Moderate

1.5.2 The Relevance of Training Engineers in e-learning

The relevance of training engineers was questioned to the participants. From the discussions and posts made by them, certain key elements were identified. They are listed below. A Likert scale was used to evaluate the options: strongly disagree-1, disagree-2, agree-3 and strongly agree-4 (Table 2).

The means were as follows: ability to set up a training course: 3.58; ability to use learning platforms: 3.64; potential of developing hybrid programmes (face-to-face and distance learning): 3.76; ability to meet different types of learner needs: 3.53; capability of developing tailor-made programmes: 3.56; capacity of assessing performance and evaluating learning: 3.53; ability to develop independent learners: 2.88; capacity of creating inverted classrooms: 2.82; potential of developing creative and stimulated learning: 2.94. The mean was 3.36, variance was 0.13, and standard deviation was 0.36. The sampling mean most likely followed a normal distribution. In this case, the standard error of the mean (SEM) calculated using the following

Table 2 Relevance of training engineers in e-learning

Relevance of training engineers	SD	D	A	SA
Ability to set up a training course	-	-	7	10
Ability to use learning platforms	-	-	6	11
Potential of developing hybrid programmes (face-to-face and distance learning)	-	-	4	13
Ability to meet different types of learner needs	-	-	8	9
Capability of developing tailor-made programmes	-	-	7	10
Capacity of assessing performance and evaluating learning	-	-	8	9
Ability to develop independent learners	-	2	8	7
Capacity of creating inverted classrooms	-	3	6	8
Potential of developing creative and stimulated learning	-	1	8	8

equation was 0.12. A confidence level of 95% (or statistical significance of 5%) is typically used for data representation. The margin of error was $3.36 \pm 0.24 (\pm 7.15\%)$.

There was high interest regarding the ability to set up a training course, using online learning platforms, developing hybrid course for students as well as assessing performance and evaluating learners. Weaker scores came from developing independent learners, the capacity of setting up inverted classrooms where learning is managed by students including the development of creative and stimulated learning from the student. The problems might come from the present style of teaching that was top-down, directive and mainly one-sided teaching approach that could interest independent learners. It is only when learners see the connection of the learning initiative with the big picture objectives; initiatives are much more likely to gain commitment [15].

1.5.3 Approaches that Training Engineers Could Adopt in Developing Online Courses

The third key question addressed the approaches that potential training engineers would have in relation to the development of online classes at the UdM. A positioning map was created to find out how the approaches varied in terms of difficulty of implementing them and the degree of involvement of lecturers or training engineers involved. These diagrammatic representations were reflected by the opinions formed by the groups of lecturers under study. Key findings were based on the feedback that they provided during the forums and discussions done on and offline with the trainers. The figure below provides an estimate of the feasibility of approaches to be adopted by training engineers (Table 3).

The evaluation is then supported by a positioning map Fig. 2.

The most challenging approach is provocative learning which requires the trainer to make the student accept the new mode of learning which is not simply distance learning but that demands a high level of commitment from the student. It is a

Table 3 Approaches adopted by training engineers in e-learning

Approach	Colour	Code	Magnitude	Difficulty	Involvement
Accompaniment of the student	Green	A	High	Moderate	High
Provocative learning	Orange	P	High	High	High
Inverted classroom	Purple	I	Moderate	Moderate	Moderate
Creative classes	Blue	C	High	Weak	High
Identity representation	Red	R	Weak	Weak	Weak

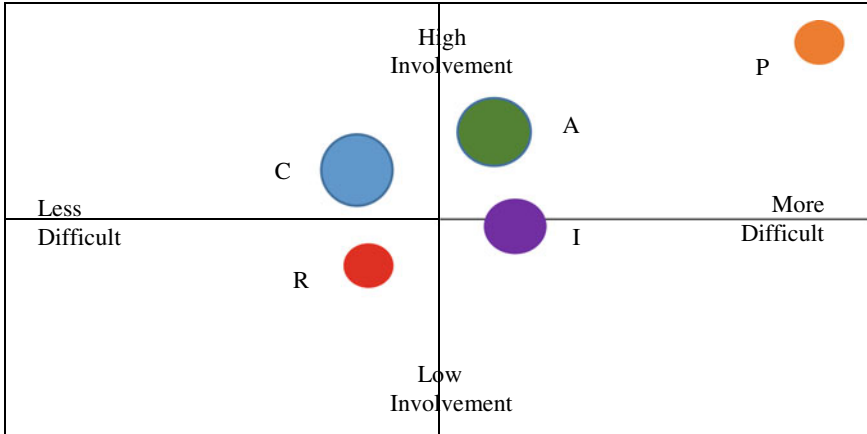


Fig. 2 Positioning map to represent learning approaches by UdM staff

compelling effort from the trainer to make the student accept learning and communicating differently. The weakest approach would be identity representation which requires the student to place himself in the learning context. There should be a certain level of involvement from his part and the difficulty remains low. Creative classes, on the other hand, are not too difficult to create because they require the training engineer to set a course that motivates, stimulates and improves learning. It was important to see the level of participation in exchanges within a community of learners, the richness and quality of interactions between learners are recognized as beneficial factors to the engagement, satisfaction and success of online learners [16].

The need for rapid results and quick turnaround of fully trained workers is driving corporate trainers to include innovative training strategies in their programs [17]. They, however, require a certain degree of creative involvement which is high from the trainer.

The inverted classroom looks challenging for the student where he/she will have to create and develop material that he/she will present to the class. Here, the level of involvement of the trainer is moderate, but the inverted classroom is important in terms of magnitude with reference to the learning process. In the context of an inverted classroom setting, it is common for students to experience a culture shock

in the way they learn as they are confronted with self-learning explicitly supported by digital media [18]. The accompaniment of the student is of high magnitude as it comes to both encouraging the student to participate in the learning process while being independent as a learner. At the same time, the level of involvement remains quite high with regards to the trainer. Changing the focus in the classroom from the faculty teaching to the students actively learning may prove to be challenging to the instructor used to actively teaching [19].

1.5.4 Sustaining e-learning at the UdM

The final question related to sustaining e-learning at the UdM. Basically, all respondents were in favour of initiating e-learning courses since they were gaining competencies as training engineers. Just implementing an e-learning programme was not substantive in that the long-term viability was sought. Key options were developed. To sustain e-learning, dedicated learning platforms, capacity, trainer motivation and evaluation and reward should be duly considered (Table 4).

The findings are represented in the chart below. The scales, following the responses, were rated as follows: very high 1, high 0.8, moderate 0.5 and weak 0.3. These were regression values (R range 0.1–1). The p value was ranked as less than 0.05 under a 95% confidence interval limit (Fig. 3).

From the data obtained and the scaled values presented, dedicated learning platforms were of highest consideration regarding the sustainability of e-learning at the UdM and that was the area where training engineers were being trained to acquire the skills and competences in developing courses, mounting them on platforms and making them practical for e-learning. Student capacity, as in any institution or course,

Table 4 Options to sustain e-learning by training engineers

Options	Funding	Importance	Sustainability
Dedicated learning platforms	High dependence [0.8]	Very high [1]	Very high [1]
Student capacity	High dependence [0.8]	High [0.8]	Very high [1]
Trainer motivation	Moderate dependence [0.5]	Very high [1]	Very high [1]
Course and training evaluation	Weak dependence [0.3]	High [0.8]	Moderate [0.5]
Reward	Weak dependence [0.3]	High [0.8]	Moderate [0.5]

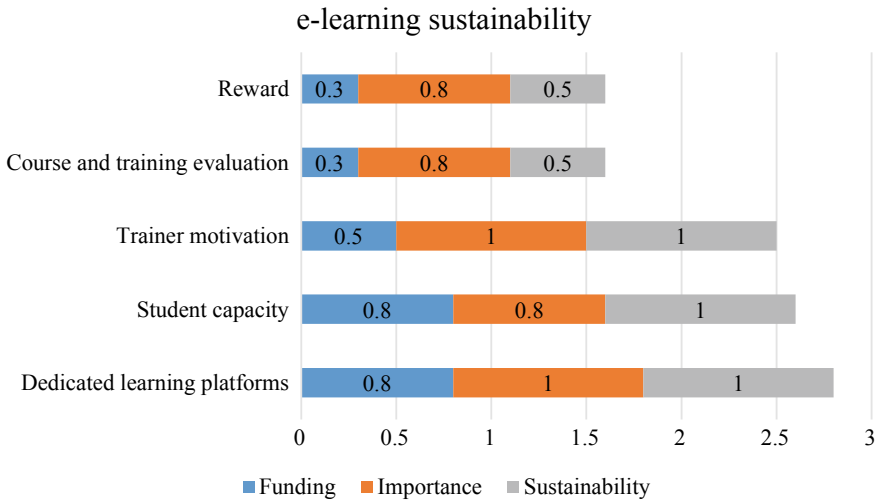


Fig. 3 Assessment of e-learning sustainability by UdM staff

remains of high importance regarding its viability. The same goes for trainer motivation who will need to be well supported and motivated to promote e-learning. Earlier findings have argued that the right level of commitment must be the subject of collective learning [20]. Course and training evaluation were of lesser importance reading the long-term viability of the e-learning project. Financial reward mattered less as lecturers are full-time staff and are already remunerated while there is no such research on the value of intrinsic and extrinsic rewards.

2 Conclusion

e-learning is positioned as a future mode of learning at the UdM. Although there have been already existing options of providing courses using computer technology and formats like Intranet, e-mail and new platforms like Zoom, Google class, etc., little attention has been so far focused on the development of e-learning. The positive consideration is that there has been a first cohort of lecturers being formed as training engineers at the university. They will obtain a diploma certifying their competence as such through a collaboration with Université Caen in France. The key questions concerned the need for enrolling in the e-learning training where the staff had strong views on its significance. The relevance of training engineers was assessed which stated that student accompaniment, provocative learning and the creative classes mattered. Concerning the sustainability of e-learning as an alternate form of e-learning, training engineers considered that dedicated and well-structured learning platforms, student capacity and trainer motivation were key success factors ensuring the success of the learning mode.

References

1. New Scientist (2019) Covid-19, The disease caused by a kind of coronavirus which first originated in Wuhan, China in late. <https://www.newscientist.com/term/covid-19/#ixzz6JgYIZ8IU>
2. Reimers, F., Schleicher, A.: A framework to guide an education response to the COVID-19 Pandemic of 2020, OECD (2020)
3. Betchoo, N.: Sub-saharan Africa's perspective of distance learning. *Int. Lett. Soc. Human. Sci.* 185–191 (2015)
4. Guillaumin, C.: Reflexivity as a skill: the challenge of new training engineering. *Cahiers de Sociolinguistique*, 85–101 (2009)
5. Gorz, C.: Reclaiming work: beyond the wage-based society. *Polity* (2001)
6. Training engineer, Definition of the term. <https://www.orientation.com/metiers/ingenieurs-de-la-formation.html>
7. Paul, M.: Support. Research and training [Online], 62 2009, released on 01,013. URL <https://journals.openedition.org/rechercheformation/435>
8. Ibid
9. De Ketele, J.: Accompanying students in higher education: an attempt at modelling, *Research and Training* [Online], 77 (2014), released on 31 December 2014 <https://journals.openedition.org/rechercheformation/2321>
10. Jacquinot, G.: Taming distance and overcoming absence. *French Pedagog. Rev.* 102, 55–67 (1993)
11. Charlier, B.: Actors: From Audience to Provider. *American Journal of Distance Education* 25(4), 226–237 (2011). <https://doi.org/10.1080/08923647.2011.622677>
12. Ibid
13. New Scientist: Covid-19, The disease caused by a kind of coronavirus which first originated in Wuhan, China in late 2019. <https://www.newscientist.com/term/covid-19/#ixzz6JgYIZ8IU>
14. Sharma, V.: E-Learning in the era of COVID-19 pandemic, *Daily Excelsior* (2020)
15. Mehta, S., Downs, H.: Strategies for digital learning success. *Centre for Creative Leadership* (2019)
16. Poellhuber, B., Roy, N., Bouchoucha, I., Anderson, T.: MOOC research initiative final report: the relationship between the motivational profiles, engagement profiles and persistence of MOOC participants (2014)
17. Andriotis, N.: Experiential learning: experiencing a growing trend in corporate learning, *efront* (2017)
18. Talbert, R.: Inverted classroom. *Colleagues* 9(1), 7
19. Hall, M.: Flipping your class, *The Inverted Instructor Blog* (2013)
20. Loriol, M., Sall, D.: Stress management at work. *La Revue des Conditions de Travail* 1(1), 56–63 (2014)

Using Scratch Software as a Teaching-Learning Tool in French Language Classes: A Case Study at Université Des Mascareignes (Mauritius)



Neelam Pirbhai-Jetha

Abstract In the teaching-learning of languages, digital tools are said to allow learners to be more motivated, and to produce a more elaborate and better-quality work. French language, compulsory in the Mauritian school curriculum, from primary school to secondary school (Cambridge O-Level), is a subject that worries learners who have difficulties in expressing themselves in writing. In this study, Scratch, a free software developed by the MIT, was used among students enrolled in the first year at Université des Mascareignes (Mauritius) to revise the basics of grammar in a fun way. Learners' comments at the end of the sessions were analysed to identify not only the positive aspects of this learning method, but also to review its shortcomings and the problems encountered. In other words, we pose the problem from the point of view of the learners, in order to know if it is possible to strengthen one's knowledge and skills in French language by moving away from the traditional method of teaching and by using Scratch software in groups.

Keywords Scratch software · Mauritius · French language · Teaching-learning · Collaborative work

1 Introduction

According to several researchers [1, 2] the use of computer tools in learning a language allows the learners to be more motivated: they thus produce a more elaborate and better-quality work. For Glen Bledsoe, what learners fear most is writing a text: finding a start, developing the ideas and coming with a logical conclusion is often very difficult for them [3]. Digital comics, therefore, can be used to arouse the interest of learners [4], as images, unlike a blank sheet of paper, is a crucial medium for developing and communicating ideas.

French language, compulsory in the Mauritian public/government school curriculum, from primary school to secondary school (Cambridge O-Level), is

N. Pirbhai-Jetha (✉)

Centre for Digital Humanities, Université des Mascareignes, Rose Hill, Mauritius

e-mail: npirbhajjetha@udm.ac.mu

a subject that worries learners who have difficulties in expressing themselves in writing. In this activity, the objective of using Scratch, a free software developed by the Massachusetts Institute of Technology (MIT), was to improve the teaching-learning of French language among students enrolled in the first year at Université des Mascareignes (Mauritius) and thus revise the basics of grammar in a fun way. In order to identify and work on the ‘errors’ that students usually leave, the first step was to form groups of three or four students. Working in a team can, according to some researchers, reduce this intimidating learning experience [5]. The group then had to find and transcribe in French an original tale, of oral tradition from their locality; thus, allowing a preservation of the cultural heritage. Following this, an animation was created using Scratch software: learners could thus put into images their text, and elaborate their story.

This study analyses the learners’ comments at the end of the sessions to identify not only the positive aspects of this learning method, but also to review its shortcomings and the problems encountered. In other words, we pose the problem from the point of view of the learners, in order to know if by moving away from the traditional method of teaching and by using Scratch software, it is possible to strengthen one’s knowledge and skills in French language.

2 Brief Literature Review

2.1 *Teaching-Learning with “Digital Natives”* [6]

The National Education Technology Plan by the U.S. Department of Education posits that “the conversation has shifted from whether technology should be used in learning to how it can improve learning to ensure that all students have access to high- quality educational experiences (2016, p. 5)” [7]. With the increasing accessibility to technology, educators have to face the “challenge of adapting their teaching styles to accommodate a new generation of digital learners” [8], known as Generation Z. Used to an infinite amount of information, mainly given in media/image form, digital natives’ learning expectations, styles, and needs are completely different from our past students. Traditional classroom teaching is obsolete with them, and there is no doubt that to sustain student engagement, educators and the education system must “adapt their teaching strategies to accommodate the digital learners, in light of their preferences for digital literacy, experiential learning interactivity, and immediacy” [8].

Often referred to as “mobile nomads”, the modern learners are always “connected” [8] via instant messaging (cited in [8]). Literature in the research of MOOCs also reveals that social media such as Facebook is important “to enhance the sense of community” and “instructors use social media as a second channel to communicate with MOOC students”. Social tools, therefore, make it easier to “engage students and improve their retention during the course” [9].

Furthermore, for digital natives, technology creates emotional connections between people [10]. Since their every-day activities are connected, learning becomes “a social activity [11]”, which cannot be “fulfilled in isolation” [8]. Indeed, most research done on digital learners reveal that they “gravitate toward group work [12]” [8] and that “they construct their knowledge” [8] when they work with others. The “TTT (talk, text, test) approach”, as a traditional teaching style, is not valued by the digital learners as the latter prefer to work in teams and participate in peer interactions [8].

This positive attitude towards learning must be developed, as collaboration is considered an important “contributor to students’ higher order of thinking” [13], and it encourages commitment, engagement and problem-solving skills.

2.2 Developing Twenty-First Century Skills in French Language Classes

The social constructivist approach to teaching and learning classroom activities also gives a great importance to active participation of learners through activities, such as discussions, debates, and collaborative work. In fact, collaboration and creativity are some of the skills needed in the workplace. An OECD report mentions that “skills are extremely important in the modern economy and in everyday society” [14]. Debates on the twenty-first century skills are numerous and ongoing, but all agree that the education system must address the challenges that workplace present. Twenty-first century skills are “not new, just newly important” [15]. Different frameworks have been conceptualised, since “more and more employment opportunities are requiring people who can think adroitly—and often think on their feet”. With “knowledge itself growing ever more specialized and expanding exponentially” and data and information being constantly updated in this high-tech society, there’s a need for people with different skills and competences [16]. But to sum up, we can refer to Kereluik’s framework (2013, p. 7), which categorise skills as “Foundational Knowledge” (high academic standards, core curriculum, disciplined mind, quantitative literacy, mathematical and scientific competence...), “Meta Knowledge” (inventive thinking, creativity, innovation, critical thinking, ability to solve complex issues/problem solving, decision making...), and “Humanistic Knowledge” (ethical reasoning, empathy, emotional intelligence, cooperation and connection with others) (cited in [17]). The last two sets of knowledge are, however, multidimensional and complex nature of constructs as they draw on different skills [18].

