

Neha Sharma
Amlan Chakrabarti
Valentina Emilia Balas *Editors*

Data Management, Analytics and Innovation

Proceedings of ICDMAI 2019, Volume 1

Advances in Intelligent Systems and Computing

Volume 1042

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.

**** Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink ****

More information about this series at <http://www.springer.com/series/11156>

Neha Sharma · Amlan Chakrabarti ·
Valentina Emilia Balas
Editors

Data Management, Analytics and Innovation

Proceedings of ICDMAI 2019, Volume 1

 Springer

Editors

Neha Sharma
Society for Data Science
Pune, Maharashtra, India

Valentina Emilia Balas
Department of Automatics
and Applied Software
Aurel Vlaicu University of Arad
Arad, Romania

Amlan Chakrabarti
A.K. Choudhury School of Information
Technology
University of Calcutta
Kolkata, West Bengal, India

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-32-9948-1

ISBN 978-981-32-9949-8 (eBook)

<https://doi.org/10.1007/978-981-32-9949-8>

© Springer Nature Singapore Pte Ltd. 2020

This work is subject to copyright. All rights are reserved 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

These two volumes constitute the proceedings of the International Conference on *Data Management, Analytics and Innovation* (ICDMAI 2019) held from 18 to 20 January 2019. ICDMAI is a flagship conference of Society for Data Science, which is a non-profit professional association established to create a collaborative platform for bringing together technical experts across industry, academia, government laboratories and professional bodies to promote innovation around data science. ICDMAI 2019 envisions its role towards data science and its enhancement through collaboration, innovative methodologies and connections throughout the globe. The conference was hosted by Lincoln University College, Kuala Lumpur, Malaysia, and was supported by the industry leaders like IBM. Other partners for the conference were Wizer and DSMS College of Tourism & Management, West Bengal, India. The conference witnessed participants from 20 countries, 12 industries, 31 international universities and 94 premier Indian universities. Utmost care was taken in each and every facet of the conference, especially regarding the quality of the paper submissions. Out of 418 papers submitted to ICDMAI 2019, only 20% (87 papers) are selected for an oral presentation after a rigorous review process. Besides paper presentation, the conference also showcased workshops, tutorial talks, key-note sessions and plenary talks by the experts of the respective field.

The volumes cover a broad spectrum of computer science, information technology, computational engineering, electronics and telecommunication, electrical, computer application and all the relevant disciplines. The conference papers included in these proceedings are published post-conference and are grouped into the four areas of research such as data management and smart informatics; big data management; artificial intelligence and data analytics; and advances in network technologies. All the four tracks of the conference were very relevant to the current technological advancements and had Best Paper Award in each track. Very stringent selection process was adopted for paper selection, and from plagiarism check to technical chairs' review to double-blind review, every step was religiously followed. We compliment all the authors for submitting high-quality research papers to ICDMAI 2019. We would like to acknowledge all the authors for their contributions and also the efforts taken by reviewers and session chairs of the conference,

without whom it would have been difficult to select these papers. We appreciate the unconditional support from the members of the National and International Program Committee. It was really interesting to hear the participants of the conference highlight the new areas and the resulting challenges as well as opportunities. This conference has served as a vehicle for a spirited debate and discussion on many challenges that the world faces today.

We especially thank our General Chair, Dr. P. K. Sinha; Vice Chancellor and Director, Dr. S. P. Mukherjee, International Institute of Information Technology, Naya Raipur (IIIT-NR), Chhattisgarh; other eminent personalities like Mr. Eddy Liew, Cloud and Solutions Country Technical Leader at IBM, Malaysia; Kranti Athalye, Sr. Manager in IBM India University Relations; Dr. Juergen Seitz, Head of Business Information Systems Department, Baden-Wuerttemberg Cooperative State University, Heidenheim, Germany; Mr. Aninda Bose, Senior Publishing Editor, Springer India Pvt. Ltd.; Dr. Vincenzo Piuri, IEEE Fellow, University of Milano, Italy; Hanaa Hachimi, National School of Applied Sciences ENSA in Kenitra, Morocco; Amol Dhondse, IBM Senior Solution Architect; Mohd Helmy Abd Wahab, Senior Lecturer and Former Head of Multimedia Engineering Lab at the Department of Computer Engineering, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM); Anand Nayyar, Duy Tan University, Vietnam; and many more who were associated with ICDMAI 2019. Besides, there was CSI-Startup and Entrepreneurship Award to felicitate budding job creators.

Our special thanks go to Janus Kacprzyk (Editor in Chief, Springer, *Advances in Intelligent Systems and Computing Series*) for the opportunity to organize this guest-edited volume. We are grateful to Springer, especially to Mr. Aninda Bose (Senior Publishing Editor, Springer India Pvt. Ltd.), for the excellent collaboration, patience and help during the evolution of this volume.

We are confident that the volumes will provide the state-of-the-art information to professors, researchers, practitioners and graduate students in the area of data management, analytics and innovation, and all will find this collection of papers inspiring and useful.

Pune, India
Kolkata, India
Arad, Romania

Neha Sharma
Amlan Chakrabarti
Valentina Emilia Balas

Contents

Data Management and Smart Informatics

Empirical Study of Soft Clustering Technique for Determining Click Through Rate in Online Advertising	3
Akshi Kumar, Anand Nayyar, Shubhangi Upasani and Arushi Arora	
ASK Approach: A Pre-migration Approach for Legacy Application Migration to Cloud	15
Sanjeev Kumar Yadav, Akhil Khare and Choudhary Kavita	
A Fuzzy Logic Based Cardiovascular Disease Risk Level Prediction System in Correlation to Diabetes and Smoking	29
Kanak Saxena and Umesh Banodha	
An Integrated Fault Classification Approach for Microgrid System	41
Ruchita Nale, Ruchi Chandrakar and Monalisa Biswal	
Role of Data Analytics in Human Resource Management for Prediction of Attrition Using Job Satisfaction	57
Neerja Aswale and Kavya Mukul	
A Study of Business Performance Management in Special Reference to Automobile Industry	69
Gurinder Singh, Smiti Kashyap, Kanika Singh Tomar and Vikas Garg	
Secure Online Voting System Using Biometric and Blockchain	93
Dipti Pawade, Avani Sakhapara, Aishwarya Badgular, Divya Adepu and Melvita Andrade	
An Approach: Applicability of Existing Heterogeneous Multicore Real-Time Task Scheduling in Commercially Available Heterogeneous Multicore Systems	111
Kalyan Baital and Amlan Chakrabarti	

Analyzing the Detectability of Harmful Postures for Patient with Hip Prosthesis Based on a Single Accelerometer in Mobile Phone	125
Kitti Naonueng, Opas Chutatape and Rong Phoophuangpairroj	
Software Development Process Evolution in Malaysian Companies	139
Rehan Akbar, Asif Riaz Khan and Kiran Adnan	
Automated Scheduling of Hostel Room Allocation Using Genetic Algorithm	151
Rayner Alfred and Hin Fuk Yu	
Evaluation of ASTER TIR Data-Based Lithological Indices in Parts of Madhya Pradesh and Chhattisgarh State, India	161
Himanshu Govil, Subhanil Guha, Prabhat Diwan, Neetu Gill and Anindita Dey	
Analyzing Linear Relationships of LST with NDVI and MNDISI Using Various Resolution Levels of Landsat 8 OLI and TIRS Data	171
Himanshu Govil, Subhanil Guha, Prabhat Diwan, Neetu Gill and Anindita Dey	
Automatic Robot Processing Using Speech Recognition System	185
S. Elavarasi and G. Suseendran	
Banking and FinTech (Financial Technology) Embraced with IoT Device	197
G. Suseendran, E. Chandrasekaran, D. Akila and A. Sasi Kumar	
Big Data Management	
GRNN++: A Parallel and Distributed Version of GRNN Under Apache Spark for Big Data Regression	215
Sk. Kamaruddin and Vadlamani Ravi	
An Entropy-Based Technique for Conferences Ranking	229
Fiaz Majeed and Rana Azhar UI Haq	
MapReduce mRMR: Random Forests-Based Email Spam Classification in Distributed Environment	241
V. Sri Vinitha and D. Karthika Renuka	
The Impact of Sustainable Development Report Disclosure on Tax Planning in Thailand	255
Sathaya Thanjunpong and Thatphong Awirothananon	
Clustering and Labeling Auction Fraud Data	269
Ahmad Alzahrani and Samira Sadaoui	

Big Data Security Challenges and Preventive Solutions	285
Nirmal Kumar Gupta and Mukesh Kumar Rohil	
Role and Challenges of Unstructured Big Data in Healthcare	301
Kiran Adnan, Rehan Akbar, Siak Wang Khor and Adnan Bin Amanat Ali	
Zip Zap Data—A Framework for ‘Personal Data Preservation’	325
K. Arunkumar and A. Devendran	
A Systematic Mapping Study of Cloud Large-Scale Foundation—Big Data, IoT, and Real-Time Analytics	339
Isaac Odun-Ayo, Rowland Goddy-Worlu, Temidayo Abayomi-Zannu and Emanuel Grant	
Studies on Radar Imageries of Thundercloud by Image Processing Technique	365
Sonia Bhattacharya and Himadri Bhattacharyya Chakrabarty	
Artificial Intelligence and Data Analysis	
PURAN: Word Prediction System for Punjabi Language News	383
Gurjot Singh Mahi and Amandeep Verma	
Implementation of hDE-HTS Optimized T2FPID Controller in Solar-Thermal System	401
Binod Shaw, Jyoti Ranjan Nayak and Rajkumar Sahu	
Design of Sigma-Delta Converter Using 65 nm CMOS Technology for Nerves Organization in Brain Machine Interface	413
Anil Kumar Sahu, G. R. Sinha and Sapna Soni	
Performance Comparison of Machine Learning Techniques for Epilepsy Classification and Detection in EEG Signal	425
Rekh Ram Janghel, Archana Verma and Yogesh Kumar Rathore	
Novel Approach for Plant Disease Detection Based on Textural Feature Analysis	439
Varinderjit Kaur and Ashish Oberoi	
Novel Approach for Brain Tumor Detection Based on Naïve Bayes Classification	451
Gurkarandesh Kaur and Ashish Oberoi	
Automatic Classification of Carnatic Music Instruments Using MFCC and LPC	463
Surendra Shetty and Sarika Hegde	

Semiautomated Ontology Learning to Provide Domain-Specific Knowledge Search in Marathi Language	475
Neelam Chandolikar, Pushkar Joglekar, Shivjeet Bhosale, Dipali Peddawad, Rajesh Jalnekar and Swati Shilaskar	
Identifying Influential Users on Social Network: An Insight	489
Ragini Krishna and C. M. Prashanth	
Factex: A Practical Approach to Crime Detection	503
Rachna Jain, Anand Nayyar and Shivam Bachhety	
Analysis of Classification Algorithms for Breast Cancer Prediction	517
S. P. Rajamohana, K. Umamaheswari, K. Karunya and R. Deepika	
Real-Time Footfall Prediction Using Weather Data: A Case on Retail Analytics	529
Garima Makkar	
Normal Pressure Hydrocephalus Detection Using Active Contour Coupled Ensemble Based Classifier	543
Pallavi Saha, Sankhadeep Chatterjee, Santanu Roy and Soumya Sen	
Question–Answer System on Episodic Data Using Recurrent Neural Networks (RNN)	555
Vineet Yadav, Vishnu Bharadwaj, Alok Bhatt and Ayush Rawal	
Convolutd Cosmos: Classifying Galaxy Images Using Deep Learning	569
Diganta Misra, Sachi Nandan Mohanty, Mohit Agarwal and Suneet K. Gupta	
Advances in Network Technologies	
Energy-Based Improved MPR Selection in OLSR Routing Protocol	583
Rachna Jain and Indu Kashyap	
A Novel Approach for Better QoS in Cognitive Radio Ad Hoc Networks Using Cat Optimization	601
Lolita Singh and Nitul Dutta	
(T-ToCODE): A Framework for Trendy Topic Detection and Community Detection for Information Diffusion in Social Network	613
Reena Pagare, Akhil Khare and Shankar Chaudhary	
ns-3 Implementation of Network Mobility Basic Support (NEMO-BS) Protocol for Intelligent Transportation Systems	633
Prasanta Mandal, Manoj Kumar Rana, Punyasha Chatterjee and Arpita Debnath	

Modified DFA Minimization with Artificial Bee Colony Optimization in Vehicular Routing Problem with Time Windows 643
G. Niranjani and K. Umamaheswari

Coverage-Aware Recharge Scheduling Scheme for Wireless Charging Vehicles in the Wireless Rechargeable Sensor Networks 663
Govind P. Gupta and Vrajesh Kumar Chawra

A Transition Model from Web of Things to Speech of Intelligent Things in a Smart Education System 673
Ambrose A. Azeta, Victor I. Azeta, Sanjay Misra and M. Ananya

Intrusion Detection and Prevention Systems: An Updated Review 685
Nureni Ayofe Azeez, Taiwo Mayowa Bada, Sanjay Misra, Adewole Adewumi, Charles Van der Vyver and Ravin Ahuja

Simulation-Based Performance Analysis of Location-Based Opportunistic Routing Protocols in Underwater Sensor Networks Having Communication Voids 697
Sonali John, Varun G. Menon and Anand Nayyar

A Hybrid Optimization Algorithm for Pathfinding in Grid Environment 713
B. Booba, A. Prema and R. Renugadevi

Dynamic Hashtag Interactions and Recommendations: An Implementation Using Apache Spark Streaming and GraphX 723
Sonam Sharma

Author Index 739

About the Editors



Neha Sharma is the Founder Secretary of the Society for Data Science, India. She was the Director of the Zeal Institute of Business Administration, Computer Application & Research, Pune, Maharashtra, India, and Deputy Director, of the Padmashree Dr. D.Y.Patil Institute of Master of Computer Applications, Akurdi, Pune. She completed her PhD at the prestigious Indian Institute of Technology (IIT-ISM), Dhanbad, and she is a Senior IEEE member as well as Execom member of IEEE Pune Section. She has published numerous research papers in respected international journals. She received the “Best PhD Thesis Award” and “Best Paper Presenter at International Conference Award” from the Computer Society of India. Her areas of interest include data mining, database design, analysis and design, artificial intelligence, big data, cloud computing, blockchain and data science.



Amlan Chakrabarti currently the Dean of the Faculty of Engineering and Technology, Professor and Director of the A.K.Choudhury School of Information Technology, University of Calcutta, India. He was a postdoctoral fellow at the School of Engineering, Princeton University, USA from 2011 to 2012. He has published around 130 research papers in refereed journals and conferences, and has been involved in research projects supported by various national funding agencies and international collaborations. He is a senior member of the IEEE and ACM, ACM Distinguished Speaker, Vice President of the Society for Data Science, and Secretary of the

IEEE CEDA India Chapter. He is also the Guest Editor of the Springer Nature Journal on Applied Sciences. His research interests include quantum computing, VLSI design, embedded system design, computer vision and analytics.



Valentina Emilia Balas is currently a Full Professor at the Department of Automatics and Applied Software at the Faculty of Engineering, “Aurel Vlaicu” University of Arad, Romania. She holds a Ph.D. in Applied Electronics and Telecommunications from the Polytechnic University of Timisoara. Dr. Balas is the author of more than 300 research papers in refereed journals and international conferences. Her research interests include intelligent systems, fuzzy control, soft computing, smart sensors, information fusion, modeling and simulation. She is the editor-in-chief of the International Journal of Advanced Intelligence Paradigms (IJAIIP) and the International Journal of Computational Systems Engineering (IJCSysE), and is an editorial board member of several national and international journals.

Data Management and Smart Informatics

Empirical Study of Soft Clustering Technique for Determining Click Through Rate in Online Advertising



Akshi Kumar, Anand Nayyar, Shubhangi Upasani and Arushi Arora

Abstract Online advertising is an industry with the potential for maximum revenue extraction. Displaying the ad which is more likely to be clicked plays a crucial role in generating maximum revenue. A high click through rate (CTR) is an indication that the user finds the ad useful and relevant. For suitable placement of ads online and rich user experience, determining CTR has become imperative. Accurate estimation of CTR helps in placement of advertisements in relevant locations which would result in more profits and return of investment for the advertisers and publishers. This paper presents the application of a soft clustering method namely fuzzy c-means (FCM) clustering for determining if a particular ad would be clicked by the user or not. This is done by classifying the ads in the dataset into broad clusters depending on whether they were actually clicked or not. This way the kind of advertisements that the user is interested in can be found out and subsequently more advertisements of the same kind can be recommended to him, thereby increasing the CTR of the displayed ads. Experimental results show that FCM outperforms k-means clustering (KMC) in determining CTR.

Keywords Advertising · Click Through Rate · Clustering

A. Kumar

Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

e-mail: akshi.kumar@gmail.com

A. Nayyar

Graduate School, Duy Tan University, Da Nang, Viet Nam

e-mail: anandnayyar@duytan.edu.vn

S. Upasani

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

e-mail: shubhangi.upasani@gmail.com

A. Arora (✉)

Department of Electrical Engineering, Delhi Technological University, Delhi, India

e-mail: aroraarushi1997@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_1

1 Introduction

Online advertising has become an important source of revenue for a wide range of businesses. With the tremendous growth in online advertising each year, it has also taken over a major area of research. Most of the revenues generated by widely used search engines as well as prevalent websites come from advertisements. It is therefore important to display relevant advertisements (ads) to the users and avoid the advertisements that are often disliked by them.

Online advertising is more economical than traditional ways of advertising like mass markets and niche media. Internet ads have a wider audience and can be viewed for days and nights altogether, in contrast to ads on television and radios that last for shorter durations and are displayed with limited frequency. Market segmentation is much more effective over the Internet than in any other medium. Thorough study of markets, customer preferences and habits and segmenting consumers into cohesive groups can be done efficiently through online advertising. Online advertising also offers small businesses numerous benefits like robust targeting, consumer insights and more effective return on investment.

Advertisements fall into two broad categories—sponsored search advertisements and contextual advertisements. Sponsored search ads are displayed on the same web pages that show results of search queries entered by users. The core purpose behind sponsored advertising is to enhance the advertiser's brand image as the ads displayed have the same form and qualities as the advertiser's original content. Contextual advertising, on the other hand, uses automated systems that display ads relevant to the user's identity and website's content. Google AdSense is one of the many well-known examples of contextual advertising. Google robots display only those ads that the users find relevant and useful. When a user visits a website, features like ad size, ad placement, etc. are extracted from the search query and sent to a server. Relevant ads are selected based on user's past history, CTR and other data. Increasing the number of ads is not a good idea as it will shoot up the earnings only for a short time before the user switches to other search engines due to poor user experience. To maximize revenue, precise placements of ads are therefore required.