The labour force of the twenty-first century must possess more analytical thinking, digital skills, and sophisticated communication skills [12]. And it is to be noted that success is not limited to “educational achievement and attainment” and job security; but also encompasses everyday life and “other areas of adult responsibility” [19]. However, the education systems have not evolved in parallel, and do not provide the

“infrastructure, pedagogical methods, or actual curricular material that will maximally prepare students for the current and future world in which they will enter and lead in their future” [15]. For years, many researchers such as Prensky have criticised the amount of money which is “being spent on trying to fix the educational ‘system.’ According to him, “it’s not the ‘system’ that we need to get right; *It’s the education that the system provides*” [20].

In order to cater for the market demands, it is, therefore, crucial to have a workforce with the twenty-first century skills, and this will be only possible if there is a shift from the traditional system to a more modern classroom, where the learners are encouraged to develop the major learning and innovative skills such as critical thinking, communication, creativity, and collaboration [13]. The use of technology in language classes can be a crucial tool in developing some of these skills.

3 Research Methodology

The French language course at Université des Mascareignes was taught over seven weeks, and contact hours were about 3.5 h per week. During the first three weeks, the teaching-learning method was mainly traditional, and consisted of revising the basics of grammar through group presentations done by students, reading of short texts, dictation, some exercises on grammar and planification of the traditional tale. Only about 14 h were left to use Scratch, and present the final work.

It was also important to know the level of French language of students before the start of the work. There were different cohorts of year one students, who were of different nationalities (but mainly Malagasy and Mauritians). Aged between 17 and 25 years old, they were enrolled in different fields of studies such as Marketing, Human Resource Management and Banking and Finance. A questionnaire was distributed to the students to learn about their linguistic background. 65 students took French language for their Cambridge Higher School Certificate; meaning that the others (about 35%) dropped French after their Cambridge School Certificate. 21 students did an additional degree in French language after their secondary school studies: one student did DALF (Diplôme approfondi de langue française) and 20 were enrolled in the DELF (Diplôme d’Etudes en langue française). Furthermore, Table 1 shows languages spoken by the students at home.

We were also interested to know whether the students had a great exposure to French language during their secondary school studies, before joining university, and we questioned them about their exposure to French language (Table 2).

The first step during week three was to set up teams of three-four students maximum. A small group is beneficial in the sense that it allows a maximum of interactions and active listening among the members. It is also easier to develop cooperative learning with a smaller group. Groups were also built according to affinity. In other terms, students could work with anyone they wanted, as the objective was to create a climate of complicity and trust among the participants. The only instructions

Table 1 Mother tongue of students

Mother tongue	Students (%)
French language	7
Mauritian Creole	43
Malagasy	21
Malagasy and French language	11
Mauritian Creole and French language	8
English (and/or Mauritian Creole; and/or hindi; and/or French language)	3
Other languages (Swahili, bamileke of Cameroun, bhojpuri...)	7

Table 2 Exposure to French language

	Low frequency	Neutral	High frequency
Reading of magazines/books/newspapers in French	12	40	43
Watching movies/documentaries in French	5	26	65

given were to create a group, of diverse nationalities, aptitudes and talents. Furthermore, a group built by field of interest develops student motivation and engagement to work on the theme given.

Step two was an introduction to Scratch. About one hour was spent to introduce the interface, that is the command panel, the programming panel and the visualization panel; and everyone worked together with the facilitator on their respective cover page. A tutorial guideline was also prepared by the facilitator and sent to all groups by email.

4 Findings and Analysis

4.1 Student Motivation and Engagement

“When teachers teach in ways that students learn in today’s digital age, students are much more engaged in the lesson content and are more interested in the information” [21].

A few positive opinions on the use of Scratch as an animation tool were noted in the French language class. Students found the course more interesting, as it helped them retain some of the grammatical rules better and it also developed their writing skills in French. One student mentioned that it was more amusing to create a story than to read a book in French. Another student was motivated as the course was different from the usual traditional way of teaching-learning. For one student, using Scratch was a good initiative as it motivated him/her to learn French. One student was also keen to do the class as it was something new. Another student found the

Table 3 Level of motivation of students

Level of motivation	Number of students
0	9
1	12
2	15
3	26
4	30
5	4
Too Complex	1

use of the software difficult at first, but with the resources made available, it became easier. The introduction of ICT tools in the teaching of a language seems to allow learners to reveal their intrinsic motivation.

However, for many students, Scratch seems to have been the wrong ICT tool that was used in the French language class. Respondents' diverse comments were on the software being too complicated and time consuming. According to one student, the idea of learning French with digital comics was interesting, but there was a need of a simpler tool. For one student, Scratch is a tool which is reserved for students in IT, and not for other fields. However, this is completely untrue as Scratch is used in primary schools in many countries in order to develop digital literacy and analytical skills in children at a young age [22]. This shows the difficulties some students have in leaving their comfort zone in the teaching-learning paradigm, and in putting in more effort. For other students, the tool made them spend more time on creating the animation, and they did not spend enough time in learning the language. One student also thought that a tool completely in French would have been better. Table 3 shows the level of motivation of students in using Scratch as a tool to learn French (Zero (0) being not all motivated and five (5) being very motivated). One student added a new row (last row) in the questionnaire and commented that s/he found the use of Scratch "too complex". It can be deduced that more than 37% of the cohort, without counting those who selected "neutral—3", was not motivated with the use of the software.

In her book on "Mauritian Education in a Global Economy" (1994), S. Bunwaree criticised the traditional pedagogical approach that has permeated the educational system in Mauritius. Her main criticisms were around the teacher whose main job was to dictate lessons which the learners had to learn by heart. Learning by rote seems to be at the heart of this pedagogy as teachers tend to think that it's the only way learners can learn and pass their exams [23]. This is developed by Boulton who posits that the teacher is often seen as "the ultimate expert" and his/her role is to "dispense that knowledge, while the learner is a passive unit to be taught, to memorise and reproduce knowledge" as "at the end of the day, the only question is, "how can I pass?" [24].

This traditional method of teaching-learning seems to still remain important to learners even in the higher education. When asked about their opinions on using a software in the French class, there were many who replied that they preferred

learning French in the traditional way. Another learner said that even if using ICT tool was interesting, s/he preferred the use of “archaic methods” such as dictation to learn French. The question we ask ourselves is how do we integrate ICT in an environment which is conditioned by the ‘spoon-feeding’ system.

Another question which is raised is whether the Mauritian or African workforce is ready to face the demands of the industry in the twenty-first century. Indeed, very few students found that using Scratch in the French class developed creativity and innovation. Two examples of negative comments which were mainly given were that “using Scratch can help develop our creativity but I do not think that we will use this in the workplace” and “the tool develops creativity but not the learning of French language”. However, a few Respondents seem to have understood the importance of developing creativity during teaching-learning, and in the workplace. This is found in a few responses of some Respondents. Despite the fact that one Respondent found using Scratch difficult, s/he also found it a good tool to express oneself and develop one’s creativity. Another Respondent revealed his/her passion for Scratch as s/he has decided to use it to create his/her own stories”. For some, using Scratch was “pleasant and innovative”. To sum up, we can cite researchers who posit that “Scratch helps young people learn to think creatively, to reason systematically, and to work collaboratively – essential life skills in the twenty-first century” [25].

4.2 Teamwork and Cooperation

Working in groups tends to reduce the intimidating experience of learning [5] and allows the learners to manage their “linguistic anxiety” [26]. According to a comparative study of David Johnson, Roger Johnson and Karl Smith on cooperative learning and competitive or individual learning, it has been illustrated that cooperative learning is more positive as it develops the learners’ competences, brings in student engagement and helps students to succeed in their academic studies and allows positive interdependence (cited in [27]). Concerning the use of Scratch in French classes, there were two main positive comments from students. For some students, working in groups were more interesting and more interactive. Some even added that when working in groups they worked harder and were more motivated. The Mauritian system, however, does not seem to have developed cooperative learning in most students. Not only learning is done by heart and by memorisation of content only, but everything is given to the student – the explanation and the interpretation, thus leaving very little freedom in autonomous learning. According to a survey done by Saurty in 2017, there is a correlation between the spoon-feeding pedagogy which prevails in the Mauritian education system and the fact that learners prefer to work individually (cited in [23]). In this study, the main complaints in group work was that “Members of the group did not participate at all” or the “group members were useless” and therefore learners “did not like working in groups”. Below are some of the comments of the students on group work (Zero (0) being “do not agree at all” and five (5) being “strongly agree”) (Table 4).

Table 4 Comments of Students on Group Work

Comments on group work	0	1	2	3	4	5
Different rhythm of work of members of the group	12	11	18	34	16	4
Communication problem and impossibility of dealing with differences	10	9	16	25	26	9
Group Members showed selfishness and individualism	18	8	12	23	20	12
I had feelings of solitude in the group	11	6	18	19	21	19
I felt ill at ease in the group	14	12	18	19	20	7
There is a suppression of one's personality in group work	16	7	10	23	26	13
I prefer to work alone	9	10	17	27	17	12
Members of the group were unavailable/lazy/lacked maturity	9	13	21	13	21	17
Members of the group were thick-headed/obtuse	9	11	18	29	14	2
Members of the group had difficulties to understand the instructions etc.	6	11	15	18	22	9

5 Conclusions and Recommendations

According to Alvin Toffler, “the illiterate of the twenty-first Century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn (1970)” (cited in [28]). Our main aim in this paper was to learn French language with an ICT tool by transcribing a tale from the oral tradition. In fact, many researchers have shown the implication of how and why technology enhances our instructional design and help learners achieve higher proficiency [7]. However, the spoon-feeding teaching style of the Mauritian education system has made it difficult for many to work in groups. Since many Respondents have found difficulties in using Scratch software in such a short period of time, one main recommendation is to use other ICT tools such as blogs or digital comics softwares such as Storyboard, Storybird or Pixton to develop the writing skills of students in French. Comparative studies between Scratch and other ICT tools can also help in coming up with new ways to make learning French more interesting and thus develop some twenty-first century skills such as collaboration and creativity.

Acknowledgements A special thanks to MRIC and AUF for funding the research work of the Centre for Digital Humanities; Uzma Peeroo and Angheene Busgeeth-Dhotah for their support in data collection; and all Respondents for answering the questionnaires.

References

1. Norlida, A., Supyan, H.: Learning in a writing course at Tenaga National Universiti. TEFL Web J. (2002)

2. Rafiza, A., Adelina, A.: The use of dialogue journal through e-mail technology in developing writing interest and skills. *Malays. Online J. Instruct. Tech.* **1**(2) (2004)
3. Bledsoe, G.: *Comic life*. Comput. Softw. Comiclife.com (2010)
4. Faulkner, G.: Digital comics spur students' interest in writing. s.l.: s.n (2009)
5. Lavy, S.: Who benefits from group work in higher education? *An Attach. Theory Persp.* **3**(2), 175–187 (2017)
6. Prensky, M.: Digital natives, digital immigrants. *On the Horizon* **9**(5) (2001)
7. Navarre, A.: *Technology-enhanced teaching and learning of Chinese as a foreign language*. Routledge, London and New York (2019)
8. Ramlee, M., Kashefian-Naeeni, S.: Moving teaching and learning into the digital era. *Int. J. English Lang. Transl. Stud.* **5**(3), 27–36 (2017)
9. Koutsakos, P., Syritzidou, E., Karamatsouki, A., Karagiannidis, C.: Exploring the role of facebook as collaboration platform in a K-12 MOOC, pp. 31–48. Springer, Switzerland (2018)
10. Elliot, A.: *The culture of AI—everyday life and the digital revolution*. Routledge, London and New York (2019)
11. Tapscott, D.: *Growing up Digital: the rise of the net generation*. New York: McGraw-Hill (1998)
12. Howe, N., Strauss, W.: *Millennials Rising: the next great generation*. New York: Vintage Books (2000)
13. Linuma, M.: *Learning and teaching with technology in the knowledge society new literacy, collaboration and digital content*. Springer, Singapore (2016)
14. Bourn, D.: *Understanding global skills for twenty first century professions*. Palgrave Macmillan, Switzerland (2018)
15. Chu, S.K.W., et al.: *Twenty first century skills development through inquiry-based learning—from theory to practice*. Springer, Singapore (2017)
16. Conklin, W.: *Higher-order thinking skills-to develop twenty first century learners*. Shell Education, California (2012)
17. Kennedy, I.G., Latham, G., Jacinto, H.: *Education skills for twenty first century teachers: voices from a global online educators' forum*. Springer, New York (2016)
18. Care, E., Griffin, P., Wilson, M.: *Assessment and teaching of twenty first century skills: research and applications*. Springer, New York (2018)
19. Pellegrino, J.W., Hilton, M.L.: *Education for life and work-developing transferable knowledge and skills in the twenty first century*. The National Academies Press, Washington (2012)
20. Prensky, M.: *From digital natives to digital wisdom: hopeful essays for the twenty st century learning*. Corwin-Sage Ltd, London (2012)
21. Lutz, C.L.: *A study of the effect of interactive whiteboards on student achievement and teacher instructional methods*. University of North Carolina, s.l. (2010)
22. Costa, S., Gomes, A., Pessoa, T. (2016) Using scratch to teach and learn english as a foreign language in elementary school. *Int. J. Educ. Learn. Syst.*
23. Saurty, K.: *Analyse de l'intégration des technologies de l'information et de la communication (TIC) dans un système éducatif conditionné par la pédagogie du 'spoon-feeding': Le cas*. Université de Limoges, France (2017)
24. Boulton, A., Azzam-Hannachi, R., Pereiro, M., Chateau, A. (2008) Learning to learn languages with ICT—but how? *CALL-EJ On Line* **9**(2)
25. Plaza, P., et al.: *Scratch as educational tool to introduce robotics*. Springer, Switzerland (2018)
26. Bernard, I.: *Pratique théâtrale et insécurité linguistique—un exemple d'enseignement du FLE en Jordanie*. *Synerg. Algérie* **10**, 225–231 (2010)
27. Brame, C., Biel, R. (2015) Setting up and facilitating group work: using cooperative learning groups effectively
28. Marshall, S.: *Shaping the university of the future: using technology to catalyse change in university learning and teaching*. Springer, Singapore (2018)
28. Levy, F., Murnane, R.J.: *The New Division of Labor: how computers are creating the next job market*. New York: Russell Sage Foundation (2004)

Improving Fraud Detection Mechanism in Financial Banking Sectors Using Data Mining Techniques



Hanan Hamdan AL-Abri, Basant Kumar, and Joseph Mani

Abstract The banking sector is facing increasing risks of fraud and malpractices. Thus, adopting new methods and approaches to detect, prevent, and predict frauds is essential. Data Mining (DM) techniques are the latest methods which are primarily responsible for ensuring data integrity, when are used in the way of machine learning approach. This paper essentially discusses how can the banking sector takes the advantage of one of these methods is a logistic regression. It is an effective classification algorithm that gives high accuracy for detecting binary problems and through the outcomes it can predict irregular transactions. The paper offers this algorithm effectively by presenting suggested developed model, as a part of a framework for the development of a fraud detection mechanism. Finally, the paper will test the suggested model by using a LR algorithm for proving its performance.

Keywords Data mining · Machine learning · Logistic regression · Fraud detection · Cross-industry process for data mining (CRISP-DM) methodology · Agile approach · Confusion matrix

1 Introduction

Fraud is an extensive practice in the financial sector that entails misapplication of accounting policies, falsified records of transactions, omissions, misappropriation of assets, and alteration of documents. It leads to a negative impact on financial institutions hamper performance, destroy customer relations and experience, particularly

H. H. AL-Abri (✉) · B. Kumar · J. Mani
Information Technology Department, Modern College of Business & Science, Muscat, Sultanate of Oman
e-mail: h.h.alabri@moe.om

B. Kumar
e-mail: basant@mcbs.edu.om

J. Mani
e-mail: drjosephmani@mcbs.edu.om

financial losses [1]. It reduces the competitiveness of a business, or limits the level of professionalism within the management hierarchy [2] and reduces organizational assets and increases liabilities, so resulting in the loss of stakeholders' trust and confidence [3]. The fraud activities will increase, especially with the great technological development that the world has witnessed lately, so the growth of online businesses is making them more vulnerable to manipulation and making fraud management a challenging task [4]. Particularly in international transactions, will continue to enhance the vulnerabilities of businesses and customers as well as is raising the average expenditure on fraud control with individual merchants aiming to retain customer trust and confidence [4]. The enhanced popularity of internet banking is promoting a lot of malware campaigns and cyber-attacks. Carminati et al. indicate that financial malware increased by 40% in 2016, where they successfully infected 2.8 million personal devices [5]. Therefore, the enhanced rate of frauds is encouraging financial institutions to employ state-of-the-art solutions that detect and address them before they aggravate the problems that lower the integrity of transactions. This study includes reviewing to the most recent methodologies and tools that are used in business and fraud detection field. As a result of reviewing, the study suggests a developmental framework of a mechanism of fraud detection and using the python language for the implementation of the developed model. Finally, the paper shows the results of the run model and measures the performance of the LR method using confusion matrix functions.

2 Literature Review

DM and machine learning are the most common ways to combat fraudulent transactions. DM is the process of extracting the required information or knowledge from big data [6]. It is a computerized analysis process is used to explore a large amount of data and search for consistent patterns and systematic relationships between variables that are not known in advance, and then to validate the findings by applying the detected patterns to form new subsets of data [7]. While, the machine learning classification is the learning process of assigning instances of predefined classes that they belong. It is used to predict class labels [8]. Thus, when the DM techniques follow machine learning approach that will improve quality of experience and help reduce unmanageable data volumes down to relatively few informed indices.

Meanwhile, banks need to apply these techniques and their methodologies according to the business objectives. Therefore, this study has reviewed the methodologies that relied on the suggested mechanism of detecting fraud and the integration between them and data mining techniques.

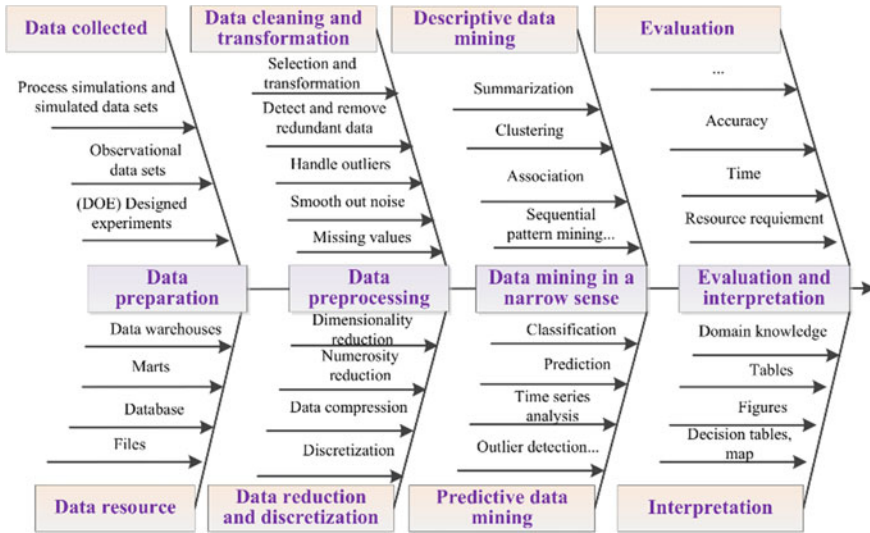


Fig. 1 The flowchart of datamining process [7]

2.1 Data Mining Methodology

Most of Data mining is carried out in four phases, to obtain the required information. In the beginning, companies collect data from different resources and upload these data on their data repositories. After that, the data get sorted out and filtered before starting the analysis to improve quality of the repository. Then, they work on the storage data and start analyzing them using business analysis applications to obtain results and then display them visually easy to share. Finally, these outputs can use by management to take advantage of them in decision-making [7] (Fig. 1).

2.2 CRISP-DM Methodology

Fraud detection is a multimillion investment that attracts numerous techniques and algorithms. Banks are one of the leading enterprises in the venture to protect access to internet banking services, avoid problems of electronic fraud, and guarantee the security of computer that customers use to complete monetary transactions. Meanwhile, the Cross-Industry Process for Data Mining (CRISP-DM) is a robust and well-proven methodology with practicality, with the flexibility that makes it the complete DM tool for meeting the needs of industrial projects [9]. The model identifies six stages of implementing a data mining project that includes business understanding for promoting awareness of a firm’s rules and business objectives and understanding the process of collecting data. More so, data preparation organizes data into a suitable

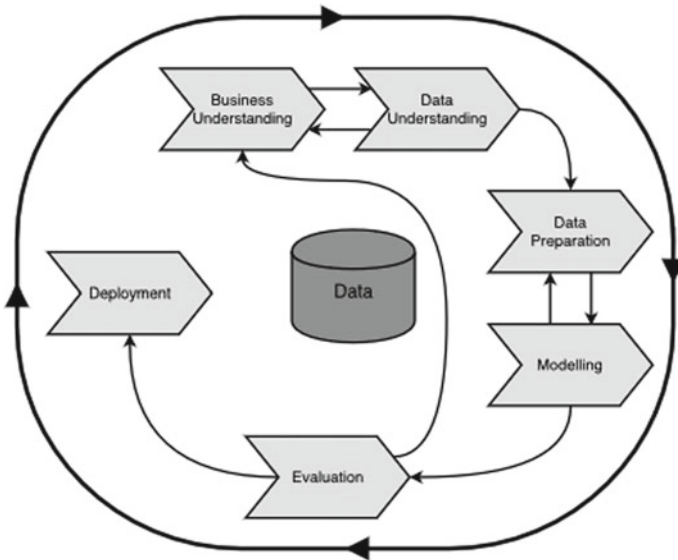


Fig. 2 The relationship between different phases of a CRISP-DM process [9]

format before importation into the software; modeling selects the most suitable data modeling technique; evaluation assesses the process to determine whether it solves the problem, and deployment installs the system and trains the users (Fig. 2).

2.3 Suitable Data Mining Methods of Financial Fraud Detection

There are several smart algorithms can help to detect the frauds in banking sector applications especially for credit card frauds. The most popular methods are decision tree, logistic regression (LR), support vector machine (SVM), artificial neural network (ANN), and random forest.

LR is a multivariate method employed to model the relationship between multiple independent variables and a dependent variable [10]. The LR estimates the probability of occurrence of an attribute by fitting all the data on the logistic curve. The model predicts the probability of an outcome with two values, such as zero or one [11, 12]. Meanwhile, LR is a binary logistic regression when the predictor value is dichotomous and the independent variable is incessant [13]. LR expresses attributes in odds where the odds of the initial attribute are given by $p/(1-p)$ where p is the probability of occurrence and $1-p$ is the probability of not occurring. However, LR should compute the probability of occurrence, meaning that variable x is introduced to relate p and x using the equation $p = \alpha + \beta x$. The value $\alpha + \beta x$ does give a suitable

model because the result does not lie within the acceptable range between 0 and 1. The application of natural logarithm transforms the LR solution to the required odd given by:

$$\text{logit}(y) = \ln(\text{odds}) = \ln\left(\frac{P}{1-p}\right) = a + \beta x \quad (1)$$

where α and β are parameters of logistic regression, x is the explanatory variable, and p is the probability of the interested outcome. The interested outcome can be obtained by the computation of antilog of the above equation.

$$p = P\left(Y = \text{interested} \mid \frac{\text{outcome}}{X} = x, a \text{ specific value}\right) = \frac{e^{a+\beta x}}{1 + e^{a+\beta x}} = \frac{1}{1 + e^{-(a+\beta x)}} \quad (2)$$

Thus, the odds of a given attribute are computed using the antilog equation with α and β as the parameters of LR and x as the explanatory variable. The detection of fraud and other intrusions into telecommunications systems are dynamic concepts that change with technological advancement. The LR dominates the fraud detection market compared to the decision tree and neural network techniques due to its ability to work with linear data for credit card fraud detection in addition to producing a simple probability formula for classification [14]. Khare and Sait's experiment involved the assessment of accuracy in LR, (SVM) model, decision tree, and random forest. The results illustrated an accuracy of 97.7% for LR, 97.5% for SVM, 95.5% of the decision tree, and 98.6% of the random forest. The researchers note that the logistic regression is the most accurate and practical model because the random forest was only effective with a large number of training data, leading to reduced speed in testing and application [12]. Therefore, LR is the most suitable model that utilizes minimal resources, with reliable outcomes and high an accuracy score.

3 Research Methodology

This study provides a proposed methodology of fraud detection in an application of banking as a use case of credit cards. The developed model based on the merger of an agile approach with ideas of the CRISP-DM methodology. It applies the LR algorithm to extract invalid cards intelligently. Where, the agile software development approach which is highly suitable for fraud detection, which is dynamic and requires continuous adjustments to meet the customer's demand [15]. The use of static models will render the final product obsolete on completion due to the rapid evolution of fraudulent activities in the banking sector. Thus, agility is essential in the development of a fraud detection system without adversely impacting cost and production schedules. The technique supports evaluation of different algorithms for the employment of the most suitable one in fraud detection and control.

The methodology that combines the agile approach with CRISP-DM methodology takes into consideration the importance of the analysis of business problems to understand them and identifying the goals well before collecting and processing the required data and building the appropriate model. Also, it can return back in these objectives to modify them based on the new results and repeat or develop the rest of the operations to reach the target goals. Therefore, this methodology will be able to build the appropriate models and use techniques that give high accuracy according to the goals set and the required data. Furthermore, it will raise the speed of deployment and reduce the time and costs, but working on such these methodologies requires working in a team rather than individuals to be more successful. Thus, the suggested mechanism of detecting fraud in the study is built by this methodology illustrated in Fig. 3.

The framework consists of successive iterations and continuous development, so that each iteration of development involves the stages of CRISP-DM starts from

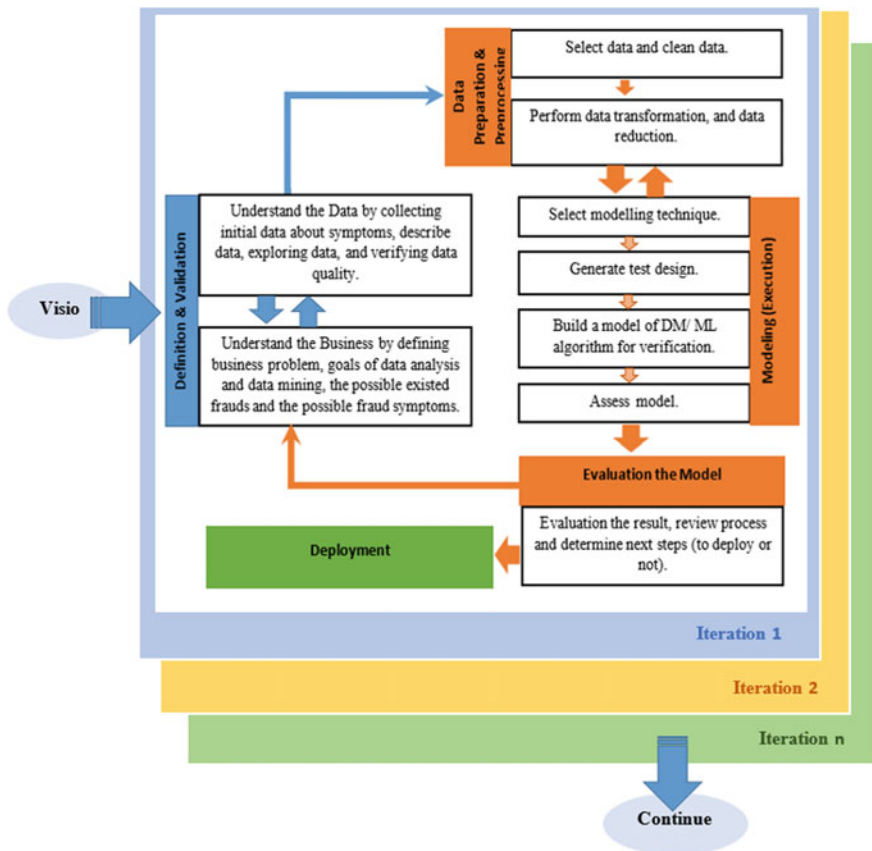


Fig. 3 The theoretical mechanism of fraud detection development in this study

the phase of determining the objectives to be achieved from building this model. Then, the required data is collected from the various data storage sources and then processed in such a way as to ensure that they are free from errors. After that, comes the modeling phase, in which the agile approach is developed in which the model is developed and an optimal algorithm is applied in the detection of anomalies transactions and their integrity from fraud through by dividing the development into successive and incremental iterations, each takes a short time to build, design and test the model to arrive at the most appropriate and efficient algorithm that gives better results in the detection of anomalous credit cards. Then applies the stage of evaluating or measuring performance and interpreting the results before they are deployed (Fig. 4).

The above model took the principle of design and work from the framework in Fig. 3. It focuses on the practical side through the application of the stages: collecting the data required, reading them from the data set and applying the processing processes. Then, modeling phase to apply the LR algorithm in order to classify transactions based on a binary matrix and evaluation of measures such as accuracy, sensitivity and F1-Score (Table 1).

4 Results and Evaluation

In this experiment of the model before the implementation of the algorithm LR. The data was preprocessed for aiming to improve the performance of the classifier and reduce its training and operating time. The preprocessing involves verification of the dataset feature space and handle the imbalanced nature of the data in the dataset. This resulted in a significant improvement in the results of the performance metrics of the LR algorithm used. The accuracy of it reached in the final result of the test set to (1.0) score. The run model in the Python compiler was invoked by the source code in LR. It took 70% of the database as training data and 30% testing data. The function trained the LR algorithm and used the test data to run prediction for the generation of the confusion matrix with two classes: Non-fraud and Fraud. As a result of all, the LR algorithm in the analysis was highly reliable in the detection of fraudulent transaction due to low misclassification of 1.0 without cases of mislabeling valid transactions (Fig. 5).

5 Conclusion

The assessment of strategies to improve fraud detection mechanism using data mining techniques provides critical insight into learning algorithms. The study illustrates that there is a correlation between DM techniques which are form the core of ML algorithms and smart search algorithms. This increases the efficiency and reliability in fraud detection and interprets available information to facilitate its exploitation

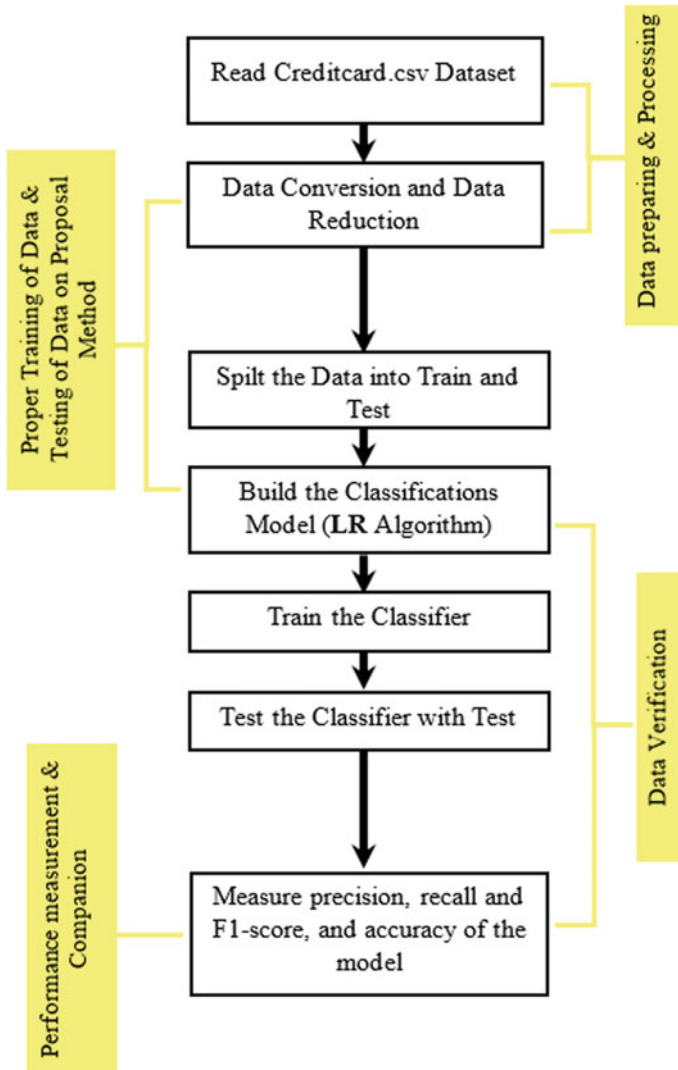


Fig. 4 The developed model methodology [8]

by market players in decision-making and assessment of fraud in the banking sector. Therefore, the study provides a new perspective on the evaluation of fraud. The use of DM in detecting fraud and predicting vulnerability presents a new approach for harnessing existing databases, the scholarly world can assess the role of different databases held by a financial enterprise. In this context, it is highly likely that the analysis of each dataset can lead to the reduction of certain type of fraud. Therefore, the study provides insight into an in-depth analysis of the problem for enhancing fraud detection and introduces a clear mechanism that can be used in the actual application

Table 1 Work requirements of implementing developed model methodology of this study

Requirement	Item	Description
Name of data set	Credit card.csv	The data set consists of 4499 rows of credit card transactions [16]
Tool of building the modeling	Python	It is one of programming languages classified within the high-level languages and object-oriented. It is characterized as being easy to learn and use [17]
Method of extracting fraudulent transactions	Logistic regression (LR)	LR is the standard method of DM and ML for a binary target variable with multiple features [14]
Method of measuring performance	Confusion matrix	It is a table that is used for describing the performance evaluation of a classification model or “classifier” on a set. The set used for evaluation is called a test dataset [18]

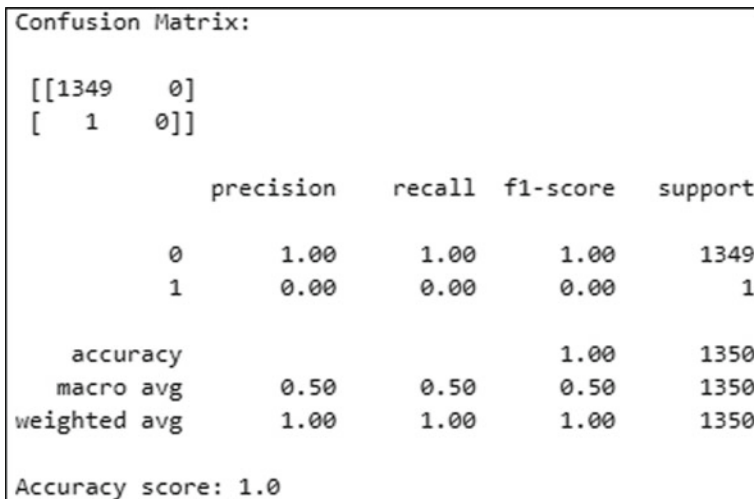


Fig. 5 The output of LR algorithm

when resorting to the use of DM techniques in finding quick solutions that produce results with high quality and accuracy ensure the stability of banking security. In future work, it is essential to assess social settings and cultural environments that motivate individuals to engage in fraud. The identification of motivation will be

critical in enhancing the effectiveness of data mining and give a clear vision in initial stages of fraud detection mechanism.