CTR refers to the number of times the advertiser's ad has been clicked (clicks) divided by number of times the ad appears on the screen (impression). Relevant placement of ad is a precondition for increasing the CTR for it. Ad performance can also be measured using CTR. CTR determination has several issues associated with it. The advertisers need to pay every time whenever an ad is clicked. It is on the basis of CTR that the search engine decides what ads are to be displayed and in what order of appearance. This is ensured by combining together the likelihood of an ad being clicked and the cost of the ad per click to create a display format that will yield maximum return. Many algorithms based on the supervised methodology have been designed to predict the CTR of advertisements like support vector machines, decision trees, Naïve Bayes, etc.

Motivated by this, the goal is to analyse and assess the application of unsupervised approach namely FCM for determining the CTR of an advertisement. This

clustering algorithm divides the total ads in the dataset into broad clusters based on whether they have been clicked or not by the user. This classification of ads helps in assessing the kind of advertisements that the user is really interested in and hence predict more of these kinds of ads. This would, in turn, result in a rise in the CTR of the displayed ads because of a greater number of clicks. The results obtained from FCM have been contrasted with KMC which has been used as a baseline model. The two techniques are assessed based on metrics like precision, recall and accuracy.

The following content is compiled as follows: Sect. 2 reviews work done by various researchers along with the brief idea of the algorithm used by them and their results. Section 3 elaborates the dataset taken, methods employed for preparing the data before implementing the algorithm and the algorithms used. The scrutiny of the experimental outcomes is done in the Sect. 4. Section 5 culminates the results along with the suggestions for future research prospects.

2 Related Work

After a thorough assessment of various studies in the past, it was found that a lot of algorithms have been proposed for the prediction of relevant ad to be displayed to the user. The author Avila Clemenshia et al. [1] in 2016, proposed a CTR prediction model using Poisson's regression, linear regression and support vector regression algorithms and displayed the ads accordingly. The dataset was provided by a digital marketing agency. Their results stated that support vector regression performed best among the three. Evaluation of the results was done on the basis of root mean squared error (RMSE) and correlation coefficient.

Authors Graepel et al. [2] in 2010 devised a new Bayesian CTR algorithm. The ad predictor presented showed better outcomes when compared to the baseline Naïve Bayes in terms of relative information gain (RIG) and areas under the curve (AUC). Hillard et al. [3] in 2010, implemented a model for estimating ad relevance. They refined it by including indirect feedback after consolidating the basic features of text overlap. In case of presence of adequate observations, click history was used. In case of no or few observations, a model was learned that could also be used for unpredicted ads. The precision, recall and f-score values were noted and improvement was observed with the new model.

A logistic regression approach was suggested by author Kondakindi et al. [4] in 2014, to predict whether an ad will be clicked or not. The dataset used for this purpose was from Avazu provided as a part of Kaggle competition. They started off with simple Naive Bayes followed by Vowpal Wabbit and finally got the best scores with logistic regression together with proper data preprocessing. Shi et al. [5] in 2016, designed a framework for prediction of CTR and average cost per click (CPC) of a keyword using some machine learning algorithms. The performance data for the advertiser's keywords was gathered from Google AdWords. The author had applied different machine learning algorithms such as regression, random forest

and gradient boosting to evaluate the prediction performance on both CTR and CPC. Results concluded that random forest transpires to be the best for both the metrics while gradient descent results are least reliable.

Wang et al. [6] in 2011, implemented a model for learning of user's click behaviours from advertisement search and click logs. Decision tree (DT), CRF, SVM and backpropagation neural networks (BPN) are the algorithms which were employed to carry out the imposition. The experimentation finally led to proving that CRF model outperformed the two baselines and SVM remarkably. Chakrabarti et al. [7] in 2008, developed a model for contextual advertising in which the revenue accrued by the site publisher and the advertising network depends upon the suitability of the ads displayed. This was followed by mapping the model to standard cosine similarity matching.

Cheng et al. [8] in 2010, developed the model for customization of click models in sponsored search. The results demonstrated that the accuracy of CTR could significantly be improved by personalized models in sponsored advertising. Edizel* et al. [9] in 2017, proposed the use of that deep convolutional neural networks for CTR prediction of an advertisement. One approach involved query-ad depiction being learned at character while the second entail word-level model by pretrained word vectors. The conclusion signified better outcome than the standard machine learning algorithms trained with well-defined features.

The analysis of the background work clearly connotes that most of the algorithm used in the past are supervised classification algorithms that involves two classes namely clicked and not clicked-demonstrating whether advertisement was clicked or not. The study of unsupervised clustering is still not much discovered in this domain to the extent of our understanding. This research paper is an endeavour to compare FCM and KMC algorithm using the dataset.

3 Data Characteristics

The following Fig. 1 demonstrates the system architecture of the research.

The following subsection explains the details.

3.1 Data Collection

The dataset is acquired from Avazu (Kaggle) for the purpose of writing this paper. It contains 11 days' worth of data in order to build and test prediction models using various machine learning algorithms. As the given data is approximately 6 GB, the data taken was 10 h of data for training and 2 h of data for testing. The data is ordered chronologically and the clicks and non-clicks are sampled according to different strategies.

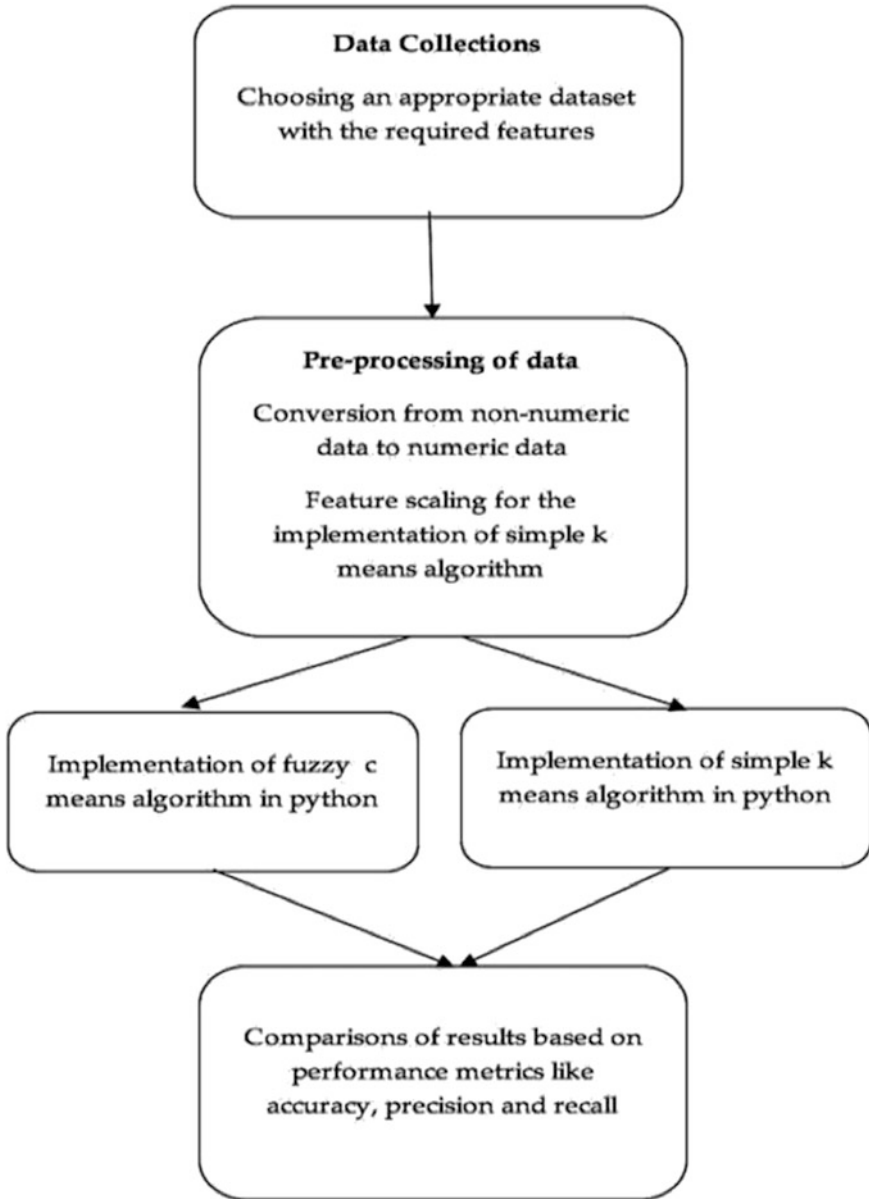


Fig. 1 System architecture

The following features are included in the dataset:

- id (unique id to identify advertisement)
- click (0 for not clicked and 1 for clicked)

- hour (in format of YYMMDDHH; it refers to the date and time the data is recorded)
- banner_pos (0 for top and 1 for bottom)
- site_id (refers to the website id)
- site_domain (website domain id)
- site_category (category to which website belongs)
- app_id (app id)
- app_domain (app domain)
- app_category (class to which app belongs)
- device_id (id of device from which ad was clicked)
- device_ip (ip address of the device)
- device_model (model number of device)
- device_type (mobile/laptop/desktop)
- device_conn_type(internet connection type-wifi/mobile data)
- C1 and C14–C21 (anonymized categorical variables).

3.2 *Preprocessing of Data*

The class attribute is removed owing to the fact that unsupervised clustering is the technique applied which will learn from the data and classify it into two clusters. The dataset contains features in both integer and string format. In order to carry out the implementation, conversion of strings into integer is done. Pandas package in Python is used to load the csv file, i.e. the raw dataset into memory, identify the columns with string values and convert them into integer values using python standard hash function.

Feature scaling is done with the help of scikit-learn library in python in the case of KMC. It is done in order to ensure fast convergence to the optimal solution and normalize the range of variables.

3.3 *Implementation of Machine Learning Algorithms*

The following section explains the clustering algorithm used.

K-means Clustering: It is one of the most straightforward learning algorithms available for unsupervised learning [10]. The procedure basically follows two main steps to classify the dataset into k clusters. The number of clusters are fixed a priori. The algorithm starts by defining k cluster centers. Each value in the dataset is then associated to its nearest cluster. This is followed by a recalculation of the k cluster centers. A new binding is done between discrete units of information in the dataset and the new cluster centers and these steps are repeated till all units have been assigned to their respective clusters.

The pseudocode is given as follows (Fig. 2):

$X = \{x_1, x_2 \dots x_n\}$ be the collection of observations and $u = \{u_1, u_2 \dots u_c\}$ are the cluster centroids. We arbitrarily select 'c' cluster centers (1). Then, the distance

Given: a training set $x(1), \dots, x(m)$ and feature vectors for each data point $x(i) \in \mathbb{R}^n$ with no labels $y^{(i)}$

(unsupervised learning problem).

Objective: Predict k centroids and a label $c^{(i)}$ for each datapoint

1. Initialize k points at random as cluster centers.
2. Assign data point to their closest cluster center according to the Euclidean distance function.

For every i , set

$$c(i) := \arg \min(j) ||x^{(i)} - \mu_j||^2$$

3. Find the new cluster centroid by calculating mean of all data points that belong to the cluster.

For each j , set
$$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)}=j\}}$$

4. Repeat steps 2 and 3 until the same points are assigned to each cluster in consecutive rounds.

Fig. 2 Pseudocode for k-means clustering

between each observation in dataset and center is calculated. Following this, each observation is assigned to that cluster for which the distance between the observation and the center is least from all the available centers (2). This assignment step is followed by a recalculation of the cluster centers (3) by executing step 1 again. If all observations belong to the clusters calculated in the previous iteration, then the algorithm is terminated.

1. Initialize $U=[u_{ij}]$ membership matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c(j) = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

3. Update $U^{(k)}, U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$ then STOP; otherwise return to step 2.

Where u_{ij} is the degree of membership of x_i in the cluster j ,

x_i is the i th of d -dimensional measured data,

c_j is the d -dimension center of the cluster

ε is a termination criterion between 0 and 1.

Fig. 3 Pseudocode for fuzzy c-means clustering

Fuzzy C-Means Clustering: FCM is also referred to as soft clustering [11]. Each observation can be a part of more than one cluster at the same time. The distance between each cluster center and value in the dataset is evaluated and based on this distance, the degree of membership is assigned to each observation. A higher degree of membership towards a cluster means that the observation is closer to that cluster center than compared to the rest of the clusters. The summation of the degree of membership of each point in dataset equals to one. Data points belong to distinct clusters in hard or non-fuzzy clustering. The degree of membership assigned to each observation plays a pivotal role here. This denotes the extent to which an observation belongs to each cluster. As we move from center to the boundary of the cluster, the degree of belongingness of the observation decreases.

The pseudocode for FCM is as follows (Fig. 3).

Let $X = \{x_1, x_2, x_3 \dots, x_n\}$ be the collection of discrete units in dataset and $V = \{c_1, c_2, c_3 \dots, c_k\}$ be the centers. Select 'k' cluster centers and initialize the membership matrix in step (1). Compute the centers 'c_j' in step (2). Calculate the membership 'μ_{ij}' in the membership matrix using step (3). Repeat step (2) and (3) until the smallest value of J is obtained.

4 Results and Analysis

The results were analysed using the accuracy, precision and recall [12] as an efficacy criterion. Accuracy is the measure of the closeness of the predicted observations to the actual value. It is the ratio of rightly predicted inspections to the total inspections made. Higher value of accuracy indicates more true positives and true negatives. Precision refers to the correctness of a model. It is simply the ratio of all the precisely predicted positives to the total number of positives predicted. More the number of true positives implies more precision. Recall is a measure of responsiveness or sensitivity of a machine model. It is the ratio of the correctly predicted positives to the count of values that belong to the class 'yes'. Recall and precision are inversely related. The following Table 1 depicts the performance analysis of FCM and KMC.

As Table 1 shows the accuracy, recall and precision values obtained for the two clusters, i.e. clicks/non-clicks predictions for the users is higher for FCM than for

Table 1 Comparison of results for k means and fuzzy c-means clustering

	k-means		Fuzzy c-means	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Accuracy	43.85438544	43.85438544	76.61767177	76.61767177
Precision	15.62280084	81.0428737	17.12049388	83.51293103
Recall	52.05182649	42.16809357	91.03165299	9.345230918

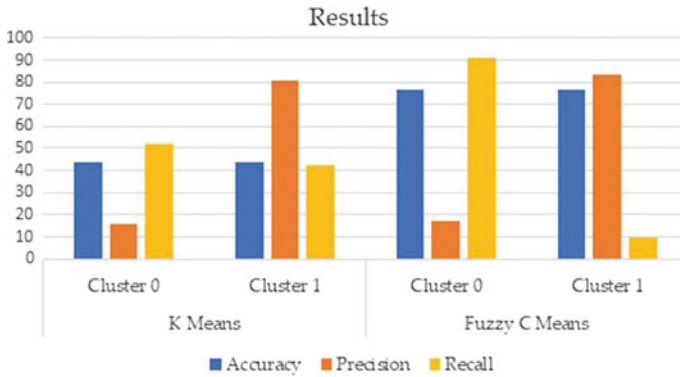


Fig. 4 Results

KMC. This means that the rightly predicted true positives and true negatives (accuracy), the degree of exactness (precision) and degree of completeness (recall) are relatively higher for FCM than for KMC.

Both the algorithms divide the data into two clusters, Cluster 0 and Cluster 1, which in turn represent a division of the dataset into two categories (Fig. 4). These categories are based on click/non-click classification for the ads, i.e. one cluster has all the ads clicked by the user. The other cluster has all ads which were not clicked by him. This gives a clear idea of the type of ads the user is inquisitive about and willing to click on.

After the classification of the ads in the dataset into categories—‘Click’ and ‘Non-Click’, we get a fair estimate of the kind of advertisements the user really wants to see. We can thereby predict more of such ads, hence increasing the CTR for the same.

The number of ‘clicks’ divided by the total number of times the ad is displayed, i.e. ‘impressions’ gives the CTR.

5 Conclusion

This paper compares the CTR as assessed by KMC and FCM algorithms. The outcome shows that FCM achieves better values of performance metrics than KMC. The results are better because it considers the fact that each data point can lie in more than one cluster and involves complex calculation of membership matrix. Conventional KMC, on the other hand, relies on hard clustering and definitively assigns each data point to the clusters.

The accuracy of the model can further be improved by assessing the shape of the clusters, including other more refined attribute selection and extraction methods that could assist in better modelling of the present system. Attribute selection is

concerned with selecting a subgroup of valid features for model selection whereas attribute extraction means deriving attributes from the already existing ones for subsequent learning. Superior results could be obtained by enhancing fuzziness coefficient. The degree of overlap between clusters is determined by the fuzziness coefficient. Higher value of m means larger overlapping between clusters. There exists a vast scope of application of other soft computing methodologies also, like swarm optimization, etc. that can be applied for determining CTR for other datasets as well.

References

1. Avila Clemenshia, P., Vijaya, M.S.: Click through rate prediction for display advertisement. *Int. J. Comput. Appl.* **975**, 8887 (2016)
2. Graepel, T., Candela, J., Borchert, T., Herbrich, R.: Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In: *IJCA* (2010)
3. Hillard, D., Schroedl, S., Manavoglu, E., Raghavan, H., Leggetter, C.: Im-proving ad relevance in sponsored search. In: *ACM* (2010)
4. Kondakindi, G., Rana, S., Rajkumar, A., Ponnekanti, S.K., Parakh, V.: A logistic regression approach to ad click prediction (2014)
5. Shi, L., Li, B.: Predict the click through rate and average cost per click for key-words using machine learning methodologies. In: *IEOM* (2016)
6. Wang., C.J., Chen, H.H.: Learning user behaviors for advertisements click predictions. In: *ACM* (2011)
7. Chakrabarti, D., Agarwal, D., Josifovski, V.: Contextual advertising by combining relevance with click feedback. In: *WWW* (2008)
8. Cheng, H., Cantú-Paz, E.: Personalized click prediction in sponsored search. In: *ACM* (2010)
9. Edizel, B., Mantrach, A., Bai, X.: Deep character-level click-through rate prediction for sponsored search. In: *Stat.ml* (2017)
10. Yadav, J., Sharma, M.: A review of K means algorithm. *Int. J. Eng. Trends Technol.* **4**(7), 2972–2976 (2013)
11. Yanyun, C., Jianlin, Q., Xiang, G., Jianping, C., Dan, J., Li, C.: Advances in research of fuzzy c means algorithm. In: *International Conference on Network Computing and Information Security* (2011)
12. Bhatia, M.P.S., Kumar, A.: Information retrieval and machine learning: supporting technologies for web mining research and practice. *Webology* **5**(2), 5 (2008)

ASK Approach: A Pre-migration Approach for Legacy Application Migration to Cloud



Sanjeev Kumar Yadav, Akhil Khare and Choudhary Kavita

Abstract Legacy application migration is a mammoth task, if migration approach is not well thought at the very start, i.e. pre-migration, and supported by robust planning especially at pre-migration process area. This paper proposes a mathematical pre-migration approach, which will help the enterprise to analyse existing/legacy application based on the application's available information and parameters an enterprise would like to consider for analysis. Proposed pre-migration assessment will help in understanding the legacy application's current state and will help in un-earthing the information with respect candidate application. Proposed pre-migration approach will help to take appropriate well-informed decision, whether to migrate or not to migrate the legacy application. As it is said that application migration is a journey, if kick-started once, needs to reach its destination else it can result into a disaster hence pre-migration is one of the important areas of migration journey.