References

1. Bartsiotas, G.A.: and Gopinathan Achamkulangare. Detection, and Response in United Nations System Organizations. United Nations, Fraud Prevention (2016)
2. ACFE: Impact of Fraud: the typical organization loses 5% of its revenues to fraud each year. ACFE, www.acfe.com/uploadedFiles/ACFE_Website/Content/documents/cfe-employer-brochure.pdf. Accessed 26 May 2019
3. Gitau, Wamboi, E., Samson, N.G.: Effect of financial fraud on the performance of commercial banks: a case study of tier 1 banks in Nakuru Town, Kenya. *Int. J. Econ. Commer. Manag.* **4**(12), 142–157
4. Javelin (2017) Financial impact of fraud study: exploring the impact of fraud in a digital world. Javelin Strategy & Research
5. Michele, C. et al.: Security evaluation of a banking fraud analysis system. *ACM Trans. Privacy Secur.* **1**(1), 1–30 (2018)
6. e Sousa, L.R., Miranda, T., e Sousa, R.L., Tinoco, J.: The use of data mining techniques in rockburst risk assessment. *Engineering* **3**(4), 552–558 (2017)
7. Lv, S., Kim, H., Zheng, B., Jin, H.: A review of data mining with big data towards its applications in the electronics industry. *Appl. Sci.* **8**(4), 582 (2018)
8. Yee, O.S., Sagadevan, S., Malim, N.H.A.H.: Credit card fraud detection using machine learning as data mining technique. *J. Telecommun. Electr. Comput. Eng. (JTEC)* **10**(1–4), 23–27 (2018)
9. Fernando, M.P. et al.: Context aware standard process for data mining. pdfs.semanticscholar.org/0982/62df29c34e35dc3bf1fa03ffa67d62a66956.pdf. Accessed 11 June 2019
10. Hyeoun-Ae, P.: An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *J. Korean Acad. Nurs.* **43**(2), 154–164 (2013)
11. Chaudhary, K., Yadav, J., Mallick, B.: A review of fraud detection techniques: Credit card. *Int. J. Comput. Appl.* **45**(1), 39–44 (2012)
12. Khare, N., Yunus, S.: Credit card fraud detection using machine learning models and collating machine learning models. *Int. J. Pure Appl. Mathem.* **118**(20), 825–838 (2018)
13. Serrano, Manuel Rubio A., et al.: Neural network predictor for fraud detection: a study case for the federal patrimony department, pp 61–66, Doi <https://doi.org/10.5769/C2012010>
14. Patel, T., Ompriya Kale, A secured approach to credit card fraud detection using hidden markov model. *Int. J. Adv. Res. Comput. Eng. Tech. (IJARCET)* **3**(5), 1576–1583
15. Raj, N.: CRISP-DM the scrum agile way. Why not!" <https://www.elderresearch.com/consulting-services/agile-data-science>, Accessed 15 March 2018
16. Goldbloom, Anthony, et al. "Credit Card Fraud Detection", [https:// www.kaggle.com/mlg-ulb/creditcardfraud](https://www.kaggle.com/mlg-ulb/creditcardfraud), 2019.
17. van Rossum, G. et al.: Python tutorial: release 3.6. 4 (2018)
18. Shung, K.P.: Accuracy, precision, recall or F1. Confusion Matrix. <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>, Accessed 15 March 2018

Digital Learning for Millenials: IT in Education and/or IT for Education



Neelam Pirbhai-Jetha, Pascal Boncoeur, and Normada Bheekharry

Abstract Digitalisation has transformed organisations (existing or new ones), social systems and the global economy. One of the major drivers and emergence of digital platforms in the education sector is technology. Since we are saturated with technological devices, how can we, therefore, neglect or refuse the role technology plays or must play in education? For many researchers and practitioners, the main topic of discussion remains whether to consider the learner or the technology. In this paper, two main trending issues are addressed, the first one accounts on how and why digital technology transforms teaching–learning process. The second one focuses on the evolution brought about by information technology in the higher education sector, bearing in mind that the twenty-first century is the century of ‘digital natives’, term coined by Marc Prensky in [1] to refer to those who have grown up with digital technology. This theoretical/conceptual paper intends to investigate the different tools and artefacts which are used in the teaching–learning environment, in order to design new research data for future use.

Keywords e-learning · Digital tools · Digital platforms · Virtual classroom · Flipped classroom

1 Introduction

Whether we like it or not, technology has brought and will continue to bring about some changes in the teaching–learning process. In 2016, the National Education

N. Pirbhai-Jetha

Centre for Digital Humanities, Université Des Mascareignes, Pamplemousses, Mauritius
e-mail: npirbhajetha@udm.ac.mu

P. Boncoeur (✉)

Department of Software Engineering, Université Des Mascareignes, Pamplemousses, Mauritius
e-mail: pboncoeur@udm.ac.mu

N. Bheekharry

Department of Management, Université Des Mascareignes, Pamplemousses, Mauritius
e-mail: nbheekharry@udm.ac.mu

Technology Plan by the U.S. Department of Education posits that “the conversation has shifted from whether technology should be used in learning to how it can improve learning to ensure that all students have access to high-quality educational experiences” [2]. There is no way out: we have to rapidly move from traditional teaching–learning to digital teaching or a blend of both which seems to favour the social-constructivist approach to teaching–learning. Reasons behind this change in the teaching paradigm seem to have been brought about by the cultural changes in society. Owing to technological advances, distance learning has moved from postal services to personalised and adaptive learning, brought by the evolution in web. Figure 1 shows the evolution of distance learning throughout the centuries with the advancement in technology and telecommunication.

Indeed, many researchers in the educational fields put forward that the uninterrupted contact with technology has changed the mind-set of modern learners, who think and work differently, and have become more emancipated. In addition to this, the fourth industrial revolution—more linked to the digital age—demands a workforce with new skills. With the increasing accessibility to technology, educators have to face the “challenge of adapting their teaching styles to accommodate a new generation of digital learners” [4] and ascertained that to sustain student engagement, educators and the education system must therefore “adapt their teaching strategies to accommodate the digital learners” [4]. Literature in the research of MOOCs, for instance, supports that social media such as Facebook is important “to enhance the sense of community” and that social tools also make it easier to “engage students and improve their retention during the course” [5]. Technology seems to create emotional

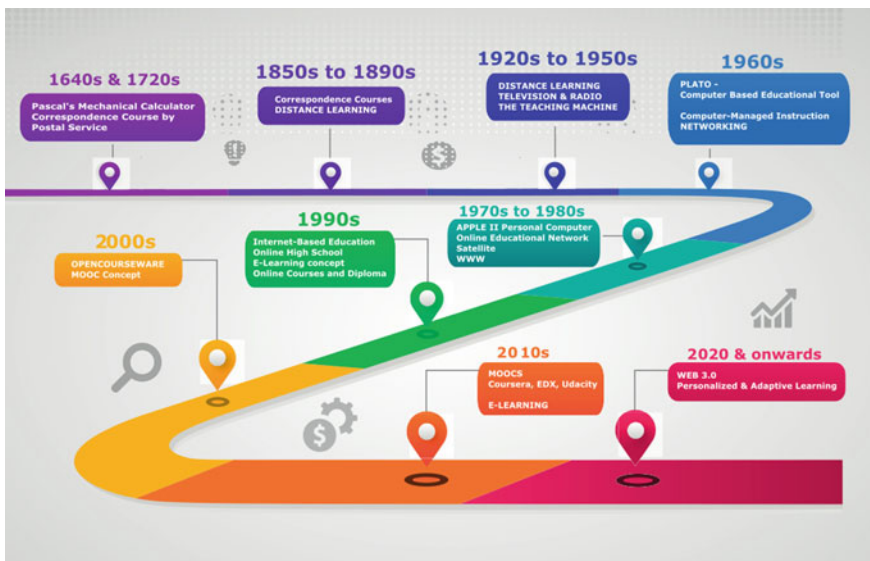


Fig. 1 Timeline of distance learning according to Demigray and Isman [3] and other sources, created with Gradient Timeline Infographic (2020)

connections between people, making teaching–learning “a social activity” [6], which cannot be done all alone. Indeed, most research done on digital learners reveal that they “gravitate toward group work” and that “they construct their knowledge” when they collaborate with others [4]. According to Tapscott [6], digital learners are “enmeshed in an interactive culture” and they have developed ten main characteristics: a fierce independence/autonomy; an emotional and intellectual openness; a desire for inclusion (with the whole world); free expression and strong views due to their confidence; innovation; maturity; curiosity, discovery, and exploration; immediacy and a wish for fast action and processing; a sensitivity to corporate interest, and authentication and trust [4]. This conceptual paper shows the importance of how the use of digital tools and e-learning can become a development tool for both the teacher and the student. In contrast to ‘traditional’ papers, where arguments are derived directly from data, conceptual papers focus on “problematizing”, “theorizing and theory building”. Hence, the need of a “theory synthesis” research in order to reduce “what is known to a manageable whole”, and thus creating a certain coherence among the different concepts and theories [7]. In the first section, we will analyse the new education paradigm, especially the evolution in higher education and the market needs. In the second section, we will look into the need to re-adapt the curriculum and to introduce new tools in order to prepare the twenty-first century labour force to face the changing needs of Industry 4.0.

2 Theoretical or Conceptual Discussion

2.1 *New Education Paradigm and Twenty First Century Skills*

An OECD report notes that ‘skills are extremely important in the modern economy and in everyday society’ [8]. There are numerous debates on the twenty-first century skills, but all agree that it is the education system that must address the challenges that workplaces present. Referring to Kereluik’s framework, skills and competences are broadly divided categories [9], and the framework was further developed by Kennedy et al. [10]. Skills, which are an integral component for constructive learning, are categorised as shown in Fig. 2:

However, very few education systems provide the “infrastructure, pedagogical methods, or actual curricular material that will maximally prepare students for the current and future world in which they will enter and lead in their future” [11]. In order to cater for the market demands/Industry 4.0, it is crucial to have a workforce with the twenty-first century skills, and this will be only possible if there is a shift from the traditional system to a more modern classroom, and if learners are encouraged to develop the major learning and innovative skills [12]. Here, technology can be a crucial tool in developing certain of these skills (communication, creativity, collaboration...). For instance, in some research on education, it has been noted

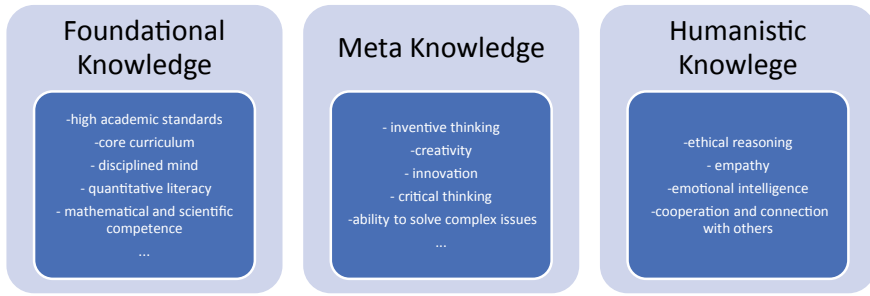


Fig. 2 Twenty-first century skills framework

that in discussion forums, learners have two roles: (i) that of learners, and (ii) the moderator which they take voluntarily, thus liberating the learner from traditional power structures in education. In other terms, they are free from the dominance of the institution and a top-down educational approach, where the teacher controls and determines the nature of the experience within the tightly controlled guidelines of the accrediting institution [12].

2.2 *Evolution in Higher Education*

2.2.1 **Role of the Facilitator**

In the previous paragraph it is clearly stated that skills needed in the future job market are changing. Even our concept of workplace is no longer “connected with a fixed physical place designed to put people together and hold them together to provide a specific service or output” [13]. Such improvements cannot remain unnoticed by higher education at large. Even if the “role of the teacher in the learning process remains central [13], the nature of teacher presence is changing due to an increase in virtual and asynchronous interactions online” [14]. Indeed, the traditional role of teachers/academics as knowledge experts is being challenged; and the academic is becoming more a facilitator, one who guides students through their learning. The sociologist, Bauman, developed the concept of “liquid times” in 2005 which was further researched in 2012. He stated that “old ways of doing things are no longer working, but new ways of doing have not yet been found and there is no clear image of what we are moving towards”. According to him, “our practices therefore need to become liquid too”; that is “we need to be agile, flexible, in movement and interdisciplinary” [15], especially in the pedagogies that are being practiced.

2.2.2 The ‘modern’ University

According to the OECD modelling, there will be a great increase in university enrolment rate and that “within the next decade a third of the adult population of the OECD will possess at least one degree” [16]. This also means that the student population is very diverse, and not all are enrolled in a full-time course. Universities will have to cater for students “who are less able to devote their time completely to study, intellectually less well prepared or able, and without the personal resources and support that characterise the more elite model of the past” [16]. The growth in student numbers also means larger classes. Technology seems to bring about some solutions to university stakeholders and this can be understood through Bloom’s Taxonomy, where knowledge (skills and abilities), comprehension, application, analysis, synthesis, and evaluation are put into practice. In order to cater for the ‘modern’ university, IT for Education seems to be the solution.

2.3 *IT for Education*

2.3.1 Virtual Classrooms and Virtual Communities

The change in student population such as working students can mean that many might not attend classes on a regular basis. The sanitary conditions and total lock-down of all life activities, which the whole world faced with COVID-19, have also made us realise the need to re-vamp the education sector. Many universities are offering distance learning classes, where students interact in virtual classrooms and in virtual communities. Online short courses or Moocs are also becoming popular, and many internationally ranked universities are offering short courses to cater for life-long learners, who wish to brush up certain of their skills [17]. E-learning platforms have therefore led to new possibilities and innovative ways in teaching–learning. Virtual classrooms are beneficial when face-to-face courses are not possible either due to low enrolment in the course or difficulty in finding certified teachers for specific subjects.

The social constructivist approach to teaching and learning classroom activities gives a great importance to active participation of learners through activities, such as discussions, debates, and collaboration. Indeed, digital learners like “to express their views and incorporate their experiences into their learning” [6]. Since they learn by doing and working with others, some researchers have called them “the Nintendo Generation” as “the key to winning Nintendo is the persistent trial and error to discover the hidden doors” [18]. Brown referred to the learners as “digital bricoleurs” and he noted that this generation collects bits of information, objects, or tools to create something new visualizations, simulations, case analyses, and other methods of participatory learning such as fieldwork are all part of the learning repertoire [18]. However, even if students are at ease with the use of technology, does it necessarily mean that they will be fully motivated and engaged during the online classes? An

article on 4 May 2020, in *The Guardian*, looks into the struggles of learners during the COVID-19 pandemic: WIFI problems, lack of equipment, lack of conducive learning environment, reduced support (technological, psychological...). There is, therefore, a dire need, for each country to look into the integration of technology into their teaching–learning environment as just uploading notes on a platform does not guarantee engagement and motivation of students.

2.3.2 Review of Curriculum Design and Assessments

With the change in the academic’s and learner’s roles, the curriculum and assessment modes have to be reviewed. Technology can be used “to tailor learning experiences to individual learners’ needs; as it is easier for teachers to design learning experiences that appeal to learners’ different proficiency levels, learning styles, and emotional or social needs [19].

Curriculum, therefore, has to be reviewed: from acquisition-based (knowledge transmission), it has to become competency-based (knowledge building). Favouring a “knowledge building education” will mean that the “classroom can become a place of sharing new thoughts and building common ground”, where one can “learn how to handle information and build knowledge” and not merely “a place to merely acquire new skills and knowledge” [12]. Another change will be in assessment mode. Technology- assisted assessments, in contrast to paper assessments, also have many advantages as setting up and correction of papers save time and resources, and allow immediate feedback, “which is particularly useful in the case of formative assessments” [20].

2.3.3 Flipped Classrooms

Bergmann and Sams [21] defines flipped learning as “that which is traditionally done in class is now done at home, and that which is traditionally done as homework is now completed in class”. To be more clear, “a typical practice of flipped learning usually involves learners viewing a short video explaining a concept at home and then applying the concept to problem- solving tasks when they come to the next class” [22]. In flipped classroom, what changes is the timing: instead of giving a more traditional lecture in class and activities as homework, the lecture is recorded and becomes the homework, and students engage in activities during classroom period [23]. This type of teaching–learning also allows students to be more engaged.

Flipped classroom was administered in order to diminish time for lecture in the classroom, to focus the class time for understanding and application of what one learned through the lecture, to have the students actively think for problem solving, and to provide opportunities for the students to learn from other members of class. As a result, Baker [23] maintains that motivation of the students increased. They were also better able to think more critically. Similar study was done by Lage et al. [24] where flipped classroom was conducted in an economics class. In this study, a

10-min video lecture was given for students to watch as homework, and during the class period, students participated in group work and experiments. According to Lage et al. [24], when they administered a survey, students preferred flipped classroom as opposed to a regular class.

3 Conclusion

To finalise emphasis should be laid on the enablers which promote the knowledge creation process: creating the correct context; managing conversations through engagement and participation without forgetting globalizing local knowledge. This paper entails a theoretical overview and the implication of how technology enhances our instructional design and help learners achieve higher proficiency and at the same time empowering learners with twenty-first century skills. From a social constructivist viewpoint, learners interact not only with others but also with various artefacts (language, images, new tools, and technology). It is apparent that learning and knowledge cannot be constructed without the mediation of such artefacts. Learners need to connect with “the other” and it is in this process, that the learner shares and generates new knowledge. However, despite the universality of technological tools, each student/institute/country is different, and more studies have to be done on the integration of technological tools in the education system. For future studies, a post-covid research on the positive aspects and shortcomings of the learning platforms and artefacts can be carried out among students, and data collected can be used offer a more conducive digital learning experience to millennials. In fact, as shown in Fig. 3, it

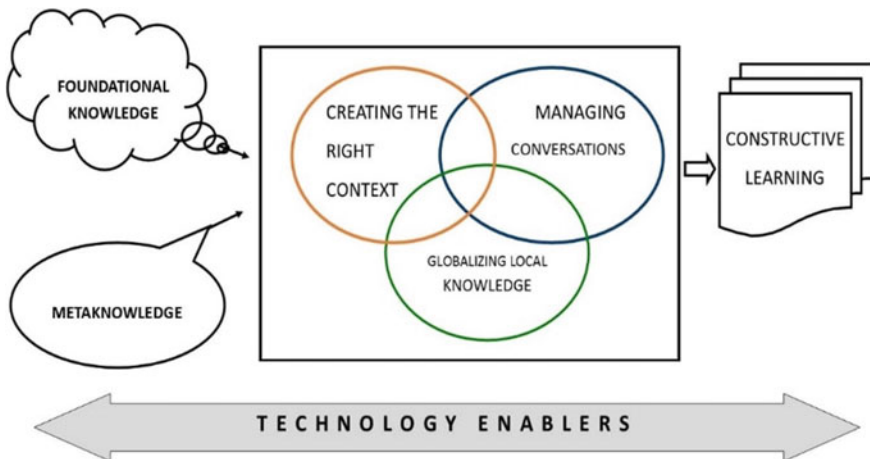


Diagram showing: Sharing & Managing Knowledge constructively

Fig. 3 Sharing and managing knowledge constructively (Bheekharry, Boncoeur, Pirbhai-Jetha)

is not in the technology but in learning how to use information and technology to create a new value, that one becomes empowered.