Keywords Application migration · ASK approach · ASK confidence level (ACL) · LAMP2C · Legacy application · Migration process · Pre-migration

1 Introduction

Technology paradigm is changing at fast pace, so is the customer requirement. In order to serve the fast-changing customer requirement is to keep abreast with ever-changing technology. Important is, if enterprise does not embrace or adopt

S. K. Yadav (✉) · C. Kavita
Jayoti Vidyapeeth Women's University, Jaipur, India
e-mail: sanjeevyadav@yahoo.com

C. Kavita
e-mail: kavita.yogen@gmail.com

A. Khare
CSED MVSR Engineering College, Hyderabad, India
e-mail: khareakhil@gmail.com

new or up-coming technology then enterprise may face challenges such as unavailability or shortage of skilled manpower as they would have moved to other latest technology, which will impact software and hardware maintenance cost (complex patching and modifications/enhancement) because scarcity of people in market with the required technology will increase the manpower cost and at the same time even the hardware will also become costly. All this may put application scalability, business continuity, customers' experience and enterprise image [1] in jeopardy and may result in inability to meet current business needs.

Each of the challenges identified may have direct or indirect impact on one or more factors. At times these identified factors seem to be interrelated, refer Fig. 2. There may be other factors, which may be enterprise or legacy application specific.

Enterprises may be ardent and ready to adopt technology such as cloud computing, which delivers resource as virtualize service [2] and offers so many other benefits. However, the biggest challenge for an organization is that enterprise can't discard the investment made in the past in terms of existing business critical and in-use applications including its infrastructure. These business-critical applications/systems are referred as an asset. These assets are built or might have evolved over a span of 10–20 years; and have become core to business process, might have become complex and large enough to be understood by a person.

Cloud computing offers multiple benefits which any enterprise will be interested in such as lower operating cost, no up-front investment and ease of maintenance. Fact is that cloud reduces capital expenditure by leveraging cloud business model of pay-as-you use, which turns capital expenditure to operation expenditure.

As exiting applications cannot be discarded, so to embrace the new technology enterprise has below-mentioned three options for lasting benefits/goal, refer Fig. 1:

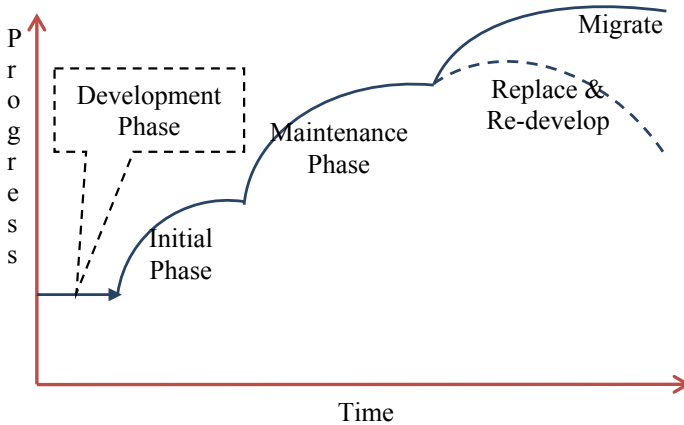


Fig. 1 Software life cycle phases and modernisation options©

- i. Application Migration—Migrate legacy application migration to new platform.
- ii. Application Replacement—Replace legacy application with off-the-shelf, available in market, application.
- iii. Application Redevelopment—Redevelop the existing application all over again.

Two options, Option ii, i.e. Application replacement or option iii, i.e. Application redevelopment, does not seem to be feasible options because the existing business-critical application might have got developed with lot of effort, time and cost. Putting similar effort, cost and time all over again may not be feasible, so feasible option available with enterprise is option i, i.e. Application migration, which will help to migrate current application to new technology/platform and embrace new technology, i.e. Cloud at the same time, will keep the business-critical application intact.

Paper is organized in six sections. Section 2 provides the literature review, which provides background of related work. Section 3 defines the problem. Section 4 provides the detail about the proposed framework. Section 5 details the ASK approach. Finally, the paper concludes with Sect. 6, which presents the concluding comments and future scope.

2 Literature Review

Cloud computing acceptance has increased over a period of time because of the benefits it brings in the area of operational and efficiency. Migration frameworks are proposed and are compared in the selected literature but none provides an end-to-end migration framework, especially in the pre-migration stage of the migration. The focus of literature review is to bring out any available pre-migration method/approach.

Summary of literature studied and identified research gap in context of chosen topic of the paper are detailed below:

Shoaib et al. [3] highlighted that due to technology advancement enterprises need to migrate from one platform to other, so existed framework was examined and reviewed. Based on this, the author is of view that multiple methods exists for migration and at the same time there are multiple risks involved during migration all these risks can be taken care if migration is performed properly. In this paper, the author describes different methods and frameworks which provides guideline for developers to enhance software migration process. As per the author, the focus area for migration is framework along with process and activities. The author concludes that there is research gap and need to be explored further.

Musale et al. [4] provided basic knowledge about the migration concept and initial factors required to migrate to cloud. The paper also highlighted that cloud computing is an emerging technology and there is substantial scope of operational

and efficiency improvement. If IT application is migration to cloud or some of the processes are migration to cloud environment, off premises. The author puts forward the cloud migration concept and discusses the essential points for migration and further discusses the after migration essential. The author is of view that migration to cloud is an art than it is a science.

Gholami et al. [5] highlighted that research around migrating legacy application to cloud has increased because of the adoption of cloud computing as an outsourcing strategy has grown in recent past. The author is of view that there is no integrated overarching cloud migration exists in current literature. In order to fulfil this gap, the author tries to propose cloud platform-independent conceptual view and reusable migration process, which has phases, activities and task. Validity of the proposed metamodel is demonstrated by analysing three existing cloud migration process models. Model can be extended by adding new constructs and can be customized for other migration paradigm.

Sabiri et al. [6], the authors said that enterprises adopting cloud is increasing because of the significant advantage cloud computing brings in; however, the migration (move) to cloud has challenges. The authors studied and compared the existing cloud migration methods to understand the strengths and weaknesses of each model based on Model-Driven Engineering (MDE) approach. The author proposed and explained a cloud migration method based on Architecture Driven Modernization (ADM), which focuses on architectural modes instead of code artefacts.

Gouda et al. [7] highlighted that cloud migration can be of two types; either from on-site premises to the cloud or moving them from one cloud environment to another, where the information can be accessed over the Internet. The authors also highlighted that cloud migration is broadly classified into two categories, i.e. Big Bang Migration and Trickle Migration. The authors have suggested multiple cloud migration patterns or categories as their requirement. The authors also suggested migration to cloud-step wise approach or procedure which will help the user to successful and fruitful migration of its IT resources to cloud.

Shawky [2] highlighted that in order to utilize cloud's capability, the system should be migrated to cloud. The author also highlights that application migration to cloud leads multiple technical challenges and these challenges cannot be ignored. According to the author, the main challenge is mapping application to cloud services as multiple aspects need to be considered, which includes other system component dependencies.

Rashmi and Sahoo [8], 'A five-phased approach for cloud migration'; highlighted that cloud computing is the rapidly growing segment of IT. Organization is moving to cloud due to its long-term benefits but they are unsure about the migration process. The authors proposed a five-phased waterfall-based migration model to cloud having feedback path to earlier phase of the process but have not dwelled into the details of each phase of the proposed model.

Pahl et al. [9] mentioned in this paper that though cloud computing has gained the attention, adoption of the technology is yet to gain acceptance as how to migrate to cloud is unanswered for many. Point in case is highlighted with the help of three provider-driven case studies and tries to extract common layer-specific migration

process. Highlights that no common procedures and tools are existing, however, establish migration core elements such as activities and steps.

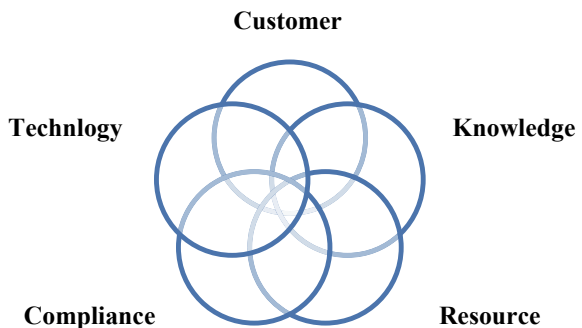
Stavru et al. [10] acknowledged the fact that enterprises are unable to take full strategic and operational advantage as they are not able to migrate legacy application to cloud due to non-existence of mature process. The authors identified and extracted challenges with the help of systematic literature review. Using expert judgment, evaluates how different agile techniques, taken from Scrum and Extreme Programming (XP), could address or overcome the identified challenges in software modernization projects in specific context.

3 Problem Specification

Challenge can be in terms of skilled resource's scarcity, software maintenance cost, lack of software scalability, business continuousness, customers experience, reduced efficiencies, knowledge retention, enterprise image [1] and inability to meet current business needs. Each of the challenges identified may have direct or indirect impact on one or more factors, so identified impacted factors seem to be interrelated. Tried categorizing these challenges into the below category; refer Fig. 2. There may be other factors, which may at times be enterprise specific or legacy application specific.

- Customer
 - (a) Application inefficient in catering to business requirement will impact the enterprise interest.
 - (b) Applications may become inefficient in catering to any area of customer requirement (customer can be internal and/or external) over time; in terms of functionality/feature(s) and better service. Inefficient application will provide an edge to competition.
- Knowledge
 - (a) Lack of in-depth knowledge in terms of understanding the existing system and its related process will or may have direct impact on the system maintenance, i.e. developing new functionality and enhancements.

Fig. 2 System maintenance and enhancement challenges



- (b) Due to non-availability of knowledgeable resources will have difficulty in getting best practices shared with respect to these legacy applications, i.e. knowledge sharing. It will increase the risk of knowledge concentration amount small set of people.
- Resources
 - (a) Reducing resource pool (attrition) or resource pool desire to move onto market relevant technology will have direct impact on resource availability for the maintenance/enhancement of legacy application.
 - (b) As legacy application knowledge gets concentrated into few resources, it will lead to increase in people retaining cost; related aspect will be increase cost of gaining resource in desired technology.
 - (c) Contracted resource pool will have risk of high dependency, in terms of understanding the existing system and its related process on to a very small set of people in an organization.
- Technology
 - (a) Enterprise may have multiple technologies in use for existing business-critical applications, so maintaining them will have high cost because the legacy application might have become unsupported due to non-existence of support from the product vendor as they might have stopped supporting the in-use version or vendor might release new version, i.e. vendor moved ahead in technology.
- Compliance
 - (a) Legacy applications may become non-regulatory compliance over a period of time if no enhancement is made due to multiple reasons mentioned above.
 - (b) Non-compliance to regulatory norms may have bigger impact than adopting new technology and moving legacy application to the technology.

Challenges highlighted above are few of the challenge, many other challenges will come which will be application or enterprise-specific challenges. These challenges motivate to come up with a framework, which will help the enterprise to migrate legacy application to cloud in an efficient manner.

4 Proposed Framework

Legacy application migration is a journey which requires a careful detailed planning irrespective of the size of the application identified for migration because no one-size-fit-approach will work in application migration. One of the biggest challenges an enterprise faces is the uncertainty of where to begin the migration process. If enterprises do not carefully plan, execute and monitor the transformation using

established processes [11] then migration can be fraught with pitfalls. Before we proceed further, it is essential to understand the reasons or trigger points for migration. At high level there are only two factors which trigger migration and these factors are either: Business Driven, Technology Driven or both. Whichever may be the reason for migration, i.e. business and/or technology; one thing which will happen first is current portfolio analysis.

Cloud computing concepts and approach itself is very different from the traditional software concepts and approach. In fact, moving to cloud is a serious step to take in a holistic manner [1], which required a careful elaborate planning for successful migration because cloud computing brings changes such as economic, legal, privacy/security, [1] etc. especially the pre-migration area of the migration process.

Cloud covers all application layers, i.e. application (SaaS), platform (PaaS) and infrastructure (IaaS), which makes legacy application migration not a straightforward task as it looks like. Rashmi et al. [8], especially when migrating a legacy application, which is tightly coupled [12] as compared to cloud application.

In order to address this issue, a framework is proposed called ‘Legacy Application Migration Process Framework to Cloud’; referred as LAMP2C, Fig. 3. LAMP2C migration framework acts as a guiding principle for migrating application to cloud and process to be followed while migrating legacy application. Scope of the proposed LAMP2C framework covers entire legacy application as shown in Fig. 3 and pre-migration is highlighted as ASK Approach is part and limited to the pre-migration area of the LAMP2C framework.

Framework helps to plan the migration in an effective and efficient manner and it also helps in speeding up migration execution process with clear and hazel free migration path.

LAMP2C, address pre-migration, migration and post-migration area of migration with an umbrella of governance area across three cloud layers. Framework is divided into area and area is further sub-divided into phases. The phases will have activities under it. LAMP2C—high-level framework is shown in Fig. 4.

Fig. 3 High-level framework

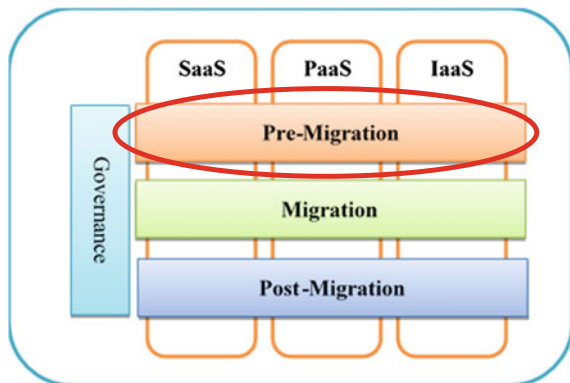
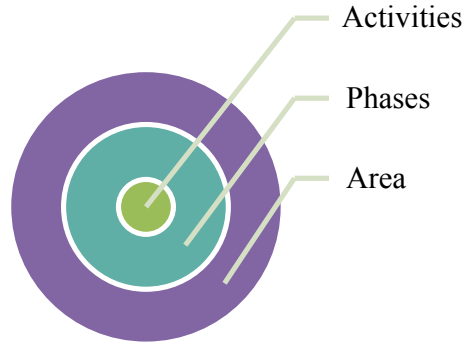


Fig. 4 LAMP2C constituents



LAMP2C is a flexible framework, which can be tweaked as per the enterprise's requirement as proposed framework does not force to follow the process end-to-end rather it suggests to go phased manner.

One of the important parts of LAMP2C is the pre-migration area, which helps in evaluating existing portfolio or application which in turn helps an enterprise in understanding the complexity, challenges and efforts that will be required for application migration. It will also address the uncertainty of 'where to begin' and other than whether to migrate or not to migrate to the cloud.

LAMP2C, Figs. 1 and 3 is copyright of the authors: Registration Number L-60550/2014.

5 Ask Approach

It is assumed that at high level there are only two factors which can trigger migration and they are either technical or non-technical (business) parameter, so assessment and its parameters are also clubbed into these two categories only. Another assumption is that a) technical assessment should consider all the three cloud's service model, i.e. IaaS, SaaS and PaaS; however, enterprise is free to choose one or more for assessment and their respective parameters. Parameters which cut across three models should be grouped into one as 'Common Parameters', as shown in Fig. 5.

Common parameters are to avoid the consideration of same parameter(s) across three layers of service model, however it will depend upon individual, how they would like to perform the pre-migration analysis. b) Non-technical, i.e. business parameter are those parameters, which are not technical in nature should be considered as one group or as suitable.

ASK Approach—a pre-migration assessment methodology is proposed to perform portfolio or application analysis under pre-migration area of LAMP2C framework. ASK Approach is an important constituent of the LAMP2C framework and its scope is limited to the pre-migration area of the LAMP2C framework as it

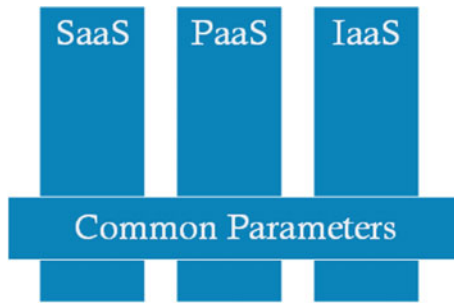


Fig. 5 Parameter identification area



Fig. 6 Pre-migration approach

provides inputs and helps in taking GO/No-Go decision with respect to legacy application migration to cloud. Proposed pre-migration assessment methodology is parameter based mathematical approach, which will be appropriate for any application identified for migration and will help enterprise to view and analyse the enterprises IT application portfolio in details based on parameters which are important for organization and application so assessment is carried out on any set of identified parameters under the mentioned category or both, i.e. technical and non-technical (business) aspect.

Pre-migration approach can be divided into five steps, as shown in Fig. 6 and detailed in subsequent paras:

As approach is parameter based, as a first step, all parameters and its sub-parameters needs to be identified. Each identified and considered parameter needs to be assigned the importance. Importance is rated between the range of 0.1, being the lowest to 5.0, being the highest, based on the parameter’s impact, as tabled below (Table 1):

Table 1 Impact rating range

Importance	Impact rating range
Very high	4.1–5.0
High	3.1–4.0
Neutral	2.1–3.0
Low	1.1–2.0
Very low	0.1–1.0

Assuming that there are n number of parameters namely C_1, C_2, C_3 and so on till C_n . Based on parameters' importance an impact rating between the range of 0.1–5.0 will be assigned for each parameter. Adding all these integrals and dividing them by the total number of parameters (in this case n) considered for analysis will give us the mean value of the data, which is being referred as '**ASK Point**' (ASKP), i.e. μ_c .

$$ASKP = \mu_c = \frac{\int C_1 + \int C_2 + \int C_3 + \int C_4 \dots + \int C_n}{n}$$

$ASKP = \mu_c = \text{Parameters' Mean}$

$n = \text{Number of parameters}$

$i = \text{Importance of parameter}$

In reduced form the above formula can be written as follows:

$$ASKP = \mu_c = \sum_1^n \frac{\int C_n}{n}$$

Considering the same data calculate the standard, which will be referred as **ASK Confidence Level (ACL)**. ACL will be calculated by adding the squares of the values found from subtracting the ASKP from each parametric value. This summed value is divided by the total number of parameters. This will give us the standard deviation σ_c , i.e. ACL.

$$ACL = \sigma_c = \frac{(\int C_1 - \mu_c)^2 + (\int C_2 - \mu_c)^2 + (\int C_3 - \mu_c)^2 + \dots + (\int C_n - \mu_c)^2}{n}$$

$ACL = \sigma_c = \text{Parameters' Standard Deviation}$

$ASKP = \mu_c = \text{Parameters' Mean}$

$n = \text{Number of parameters}$

In reduced form the above formula can be rewritten as follows:

$$ACL = \sigma_c = \sum_1^n \frac{(\int C_n - \mu_c)^2}{n}$$

In order to generate boundary for the identified parameter, which is going to be referred as **ASK Belt** (ASKB). ASKP and ACL will help to build ASK Belt. ASKB will have two components a) ASKB Right, i.e. $ASKB_R$ and b) ASKB Left, i.e. $ASKB_L$. $ASKB_R$ and $ASKB_L$ can be calculated as shown below:

$$ASKB_R = ASKP + ACL$$

$$ASKB_L = ASKP - ACL$$

Once, ASKP, ACL and ASKB (ASKB_R and ASKB_L); bell curve can be calculated using the below formula (Gaussian equation)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(x-\mu_c)^2}{2\sigma_c^2}}$$

The following are the value for the above formula:

- $\pi = 3.14$
- $e = 2.71$ and
- $x =$ Parameter Impact
- $\mu_c =$ ASKP
- $\sigma_c =$ ACL

As fifth and the last step; calculate various other values required in the Gaussian function. The parameters have been arranged in the ascending order of the impact rating for facilitating the plotting process. Plotting the above $f(x)$ values versus x values, we get the following bell curve (as defined by the Gaussian function), refer Fig. 7.

Points which falls between ASKB_R and ASKB_L can be safely considered for legacy application migration, however, points which fall outside the belt need to be re-looked and considered with caution.

These inputs will provide enough data points to conclude as to Go ahead with the legacy application migration or still there are points/parameters, which need to be re-looked at. And will provide inputs to come up with migration plan.

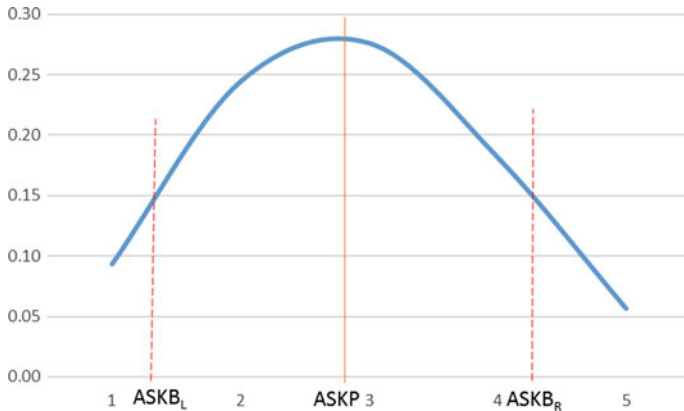


Fig. 7 Bell curve

6 Conclusion

Proposed mathematical pre-migration methodology called ‘ASK Approach’ under the LAMP2C migration framework. Important is, impact rating between the range of 0.1–5.0 is assigned to each identified and considered parameters. 0.1 being the lowest and 5.0 being the highest. Assigning an impact rating to each parameter will eliminate any guesswork among any of the person associated with respective application migration. Impact rating will help to drive ASKP, i.e. μ_c , which in turn will help to drive ACL, i.e. σ_c . ASKP and ACL will help to drive ASKB_R and ASKB_L which in turn will help to calculate $f(x)$ for each considered parameter. Bell curve can be plotted to identify the parameters, which helps to provide a view about, which legacy application parameters should be considered with confidence and which parameter needs more attention. It requires ultimate help to take informed decision of GO or NO-GO with respect to application migration. With the help of this pre-migration approach enterprise can also make phase or staggered legacy migration approach. Phased manner approach provides the required feel confident factor and return on their investment can be seen as migration project progress. Big bang migration approach may not provide the required feel good factor and always have a project failure risk.

Proposed ASK Approach is focused on parameter which falls with the ASK Belt, but does not provide any inputs as to what needs to be done with parameters, which are outside the ASK Belt, so ASK Approach can be further enhanced to come up with mathematical formula or otherwise, as to how to deal with the parameters, which falls outside the ASK Belt, i.e. ASKB_R and ASKB_L.

References

1. Khajeh-Hosseini, A., Sommerville, I., Sriram, I.: Research challenges for enterprise cloud computing (2010)
2. Shawky, D.M.: A cost-effective approach for hybrid migration to the cloud. *Int. J. Comput. Inf. Technol.* **2**(1), 57 (2013)
3. Shoaib, M., Ishaq, A., Ahmad, M.A., Talib, S., Mustafa, G., Ahmed, A.: Software migration frameworks for software system solutions: a systematic literature review. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **8**(11), 192–204 (2017)
4. Musale, A.D., Khot, P.G.: A roadmap towards cloud migration. *IOSR J. Comput. Eng. (IOSR-JCE)*. e-ISSN: 2278-0661, p-ISSN: 2278–8727
5. Gholami, M.F., Low, G., Beydoun, G.: Conceptualising cloud migration process. In: Twenty-Fourth European Conference on Information Systems (ECIS), Istanbul, Turkey (2016)
6. Sabiri, K., Benabbou, F., Hain, M., Akodadi, K.: A survey of cloud migration methods: a comparison and proposition. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **7**(5), 598–604 (2016)
7. Gouda, K.C., Dwivedi, D., Patro, A., Bhat, N.: Migration management in cloud computing. *Int. J. Eng. Trends Technol. (IJETT)* **12**(9), 466–472 (2014)
8. Rashmi, M.S., Sahoo, G.: A five-phased approach for cloud migration. *Int. J. Emerg. Technol. Adv. Eng. (IJETA)* **2**(4), 286–291 (2012)

9. Pahl, C., Xiong, H., Walshe, R.: A comparison of on-premise to cloud migration approaches—a tale of four cloud migration processes. In: European Conference on Service-Oriented and Cloud Computing (2013)
10. Stavru, S., Krasteva, I., Ilieva, S.: Challenges for migrating to the service cloud paradigm: an agile perspective (2012)
11. PMI White Paper: Cloud computing: the new strategic weapon (2012)
12. Bisbal, J., Lawless, D., Wu, B., Grimson, J., Wade, V., Richardson, R., O’Sullivan, D.: An overview of legacy information system migration. In: Asia-Pacific Software Engineering Conference (1997)
13. Yadav, S.K., Khare, A.: Legacy applications migration process to cloud—an approach framework. *Int. J. Comput. Technol. Electron. Eng. (IJCTEE)* **4**(2), 11–14 (2014)

A Fuzzy Logic Based Cardiovascular Disease Risk Level Prediction System in Correlation to Diabetes and Smoking



Kanak Saxena and Umesh Banodha

Abstract The cardiovascular disease (CVD) is one of the major causes of death among the people having diabetes in addition to smoking habits. It will create tribulations for every organ of the human body. Smoking becomes fashion among the youth from their childhood which results in premature death. The intention of this paper is to explain the impact of diabetes and smoking along with high BP, high pulse rate, angina affect, and family history on the CVD risk level. The concept used is based on the knowledge-based system. We have proposed a fuzzy-logic-based prediction system to evaluate the CVD risk among the people having diabetes with smoking habits. The aim is to facilitate the experts to provide the medication as well as counsel the smokers well in advance. This will not merely save the individual but also an immense relief to concern. The data set is used from UCI (Machine Learning Repository). Most of the researchers worked on diabetes or smoking impact on CVD separately, but the proposed system demonstrates how drastically it will affect ones' health condition.

Keywords Approximation function · Cardiovascular disease · Decision-making · Diabetes · Fuzzy logic · Smoking

1 Introduction

Cardiovascular Diseases (CVD) are very widespread disease that becomes the ground of death and morbidity among the humans [3, 5, 19]. The data from the national heart association (2012) showed that 65% of people with diabetes die from some sort of heart disease or stroke. The Framingham study was the earliest sub-

K. Saxena (✉) · U. Banodha
Department of Computer Application, Samrat Ashok Technological Institute,
Vidisha, India
e-mail: ks.pub.2011@gmail.com

U. Banodha
e-mail: ub.pub.2011@gmail.com

stantiation to demonstrate that people with diabetes are more (near about 2–4 times) susceptible to CVD disease than those who did not have diabetes [12]. It is very clear that diabetes if not controlled then it becomes the unending cause of not only the CVD but many more diseases [16]. As the WHO the percentage of the affected group is of 30 million in the year 1985, which increased to 135 million in the year 1995, and in the year 2005 it became 217 million. The prediction is near about 366 million people will suffer from diabetes in 2030, either they belong to developing or developed countries. On the other hand, smoking is the most imperative avertable cause of death which also accelerates the chances of occurrence of CVD among people, i.e., enhances the risk level [4]. It is also stated that smoking is the second leading cause after the high BP for CVD mortality. The WHO reported that more than one billion people smoke and the frequency rising surprisingly [14]. According to [8] the risk level of CVD estimated 15- year long period ahead for a smoker in comparison to a nonsmoker. Nowadays, the health care domain is one of the key domain where research is going on in the area of knowledge base to predict the impending risk in a patient and future of a healthy person with bad habits. The importance of the knowledge base is that it analyzes the data using various techniques. The experts took advantage of using the computerized diagnosis system where the information draws together in direct as well as indirect form. Earlier the work is done manually, though the decisions were taken by the help of the computer-based decision support system. With the improvement in the technology, the manually collected data problems have solved and now the system acquires the data which are directly or indirectly related to the physical condition. Today, there are many health care apps which assist the users as well as the experts for quick evaluation, thus avoiding uninvited events in ones' life. The evaluation varies on the input parameters and the techniques applied for the evaluation. Currently, the focus is on the flexibility of the conditions with intelligent evaluation. For the proposed system the highly suitable basis is the fuzzy set theory and the logic. The concept helps us in the interpretation of the facts with its findings and enhances the system with the cross diagnosis and estimation of the risk level.

In the proposed prediction system the main focus is on the CVD in relation to diabetes and smoking (urine nicotine). For testing purpose, the data set from UCI is used with major eight attributes as shown in Fig. 1. Among these, the major inputs are eight (inclusion of smoking) with adding together and one is used for the output purpose.

2 Literature Survey

Phuong et al. [13] illustrated the use of fuzzy logic and its application in the field of medicine. Oad and DeZhi [10] described the fuzzy-rule-based system for prediction of heart disease. Neshat et al. [11] proposed a fuzzy expert for liver disorder. Anooj [1] developed the weighted fuzzy clinical decision support system for the heart disease risk level prediction. Alessandra Saldanha deMattosMatheus et al. discussed

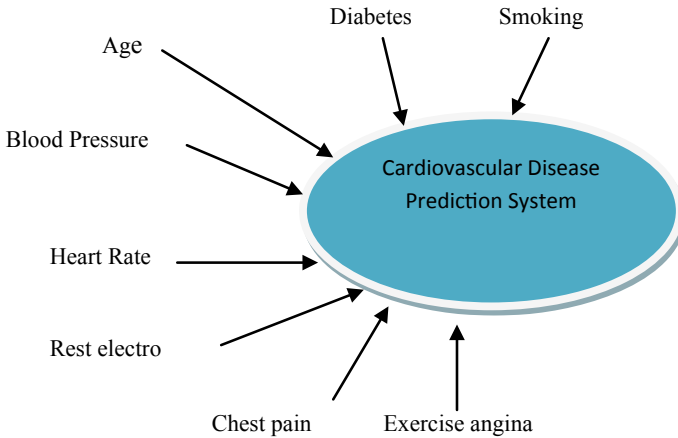


Fig. 1 Inputs of the cardiovascular disease risk level prediction system

the impact of the diabetes on the cardiovascular disease with obesity property in the patient. Rajeswari and Vaithyanathan [15] developed heart disease diagnosis system using fuzzy logic and genetic algorithm. Anbe et al. [9] proposed the fuzzy set diagnosis system for the valvular heart disease. Kumar and Kaur [17] exemplified the fuzzy rules based on the six parameters to predict the heart disease. Phongsuphap and Pongsupap [18] evaluated the health care policies to assist the authorities on the basis of the survey data which consists of qualitative as well as qualitative. The authors surveyed on the epidemiology of the use of tobacco smoking in the United States. They also emphasized the study of the characteristics of the nicotine and the clinical applications of medication [6]. There is the correlation with the smoking and the cardiovascular disease as it may be one of the causes of altered lipid metabolism, increased demand for myocardial oxygen and blood and many more [20, 21]. In [2, 7], the group of the smokers was studied with the impact of the serum cotinine levels in smokers as well as in nonsmokers.

3 Cardiovascular Disease Prediction System

The proposed prediction system is based on the fuzzy logic technique which may feasibly be used to predict the risk level of the CVD in the patients with diabetes and smoking habits. The preference of the fuzzy techniques is through on the uniqueness of self-explanatory assessment making procedure. In fuzzy logic, its effectiveness depends on the formation of the set of acts. The biased predicament can be unconcerned if the set of acts be capable to form automatically without the intervention of the experts. According to Table 1 the stages and steps of fuzzy logic are illustrated, which commence with the preprocessing of the data followed by the

Table 1 Stages and Steps of Fuzzy Logic

Stages	Steps
Preprocessing	Mining of the parameters
Fuzzification	Define the linguistic variables with their ranges
	Construction of the membership functions for input as well as for output variables
	Conversion of crisp data into the fuzzy data sets with the help of membership functions
Fuzzy rule base	Construction of knowledge base rules
Defuzzification	Mapping to quantifiable risk
Decision-making	Selection of the choice on the basis of the relevant information available

fuzzification and the rules generation. Thus the prediction system instigates with the symptoms performance-based analysis. Further, it will also depict the risk level in the midst of investigation and lend a hand to the experts to take the suitable decision at suitable point in time.

Figure 2 demonstrates the steps used in the CVD prediction system. The pre-processing of the system that consists of patient clinical data pattern to facilitate the input into the system by converging the data into accessible form, by performing the data mining to find the missing values (if any), convert it into the permissible range and significant task is the selection of the attributes which directly or indirectly fatally affect the prediction system. As it’s a nontrivial fact that the data

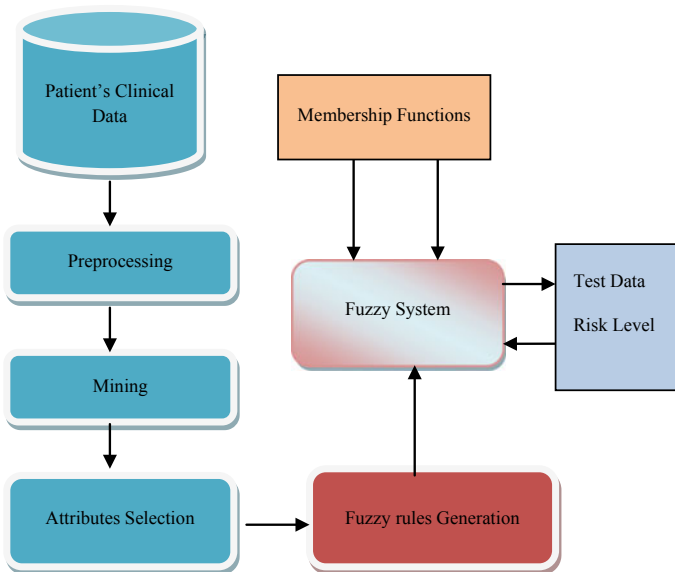


Fig. 2 Proposed fuzzy-based prediction system

which are not normalized cannot envisage the system procedure. The next step is to classify the data set in two categories, either suffering from the CVD or not.

The behavior of prediction system is uncertain. The fuzzy system is more appropriate to handle such type of formless medical events. The fuzzy system consists of the fuzzifier, fuzzy rules, fuzzy inference engine, and defuzzifier. In the fuzzification procedure, the crisp data set is to converted into a fuzzy value and the fuzzifier is the mapping from the data set to the fuzzy linguistic data value—a realistic approximation which is attained by the membership function. The fuzzy rule is to get fractioned into two parts as (1) condition and (2) conclusion. The inference engine interprets the rules logic with the appropriate reasoning, thus results in the form of the output—risk factor. Finally, the defuzzification is the reverse procedure of the fuzzification, where the mapping is to done from fuzzy value to the crisp data value. The system works with the identification of the input and output variables. Member function is used to characterize the fuzzy system and the suitable and succinct way is to characterize the member function mathematically. The significance of the mathematical formula is that it establishes contribution of the input variable to fuzzy set. The output of the prediction system has to be defuzzified, that is it has to be converted into a precise quantity, which can be the logical union of two or more fuzzy membership functions defined on the cardiovascular affect world.

The following process is to fuzzify all inputs and outputs and bring close to the contribution to which these inputs and outputs fit into each of the appropriate fuzzy sets. In the prediction system, the subsequent process is to discover the weighted fuzzy rules which have to engender from the various attributes resolution weighted rules. The method used for the automatic engenders the set of acts that are based on the structure of the state of fuzzy modals. The decision rules were attained from IF and THEN fractions of the specific class. These were acquired by converting into linguistic variables on the basis of the fuzzy membership functions. The first attribute used is age which is classified into four fuzzy sets (minor, young, adult, and old). The triangular as well as trapezoidal membership functions are used for the fuzzy sets. The same place into practice for blood pressure, chest pain, heart pain, blood sugar, rest electro, affected, and urine nicotine. In this, the fuzzy sets were separated into three or four levels that as follows.