References

1. Prensky, M. (2001) Digital natives, digital immigrants. *On the Horizon* **9**(5)
2. Navarre, A.: *Technology-Enhanced Teaching and Learning of Chinese as a Foreign Language*. Routledge, London and New York (2019)
3. Demiray, U., Isman, A.: *History of Distance Education*. Online Distance Education Book, TOJET (2003)
4. Ramlee, M., Kashefian-Naeeni, S. (2017) Moving teaching and learning into the digital era. *Int. J. English Lang. Transl. Stud.* **5**(3)
5. Koutsakas, P., Syritzidou, E., Karamatsouki, A., Karagiannidis, C.: *Exploring the Role of Facebook as Collaboration Platform in a K-12 MOOC*, pp. 31–48. Springer, Switzerland (2018)
6. Tapscott, D.: *Growing Up Digital. The Rise of the Net Generation*. McGraw Hill, New York (1998)
7. Jaakkola, E.: *Designing Conceptual Articles: Four Approaches*. Springer Nature, AMS Rev (2020)
8. Bourn, D.: *Understanding Global Skills for Twenty First Century Professions*. Palgrave Macmillan, Switzerland (2018)
9. Kereluik K., Mishra, P. Fahnoe, C., Terry, L.: What knowledge is of most Worth: teacher knowledge for 21st century learning. *J. Digit. Learn. Teacher Educ.* **29**(4) (2013)
10. Kennedy, I., Latham, G., Jacinto, H.: *Education Skills for 21st Century Teachers: Voices From a Global Online Educators' Forum*. Springer, London (2016)
11. Chu, S.K.W., et al.: *Twenty First Century Skills Development through Inquiry-Based Learning*. Springer, Singapore (2017)
12. Linuma, M.: *Learning and teaching with technology in the knowledge society new literacy, collaboration and digital content*. Springer, Singapore (2016)
13. Guri-Rosenblit, S.: Distance education in the digital age: common misconceptions and challenging tasks. *J. Dist. Educ.* **23**(2) (2009)
14. Trede, F., Markauskaite, L., McEwen, C., Macfarlane, S.: *Education for Practice in a Hybrid Space Enhancing Professional Learning with Mobile Technology*. Springer, Singapore (2019)
15. Bauman, Z.: *Liquid times: living in the age of uncertainty*. Polity Press (2007)
16. Marshall, S.: *Shaping the university of the future: using technology to catalyse change in university learning and teaching*. Springer, Singapore (2018)
17. Sammons, M.: Collaborative interaction. In: Moore, M. (ed.) *Handbook of Distance Education*. Lawrence Erlbaum Associates, New Jersey (2007)
18. Skiba Diane, J., Barton Amy, J.: Adapting your teaching to accommodate the net generation of learners. *Online J. Issues Nurs.* **11** (2006)
19. Gabbard, W.J., Starks, S.H., Jagggers, J., Cappiccie, A.C.: *Effective strategies for teaching cultural competency to MSW students in a global society*. *World Acad. Sci. Eng. Tech.* (2011)
20. Dodds, P., Fletcher, J.D.: Opportunities for new “smart” learning environments enabled by next-generation web capabilities. *J. Educ. Multim. Hyperm.* **13**(4) (2004)
21. Bergmann, J., Sams, A. *Flip Your Classroom: Reach Every Student in Every Class Every Day* (pp. 120–190). Washington DC: International Society for Technology in Education (2012)
22. Stone, B.: Flip your classroom to increase active learning and student engagement. In: *Proceedings from 28th Annual Conference on Distance Teaching and Learning*, Madison, WI (2012)

23. Baker, J.W. The “Classroom Flip”: Using web course management tools to become the guide by the side. In J. A. Chambers (Ed.), *Proceedings of the 11 th International Conference on College Teaching and Learning* (pp. 9–17). Jacksonville, FL: Florida Community College at Jacksonville (2000)
24. Lage, M.J., Platt, G.J., Treglia, M. Inverting the Classroom: A Gateway to Creating an Inclusive Learning Environment. *J Econ Edu* **31**(1):30 (2000)

Supply Chain Management—Marketing Integration a Key Element in the Digital Era



Normada Devi Bheekharry

Abstract The purpose of this paper is to study the development and influence of digitalization in the marketing processes and supply chain management and how an integration of both can be a useful tool in marketing decision making. Many research has been conducted on supply chain management and marketing; however, little attention has been drawn to the combination of these two functions and the potential impacts of information technology in decision making and value-creation. Data generated from Internet of things, cloud computing, big data analytics and customer profiling have created new opportunities for businesses to supplement their business intelligence for better understanding customer demand and bargaining power of suppliers.

Keywords Marketing · Supply chain management · Big data analytics · Supply chain management—marketing integration · Customer intelligence

1 Introduction

The integration of supply chain management (SCM) and marketing (M) has been a subject of interest to academicians, marketers and practitioners; consequently, many papers have been published in that context. SCM has always been associated with supply in terms of logistics and operations by definition and M as a management function for creating, developing and delivering demand by understanding the needs of customers. In both functions (M and SCM), the time element is very important. If time is not respected in M, that is products are not produced on time, placed within customer's reach and not delivered when needed, the customers will not be satisfied and will ultimately turn toward competitors' products and services. That is why much reference are taken from the definition of M from the Chartered Institute of Marketing, UK, where M is defined as: "... It is all about getting the right product or service to the customer at the right price, in the right place, at the right time." The same

N. D. Bheekharry (✉)

Department of Management, Université Des Mascareignes, Pamplémousses, Mauritius

e-mail: nbheekharry@udm.ac.mu

importance of time is attached to SCM, where if products and/or raw materials are not distributed to supplier, manufactures, distributors and customers, organizations will incur massive lost. As a result, products may perish or organizations should invest more in warehousing and inventory cost. The SCM function is mainly focused with operations and inbound/outbound logistics. By simple definition, put forward by practitioners: “It is getting the right product in the right place at the right time”.

Hence, the interface between SCM and M can easily be understood and why it is a research topic that has long interested researchers. Nonetheless, in 2010, in a research paper published by Mentzer and Gundlach, entitled “Exploring the relationship between marketing and supply chain management,” it was declared that the topic of SCM and M has not been explored in great depth [1]. Innovation in information technology has brought massive contribution to improve the flow in SCM, and in today’s competitive environment, companies are searching ways to stay ahead, improve their competitive position and achieve superior financial performance [2]. In this new millennium business practices have altered and are moving toward a more collaborative commerce among suppliers, manufacturers and distributors [3]. This collaboration of business processes has helped to reduce time-to-market, improve responsiveness to demand and increase customer service levels [4] (Fig. 1).

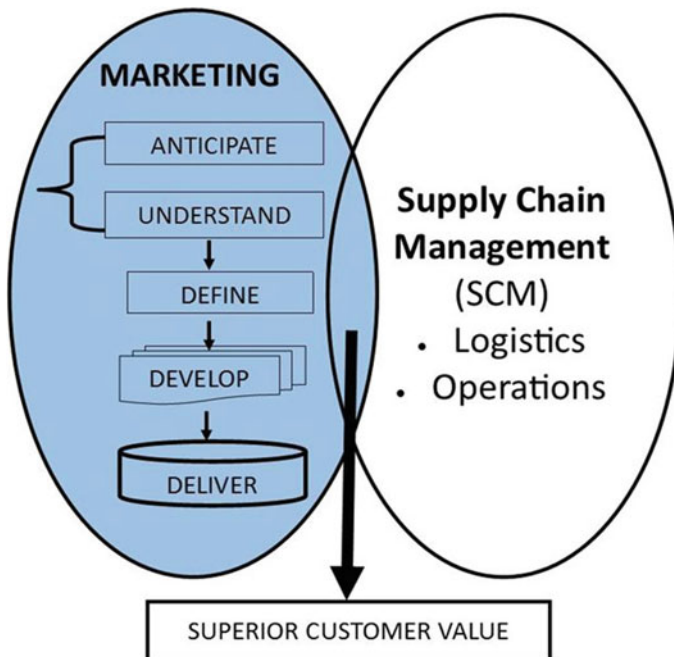


Fig. 1 SCM–M integration

This paper is a conceptual one and is aimed to elaborate on the importance of SCM–M in creating superior customer value which is pivotal for business survival and growth and achievement of superior financial performance [2].

2 Literature Review

2.1 *Digital Era*

Digitalization has revolutionized the information technology and forced businesses to move toward the Internet as a platform for several operational functions. The Internet is evolving and over the past twelve years has become the major source of communication worldwide. It is a global network of interconnected networks. Three important types of networks form part of the Internet: intranet, extranet and web [5]. Creating customer value is an important element in the value chain [6] and on the other hand is pivotal for firm survival [7]. Communication is vital and seen as a medium to control the relationships within the chain together and the Internet provides a great opportunity to do so [8]. In a research conducted by Sharma in 2002, it was clearly stated that firms need to develop an information platform that integrates all business processes and functions, including suppliers and customers [9]. This reinforces and maintains value along the chain that forms the business environment and allows an increase in both efficiency and effectiveness [6]. It has been argued that supply chain network is precisely about inter and intra communication and information flow within an organization and its key stakeholders [10]. The responsibility of the organization will lie in building and maintaining effective information flow to avoid friction and enhance collaboration and co-operation among different members along the chain.

2.2 *Technology Derivatives and Tools*

Digital technology enables firm to react positively and proactively to customer needs and at the same time improve customer-side operations. Easy, inexpensive and quick access to digital information transforms individuals, businesses, economies and societies. Technology allows for connections between individuals and communities and fosters relationship among businesses and societies. Research has proven that digital technologies play a significant function in almost every organization [11].

- (1) Websites: A dynamic website can provide information, knowledge and social interaction which facilitates communication among organization's stakeholders (employees, customers, suppliers, media, investors and distributors among others) and at the same time promotes the creation of value [12]. Websites are easy to access and not costly. Furthermore, each channel member not only provides but is provided with information thus forming a web connectivity [13].

Various audiences have different access points: for example, the public through the Internet, extranet is available to customers, suppliers, partners, whereas the intranet is solely for communication inside the organization [14].

- (2) Web 2.0 tools: Innovation and development in information technology particularly Web 2.0 technologies have given rise to social networks platforms, for example, [Facebook, Twitter, blogs and YouTube]. These different platforms contribute to advancement in communication like sharing ideas, improved access to knowledge, reduce in costs of communications, travel and operations [15]. Arora [16], states that a good relationship between digital technologies and its users does not only improve user satisfaction but also promotes effective communication. Customer relationship management (CRM) has been improved further into CRM 2.0 or as social CRM [17]. CRM 2.0 is a powerful management tool for employee interaction and strengthening relationship with customers, suppliers and outside partners. Web 2.0 has facilitated the sharing of knowledge and establish co-operation in the form of virtual group [15]. According to Strauss [18], customers are given the facility to upload or share content online. This interaction has encouraged customers to participate, collaborate and give their feedback on product, service or the organization. Customers engagement online can increasingly be observed favoring a win–win relationship between organization and customers. Another source of communication which Web 2.0 has accentuated is electronic word-of-mouth, which may positively or adversely influence existing or prospective customers [19]. Watson [20] outlines that information quality can be transmitted through digital technology from the management of these information contributes to the growth in sales and increase efficiencies by reducing costs.

3 Clouds, Big Data and Internet of Things (IoT)

The cloud is used as platform to collect, store and process data. It is also used to analyze and share information. The cloud allows business to lease computer power as and when required instead of having to buy expensive equipment [18]. Digital platforms as mentioned in the previous section, that is technology derivatives and tools, bombard organizations daily and by seconds by huge and different forms of data. This phenomenon is known as big data. The major drivers for linking big data to cloud are cost reduction, reduction in overheads and flexibility [21].

Digitizing the supply chain with sensors and IoT can also improve demand forecasting and fulfillment. In today's retail world where consumers increasingly bounce across channels to research, select and purchase products, understanding demand can be a challenge.

There is a lot of noise and confusion around IoT in retail, but it is clear that retailers are hoping to improve inventory accuracy and management while providing a better customer experience [22].

4 Applying the Concept

The following section will consider the effectiveness and efficiency of proper technology, supply chain management and marketing blend. The case study is taken from SAS Blog whereby the value of managing information and integrating SCM and M is seen to promote organization development and growth.

Carrefour is a multinational food retailers network in the world with a distribution of 12,000 stores in over 30 countries. Carrefour serves 105 million customers worldwide and posted sales of 88.24 billion euros in 2017. The group has more than 380,000 employees who contribute to making Carrefour the world leader in the food transition for all, offering quality food every day, accessible everywhere and at a reasonable price. It is the first French retailer using AI to increase availability of merchandise in-store and online and relies on up to date technology to optimize supply chain management and reduce wastage. The multinational giant aims and objectives are to optimize its supply chain as part of the company's global transformation plan to better meet customer expectations through advanced technologies. Carrefour aims to create a unique online and in-store shopping universe where the most suitable merchandise for recognized loyalty customers is guaranteed any time, any place. The deployment of artificial intelligence (AI) to optimize the supply chain is a first for the French retail sector.

Carrefour is meeting the challenge of Omni channel distribution and inventory optimization by using effective means to collect and process data from stores, warehouses and e-commerce sites. Analysis of the data will improve downstream forecasting and upstream ordering with suppliers—reducing waste and overstocks.

5 Conclusion

The central argument of this conceptual paper is to show how the business environment is changing and how organization's processes should be aligned to deliver better customer service and competitive advantage. The business world is moving toward Industry 4.0 where organizations are adopting digital technologies to create connections between operational processes in order to gather and share real-time market and operational information. Kozinets et al. [23] stated that the diffusion of Internet technologies and their associated characteristics have greatly influenced the emergence and evolution of consumer empowerment and this is an area of future research where the relationship between consumer empowerment, marketing and supply chain management can be explored. With the evolution of the web technologies and digital platforms, organizations are bombarded with data. Analyzing this information firms can use big data and big data analytics to better understand who their customers are, what they buy, when they buy and how they buy? However, much research has not been done in this area and an interesting area of future research can be the integration of SCM-M and online customer behavior.

The major purpose in any business whether be public, private or NGOs is to find an equilibrium between demand and supply. Understanding the demand of the market and supplying the product and or service at the right time is the principal rule for competitive advantage. Integrating demand and supply that is marketing and SCM definitely makes organization more efficient and effective. For organization to survive in this competitive and dynamic business, they need to use innovative technological devices to survive.

References

1. Mentzer, J., Gundlach, G.: Exploring the relationship between marketing and supply chain management: introduction to the special issue. *J. Acad. Mark. Sci.* **38**(1), 1–4 (2010)
2. Woodruff, R.B.: Customer value: the next source for competitive advantage. *J. Acad. Mark. Sci.* **25**(2), 139 (1997)
3. Harreld, H.: Supply-chain collaboration. *InfoWorld* **23**(52/53), 22–25 (2001)
4. Favilla, J., Fearn, A.: Supply chain software implementations: getting it right. *Supply Chain Manage. Int. J.* **10**(4), 241–243 (2005)
5. Strauss, J., Frost, R.: *E-marketing*, 5th edn. University of Nevada at Reno, Pearson (2005)
6. Foster, T.: Into the depths of the I-E-I framework: using the internet to create value in supply-chain relationships. *Supply Chain Manage. Int. J.* **12**(2), 96–103 (2007)
7. Ardito, L., Messeni Petruzzelli, A., Panniello, U., Garavelli, A.: Towards industry 4.0: mapping digital technologies for supply chain management-marketing integration. *Bus. Process Manage. J.* (2018)
8. McGuffog, T.: The obligation to keep value chain managements simple and standard. *Supply Chain Manage.* **2**(4), 124–133 (1997)
9. Mohamed, S.: Web-based technology in support of construction supply chain networks. *Work Study* **52**(1), 13–19 (2003)
10. Foroudi, P., Gupta, S., Nazarian, A., Duda, M.: Digital technology and marketing management capability: achieving growth in SMEs. *Qual. Mark. Res. Int. J.* **20**(2), (2017)
11. Setia, P., Venkatesh, V., Joglekar, S.: Leveraging digital technologies: how information quality leads to localized capabilities and customer service performance. *Mis Quarterly* **37**(2), 565–590 (2013)
12. Baker, M.J., Buttery, E.A., Richter-Buttery, E.M.: Relationship marketing in three dimensions. *J. Interact. Mark.* **12**(4), 47–62 (1998)
13. Hoey, C.: Maximizing the effectiveness of web-based marketing communications. *Mark. Intell. Plann.* **16**(1), 31–37 (1998)
14. Chaffey, D., Mayer, R., Johnston, K., Ellis-Chadwick, F.: *Internet Marketing: Strategy, Implementation and Practice*. Prentice Hall, Harlow (2003)
15. Sharma, G., Baoku, L.: Customer satisfaction in Web 2.0 and information technology development. *Inf. Technol. People* **26**(4), 347–367 (2013)
16. Arora, N.K.: Interacting with cancer patients: the significant of physicians' communication behavior. *Soc. Sci. Med.* **57**(5), 791–806 (2003)
17. Greenberg, P.: CRM at the speed of light. In: *Social CRM 2.0 Strategies, Tools, and Techniques for Engaging Your Customers*, 4th edn. McGraw-Hill Osborne Media, New York, NY (2009)
18. Strauss, J., Frost, R.: *E-marketing*. Pearson Education International, Upper Saddle River, NJ (2009)
19. Sen, S., Lerman, D.: Why are you telling me this? An examination into negative consumer reviews on the web. *J. Interact. Mark.* **21**(4), 76–94 (2007)
20. Watson, D.: Understanding the relationship between ICT and education means exploring innovation and change. *Educ. Inf. Technol.* **11**(3/4), 199–216 (2006)

21. Pasalapudi, S.K.: Trends in cloud computing: big data's new home. Oracle (2014). Available at: www.oracle.com/us/corporate/profit/big-ideas/012314-spasalapudi-2112687.html (accessed 21 January 2019)
22. Heindrick G: IoT application in retail: the supply chain (accessed on 27 February 2019)
23. Kozinets, R.V., de Valck, K., Wojnicki, A.C., Wilner, S.J: Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities. *J. Mark.* **74**(2), 71–89 (2010)
24. Kozinets, R.V.: E-tribalized marketing? The strategic implications of virtual communities of consumption. *Eur. Manage. J.* **17**(3), 252–264 (1999)
25. Phairor, K., Hanmer-Lloyd, S.: Rethinking channel communications: an emerging role for the extranets within distribution channels. In: *Marketing Theory and Applications*, American Marketing Association Winter 2002 Educators' Conference, vol. 13, pp. 16–22. American Marketing Association, Chicago, IL (2002)
26. Prahalad, C.K., Ramaswamy, V.: Co-opting customer competence. *Harvard Business Review* (2000)
27. Tian, X.: Big data and knowledge management: a case of déjà vu or back to the future? *J. Knowl. Manage.* **21**(1), 113–131 (2017)

Bank Customer's Credit Score Prediction Using Feature Selection and Data Mining Algorithm



Durgesh Kumar Singh and Noopur Goel 

Abstract Data mining techniques of classification and prediction model are used for analyzing the customer of bank through their real data, customer is a performing as asset or non-performing asset for the bank. Every year each bank faces a similar problem that most of them customers do not refunding loan installment on time and many precisely become defaulter for the bank. So every bank needs a system that can predict in future any customer in future will be profitable or not profitable asset for the bank. If any customer found as non-performing asset means, it has bad credit score. In such case if a customer further requests for new loan, bank can easily identify him as defaulter and reject his/her request on the basis of our new proposed models. In this way, bank extracts their non-performing assets. Besides, bank can also identify their new customers for having good credit score and may offer them other services for the beneficial of bank. There are many as such models that are used for prediction. This paper compares different classification and prediction algorithms to developed best suitable model to analyze loan requesting customer's data set using their credit score. Mainly in this paper, there is a comparison between random forest algorithms and logistic regression which is best suitable for prediction of credit scores of customers with apply k best feature selection on data set. The aim of proposed model is to reduce bankruptcy, non-performing assets and the losses of bank.