Age (separated into 4 levels)			
Minor	Young	Adult	Old
$\mu_{\text{minor}}(x) = 1,$ $x \in [1,15]$	$\mu_{\text{young}}(x) = (x - 15)/$ $(24 - 15),$	$\mu_{\text{adult}}(x) = (x - 33)/$ $(40 - 33)$	$\mu_{\text{old}}(x) = (x - 54)/$ $(54 - 40),$
$\mu_{\text{minor}}(x) = (18 - x)/$ $(18 - 15),$ $x \in [15,18]$	$x \in [15,24]$ $\mu_{\text{young}}(x) = (42 - x)/$ $(42 - 24),$	$x \in [33,40]$ $\mu_{\text{adult}}(x) = (61 - x)/$ $(61 - 40),$	$x \in [40,54]$ $\mu_{\text{old}}(x) = (75 - x)/$ $(75 - 54),$
$\mu_{\text{minor}}[x] = 0,$ otherwise	$x \in [24,42]$ $\mu_{\text{young}}(x) = 0,$ otherwise	$x \in [40,61]$ $\mu_{\text{adult}}(x) = 0,$ otherwise	$x \in [54,75]$ $\mu_{\text{old}}(x) = 0,$ otherwise

Chest pain (separated into 3 levels)

Alert	Risk	High risk
$\mu_{\text{alert}}(x) = (x - 1)/(2 - 1),$ $x \in [1, 2]$	$\mu_{\text{Risk}}(x) = (x - 2)/(3.5 - 2),$ $x \in [2, 3.5]$	$\mu_{\text{HighRisk}}(x) = (x - 3.5)/$ $(4 - 3.5), x \in [3.5, 4]$
$\mu_{\text{alert}}(x) = 0,$ otherwise	$\mu_{\text{Risk}}(x) = 0,$ otherwise	$\mu_{\text{HighRisk}}(x) = 0,$ otherwise

Blood pressure (separated into 5 levels)

Low	Normal	Alert	High	Very high
$\mu_{\text{low}}(x) = 1,$ $x \in [1, 100]$ $\mu_{\text{low}}(x) = (120 - x)/$ $(120 - 99),$ $x \in [99, 120]$ $\mu_{\text{low}}[x] = 0,$ otherwise	$\mu_{\text{normal}}(x) = (x - 110)/$ $(121 - 110),$ $x \in [110, 121]$ $\mu_{\text{normal}}(x) = (132 - x)/$ $(132 - 121),$ $x \in [121, 132]$ $\mu_{\text{young}}(x) = 0,$ otherwise	$\mu_{\text{alert}}(x) = (x - 125)/$ $(143 - 125),$ $x \in [125, 143]$ $\mu_{\text{alert}}(x) = (166 - x)/$ $(166 - 143),$ $x \in [166, 143]$ $\mu_{\text{alert}}(x) = 0,$ otherwise	$\mu_{\text{high}}(x) = (x - 158)/$ $(184 - 158),$ $x \in [125, 143]$ $\mu_{\text{high}}(x) = (207 - x)/$ $(207 - 184),$ $x \in [184, 207]$ $\mu_{\text{high}}(x) = 0,$ otherwise	$\mu_{\text{very high}}(x) = 0$ $x \in [1, 157]$ $\mu_{\text{very high}}(x) = (x - 157)/$ $(186 - 157),$ $x \in [157, 186]$ $\mu_{\text{very high}}(x) = 1,$ $x \in [207, 350]$

Blood sugar (separated into 5 levels)

Low	Normal	Alert	High	Very high
$\mu_{\text{low}}(x) = 1,$ $x \in [1, 65]$ $\mu_{\text{low}}(x) = (80 - x)/$ $(80 - 65),$ $x \in [65, 80]$ $\mu_{\text{low}}[x] = 0,$ otherwise	$\mu_{\text{normal}}(x) = (x - 71)/$ $(127 - 71),$ $x \in [71, 127]$ $\mu_{\text{normal}}(x) = (139 - x)/$ $(139 - 127),$ $x \in [127, 139]$ $\mu_{\text{normal}}(x) = 0,$ otherwise	$\mu_{\text{alert}}(x) = (x - 127)/$ $(155 - 127),$ $x \in [127, 155]$ $\mu_{\text{alert}}(x) = (167 - x)/$ $(167 - 155),$ $x \in [167, 155]$ $\mu_{\text{alert}}(x) = 0,$ otherwise	$\mu_{\text{high}}(x) = (x - 148)/$ $(155 - 148),$ $x \in [148, 155]$ $\mu_{\text{high}}(x) = (250 - x)/$ $(250 - 199),$ $x \in [199, 250]$ $\mu_{\text{high}}(x) = 0,$ otherwise	$\mu_{\text{very high}}(x) = 0$ $x \in [1, 210]$ $\mu_{\text{very high}}(x) = (x - 210)/$ $(250 - 210),$ $x \in [210, 250]$ $\mu_{\text{very high}}(x) = 1,$ $x \in [250, 500]$

Heart pain (separated into 4 levels)

Low	Normal	High	Very high
$\mu_{\text{low}}(x) = 1,$ $x \in [1, 55]$ $\mu_{\text{low}}(x) = (71 - x)/$ $(71 - 55),$ $x \in [55, 71]$ $\mu_{\text{low}}[x] = 0,$ otherwise	$\mu_{\text{normal}}(x) = (x - 66)/$ $(73 - 66),$ $x \in [66, 73]$ $\mu_{\text{normal}}(x) = (98 - x)/$ $(98 - 73),$ $x \in [73, 98]$ $\mu_{\text{normal}}(x) = 0,$ otherwise	$\mu_{\text{high}}(x) = (x - 87)/$ $(123 - 87),$ $x \in [87, 123]$ $\mu_{\text{high}}(x) = (123 - x)/$ $(146 - 123),$ $x \in [123, 146]$ $\mu_{\text{high}}(x) = 0,$ otherwise	$\mu_{\text{very high}}(x) = 0$ $x \in [1, 210]$ $\mu_{\text{very high}}(x) = (x - 172)/$ $(198 - 172),$ $x \in [172, 198]$ $\mu_{\text{very high}}(x) = 1,$ $x \in [198, 314]$

Rest electro (separated into 3 levels)

Alert	Risk	High risk
$\mu_{\text{alert}}(x) = 1, x \in [1]$	$\mu_{\text{Risk}}(x) = 2, x \in [2]$	$\mu_{\text{HighRisk}}(x) = 3, x \in [3]$

Affected

$\mu_{\text{affected}}(x) = 1, x \in [1], \mu_{\text{affected}}(x) = 0,$ otherwise
--

Urine nicotine (separated into 3 levels)		
Alert	Risk	High risk
$\mu_{\text{alert}}(x) = (x - 225) / (243 - 225),$ $x \in [225, 243]$	$\mu_{\text{high}}(x) = (x - 358) / (384 - 358),$ $x \in [358, 384]$	$\mu_{\text{very high}}(x) = 0,$ $x \in [1, 557]$
$\mu_{\text{alert}}(x) = (366 - x) / (366 - 243),$ $x \in [243, 366]$	$\mu_{\text{high}}(x) = (507 - x) / (507 - 384),$ $x \in [384, 507]$	$\mu_{\text{very high}}(x) = (x - 557) / (786 - 557),$ $x \in [557, 786]$
$\mu_{\text{alert}}(x) = 0,$ otherwise	$\mu_{\text{high}}(x) = 0,$ otherwise	$\mu_{\text{very high}}(x) = 1,$ $x \in [631, 1473]$

4 Proposed Algorithm

In this paper, we have used the conception of the fuzzy logic and establish the appropriate weights by normalizing the data set according to the derived fuzzy said rules. These normalized data sets are used to predict the risk level in any human being of cardiovascular disease with the impact of diabetics and/or smoking. These will also recognize the approximate acceleration of the disease in the case of smoking. The nicotine factors are included but a common value is considered instead of the type of smoking (cigarette, cigar, bidi, or huka). We have used the above rules and test it using the functions in SciLab.

Proposed Algorithm:

Step 1 [Preprocessing]

Normalized the given data set with respect to the input parameters fuzzy rules described in the Sect. 3. Let (F_1, D) , (F_2, S) , and (F_f, C) are the fuzzy sets, where (F_1, D) denotes the domination of the diabetics among the patients and (F_2, S) denotes the domination of smoking among the patients data sets and (F_f, D) is defined as $(F_f, D \times S)$, where $F_f(\alpha, \beta) = F_1(\alpha) \tilde{\cap} F_2(\beta), \forall \alpha \in D$ and $\forall \beta \in S$ and $\tilde{\cap}$ is the fuzzy intersection operation of two fuzzy sets.

Step 2 [cardinal set]

Discover the cardinal set from the preprocessed data in Step 1.

Step 3 [Approximation Function]

Congregate the multiple aggregation functions of the symptoms to a single fuzzy set using the following

$$\frac{1}{|S|} \sum_{x \in S} \mu_c \Gamma_c(x) \mu_{\gamma_c}(x)(p)$$

where S is the set of symptoms and p represents the patients with the membership and cardinal functions.

Step 4 [Decision-Making]

Finally, the maximum value has to be selected from the result obtained by

$$e_i = \sum_{j=1}^n p_{ij}$$

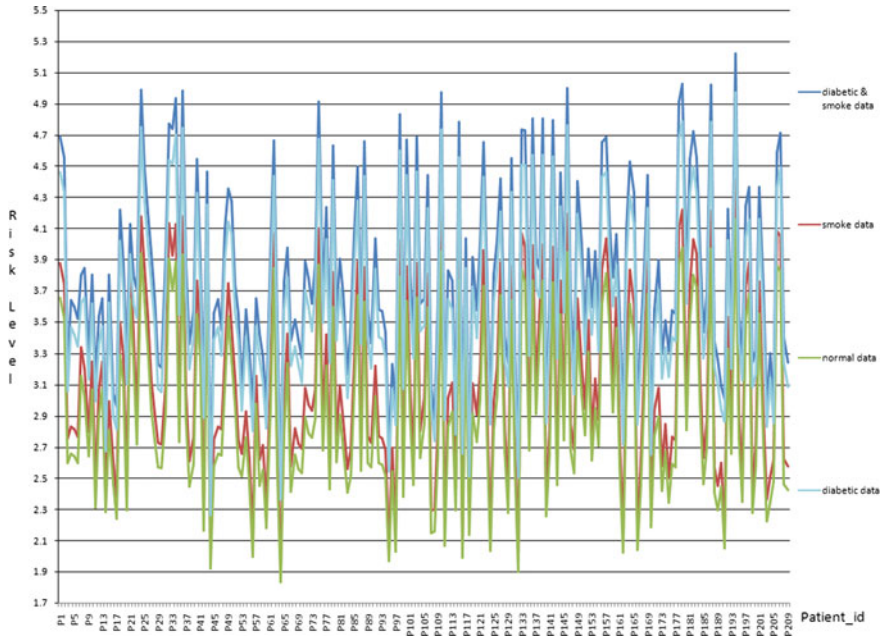
where i represents the signs and j represents the patients affected by that sign.

Thus, the sum of the i th row is represented by e_i and P_{ij} indicates the symptoms in which s_i dominates all the other patients. This will alarm the experts to take the appropriate action immediately.

The above algorithm is to apply on the data set with and without the diabetics as well as the urine nicotine (to find the consequence of smoking), in order to find out the impact of one of the cardiovascular disease. First, we applied the algorithm on the normal data set (data set of the patients with no diabetics and no smoking). Second, the algorithm is applied to the data where the patients are having sugar level (data set of the patients with the diabetics). Third, it is applied on the patients who used to smoke only (data set of the patients not having the diabetic but presence of the nicotine (smoking)) and lastly it is applied to the data of the patients with diabetic and nicotine (smoke). Thus, the algorithm determines the impact of the smoke on the cardiovascular disease with diabetics in the patients. This will not only predict the risk in the patient's life but also supports the experts to take the right and acceptable decision on time so that a person can save his/her precious life and can change the lifestyle accordingly. For this, the experts may treat them (if required) by means of counseling or proper medication.

5 Result Analysis

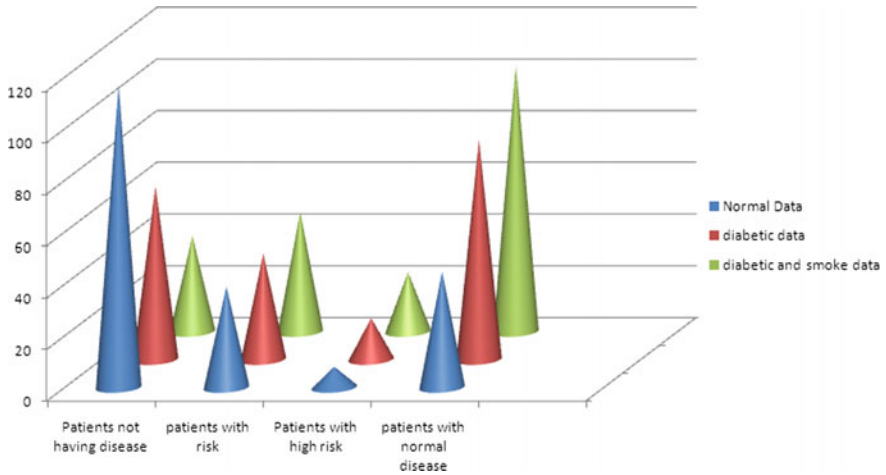
We applied the algorithm in SciLab on the normalized data set. Graph 1 represents the four responsiveness of the algorithm on the UCI data set of cardiovascular disease (in all 209 tuples in the data set, thus 209 patients are considered) which enhanced with the urine nicotine used to measure the smoking factor in the patients. The normal data stands for the input variables except the blood sugar (diabetic) and urine nicotine (smoke). Similarly, the diabetic data set contains the blood sugar, smoke data set consists of the urine nicotine, and diabetic and smoke data set comprise the blood sugar as well as the urine nicotine. Graph 1 represents the normal data (green color), smoke data (red color), diabetic data (light blue color), and smoke and diabetic data (dark blue color). By comparing the facts that diabetic data is more prominent and when it is enhanced with nicotine factor the risk level is at the highest level. The result is evidence for patients accelerating the chances of



Graph 1 Representation of the normal data, diabetic data, smoke data, and diabetic and smoke data

occurrence of the CVD who were not having the CVD initially. Thus, the diabetic and smoke factors augment the risk level of the cardiovascular disease in the patients who even do not have initially the CVD disease. For example, the patient P65 is not affected with the cardiovascular disease but when diabetic as well as smoke factors added the risk factor increased from 1.87 units to 4.26 units with the diabetics and enhanced to 4.63 units when smoking factor added with the diabetics. This illustrates the impact of the diabetic which enhanced it with the additional factor of smoking. This implies that smoking accelerates the risk factor of cardiovascular disease in any human being if he/she is not having the problem at present. On the other hand, the patient is suffering from CVD with the initial value approximately 2.7 augmented to approximately 5.3 with acceleration in the weighted values in presence of the blood sugar and nicotine.

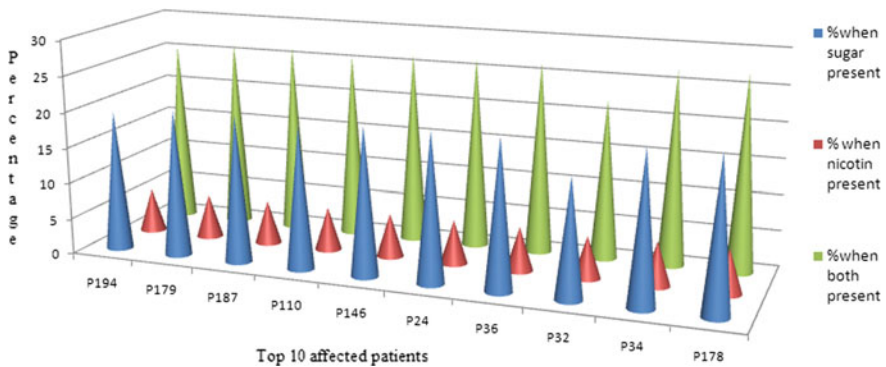
Graph 2 represents the result after the applicability of Step 3 of the algorithm. The approximation function demonstrates the enhancement in the number of patients from normal to risk and risk to high risk factor and more specific the number of the patient with CVD disease increases (earlier it was 117 which were not having the disease, but after the diabetic and with the impact of smoke it remains only 37 who were not yet affected with the cardiovascular disease). This implies that the high risk and risk patients were increased in the ratio of 17.9%, 65.21%, and more observing is 128.88% to the person who was not having the



Graph 2 Number of patients at various stages with respect to data

disease. In the above prediction, the weights of the other input factors were also considered and found that were more prominent with the inclusion of the blood sugar and urine nicotine. Hence, there is a leap in the increased percentage of patients with the risk and not of high risk. This demonstrates that the blood sugar (diabetic) and nicotine (smoke) comprise massive adverse impact on the other factors too. The result congregates the multiple aggregation functions of the symptoms to a single fuzzy set.

Graph 3 represents the patients' who were are at high risk and want an immediate medication. This is the result of Step 4 of the proposed algorithm. The detection of the enhancement in the patient status is based on the multiple-input sign data set. The Step 4 sums up the entire row, and then the difference is to calculate between the row sum and the column sum, in order to find the sign



Graph 3 The top 10 affected patients

overriding on the others. The results showed that the nicotine values were more prominent in the case of the patients with high sugar level. Thus, it is very clear that if a patient is having diabetic, the smoking accelerates the cardiovascular disease. The graph illustrates the top 10 affected patients' records only. It exhibits the amount of the percentage nicotine overrides the blood sugar in respect to the blood pressure, chest pain, angina, report of the electrocardiogram, and heart rate.

6 Conclusion

The concept is used of fuzzy theory which is very effective to handle the dynamic situations because of its characteristics. The proposed algorithm is designed to handle the multiple observations of the input variables which may or may not have the uncertain behavior. The algorithm dealt with the aggregation of the various observation functions based on the observed variables with and without the nicotine (smoke) factor inclusion and finally concludes with the recognition of critical patients who need immediate medication. This will not only facilitate the experts but also helps out in the prediction of the cardiovascular disease on the basis of the current lifestyle of the patient. At last, maybe our experiments results considered to be the starting point for the formation of the fuzzy prediction system for the evaluation of the cardiovascular risk where nicotine is one of the accelerating factors.

The limitation of the paper is of the data considered to be static. In reality, the data is of cumulative nature which augmented in due course time, if appropriately not assessed. Thus, the results will be more worst if the nicotine level amplifies in the body.

References

1. Anooj, P.K.: Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. *J. King Saud Univ. – Comput. Inf. Sci.* **24**(1), 27–40 (2012)
2. Benowitz, N.L., Bernert, J.T., Caraballo, R.S., Holiday, D.B., Wang, J.: Optimal serum cotinine levels for distinguishing cigarette smokers and nonsmokers within different racial/ethnic groups in the United States between 1999 and 2004. *Am. J. Epidemiol.* **169**, 236–248 (2009)
3. D'Agostino, R.B., Vasan, R.S., Pencina, M.J., et al.: General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753 (2008)
4. Elliott, P., Andersson, B., Arbustini, E., et al.: Classification of the cardiomyopathies: a position statement from the European Society of Cardiology Working Group on Myocardial and Pericardial Diseases. *Eur Heart J* **29**, 270–276 (2008)
5. Nomesko, (2013): Health Statistics for the Nordic Countries 100. <http://norden.diva-portal.org/smash/get/diva2:941584/FULLTEXT01.pdf>

6. Onor, I.O., Stirling, D.L., et al.: Clinical effects of cigarette smoking: epidemiologic impact and review of pharmacotherapy options. *Int. J. Environ. Res. Public Health* **14**, 1147 (2017). <https://doi.org/10.3390/ijerph14101147>
7. Iftikhar, A.R., Ahmad, M., Siddiqui, A.S.: A study on smoking problem using fuzzy matrix method. *Int. J. Math. Arch.* **7**(1), 147–152 (2016)
8. Keto, J., et al.: Cardiovascular disease risk factors in relation to smoking behaviour and history: a population-based cohort study. *Open Heart* **3**, e000358 (2016). <https://doi.org/10.1136/openhrt-2015-000358>
9. Anbe, J., Tobi, T.: Fuzzy set system application to medical diagnosis: a diagnostic system for valvular heart diseases. *Fuzzy Theory Syst* **2**, 937–956 (1999)
10. Oad, K. K., DeZhi, X.: A fuzzy rule based approach to predict risk level of heart disease. *Glob. J. Comput. Sci. Technol: C Softw. Data Eng.* **14**(3.1), 17–22 (2014)
11. Neshat, M., Yaghoobi, M., Naghibi, M.B., Esmaelzadeh, A.: Fuzzy expert system design for diagnosis of liver disorders. In: *International Symposium on Knowledge Acquisition and Modeling*, pp. 252–256 (2008)
12. Matsue, Y., Suzuki, M., Nakamura, R., et al.: Prevalence and prognostic implications of pre-diabetic state in patients with heart failure. *Circ. J.* **75**, 2833–2839 (2011)
13. Phuong, N.H., et al.: Fuzzy logic and its applications in medicine. In: *Proceedings of Asian Pacific medical informatics conference APAMI-MIC'2000, Hong Kong, Sep 27–30*, pp 1–11 (2000)
14. Ponikowski, P., Voors, A.A., Anker, S.D., et al.: ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur. J. Heart Fail.* **18**, 891–975 (2016)
15. Rajeswari, K., Vaithyanathan, V.: Heart disease diagnosis: an efficient decision support system based on fuzzy logic and genetic algorithm. *Int. J. Decis. Sci. Risk Manag.* **3**(1–2), 81–97 (2011)
16. Wild, S., Roglic, G., Green, A., et al.: Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* **27**(5), 1047–1053 (2004)
17. Kumar, S., Kaur, G.: Detection of heart diseases using fuzzy logic. *Int. J. Eng. Trends Technol. (IJETT)* **4**(6), 2694–2699 (2013)
18. Phongsuphap, S., Pongsupap, Y.: Evaluation of responsiveness of health systems using fuzzy-based technique. In: *IEEE International Conference on Fuzzy Systems*, pp. 1618–1623 (2014)
19. Mahmood, S.S., et al.: The Framingham Heart Study and the epidemiology of cardiovascular diseases: a historical perspective. *Lancet* **383**(9921), 999–1008 (2014)
20. U.S. Department of Health and Human Services. *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA, USA, pp. 1–36 (2014)
21. United States Department of Health and Human Services. *How Tobacco Smoke Causes Disease: A Report of the Surgeon General*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA, USA (2010)

An Integrated Fault Classification Approach for Microgrid System



Ruchita Nale, Ruchi Chandrakar and Monalisa Biswal

Abstract In this paper, a moving windowing approach-based integrated fault classification algorithm is proposed for microgrid system. In a microgrid system, the nonlinear operation of control devices connected to distributed generation (DG) imposes problem for identifying the exact faulty class. In order to mitigate this issue, an integrated moving window averaging technique (IMWAT) is proposed. The method utilizes current signal at the line end. In this technique, first, the decision of the fault detection unit (FDU) is analyzed and based on that fault class is detected. The FDU uses the conventional moving window averaging technique. Different logics are framed to identify the symmetrical and unsymmetrical faults. The method is tested on a standard microgrid network and obtained results for different fault cases prove the efficacy of the proposed method.

Keywords Close-in fault · Fault classification · High resistance fault · Microgrid system · Moving window averaging technique · PV system

1 Introduction

Microgrid system is emanating as a primary part of distribution network owing to the technological advancement in distributed generation (DG) system [1–6]. It consists of various renewable energy sources such as wind power, fuel cells, and

R. Nale · R. Chandrakar · M. Biswal (✉)
Department of Electrical Engineering, National Institute of Technology, Raipur 492010,
Chhattisgarh, India
e-mail: mbiswal.ele@nitrr.ac.in

R. Nale
e-mail: ruchita0119@gmail.com

R. Chandrakar
e-mail: ruchichandrakar14@gmail.com

photovoltaic cells providing a means to reduce greenhouse gas emissions and to supply clean and continuous power. These DG sources can be employed near loads for reliable power supply as it decreases the power transmission losses and enhances the energy efficiency of the power system [7–9]. Microgrids can be operated in two modes, i.e., when connected to grid and also when it is isolated from the main grid (islanded mode). The most significant feature of microgrid is its ability to operate in autonomous mode when there is fluctuation in voltage, deviation in frequency or when there is a fault in main grid. Hence, ensuring continuity of supply to critical loads. Enhanced service quality, improved reliability, and efficiency are the common advantages offered by the microgrid [10]. However, usage of different types of DG for generation of power poses challenges for operating, integrating, controlling, and protecting the microgrid [11–16].

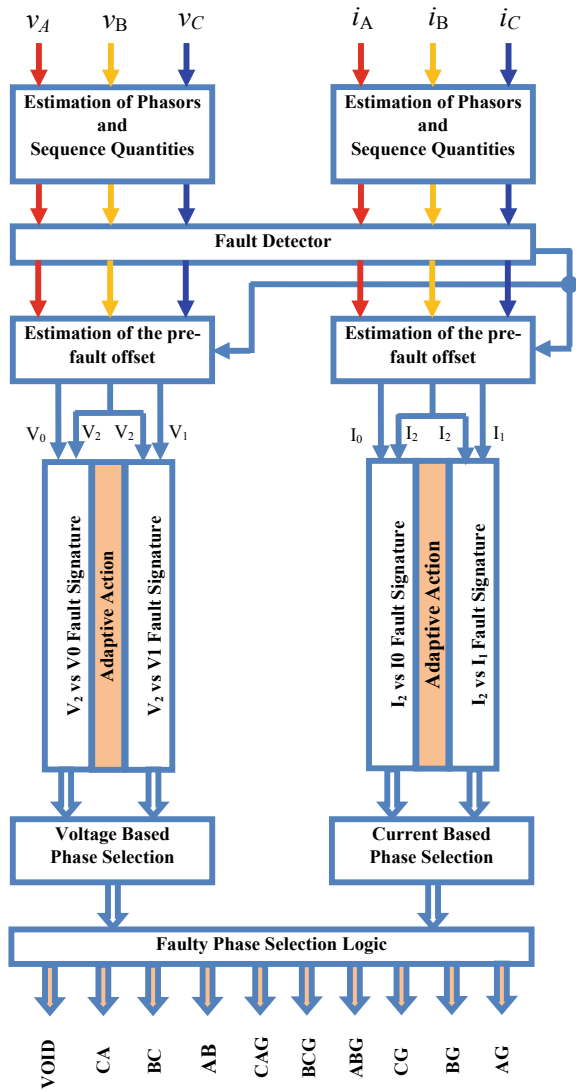
Fault detection is an essential and critical issue in microgrid system because of the nonlinear operation of control devices connected with renewable sources and due to the wide variation in fault current magnitude in different operating modes such as grid-connected mode and islanded mode. Also, the presence of different types of DGs affect the magnitude of fault current. Owing to high penetration of inverter interfaced DG protection of microgrid is a tedious task. The contribution of fault current by inverter interfaced DGs is limited to two to three times the rated current [17].

Different solutions have been provided by different researchers to detect and classify the types of fault in microgrid system. The diagram of a general fault classification algorithm used in protective relay is provided in Fig. 1. In this diagram, the basic procedure for fault classification function performed by relay is presented. With the help of fault generated voltage and current signal, different sequence components are calculated and the relative angle between two quantities are used to estimate the exact faulty phase.

In [18], two new fault classification logics are proposed to identify the accurate faulty phase in microgrid system. The first method is based on voltage angle and magnitude and the second method used voltage angle to identify the fault class. A new approach-based on optimal wavelet functions is proposed to detect the exact faulty class and reported in [19].

In this paper, a new method is proposed. The method operates on the basis of two integrated logics. This IMWAT method overcomes the drawbacks of existing moving window averaging which fails to work correctly during high resistance fault, nonlinear operation of control devices connected to wind turbines, switching of nonlinear loads, and mode changing operation of microgrid. The proposed method is tested on a standard microgrid system, i.e., Aalborg distribution system model. The power system model is developed in EMTDC/PSCAD environment. The response of the method is verified for different symmetrical faults, unsymmetrical fault, high resistance fault, close-in fault, switching of nonlinear loads, and

Fig. 1 Fault classification algorithm used in protective relay



different operating modes. From the analysis report, it is confronted that the proposed method is able to identify the exact faulty phase within half-cycle time period from the instant of fault occurrence and the decision time is fixed irrespective of the system and fault conditions.

The main body of the paper is organized as follows. In Sect. 2 fault classification based on IMWAT approach is described. Next, simulation results for different test conditions are provided in Sect. 3. The conclusion of the paper is provided in Sect. 4.

2 Fault Classification in Microgrid

2.1 Moving Window Averaging Technique

During transient phenomenon or fault scenarios, the average of one window current signal is a non-zero quantity. With suitable threshold value exact fault detection is possible. In this work, a moving window averaging technique based on fault detector unit is considered for processing the fault classification task using IMWAT technique.

Using moving window averaging technique, the magnitude of one cycle current samples can be modeled as

$$i_{MW}(s) = \frac{1}{N} \sum_{p=s-N+1}^s i(p) \quad (1)$$

where s is the sampling instant, p is the instantaneous sample, and N is the number of samples per cycle.

The relation in recursive form can be represented as

$$i_{MW}(s) = i_{MW}(s-1) + i(s) - i(s-N) \quad (2)$$

Using (2), a fault detector is developed in the proposed method. The decision of the FDU is employed in the classification algorithm to classify the exact faulty phase in microgrid system. From the several advantages, detection capability during high resistance fault is a unique feature of moving window averaging technique. Under far-end high resistance fault, the faulty phase current will be very less and can be in the same order as healthy phase. Even though the scenario is rare, but with the arrival of such an event, moving window averaging technique will fail to detect exact faulty phase within the stipulated time period. Such a case is undesirable and degrade the performance of protection function. From the study, it is observed that under faulty condition, the amplitude of affected phase moving window average value is more as compared to other phases if an observation is drawn on the basis of a fixed half-cycle calculated value just after fault inception. This unique feature can be incorporated for the detection of exact faulty phase in microgrid system. In order

to further enhance the reliability of relay, an integrated approach-based fault classification technique is devised that integrates the decision of FDU along with other computed paradigms which are described in the next section.

2.2 IMWAT Algorithm [19]

The proposed algorithm for classification of faults is as follows:

- (1) The value of i_{MW} is computed using Eq. (1) by considering current sample for one cycle.
- (2) Using recursive update formula mentioned in Eq. 2, value of i_{MW} for each phase is calculated.
- (3) For detection of faults, threshold (β_t) is chosen to prevent false operation of relaying. For any further deviation above the threshold value, the threshold counter (j) counts 1. Each counter is provided for each phase.
- (4) Check if $|i_{MW}| \geq \beta_t$

If yes, $j = 1$, otherwise 0. Then $j_{new} = j_{old} + 1$.

Then threshold counter continues to evaluate for consecutive three samples. If any time $j = 0$ comes in-between counter again reset to zero otherwise continue counting.

If $j_{new} = 3$, FDU = 1 otherwise 0.

The aforementioned process is performed for the fault detection task in three phases. FDU is reset to zero once the fault is detected. Next, the further steps of fault classification are evaluated and are explained below.

- (5) Next the residual current through the ground can be calculated using formula,

$$i_g(s) = (|i_a(s) + i_b(s) + i_c(s)|) \quad (3)$$

In (3), i_a , i_b , and i_c represent the instantaneous current for phase a, b, and c, respectively. i_g aids in distinguishing ground fault from phase fault.

- (6) Along with fault detection, the developed fault classification algorithm evaluates, $|i_{MW_a}|$, $|i_{MW_b}|$, $|i_{MW_c}|$, the summation and difference of two-phase moving sum current based on continuous half-cycle considering from the moment of occurrence of fault.

For example, the fault is detected at s th sample, so the proposed method initiates computing above data from $(s - 2)$ th to $(s + 7)$ th sample. The observation duration

is kept as half-cycle for better comparison and correct classification of different faults.

The computational procedure of consecutive phase summation and difference are provided below.

$$\begin{aligned} MS_{ab}(s) &= ||i_{MW_a}(s)| + |i_{MW_b}(s)|| \\ MS_{bc}(s) &= ||i_{MW_b}(s)| + |i_{MW_c}(s)|| \\ MS_{ca}(s) &= ||i_{MW_c}(s)| + |i_{MW_a}(s)|| \end{aligned}$$

Similarly,

$$\begin{aligned} MD_{ab}(s) &= ||i_{MW_a}(s)| - |i_{MW_b}(s)|| \\ MD_{bc}(s) &= ||i_{MW_b}(s)| - |i_{MW_c}(s)|| \\ MD_{ca}(s) &= ||i_{MW_c}(s)| - |i_{MW_a}(s)|| \end{aligned}$$

For each calculated value of i_g , $|i_{MW}|$, MS , and MD a data matrix of size 10×10 is devised. Suppose the data matrix of size $n \times n$ is denoted by D_{nn} , where n is equal to 10. D_{nn} can be expressed as,

$$D_{nn} = \begin{bmatrix} D_{11} & \dots & \dots & D_{n1} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ D_{1n} & \dots & \dots & D_{nn} \end{bmatrix}$$

where $D_{11} = i_{g1}$, D_{21} to D_{n1} are $|i_{MW_a}|_1, |i_{MW_b}|_1, |i_{MW_c}|_1$,

$$MS_{ab_1}, MS_{bc_1}, MS_{ca_1}, MD_{ab_1}, MD_{bc_1}, \text{ and } MD_{ca_1}.$$

- (7) From each column indices are selected. For example, the maximum value in column 1 is $f1$ which denotes the maximum of i_g from a data matrix of size 10. As the calculated values are not constant so direct comparison is not possible, hence index values are selected.

Similarly, $f2 = \max(|i_{MW_a}|)$, $f3 = \max(|i_{MW_b}|)$, $f4 = \max(|i_{MW_c}|)$, $f5 = \max(MD_{ab})$, $f6 = \max(MD_{bc})$, $f7 = \max(MD_{ca})$, $f8 = \max(MS_{ab})$, $f9 = \max(MS_{bc})$, $f10 = \max(MS_{ca})$. The different indices are represented as $f1$ to $f10$ that will compare among each similar group with minim, maxim function. Rules for each fault type is developed and is given in Table 1.

Table 1 Fault classification logic

Fault Type	Logic-1	Logic-2	Logic-3	Logic	Output
	$f1$	$\text{minim}(f5, f6, f7)$	$\text{minim}(f8, f9, f10)$		
a-g	$\geq \beta_t$	$f6$	$f9$	AND	1
bg		$f7$	$f10$		2
c-g		$f5$	$f8$		3
		$\text{minim}(f5, f6, f7)$	$\text{maxim}(f8, f9, f10)$		
ab-g	$\geq \beta_t$	$f5$	$f8$	AND	4
bcg		$f6$	$f9$		5
cag		$f7$	$f10$		6
		$\text{minim}(f2, f3, f4)$	$\text{minim}(f5, f6, f7)$		
ab	$\leq \beta_t$	$f4$	$f5$	AND	7
bc		$f2$	$f6$		8
ca		$f3$	$f7$		9
abc		$f2 \geq \beta_t$ AND $f3 \geq \beta_t$ AND $f4 \geq \beta_t$			

As mentioned in Table 1, if all the rules (rule 1,2 and 3) are satisfied at the same time for any fault in the line for which AND logic is applied then a particular fault case is recorded. For example, a fault occurred with $f1 \geq \beta_t$, AND $\text{minim}(f5, f6, f7) = f6$ AND $\text{minim}(f8, f9, f10) = f9$ then output is 1, i.e., a-g fault. To avoid ambiguity in decision of the relay during single-line-to-ground and double-line-to-ground fault two different functions are utilized. In double-line-to-ground fault both minim and maxim function is used, whereas in single-line-to ground fault only minim function is employed. Hence, Table 1 indicated that a secure fault classification algorithm can be devised using integrated moving sum technique.