Keywords Data mining · Classification models · F1-score · Bank credit · Feature selection

D. K. Singh (✉) · N. Goel
VBS Purvanchal University, Jaunpur, India
e-mail: durgeshsingh111@gmail.com

N. Goel
e-mail: noopurt11@gmail.com

1 Introduction

Credit enables individuals or businesses to “eager to pay” or “buy ahead of ability.” Banks give facility them to adequate loans based on costumers requirement in the corporate organization, agricultural and business sectors. In addition, at what time customer uses intrinsic intelligence just in the case of credit, it leads to financial enhancement. In recent scenario, government of country in bank decision that sometime becomes problematic for the banks and it is one the main reasons of bank’s non-profitable assets (NPA). To reduce this bankruptcy, NPA and the losses of bank, we have applied some data mining algorithm. These algorithms use as a tools that will help to determine the ability of customers to repay credit loans in a timely manner or not and categorize the customers as “good credits” or “bad credit.” A credit score is a numerical assessment that a bank uses with your credit report to assess the risk of providing you with a loan or providing credit to you [1].

One of the main purposes for organizing a bank is to lay out loans. But to maintain functioning, the bank issues these loans to those who are able to pay back, thereby minimizing the risk of outstanding lending. Nevertheless, risk management knowing who is notable of credit remains a continuing challenge in the banking industry. Ability to identify risk levels, customers based on characteristics such as job, age, marital status, salary span/net asset value, solvency status, etc. are key values that banks must examine before providing loan to customers. With the value of credit risk scores, you can help the bank to decide considerable interest to charge on the loan. Although, these risk elements from time to time fail to make an informed conclusion about the customer’s creditworthiness.

Data mining algorithms will be used to analyze the loan approval data and will find out a patterns that will help to predict non-profitable customer and thus help banks make better decisions in the future. Data set from different sources will be used for creating a framework, and then different data mining algorithms will be applied to extract the patterns and get the results with maximum accuracy.

2 Related Work

Many research already have been discussed on this problem with data mining and machine learning algorithm in the area of bank and loaning. Some examples are as such-

Hsu et al. [2] for predicting the customer credit applied support vector machine on data set firstly without feature selection and after feature selection then compare their results and found after feature selection results of accuracy rate has been improved.

Turksen et al. [3] predict credit related to loaning of bank customer using supervised and unsupervised learning algorithm; different algorithms give different accuracy rate.

The model proposed by Jafarpour et al. [4] on Iranian bank's data set to predict loaning accuracy of customers and also establishes a relation between customer's requirement and banks and formulates a equation through which bank predict credibility of customers in concern of loaning.

Hassan et al. applied neural network algorithm for loaning of customer in bank. Firstly create a supervised model of bank credit data set. This model is very useful for bank to analyzing that any customer will return their sanction loan in future or not.

Jin et al. [1] employed data mining algorithm to found non-profitable assets of the bank in concern loaning and they also do a comparative study on different data mining algorithm and they conclude that support vector machine algorithm performed best.

Morco et al. [5] proposed a data mining approach for the prediction of customer's credit of Portuguese retail bank in telemarketing. They analyze the result of accuracy of different data mining model and found that neural network data mining algorithm performance is best.

A data mining approach conducted by Li et al. [6] on using data set of customers to predict risk using attribute bagging method. They found that the performance is outstanding using two credit databases.

3 Proposed Model

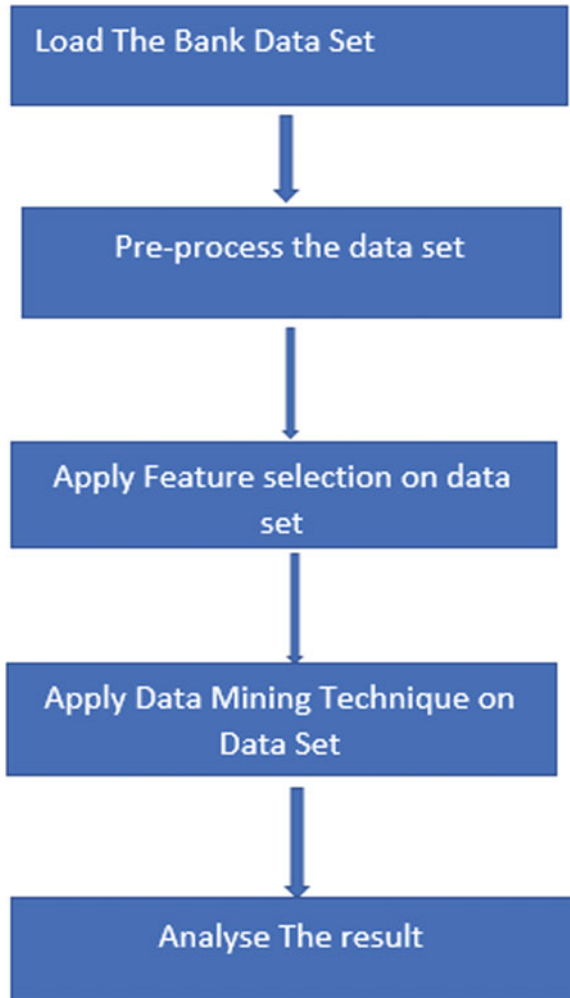
Figure 1 shows the flow of our paper. To process all the step, Python code is used on Jupiter notebook interface. The very first step is to load the data set then pre-processed to all attributes of data set. Then using very important technique that helps to improve the accuracy result of algorithm is feature selection through which redundant values attribute is eliminated. The next step after feature selection on pre-process data set apply data mining approaches to find the result and in the last analyzes the result.

4 Data Description

4.1 Data Source

The data set is used of direct marketing campaigns of a Portuguese banking institution. The data set has been available publicly at UCI machine learning repository.

Fig. 1 Proposed model



4.2 Data Description

Number of Instances: 45,211 for bank-full.csv. Number of Attributes: 16 + output attribute. Data set is separated in 70–30%. 70% for training data set and 30% for testing data set on each particular model.

Figure 2 explains the dat set attribute description and their type.

S.no	Attribute	Attribute Description	Attribute Type
1	Age	Customer’s Age	Numeric
2	Job	Type of Job	Categorical
3	Marital	Marital Status	Categorical
4	Education	Education Status	Categorical
5	Default	Customer has credit in default	Binary(yes or no)
6	Balance	Average yearly balance	Numeric
7	Housing	Has housing loan?	Binary(yes or no)
8	Loan	Has personal loan?	Binary(yes or no)
9	Contact	Contact communication Type	Categorical
10	Day	Last contact day of the month	Numeric
11	Month	Last contact month of year	Categorical
12	Duration	Last Contact Duration in sec	Numeric
13	Campaign	Frequency of contacts performed	Numeric
14	Pdays	number of days that passed from last contacted	Numeric
15	Previous	no. of contacts before campaign	Numeric
16	Putcome	outcome of the previous marketing campaign	Categorical
17	Y(Output Variable)	has the client subscribed a term deposit?	Binary(yes or no)

Fig. 2 Data set description

5 Feature Selection

Feature selection is the process of selecting best attributes from any given data set. The best attribute means that value matters most in output variable for the best prediction of accuracy.

It also reduces over fitting means to eliminate those attributes from data set that not play major role in making of decision. Due to Reduces Training time of algorithm, data set gets trained in less time. This way feature selection plays a major role in finding accuracy rate of different algorithm on different data set.

There are different feature selection method on which we used select *k* best feature selection using Python code on Jupiter notebook.

The value of attributes in which higher value attribute will play major role in data mining algorithm to finding accuracy and less value attribute will not so much affect in the accuracy result.

In Fig. 3, the least play role, in accuracy of algorithm, attribute is eliminated. Now we are using only these attribute in our data set to predict accuracy of customers.

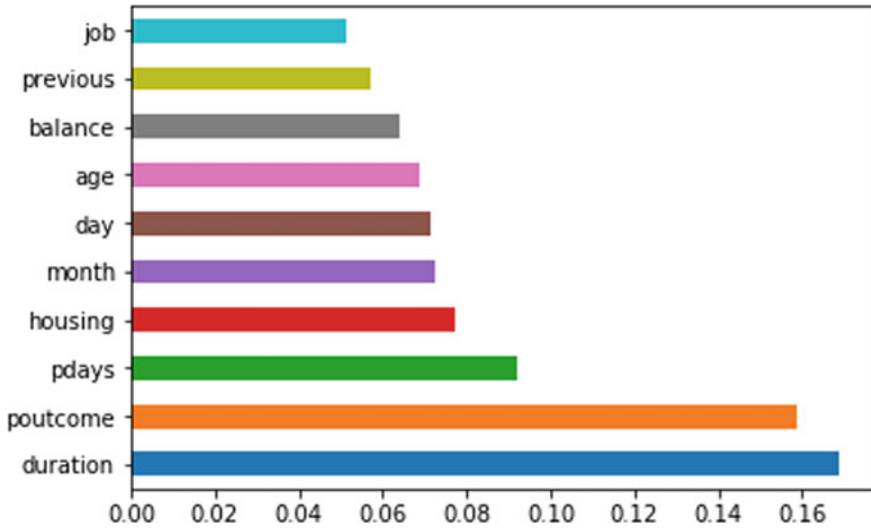


Fig. 3 Selected attribute after feature selection

6 Model Used

In this paper, two data mining models are used for prediction of bank customer behavior regarding to refund the loan. Different mining classification model gives accuracy result different on the same data set. Then create a comparison between these different mining models and suggest to bank which model will be best for evaluating the credit score of customer.

Following algorithms are used.

- Random forest algorithm
- Logistic regression algorithm.

6.1 Random Forest Algorithm

Random forest or random decision forest is algorithm that works on creating many decision tree in time of training phase. It is basically an ensemble learning method for classification, regression that operate on multiple decision tree at training time and the determination of the majority of the trees is selected by the random forest as the last finding. Random forest deals in the best way from the over fitting, that negatively impacts on the accomplishment of the model on new data. For the large data set, random forest model produces high accuracy; it runs very efficiently on large database. It also maintains its accuracy in case of missing data in large number [7].

6.2 Logistic Regression Algorithm

Logistic regression algorithm is a technique that can be used in traditional statistic as well as in data mining. Logistic regression algorithm is much similar to linear regression except that logistic regression algorithm predicts whether something true or false instead of something continuous like size.

Logistic regression is a regression analysis used to predict the results of a classification dependent variable based on one or more predictors. In logistic regression, a S-shaped curve used instead of straight line like in linear regression. The formula for a univariate logic curve is $p = \frac{e^{(c0+c1 \times 1)}}{1+e^{(c0+c1 \times 1)}}$.

To perform the logarithmic function can be applied to obtain the logistic function $\log_e \frac{p}{p-1} = c0 + c1 \times 1$.

Here p is probability in one class and $p-1$ is on another class. Logistic regression is easy to implement and simple and used for wide variety of problem with good performance.

7 Discussion of Result

Applied data mining algorithm on the data set and calculated their training and test case accuracy on different field to find bad customer and good customer for bank. The accuracy result is following in the Table 1.

On the basis of result, shown in Table 1, the accuracy of mining algorithm the logistic regression algorithm’s test case accuracy result is less than the random forest classifier’s test case accuracy result. Hence, random forest algorithm is the most suitable algorithm for the data set than logistic regression. F1 score result also show that, logistic regression accuracy is less than the random forest classifier. Random forest is the best for analyzing data set of bank and on that basis, bank can easily predict which customer retention is profitable for bank. Figure 4 shows the result of accuracy using graph.

Table 1 Summarized view of results of data mining classifier

Algorithm	Training case accuracy	Test case accuracy	Precision	Recall	F1-score
Logistic regression	0.825	0.810	0.84	0.90	0.80
Random forest classifier	0.988	0.855	0.86	0.93	0.89

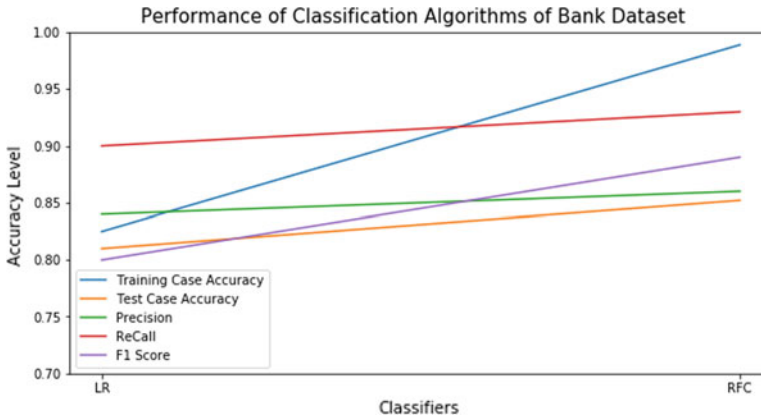


Fig. 4 Performance analysis of classification algorithm

8 Conclusion

In this paper, we used different data mining algorithm to build a model for bank on which basis bank can decide any customer as a bad customer or good customer. Then banking system can decide their future decision regarding on that particular customer. Good customer means whose credential is good and vice versa. Data mining algorithm predicts that customer is valuable or profitable asset on regarding different record as such their age, occupation, marital status, education, credit is in default, balance, loan type (as personal), last contact duration, put come (outcome of the previous campaign).

In this paper, data mining algorithm is used to create a model using Python languages with their different packages to calculate accuracy in which random forest algorithm result is best in comparison to logistic regression.

References

1. Jin, Y., Zhu, Y.: A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending. School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China
2. Hsu, C.F., Hung, H.F.: Classification methods of credit rating—a comparative analysis on SVM, MDA and RST. In: 2009 International Conference on Computational Intelligence and Software Engineering, pp. 1–4. (2009)
3. Turkson, R.E, Baagyere, E.Y., Wenya, G.E.: A machine learning approach for predicting bank credit worthiness. In: 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), pp. 1–7. Lodz (2016). <https://doi.org/10.1109/ICAIPR.2016.7585216>
4. Jafarpour, H., Sheikholeslami Garvandani, H.: New model of customer relationship management in Iranian banks. *Icbme.Yasar.Edu.Tr*, pp. 1–12, (2012)
5. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.* **62**, 22–31 (2014)

6. Li, J., Wei, H., Hao, W.: Weight-selected attribute bagging for credit scoring. *Math. Probl. Eng.* **2013**, (2013)
7. Day, E.A., Hendershott, P.H.: Household Demand for Policy Loans. *J. Risk Insur.* **44**, 411–423 (1977)

The Use of the WhatsApp Platform as an Educational Tool During the Confinement Period of the Outbreak of Covid-19



Nundini Devi Akaloo

Abstract The purpose of this research paper is to gather the views on the use of the WhatsApp platform as an educational tool during the confinement period of Covid-19. Both quantitative and qualitative methods have been used for data collection from the year one department of management students of the Université des Mascareignes (UdM). All the students owned smartphones and WhatsApp was already part of the e-routines for communication for both learners and lecturer. It was also found that during the confinement period, this platform with its numerous features helped to create a positive and cordial climate among the students and lecturer, thus creating a sense of belonging through the WhatsApp group. This platform provided the required support towards the successful completion of the set learning objectives. The survey revealed that the majority of students would like to continue using WhatsApp in their educational process post-Covid-19 period too.

Keywords WhatsApp · Communication · Online platform · Higher education · Educational tool · Covid-19 · Constructivism

1 Introduction

With the worldwide outbreak of the pandemic Covid-19, Mauritius woke up on March 20, 2020, shaken and confused in sudden national confinement. Our whole world was immobilized. As academics we asked ourselves how were we to complete our classwork and assessments. Education cannot stop, it is a fundamental human right. How would we go about that? So far, under normal conditions the contact hours comprised of on campus face-to-face interactions with the students in classes or in the university's lecture theatre. Additionally, handouts would be sent (in pdf format or as a word document, if not already given in class), references and power point presentations by email as well as communicating with the class representatives on smartphones by using WhatsApp. These have been the standard practice.

N. D. Akaloo (✉)
Université Des Mascareignes, Beau Plan, Pamplemousses, Mauritius
e-mail: nakaloo@udm.ac.mu

The objective being the continuation of classes online, it was essential to reorganize oneself, recalibrate the course material and to find the instruments effective to the situation. For the first week of lockdown, handouts and assignments were sent to the classes through their group email. How to ensure that the handouts and assignments were read and properly understood? Difficult to answer as the responses received from students by email were sparse. Email lacked the feedback that we received from the interpersonal communication in classrooms. It seemed that there was a break in the communication channel. How to be able to interact with them?

During this confinement period then, communicating and meeting the learning objectives of the 109 students of year one from the Department of Management for the next six weeks were the researcher's prime objective. After a short period of uncertainty, resiliently adjustments and adaptations were made to our pedagogy to match the orders of the day which were to complete the learning outcomes and assessments online.

Without realizing, Covid-19 enthused us to the frontiers of e-learning that is learning through the use of electronic technologies, being an aspect of Education 4.0. Education 4.0 [13] is the future of education and learning that is a result of the integration of technology which aims to meet the aims of education. Online technologies with their power to immediate connectivity were certainly the way forward. According to the latest report by UNESCO (2020), more than 1.5 billion students from 165 countries have been forced to shift from face-to-face education to online learning. Clearly, the pandemic acted as a key catalyst for redefining the educational shifting from our traditional ways of teaching towards the e-learning.

The aim of this study, therefore, is to explore the use of the WhatsApp platform as an educational tool during the confinement period of the year one students of the Department of Management.

2 Background

In such a period, therefore, it was important to assess the situation and see what was immediately available for use. Mauritius Telecom [10] has placed Mauritius as the eighth most fibred country in the world in terms of connectivity and the first in Africa on connecting people to the increasingly digital world with the deployment of fibre to homes. The network redesign and increased Internet speed of up to 100 Mbps resulted in an increase in the use of ICT services. The service provider has further extended its connectivity through the modernisation of its mobile access network to provide island-wide 4G LTE coverage and ultra-fast mobile broadband access. The mobile subscribers benefit of fibre-like wireless broadband on-the-go experience. Given this technological infrastructure in place, it meant for us that within these unique circumstances, we have the opportunity to have access to all our students wherever they are in the country to meet our teaching and learning objectives. It is to be noted too that some of our foreign students had left for their country and we did have access to them.

Wong et al. [19] posit that young adults interact widely using social media where they share, generate and exchange data in virtual communities and networks. Sham et al. [15] assert that with these new communication facilities, young adults especially are spending an increased amount of time on social media sites. 75% of young adults claim that they it would be difficult for them to be without social media even for a day (Jones 2015). As per the data collected, the UdM students are no different. It was found that the amount of time they spent online was as follows: 35.9% of students spend 3 h, 15.4% 30 min per day, 12.8% were online a whole day, another 12.8% spent 2 h per day, 10.3% spent 1 h per day, 10.3% more 4 h per day and 2.6% spent 5 h per day. Their high level of online presence would but further engage students in completing the learning outcomes.

UdM students are millennials, they are also known as digital natives. They are children born in the 1980s to mid-2000s. They are members of “[...] the first generation to grow up in a world where the Internet was always present [and who] socializes more online, downloads more entertainment media, and consults the Web for a wider range of purposes” (Herring 2008). Case [7] added that these millennials, have grown up in a culture of instant messaging, always being connected and “[...] everything comes to them, and that they become very excited about it and very addicted to it” [7].

As concerns the positive situational context from both the national technological infrastructure and the millennials social network culture, the use of instruments has been achievable. However, which ones of the instruments to choose from. A series of tools have since been used in parallel as a support to get our message effectively across during this period of confinement: email, SMS, Microsoft Team. It is, however, the academic-student collaborative experience of the use of WhatsApp on their smartphones that will be shared in this paper.

3 Rationale of Choosing the Right Instrument

3.1 The Smart Phone

Alqahtani et al. [2] advanced that the extensive use smartphones by university students in many parts of the world shows that smartphones have become an important element of their e-routines. The smart phone was chosen as instrument. Smart phones have become increasingly widespread and [9] have become omnipresent in the world today with the power of portable computing in the hands of everyone. Case [7] calls the relationship we have with our smart phones ambient intimacy, which means that “we are not connected to everybody all the time, but at any time we can connect to anyone we want” [7]. Despite that we tend to carry these online selves with us all the time, and worry about them all the time. We all use the Internet and our smart phone all the time, and we can call or text anyone around the world in an

instant. It was logical, therefore, that the smartphone was chosen as the instrument of communication.

It has been further advanced by Case [7] that the use of smartphones can be said to represent one's online self. Nyrup [11] adds that these online selves have become a sort of safety net to the world and can, therefore, be linked to Maslow's hierarchy of needs. Studies he claims found that there are three things that the modern man carries around with him: keys, money and the phone. Keys provide access to shelter, money buys food and provide sustenance. The phone has been used for calling or texting others, or connecting to emergency services in case of emergency. This original use of the phone would help justify the safety needs of the moderns.

Nyrup [11] posits that with the dawn of the smart phone, it becomes this device which we carry around, and which holds an enormous amount of information, and students are no exception. It becomes an object which can keep us from experiencing what is called the Fear of Missing out, or FoMo, which "[...] is characterized by the desire to stay continually connected to what others are doing" (Przybylski et al. 2013, p. 1841) as cited by Nyrup [11]. The desire and the possibility of always being connected, and subconsciously knowing that you might miss a chat for instance, can lead a person to experience FoMo.

3.2 *WhatsApp Platform*

The use of social network has become popular in everyday communication. The most recent popular social network is WhatsApp application. WhatsApp Messenger is a proprietary, cross-platform instant messaging application for smart phones. Its instant messaging feature for academic purposes gives students opportunities to interact together and share knowledge Chan (2005) as cited by Amry [3]. WhatsApp contains a variety of functions, such as text messages, audio files, video files, attached images and links to web addresses [4]. WhatsApp allows its users to use their Internet connection to send messages to each other. WhatsApp is like a chat program for mobile phones. Unlimited messages can easily be sent to WhatsApp friends, freely. In higher education, WhatsApp is used for the enhancement of discussions and sharing information among students and their lecturers [14]. They further added that this social network has opened up new opportunities of interaction and collaboration between teachers and learners.

Further, Gon and Rawekar [8] state that WhatsApp has become a new and convenient platform for teaching and teaching with which teachers can be present anywhere and at any time. According to Rovai (2002) as cited by Al-Omary et al. [1], students today need cooperative and collaborative learning activities to construct and share knowledge. A reason behind the continuously increasing popularity among teenagers and young adults is WhatsApp features which includes group chats and location sharing Webwise [17].

The use of WhatsApp does not need special arrangement. Indeed, all the students already own the hardware that is smartphones and they have already downloaded the

WhatsApp application software for free. WhatsApp is already part of the e-routines for communication for both learners and the academic. Therefore, the work would proceed based on the knowledge of WhatsApp, already held by all participants.

3.3 Modus Operandi Interpersonal Communication in an Online Classroom Context

During the confinement period therefore, it was decided that the smart phones would be used and WhatsApp as an online platform of communication. There were 109 students, the interaction between each one of them would be challenging. Therefore, yet again, I used the resources that was available. Each class had two class representatives, and there were three classes. It was, therefore, these six class representatives who would be my front-liners. To ease the communication with the classes, a group email was created on Monday 23 March 2020, called “2020 Yr1 Sem2 Leaders” comprising of the six class representatives. It was them who were communicating to their class at the beginning and reporting back to me. They were also monitoring that the work was getting done. After, I started communicating with all the 109 students through three Group WhatsApp one for each class: HRM1.1 WE, HRM1.2 WE and MKTG1 WE. Therefore, I interacted with the class representatives to get the message across. And I had direct contact with all 109 students. They were able to ask questions, initiate discussions on a topic to me and their friends.

4 Research Methodology

This research is grounded in the tenets of the constructivist learning theory. According to constructivism, learners construct their knowledge based on their based on their own prior and current understanding [6]. Learners are active constructors of knowledge and not as unreceptive receivers. Vygotsky [16] contends the learning environment fosters interaction between learners and their classmates as well as their instructor. The constructivists state that a learning environment should be created by the instructor where he/she is a facilitator rather than a source of knowledge. The burden of learning, therefore, should be on the students. Vygotsky [16] further upholds that learners should be given tools to support them build their knowledge. Amry (2014) posited that online asynchronous or synchronous discussion amongst students on social networks vis the WhatsApp platform have a cognitive added value which provides them with the opportunity to construct and share knowledge. The use of WhatsApp as a tool has provided students a learning environment during the confinement period, where they have been able to actively take part in the learning process. This has also provided learners with the opportunity to communicate and

collaborate among each other as a team within an educational context to get the work assigned to them done.

4.1 Analysis Techniques

Both quantitative and qualitative methods have been used for data collection. Quantitative data were collected using a questionnaire instrument to measure the students' use of WhatsApp during the confinement period of Covid-19 for their online education. The questionnaire was divided into four sections. Section A was questions referring to the use of smart phones and WhatsApp pre-Covid-19 period; Section B on the use of WhatsApp as an educational tool during the Covid-19 at the UdM; Section C on the use of WhatsApp as an educational tool after the Covid-19 at the UdM; and Section D on the demographic details of the participants. A 5-point Likert scale of "strongly disagree" to "strongly agree" was utilized to collect students' responses in Sections B and C. The data were sent and retrieved through excel using WhatsApp platform. The focus group for the qualitative data has been chosen to complement and further explain statistical information obtained from the quantitative research. The focus group as claimed by Obiozor [12] would usually provide a revealing source of data. Furthermore, selecting this qualitative study for the WhatsApp platform would take advantage of its conducive and interactive e-social environment provided to participants. The creation of transcripts was not needed as all communications which was in writing were available on the screen. The focus group consisted of the six class representatives. As posited by Breen [5], these students gave a deeper insight of the phenomenon under study by gaining the students' perceptions of the use of WhatsApp as an educational tool during the Covid-19 confinement period. Williams and Katz [18] posited that researchers wishing to enhance the results from interview or survey questions might gain more information from asking the same questions within a focus group setting.

4.2 Participants

The participants were from the year one students reading, respectively, for a Bachelor's Degree with Honors in Human Resource Management, and Marketing from the Department of Management of the Faculty of Business Management at the Université des Mascareignes. There are two classes of Human Resource Management (HRM) and one class of Marketing students. A sample of 39 students was taken out of 109 based on a voluntary basis representing 36% of the population. Table 1 shows the demographic characteristics of the participants.

Table 1 Profile of participants

Details		Frequency	Percentage
Gender	Male	11	26.0
	Female	28	72.0
Age	18–21	35	89.7
	22–25	3	7.7
	26–29	1	2.6
Country of origin	Mauritius	30	76.9
	Madagascar	7	17.9
	Congo	2	5.1
Specialisation	HRM	23	60.5
	Marketing	15	39.5

5 Findings and Discussions

The following findings were evidenced from the research. They are elaborated in the sub-sections below; types of communication, usefulness of the WhatsApp platform and respondents’ experience of using WhatsApp during the confinement.

5.1 Types of Communication

WhatsApp was used as a communication channel to interact with the various stakeholders involved. In this study, it was found that WhatsApp was used to communicate with the lecturer, class representatives and students, their group members, about course materials (Fig. 1), assessments (Fig. 2), feedback from the classes (Fig. 3). It was also found that this platform helped create a positive and cordial climate among the students and lecturer, thus creating a sense of belonging through the WhatsApp group (Fig. 4). It also allowed students to interact among themselves so as to share information and work as a group to respect deadlines which was monitored by the class representatives. The use of this platform allowed the sharing of documents with immediate accessibility relating to the learning outcomes.

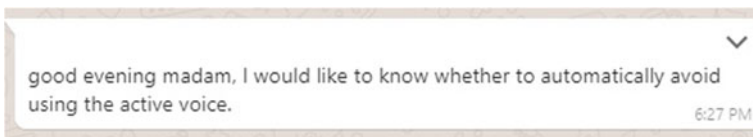


Fig. 1 Question of student on course material

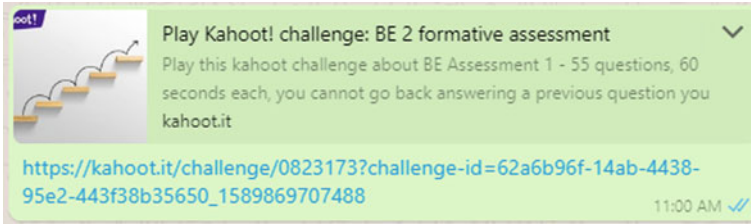


Fig. 2 Lecturer sharing assessment link on Kahoot

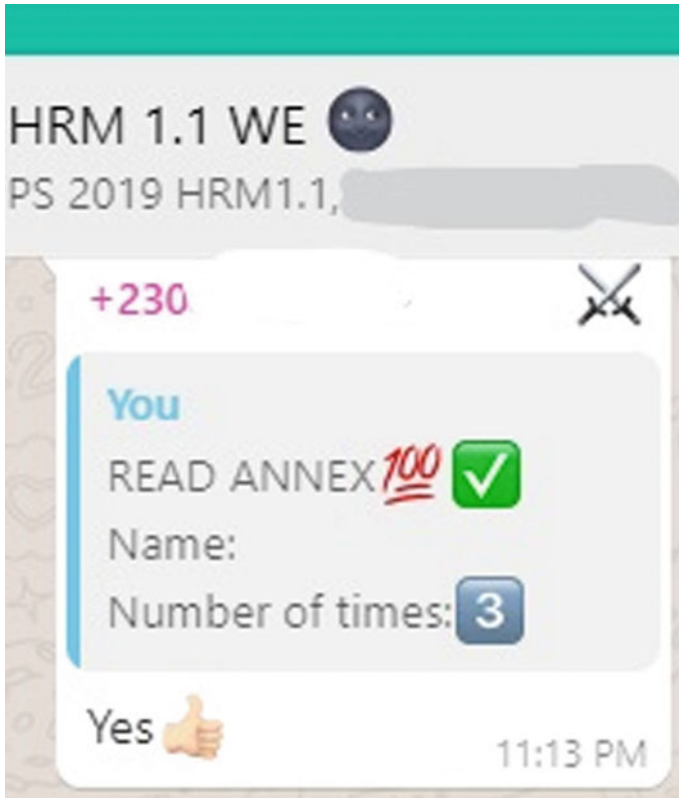


Fig. 3 Class representative reporting on work done by students

The following are a sample of some of the illustrations of the communication mentioned above. The names and phone numbers of the participants have been blotted out.

It is noted too that communication on WhatsApp being instant, students expected immediate response from the lecturer. It is to be added that messages exchanged were not time barred; see Fig. 3 where time of communication was 11.13 pm. Many

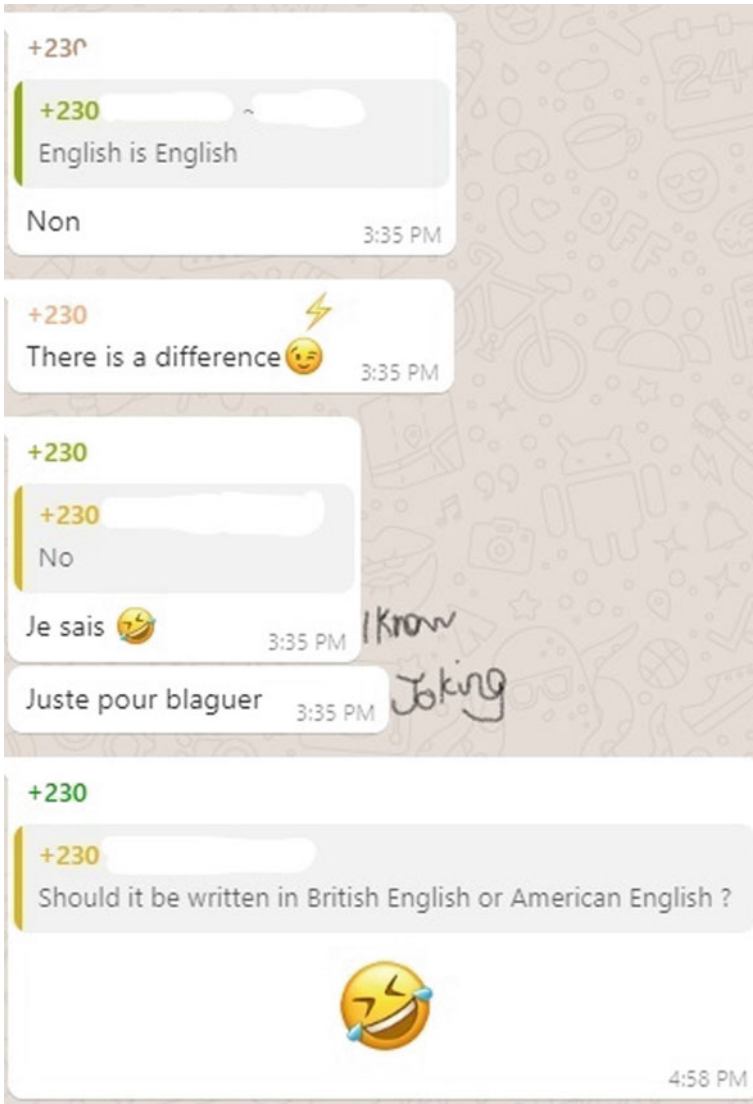


Fig. 4 Interaction among students—use of emoji

a times queries were answered immediately. There was the challenge from the lecturer to be constantly available, though challenging it did not prove that difficult as we were in a period of confinement and we were left with six weeks to complete our learning outcomes. Answering to all the queries was time-consuming for the lecturer. It required constant follow-up and more hours of work. However, through these different exchanges both parties got to know each other better which established a relationship of dependence. Also, through the group WhatsApp each and every

student could be helped individually in a tailored manner; see Fig. 1. The students were requested to post their query on the group platform so that all could benefit, still two of them were sending individual messages which was also answered individually. The question and answer was then put on the platform without mentioning the name of that person.