3 Simulation Results

The performance of the proposed technique is evaluated by considering a standard microgrid system shown in Fig. 2. It is a part of a distribution network, owned by Himmerlands Elforsyning, in Aalborg, Denmark. The system consists of three wind turbine generators and one combined heat and power plant unit. Generator data, line data, and grid data are considered from [20]. To test the efficiency of the proposed fault classification approach in the presence of various types of distributed generation units, the system is modified by replacing DG3 with photovoltaic (PV) system at bus 14. The nominal power generation of DFIG wind turbine generator and PV is

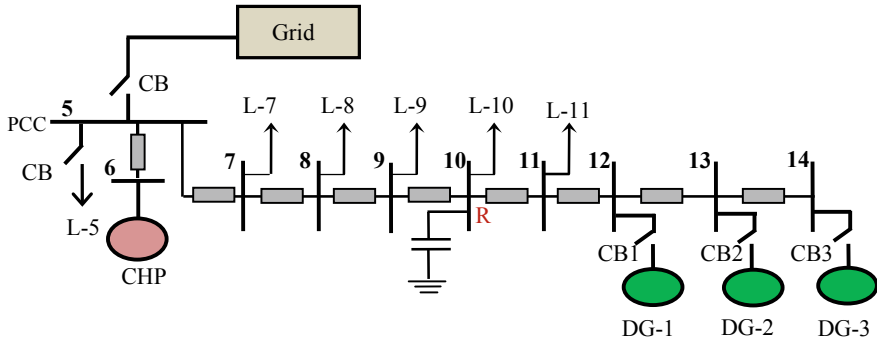


Fig. 2 Standard distribution network

2 MW and 630 kW, respectively. The simulation of the given network is performed in EMTDC/PSCAD. MATLAB software is used for logic development and the response of the technique is evaluated for different fault cases such as unsymmetrical and symmetrical faults, high resistance faults, close-in faults, nonlinear switching of load, and change in operating mode of microgrid system. The obtained results from each case are described below.

3.1 Single-Line-to-Ground Fault (LG)

In distribution system, LG faults are frequently occurring faults. To assess the potential of the proposed technique, a-g fault is simulated considering 5Ω as the fault path resistance at 0.3 s in-between bus 10–11. The response of both fault detector unit and IMWAT are presented in Fig. 3. The computed absolute values of the average of one window current signal for each phase are shown in the figure. For phase-a, the measurement for three consecutive computed values is greater than the threshold, then the Counter (j) is set to 3. Hence, the output of FDU for all the three phases is $[1, 0, 0]$ at 0.303 ms and is depicted in Fig. 3(b). So, the algorithm takes 3 ms to detect the fault using IMWAT. The proposed method further evaluates the value of i_g , MD, and MS as shown in Fig. 3. From the plots, it is noticed that $f1 \geq \beta_t$ AND $\min(f5, f6, f7) = f6$ AND $\min(f8, f9, f10) = f9$. Since all the conditions for a-g fault scenario are satisfied concurrently, the decision provided by relay is 1 which is correct.

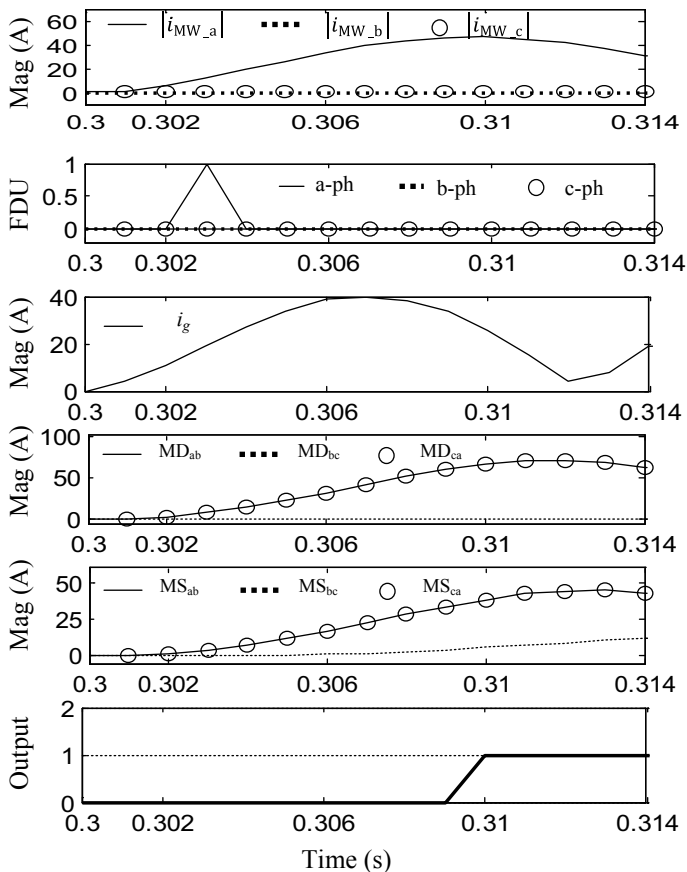


Fig. 3 Performance of FDU and proposed method during a-g fault

3.2 High Resistance Double-Line-to-Ground Fault

Next, to observe the response of proposed IMWAT technique for high resistance double-line-to-ground fault scenario, ab-g fault is created in-between bus 10–11 with a fault resistance of 150 Ohm. Fault is created at 0.3 s. The performance of relay for fault detector unit and proposed method is recorded and is depicted in Fig. 4. During far-end high resistance fault condition, the response of FDU for each phase is “1”. As a consequence, the decision interpreted by the relay is incorrect. Hence, the information provided by moving window averaging approach is not adequate for correct faulty phase selection. In order to achieve the exact

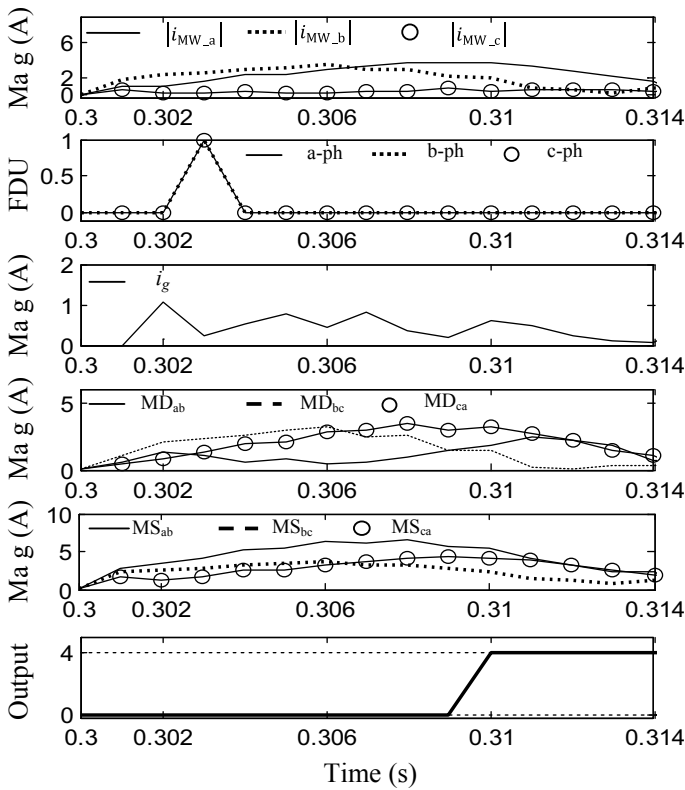


Fig. 4 Performance of FDU and proposed method during far-end high resistance a-b-g fault

classification of faults, the performance of moving window average-based FDU can be enhanced by employing the proposed IMWAT technique. So, by the integration of various proposed logics, the relay provides decision as “4”, i.e., ab-g fault which is correct.

3.3 Double Line Fault

For the simulation of double line fault to observe the response of the proposed method, a b-c fault in-between bus 10-11 is created at 0.3 s. The fault resistance is considered as 5 Ohm. The response of the FDU and the proposed method are shown in Fig. 5. From the response of FDU it is cleared that, FDU only detects the

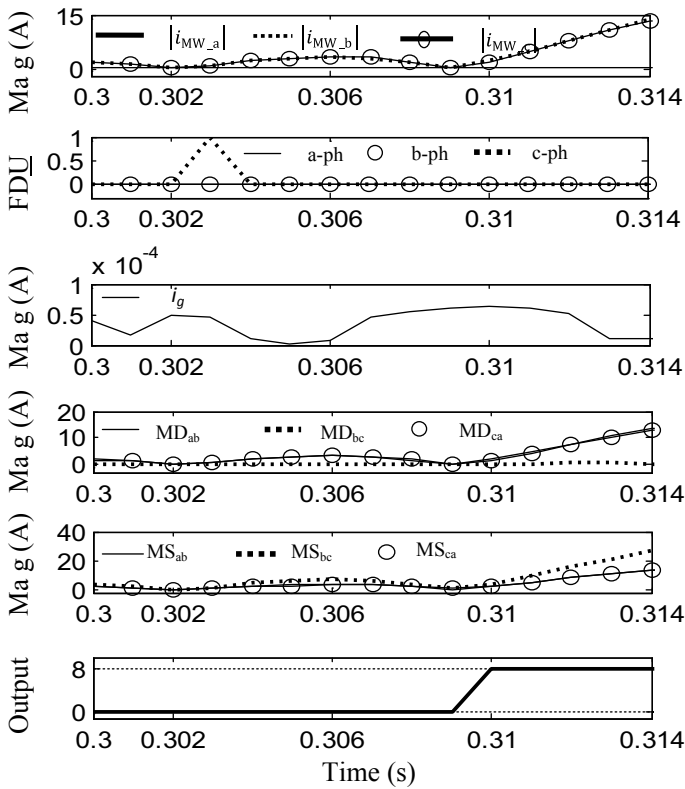


Fig. 5 Performance of FDU and proposed method during b-c fault

phase-c for b-c fault scenario. But the output of the proposed method is “8”, which indicates the occurrence of b-c fault in microgrid system.

3.4 Close-in Fault

Close-in faults are more severe because the higher fault current magnitude deteriorates the performance of current transformer. So, relay finds more challenge during close-in fault. To simulate this condition, three-phase fault is simulated in-between bus 10–11 at 0.3 s. The fault resistance is 2 Ohm. By moving sum approach fault is detected at 0.303 s I all the three phases. The output of FDU for all the three phases will be [1]. From Fig. 6, $f1 \geq \beta_t$ AND $f2 \geq \beta_t$ AND $f3 \geq \beta_t$ AND $f4 \geq \beta_t$ this provides an output of 10, i.e., three-phase fault by IMWAT approach. Therefore,

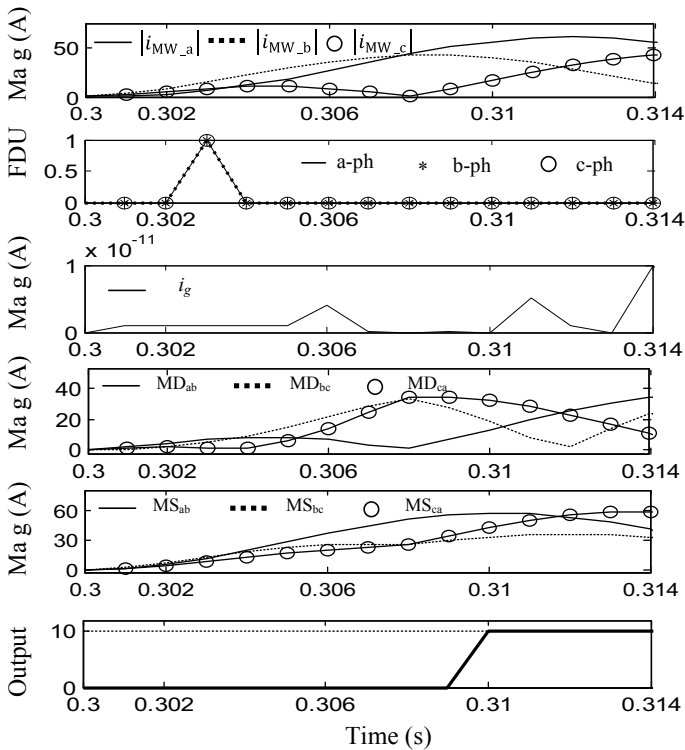


Fig. 6 Performance of FDU and proposed method during close-in three-phase fault

the algorithm can correctly classify close-in three-phase fault which is clear from the result.

3.5 Fault During Switching of Nonlinear Load

The application of electronic gadgets such as personal computer, laptop, printers, uninterrupted power supply, and music instruments, the generation of harmonics in distribution level is more. This may lead to false operation of fault classification algorithm. So, to verify the impact of the proposed method during switching of nonlinear loads, a 0.5 MW three-phase diode rectifier with resistive load is switched on at 0.3 s. At the same time, c-g fault is created in the line between bus 10–11.

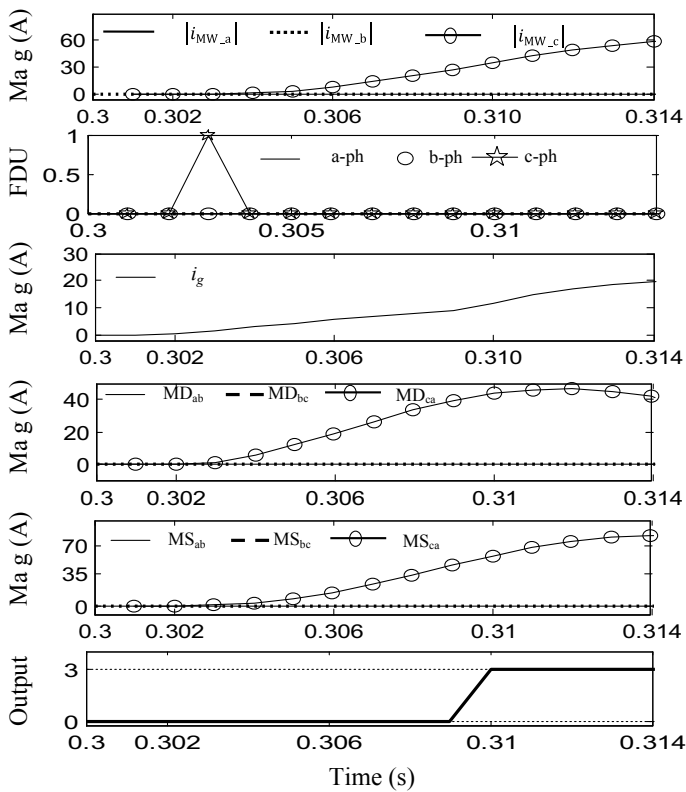


Fig. 7 Performance of FDU and proposed method during nonlinear load switching

As depicted from the output response of the proposed method shown in Fig. 7, the switching of nonlinear load has no significant effect on the proposed method.

3.6 Fault During Different Operating Modes

All the abovementioned results are provided for grid-connected mode. So, to verify the response of the proposed method during islanding mode, the PCC bus breaker, i.e., CB-4 is kept open. For this islanding case, an a-g fault is created at 0.305 s. The response of the proposed method for this case is shown in Fig. 8. From this figure, it is clear that the response of the method is also accurate for islanded mode as the final output by IMWAT technique is “1” which is correct.

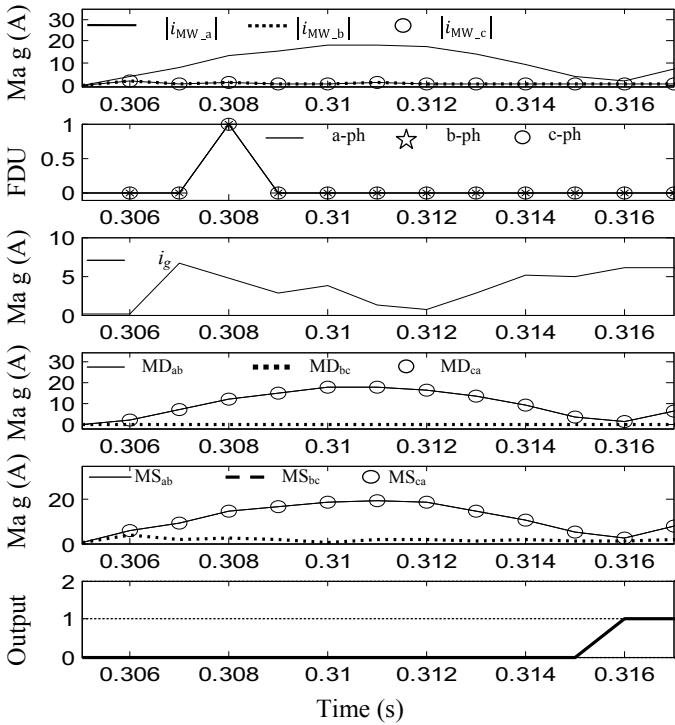


Fig. 8 Performance of FDU and proposed method for islanded mode of operation

So, the response of the method is verified for different critical cases and it is noticed that within one cycle time period accurate decision can be drawn irrespective of the critical operating conditions of fault and system.

4 Conclusion

Exact fault type selection is a challenging task in microgrid system due to the incorporation of renewable sources in the existing passive distribution network. The nonlinear operation of control devices, switching of nonlinear load and changing mode of operation from grid-connected mode to islanded mode is the typical critical operating condition of microgrid system during which relay find challenges in discriminating the fault types. The other fault conditions such as close-in fault and high resistance fault may degrade the performance of existing protective relay

algorithms. To mitigate this technical and fault generated issues, a new integrated approach is proposed in this paper. In the integrated approach, the output of moving window averaging-based FDU is incorporated in the fault classification logic to obtain accurate decision. For ground fault detection residual current through ground is computed. Different indices are calculated for the fault class selection. The response of the proposed method is verified for various cases. It is observed from the results that within half-cycle time period exact fault type selection is possible in microgrid system.