The communication among students was cordial, and at no time were there any rude comments or remarks among each other in the group WhatsApp that was used for the class. The students made use of emoji to express their feelings. The emoji added an emotional level to the text on the digital platform. It expresses in a nonverbal way what would be expressed in real life. A question asked was whether the online social presence of the lecturer influenced the behaviour of students. From the response received from the respondents, it was found that 94.9% responded positively to the query because the lecturer could oversee all communications and 5.1% replied negatively. The reasons given to answering yes were that since *“the lecturer could oversee all communications, there was a more disciplinary attitude amongst all of us which lead to appropriate way of conversation and therefore resulting in behaving well”*. The word respect recurred many times towards the lecturer and their fellow friends, *“I become more conscious about messages that I send because as a student I respect the teacher and the classmates”*. Another reaction was *“I think that we are responsible and respectful enough to behave well towards our classmates even if there is no lecturer on the WhatsApp group to oversee the communication”*. Another response had to do with the online self or digital self that each create for themselves. The feedback was *“hence I don’t want to leave any negative impact or impression while communicating for a purpose with the classmates and teacher”*. Those who answered negatively said that with or without the presence of the lecturer they made sure they behaved well with their classmates. They further added *“we are a team, we should cooperate, and we help each other”*. It is to be observed that those who responded negatively were the class representatives. They were given the leadership of their class.

5.2 Usefulness of the WhatsApp Platform

The help of the class representatives was essential to help manage and control 109 students. They have been a dedicated group of six persons committed to achieve our learning objectives. They have been the front liners and the enablers of the module through WhatsApp. Each class had two class representatives which made it easier to manage an average of 17 persons each out of a total of 103 students. The feedback was requested from them on their views of communicating through WhatsApp for the class during the confinement period were both mostly positive:

“Well, for me, I found WhatsApp very helpful in communicating your messages to the whole class of 38 students. And particularly when we created the group for the assignment so that you could answer any queries, it enabled each one of us to ask you for help which was great....” ... *“I think through WhatsApp we had the*

Table 2 Use of WhatsApp as an educational tool at the Université des Mascareignes during Covid-19

Statements	SA	A	N	D	SD
It was easy for me to use WhatsApp as a support to my learning	55.3	26.3	15.8	0.0	2.6
I felt comfortable using WhatsApp	47.4	28.9	18.4	5.3	0.0
I received quick feedback from the instant messaging feature of WhatsApp	39.5	47.4	13.2	0.0	0.0
The use of WhatsApp did not involve additional cost, only connectivity	52.6	34.2	10.5	2.6	0.0
WhatsApp has helped me complete the modules	26.3	50.0	23.7	0.0	0.0

opportunity to be connected easily with the lecturers and students where information has been distributed very quickly and received at the other end without any problem.” ... “WhatsApp is a very mode of communication just click the Wi-Fi button to get notified besides we can also see who has acknowledged the message and who has not yet seen. A good platform to get instant response from students”.

Another feedback received conveyed its positive use but highlighted a bothersome aspect:

“I think WhatsApp is a good communication tool (easy effortless and accessible) we all took great advantage of this for our assignments and question answers sessions we had. For all, Thank you! however, a disadvantage of receiving messages at odd hours (23 h...) etc..... otherwise for my part I agree that it’s been very useful”.

The fact that one could receive messages anytime could have been a source of uneasiness for some while an advantage to others of always being kept informed. However, one also has the choice of muting their phone if they do not want to be disturbed. Overall, it can be seen that the respondents positively commented on the use of the WhatsApp platform.

With regard to the input from the respondents to the ensuing statements, the following are the responses.

As per Table 2, the overall percentage of students who agreed and strongly agreed on all statements was 82.1%. As a result, it is obvious that students felt that using WhatsApp as an educational tool during the confinement was indeed a support to their learning. They were comfortable using WhatsApp. They received quick feedback, a student even commented *“she sees and responds to everything”*. WhatsApp required no additional cost, and the platform helped them to complete the modules. They added that the lecturer clarified and explained the questions they had *“thus enabling the students to better understand”*. A total of 15.6% remained neutral in their responses. An overall minority of 2% who disagreed and strongly disagreed said that they were not comfortable using WhatsApp *“because there were too much distractions while using it. Sometimes I could easily go on doing other things on my phone instead of using WhatsApp during my learning sessions”*; *“Students can easily get distracted due to multiple chats”*. Another thought that there was a *“lack of privacy”* on the Group WhatsApp which explains why few questions were directed to the lecturer

directly. The following is a table with the comments received from respondents on their experience of using WhatsApp.

5.3 Future Use of WhatsApp as an Educational Platform Post-Covid-19

As per Table 2, the overall percentage of students who agreed and strongly agreed on the statements was 60.3, 33.3% were neutral and 6.4% disagreed and strongly disagreed. This would imply that a majority would agree to continue using WhatsApp as a tool to help complete their learning objectives. Among those 6.4%, is a foreign student who says that she uses this platform to communicate with her family, these messages and calls disturbs her and prevents her from concentrating.

6 Conclusion and Future Works

It is a well-accepted fact today that the emerging technologies have played a very pivotal role in mitigating the negative impact of Covid-19 on specifically academia worldwide. Undoubtedly, the social networking platform like WhatsApp has proved to be the leader among all of them, by the virtue of its simple but robust framework. Hence, the findings of the study become very relevant in today's context.

Key findings:

- The use of WhatsApp has been successful during the confinement period as a support to complete the modules and meet learning objectives. WhatsApp already formed part of the students' e-routine which concurs with the findings of [2]. The features of instant messaging and traceability helped bond a cordial collaborative community of learners. The online social presence of the lecturer positively influenced and motivated the digital natives. The students always wanted to portray a positive digital image of themselves.
- One of the key factors of achievement, using the features of the WhatsApp platform as an educational tool has been the involvement of the class representatives acting as front liners, chasing, monitoring and controlling that the work from their class was getting done.
- A vast majority of students agreed to continue using WhatsApp in their post-Covid-19, educational process.

In this era of Industry 4.0, the academic world is looking forward to the game changing technologies, which will contribute to the human aspirations to excel beyond boundaries, thereby challenging the contemporary practices and replacing it with more flexible, cost-effective, reliable and high-quality practices wherever possible.

The limitations of this study are that it was based on the experience of students from only one department at the university.

An extended study could be carried out in the other faculties of the Université des Mascareignes and other universities too on the use of the WhatsApp platform as an educational tool during and post-Covid-19.

References

1. Al-Omary, A., El-Medany, W.M., Isa, K.J.E.: The impact of SNS in higher education: a case study of using WhatsApp in the University of Bahrain. In: Proceedings—2015 5th International Conference on e-Learning, ECONF 2015, pp. 296–300. Institute of Electrical and Electronics Engineers Inc. (2016). <https://doi.org/10.1109/ECONF.2015.72>
2. Alqahtani, S.M.M., et al.: WhatsApp: an online platform for University-level English language education. *Arab World Engl. J.* **9**(4), 108–121 (2018). <https://doi.org/10.24093/awej/vol9no4.7>
3. Amry, A.B.: The impact of WhatsApp mobile social learning on the achievement and attitudes of female students compared with face to face learning in the classroom. *Eur. Sci. J.* **10**(22), 116–136 (2014). Available at: <https://eujournal.org/index.php/esj/article/view/3909>
4. Bouhnik, D., Dshen, M.: Whatsapp goes to school: mobile instant messaging between teachers and students. *J. Inf. Technol. Educ. Res.* **13**, 217–231 (2014). <https://doi.org/10.28945/2051>
5. Breen R.L.: A practical guide to focus-group research. *J. Geogr. Higher Educ.* **30**(3), (2006). Published online: 22 Jan 2007, retrieved 22 Jun 2020
6. Bruner, J.S.: *Toward a theory of instruction*. The Belknap Press of Harvard University Press, Cambridge, MA (1966)
7. Case Amber, TEDWomen: We are all cyborgs now. Retrieved 26 May 2020, from https://www.ted.com/talks/amber_case_we_are_all_cyborgs_now/transcript (2010)
8. Gon, S., Rawekar, A.: Effectivity of e-learning through WhatsApp as a teaching learning tool. *MVP J. Med. Sci.* **4**(1), 19–25 (2017)
9. Goundar, S.: What is the potential impact of using mobile devices in education? By. *SIG Glob. Dev.* **4**(December), 1–30 (2011)
10. Mauritius Telecom: Retrieved May 28, 2020 from website: <https://telecom.mu/aboutus/our-net-work/> (2020)
11. Nyrup (2016) Smart phones as an embodied technology innovation. In: T. E. and Dits, S. O. 'IT University of Copenhagen
12. Obiozor, W.: Applying focus groups in educational research in Africa. *Rev Educ.* (2008). Available at: https://works.bepress.com/drwilliams_obiozor/3/
13. Pangandaman, H.K., et al.: Philippine higher education vis-à-vis education 4.0: a scoping review. *Int. J. Adv. Res. Publ.* **3**(3), 65–69 (2019). Available at: www.ijarp.org
14. Popescu, M.: The impact of SNS in higher education, pp. 296–300. (2013). <https://doi.org/10.1109/ECONF.2015.72>
15. Sham, M., Ali, S., Kootbodien, A.: The effectiveness of WhatsApp as an interpersonal communication medium among Abu Dhabi University students. *Int. J. Media J. Mass Commun.* **3**(1), 11–19 (2017). <https://doi.org/10.20431/2454-9479.0301002>
16. Vygotsky, L.S.: *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA (1978)
17. Webwise: Explainer: What is WhatsApp? Retrieved May 27 2020, from <https://www.webwise.ie/parents/explainer-whatsapp/> (2016)
18. Williams, A., Katz, L.: The use of focus group methodology in education: some theoretical and practical considerations. *Int. Electron. J. Leadersh. Learn.* **5**(3), (2001)
19. Wong Charlene, A., Merchant Raina, M., Moreno Megan, A.: Using social media to engage adolescents and young adults with their health Published in final edited form as: *Healthc (Amst)*. **2**(4), 220–224 (2014). <https://doi.org/10.1016/j.hjdsi.2014.10.005>. PMID: PMC4433153

Author Index

A

Abdel Hassan, Naafeera B. R., 735
Aggarwal, Naveen, 497
Agrawal, Vinam, 19
Ahmed, Eshtiak, 777
Ahuja, Bharti, 201
Ajufo, Chiamaka Ann Marie, 831
Akaloo, Nundini Devi, 899
Akter, Nasrin, 233
AL-Abri, Hanan Hamdan, 861
Alaka, Benard, 427
Alatrash, Rawaa, 123
Amic, Seeven, 469
Anzar, Tansif, 177
Arefin, Mohammad Shamsul, 255
Armoogum, Sandhya, 337
Armoogum, Sheeba, 301, 571
Arora, Indu, 515
Aumeer, Wafiik, 371
Awasthi, Anjaneya, 143

B

Babber, Karuna, 43
Bajpai, Avirag, 683
Balaji, S., 187, 485
Banerjee, Arko, 313
Behera, Pratap Kumar, 557
Bekaroo, Girish, 637, 831
Betchoo, Nirmal Kumar, 839
Bhagbut, Gulshansingh, 69
Bheda, Hitesh A., 585
Bheekharri, Normada Devi, 871, 881
Biswal, Ankita, 323
Biswas, Anupam, 269
Boncoeur, Pascal, 871

Boolakee, Roshni Vidya S., 337
Brunel, Tsiatsipy Durand, 763

C

Chaudhary, Aishwarya, 85
Chavan, Pranay, 653
Chawla, Raghav, 19
Choksi, Darshan B., 585
Chowdhury, Atiqul Islam, 777
Chowdhury, Shahnaj, 777
Chuttur, Yasser, 55, 95, 109
Cowlessur, Sanjeev K., 323, 507
Cunden, Tyagaraja S. M., 719, 735

D

Das, Abhijit, 269
Dash, Rajashree, 221
Dash, Rasmita, 221
Das Mohapatra, Subhashish, 607
Das, Sanchali, 163
Debbarma, Swapan, 163
Dookhitram, Kumar, 211
Dua, Mohit, 19

E

Ezaldeen, Hadi, 123

F

Fakra, Ali Hamada Damien, 671
Foogooa, Ravi, 337

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

C. R. Panigrahi et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 1299,

<https://doi.org/10.1007/978-981-33-4299-6>

G

Gajjar, Nidhi, 463
 Gangopadhyay, Sugata, 557
 Garg, Rajni, 515
 Garhnayak, Dibyasha, 221
 Gatina, Jean Claude, 671
 Geerish, Suddul, 293
 Ghosh, S., 387
 Goel, Chahat, 85
 Goel, Noopur, 135, 143, 889
 Gupta, Anu, 515
 Gupta, Subhash Chandra, 135

H

Halfani, Hadija Ramadhani, 637
 Hashmi, Md Farukh, 201
 Hosenkhan, Mohammad Reza, 599
 Hosen, Md Imran, 777
 Hossain, Eftekhari, 281
 Hossain, Tonmoy, 233
 Hossen, Md. Shakhawath, 437
 Hossen, Md. Sharif, 437
 Hota, Nagarjuna, 661

I

Indu, S., 85
 Islam, Ashraf, 777
 Iyer, Mohit, 3

J

Jain, Akshay, 653
 Jenifer Suriya, L. J., 187
 Jerlin, A., 187
 Jhugroo, M. Saheer, 707
 Jindal, Muskan, 463

K

Kaje, Smitha Bhat, 749
 Kasamani, Bernard Shibwabo, 427
 Kataully, M. Shahil, 707
 Kaur, Chhinder, 153
 Khedo, Kavi Kumar, 351
 Khoodeeram, Rajeev, 371
 Kotasthane, Aditya, 527
 Kumarasamy, Kunaraj, 187
 Kumar, Basant, 861
 Kumar, Nitin, 29

L

Latchoomun, Lekhram Singh, 763

Lockmun-Bissessur, Vimi Neeroo, 791
 Lollchund, Michel R., 719, 735

M

Mahadooda, Haydar, 95
 Maitra, Shithi, 233
 Majumdar, Sudipta, 85
 Mallick, Amiya Kumar, 387
 Mani, Joseph, 861
 Mantry, Unnati, 221
 Maradiya, Pratik, 653
 Maria Wensch, S., 187
 Mehra, Ritika, 3
 Mishra, Ashutosh, 221
 Misra, Rachita, 123
 Misra, Subhas Chandra, 683
 Misra, Sudip, 527
 Mohamudally, Nawaz, 301, 571
 Mohanty, Rajani Kanta, 507
 Mongia, Vikas, 413
 Mooznah Auleear Owodally, A., 805
 Moshikul Hoque, Mohammed, 281
 Mungloo-Dilmohamud, Zahra, 69
 Mungur, Utam Avinash Einstein, 351

N

Nagarajan, Pranav, 401
 Nagowah, Soukshme D., 707, 817
 Nanda, Sarmistha, 323
 Nayak, Jagadish, 749
 Nayak, Suvendu Chandan, 607, 621
 Nohur, Deojeet, 507
 Nundlall, Chitra, 817

P

Pal, Prashnatita, 387
 Panda, Sourav, 221
 Panigrahi, Chhabi Rani, 313, 323, 607, 621
 Parida, Prashanta Kumar, 449
 Parida, Sasmita, 607, 621
 Patel, Nehal, 463
 Pati, Bibudhendu, 313, 323, 607, 621
 Patra, Anannya, 485
 Pattanayak, Binod Kumar, 507, 599, 661
 Peeroo, Swaleha, 791, 805
 Phavish, Babajee, 293
 Pirbhai-Jetha, Neelam, 851, 871
 Pokhun, Leevesh, 55
 Pooloo, Nabeelah, 371
 Poray, Jayanta, 387
 Priyadarshini, Rojalina, 123

Pujari, Arun K., [313](#)

R

Rachadi, Mohamed Nasroudine Mohamed, [671](#)

Rahman, Mir Lutfur, [543](#)

Rahman, Mohammad Masudur, [777](#)

Rajoo, Ashley, [351](#)

Ramsawock, Gianeshwar, [469](#)

Randriantsoa, Ando Ny Aina, [671](#)

Ranjaranimaro, Manitra Pierrot, [671](#)

Rathod, Kiran, [653](#)

Rautray, Rasmita, [221](#)

Ravi, Foogooa, [293](#)

Rawoteea, Nandishta, [109](#)

Richa, [363](#)

Robert Rajkumar, S., [187](#)

Roy, Arijit, [527](#)

S

Sahana, Bikash Chandra, [387](#)

Saluja, Kshitij, [497](#)

Salunke, Sharad, [201](#)

Sandeep Kumar, V., [695](#)

Sandhip Laldjee, Shivianee, [831](#)

Sandhya, Armoogum, [293](#)

Sarker, Pranta, [543](#)

Satpathy, Sambit, [163](#)

Savy, David, [791](#)

Sethi, Pawandeep Singh, [19](#)

Shafiul Alam Forhad, Md., [255](#)

Shafiul Alam, Mohammad, [233](#)

Shambhavi, B. R., [485](#)

Sharif, Omar, [281](#)

Sharma, Anand, [153](#), [413](#)

Shivhare, Shiv Naresh, [29](#)

Sindhu, K., [485](#)

Singh, Durgesh Kumar, [889](#)

Singh, Jyoti, [363](#)

Soyjaudah, K. M. Sunjiv, [469](#)

Suddul, Geerish, [211](#), [337](#)

T

Tameswar, Kajal, [211](#)

Thaker, Chirag S., [585](#)

Thyagarajan, Jayavignesh, [401](#)

V

Venkatadri, M., [201](#)

Vivek, V., [485](#)

W

Wadhwa, Shruti, [43](#)

Wong Suk Hee, Ah-Kwet Rémi, [637](#)

Z

Zahan Mithila, Afrina, [233](#)