References

1. IEEE Standard for Interconnecting Distributed Resources with Electric Power Systems. IEEE, Standard 1547-2003 (2003)
2. Impact of Increasing Contribution of Dispersed Generation on the Power System. CIGRE, Working Group 37.23 (1999)
3. G59/1 Recommendations for the connection of Embedded Generating Plant to the Regional Electricity Companies Distribution Systems, Electricity Association standard (1991)
4. Xu, W., Zhang, G., Li, C., Wang, W., Wang, G., Kliber, J.: A power line signaling based technique for anti-islanding protection of distributed generators-part I: scheme and analysis. *IEEE Trans. Power Deliv.* **22**(3), 1758–1766 (2007)
5. Wang, W., Kliber, J., Zhang, G., Xu, W., Howell, B., Palladino, T.: A power line signaling based scheme for anti-islanding protection of distributed generators—part II: field test results. *IEEE Trans. Power Deliv.* **22**(3), 1767–1772 (2007)
6. Ahmad, K.N.E.K., Selvaraj, J., Rahim, N.A.: A review of the islanding detection methods in grid-connected PV inverters. *Renew. Sustain. Energy Rev.* **21**, 756–766 (2013)
7. PVPS IEA: Evaluation of islanding detection methods for photovoltaic utility interactive power systems. Report IEA PVPS T5-09 (2002)
8. Velasco, D., Trujillo, C.L., Garcera, G., Figueres, E.: Review of anti-islanding techniques in distributed generators. *Renew. Sustain. Energy Rev.* **14**(6), 1608–1614 (2010)
9. Timbus, A., Oudalov, A., Ho, C.: Islanding detection in smart grids. In: *Energy Conversion Congress on Exposition (ECCE)*, pp. 3631–3637 (2010)
10. Nale, R., Biswal, M.: Comparative assessment of passive islanding detection techniques for microgrid. In: *Proceedings of International Conference on Innovations Information, Embedded and Communication Systems, Coimbatore*, pp. 1–5 (2017)
11. Nale, R., Biswal, M., Kishor, N.: A transient component based approach for islanding detection in distributed generation. In: *IEEE Transactions on Sustainable Energy (Early Access)* (2018)
12. Ray, P.K., Kishor, N., Mohanty, S.R.: S-transform based islanding detection in grid-connected distributed generation based power system. In: *Proceeding of IEEE International Energy Conference and Exhibition (EnergyCon)*, pp. 612–617 (2010)
13. Ray, P., Mohanty, S., Kishor, N.: Disturbance detection in grid connected distribution system using wavelet and S-transform. *Electr. Power Syst. Res.* **81**(3), 805–819 (2011)
14. Mohammadzadeh Niaki, A.H., Afsharnia, S.: A new passive islanding detection method and its performance evaluation for multi DG systems. *Electr. Power Syst. Res.* **110**, 180–187 (2014)
15. Mishra, M., Sahani, M., Rout, P.K.: An islanding detection algorithm for distributed generation based on Hilbert-Huang transform and extreme learning machine. *Sustain. Energy Grids Netw.* **9**, 13–26 (2017)

16. Mahat, P., Chen, Z., Bak-jensen, B.: Review on islanding operation of distribution system with distributed generation. In: Power and Energy Society General Meeting, pp. 1–8 (2011)
17. Keller, J., Kroposki, B.D.: Understanding fault characteristics of inverter-based distributed energy resources. Nat. Renew. Energy Lab., Golden, CO, USA, Tech. Rep. NREL/TP-550-46698 (2010)
18. Abddelagayed, T.S., Morsi, W.G., Sidhu, T.S.: A new approach for fault classification in microgrids using optimal wavelet functions matching pursuit. *IEEE Trans. Smart Grid* 9(5) (2018)
19. Biswal, M.: Faulty phase selection for transmission line using integrated moving sum approach. *IET Sci. Meas. Technol.* **10**(7), 761–767 (2016). <https://doi.org/10.1049/iet-smt.2016.0081>
20. Mahat, P., Chen, Z., Bak-Jensen, B.: A hybrid islanding detection technique using average rate of voltage change and real power shift. *IEEE Trans. Power Deliv.* **24**(2), 764–771 (2009)

Role of Data Analytics in Human Resource Management for Prediction of Attrition Using Job Satisfaction



Neerja Aswale and Kavya Mukul

Abstract The reputed management publications like Harvard Business Review (HBR) have started stressing upon the emergence of data-driven management decisions. The enhancing investments in data and analytics are underlining the aforementioned emergence. According to International Data Corporation, this investment is expected to grow up to \$200 billion by 2020. In such a data lead management world collecting, managing, and analysing the human resources-related data becomes a key for any rather every organization. Human resource analytics is changing into necessary as strategic personnel designing is the need of the hour and helps organizations to investigate each side of HR metrics. HR analytics could be a holist approach. According to KPMG—India’s Annual Compensation Trends Survey 2018–19 the average annual voluntary attrition across sectors is 13.1%. This is a considerably high percentage. Hence, antecedents leading to attrition are needed to be explored in order to propose appropriate HR policies, strategies, and practices. In relevance to these facts, this study focused on proposing a data-driven predictive approach that examines the relationship between the attrition (dependent variable) and other demographic and psychographic independent variables (Antecedents). The present study found that there is a strong relationship between job satisfaction and attrition. Further, there is a higher probability that the employees having work experience between 0–5 years may leave the organizations. Such data-based outcomes may offer help to HR managers in addressing the problems like attrition which intern may increase ROI. Thus, this paper underlines the emergence and relevance of analytics with special reference to human resource management domain.

Keywords Analytics · Human resources management · Attrition · Job satisfaction

N. Aswale (✉) · K. Mukul
MIT School of Technology Management, Vishwashanti Marg, Rambaug Colony, Kothrud,
Pune, Maharashtra 411038, India
e-mail: neerjaaswale@gmail.com

K. Mukul
e-mail: kavya.mukul21@gmail.com

1 Introduction

The emergence of technologies has changed the way the business domains used to be operated. Taking these developments into consideration it has become important to accept that the various business functions have started deploying the technological advancements. The human resources management is no longer an exception to this phenomenon. Therefore, the “Human resource analytics (HR analytics) is a vicinity in the subject of analytics. The HR analytics refers to application of analytic techniques to the functions of human resource department. The major emphasis of HR department is to achieve enhanced productivity by increasing individual performance of the employee. The increased individual performance will subsequently result into a higher return on investment.” The modern philosophy of human resource management has shifted focus from Human resources to Human Capital. This has added roles and responsibilities of HR professionals. The modern HR professionals are suppose to predict futuristic standpoints to the organization using analytics. This shift demands proper exploration, examination, and evaluation of individual performances. This approach will further help for talent management practices that focus on retention and acquisition of human resources. In such cases, the data-based evidence would help HR managers in designing the HR strategies. These evidence/trends can be analyzed using analytics that includes application modern business analytics techniques like data mining to human resources data. In continuation to this introductory insights the next section details about the need for using HR analytics for strategic HR decision-making. Further, it also informs about studies conducted in HR domain for understanding the application of HR analytics.

Being a multidisciplinary method, HR analytics combines methodology that improves the people-related selections to enhance person and company’s performance. HR analytics is also known as people analytics, workforce analytics, and talent analytics. The HR analytics introduces high-end predictive modeling approaches that forecast the effect of changing the policies. Further, the HR analytics has become an integral part of human resources functions like recruitment, training, development, succession planning, compensation, benefits, retentions, and engagement. Traditionally, analytics is being used to focus on the rate of turnover and the cost of hiring. Using complex statistical analyses, HR analytics enables corporations to measure the business impact of people policies and predicts the future of the workforce. This allows managers to measure the relationship between financial performance of the organization and the human resource practices.

Muscalu and Serban [1] examined the relevance of HR analytics for strategic human resource management. The authors deployed HCM: 21 model to know the effectiveness of the department. Momin and Taruna [2] reported that the growing HR-related problems such as succession planning, recruitment, and employee retention. Sujeet et al. [3] mentioned that Human Resource Predictive Analytics (HRPA) was found useful for all HR functions. Bindu [4] conducted a review emphasizing on HR analytics skills and as a result of focus group discussions reported that bad records would lead to negative HR analytics, whereas business

analytics provides a multidimensional method toward constructing effective HR strategies [5]. Globcon Technologies Pvt. Ltd [6] reported that application of HR analytics has opened emerging niche areas in HR analytics as certain specific problems get addressed by means of HR analytics. Rajbhar et al. [7] concluded that HR analytics helps in improving workers overall performance in a team. This study also mentions that HR analytics can be used for increasing the productivity. Shweta [8] concluded that applying analytic procedures to the human resource data help in enhancing worker performance and consequently getting a higher return on investment.

2 Methodology

It was observed during literature review that the role of HR analytics is very well explored in developed countries and hardly explored for developing countries like India. Further, the practical relevance of HR analytics is hardly explored. Even though these study mentioned about relevance of HR analytics, a few studies have explored a specific phenomenon like attrition. This has created a scope for the present study that emphasized on attrition and explored the various factors related to this phenomenon. To have a systematic approach this study defined certain research objectives. These objectives are as follows:

- (a) To study the factors affecting attrition.
- (b) To predict the factors influence attrition.
- (c) To identify factors affecting job satisfaction.
- (d) To make recommendations for retention of employees on the basis of analysis.

H0: Attrition is directly co-related to job satisfaction.

H1: Attrition is not directly co-related to job satisfaction.

HR analytics is defined as “a method for evaluating and understanding the causal relationship between HR practices and overall performance results of organizations (such as client satisfaction, sales, or profit)”. It also helps in supplying authentic and reliable foundations for human capital choices for integrating enterprise strategy and performance. This integration could be achieved using statistical strategies and experimental tactics based on various metrics of efficiency, effectiveness, and competitive advantage. Therefore, this section details the relevance of using data analytics for the HR functions. The prime advantages reported by the past researchers are as follows:

- (a) HR analytics is affordable and provides the data-based evidence for both strategic and operational decisions.
- (b) The enhanced technological advances also provided the access to big data related to human resource which makes use of HR analytics relevant.

- (c) It also helps in tackling the global talent warfare by enhancing both talent acquisition and retention.

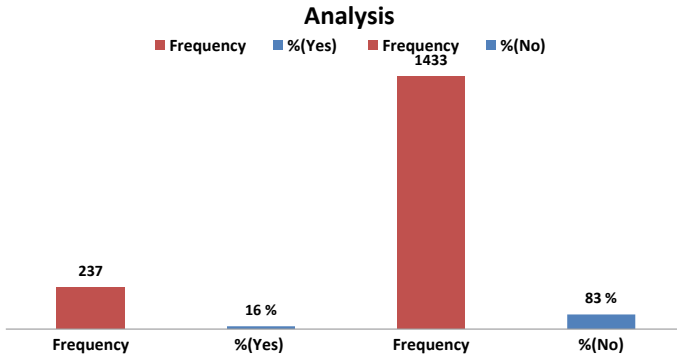
Further, there are three levels prominently used for analysis :

- I. **Descriptive:** Descriptive HR analytics reveals and describes relationships. This type of analysis compares the present scenario with historical data patterns in order to generate insights. It consists of dashboards and score-cards; team of workers segmentation; statistics mining for main patterns; and periodic reports. Traditional HR metrics are mostly successful metrics (turnover rate, time to fill, value of hiring, volume employed and trained, etc.). The predominant center of attention is on cost reduction and process improvement.
- II. **Predictive:** Predictive analysis covers an array of techniques (statistics, modeling, information mining) that use present day and historical facts to make predictions about the future policies, strategies, and practices. It deals with identifying probabilities and dealing with the elements of risk. The human capital (people) leverage intangible belongings to enhance employer performance, and hence the impact of various factors on human capital need to be explored. The predications related to such explorations may result in typical profiles and trends. These profiles and trends may be used for strategic human resources management decisions.
- III. **Prescriptive:** Prescriptive analytics is used to analyze complicated facts to predict outcomes, provide resolution options, and exhibit desire impacts. The technique deals with reporting of HR metrics and prescriptive modeling of commercial organization practices. It involves in understanding the impact of knowledge of investments on the bottom line. Such kind of scenarios is rare in case of HR domain.

Taking the aforementioned logic into consideration this study used predictive approach in order to understand the relationship between attrition and job satisfactions in a better way. This study used a sizeable data set related to HR domain (fictional) created by IBM scientists. The data set consist of 1470 data points and 35 variables. This data is specifically used for analysis of various HR-related attributes. The statistical tools like Statistical Package for Social Sciences (SPSS) and Microsoft Excel were used for analysis.

3 Data Analysis

The prime objective of this study was to identify the predictive indicators for identifying the typical profile of the employees who have left the organization. In order to develop these indicators, the data was divided into the employees left the organization (237) and still working in the organization (1433). This classification is presented as a Graph 1.



Graph 1 Classification of the employees

From the analyzed data it can be concluded that out of the total 1470 employees 16% employees have already left the organization. This number is 237 employees. The typical characteristics or behavioral pattern or the predictive indicators were identified on the basis of the employees who have left the organization.

Further, in order to identify the predictive indicators, the response or the characteristics of employees who left the organization were taken into consideration. The frequency analysis conducted for determining the predictive indicators is presented in Tables 1 and 2. Table 1 details about the factors related to an individual whereas Table 2 represents the prominent organizational factors identified by the researcher.

4 Individual Factors

The analysis presented in Table 1 indicates that the tendency to leave the organization is significantly higher among males, trainees, and unmarried employees. Further, it is moderate among sales representatives and people having life sciences as educational background. It is also there among lab technicians and people belonging to age group 31–35.

Similar kind of analysis is done for the organizational factors.

5 Organizational Factors

Researcher had observed that out of 20 factors, certain factors significantly influence the rate of attrition which are performance rating, years in current role, total years at company, years since last promotion, traveling in job, stock level, and percentage hike in salary. Further, the attrition may occur because of work–life

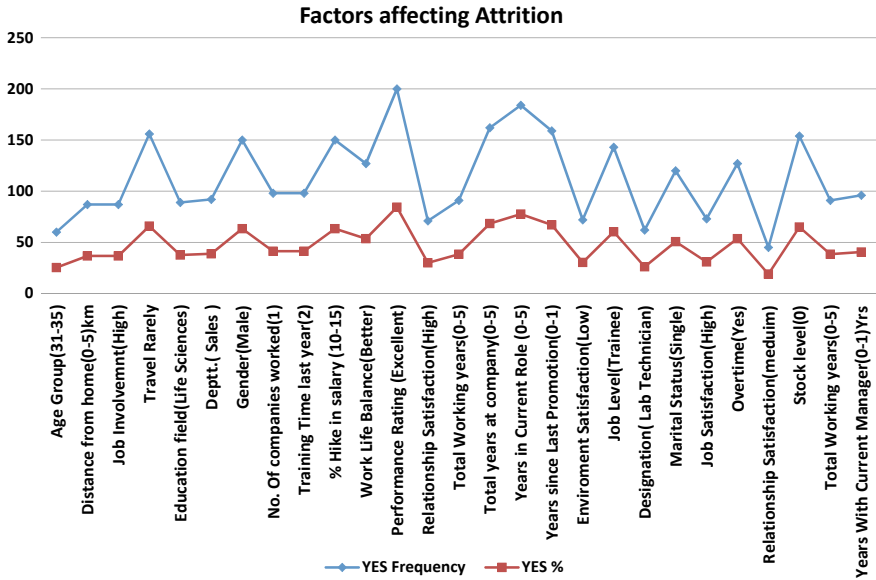
Table 1 Individual factors

Category	Frequency	Percentage
Gender (Male)	150	63
Job level (Trainee)	143	60
Marital status (Unmarried)	120	51
Department (Sales)	92	39
Education field (Life sciences)	89	38
Designation (Lab technician)	62	26
Age group (31–35)	60	25

balance, overtime, no. of companies worked, training time last year, years with current manager, total working experience, total working years, distance from home, job involvement, job satisfaction, environment satisfaction, and relationship satisfaction. Tables 1 and 2 are graphically represented as Graph 2.

Table 2 Organizational factors

Category	Frequency	Percentage
Performance rating (Excellent)	200	84
Years in current role (0–5)	184	78
Total years at company (0–5)	162	68
Years since last promotion (0–1)	159	67
Travelling in job (Rarely)	156	66
Stock level (0)	154	65
Percentage hike in salary (10–15)	150	63
Work life balance (Better)	127	54
Overtime (Yes)	127	54
No. Of companies worked (1)	98	41
Training time last year (2)	98	41
Years with current manager(0–1) yrs	96	41
Total working experience (0–5)	91	38
Total working years (0–5)	91	38
Distance from home (0–5) km	87	37
Job involvement (High)	87	37
Job satisfaction (High)	73	31
Environment satisfaction (Low)	72	30
Relationship satisfaction (High)	71	30
Relationship satisfaction (Medium)	45	19



Graph 2 Factors affecting attrition

Table 3 Results of hypothesis testing

	Attrition	Job Satisfaction
Attrition	1	
Job satisfaction	0.903481126	1

6 Hypothesis Testing

In order to test the hypotheses, correlation analysis is used. The results of hypothesis testing indicate that attrition and job satisfaction are strongly co-related. (Table 3)

Hence, from the above co-relation matrix, we can conclude that hence Ho Hypothesis is proved.

7 Multiple Regressions for Predictive Analysis

After confirming the impact of job satisfaction on attrition, the next step was to identify the parameters that impact the job satisfaction among the employees who left the organization. In order to identify these factors, the job satisfaction was

Table 4 Details of regression analysis

Regression statistics parameter	Details
Multiple R	0.339554231
R Square	0.115297076
Adjusted R Square	0.055249366
Standard error	1.086733162
Observations	237

Table 5 Results of ANOVA

ANOVA	<i>df</i>	SS	MS	<i>F</i>	Significance <i>F</i>
Regression	15	34.01409675	2.2676065	1.920091	0.022485206
Residual	221	260.9985615	1.180989		
Total	236	295.0126582			

treated as a dependent variable and 33 factors related to employees who left the organization as independent variables. The results of multiple regression are presented in Table 4.

In order to check the predictability of the factor, the multiple regression matrix was used. The threshold value for treating a factor as a predictive < 0.15 (Table 5).

From the above data, it is observed that four factors who have values less than or equal to 0.15 can be used to predict the job satisfaction. These factors are over time, stock option level, total working years, years at company, years with current manager. These predictors can be used for predicting the existing employee's tendency to leave the organization. Further, this will help the HR manager to design better strategies. The probability predications are presented in Table 6.

8 Probability Prediction for Attrition

From the above Table 7, it can be predicted that the employees who have been working for 0–5 years may leave the organization for growth, as the similar trend was observed with the employees who left the organization. This analogy can be applied for other factors like overtime years with current manager and total working experience.

Table 6 Predictors of job satisfactions

Particulars	Coefficients	Standard Error	t Stat	P value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.377592373	0.860301972	3.9260544	0.000115	1.682146867	5.0730379	1.682146867	5.073037879
Monthly Income	-3.08897E-05	3.04808E-05	-1.013415	0.31197	-9.09599E-05	2.918E-05	-9.096E-05	2.91805E-05
Monthly Rate	2.66044E-05	1.02223E-05	2.6025719	0.009879	6.45864E-06	4.675E-05	6.45864E-06	4.67501E-05
Num Companies Worked	0.004138138	0.030422756	0.1360211	0.891928	-0.0558177	0.064094	-0.0558177	0.064093975
Over Time	-0.258561383	0.1444734	-1.789682	0.074874	-0.543283242	0.0261605	-0.54328324	0.026160476
Percent Salary Hike	-0.036399008	0.031360677	-1.160658	0.247033	-0.098203258	0.0254052	-0.09820326	0.025405243
Performance Rating	-0.081905796	0.328406837	-0.249403	0.80328	-0.729115631	0.565304	-0.72911563	0.565304038
Relationship Satisfaction	0.053887991	0.064412541	0.8366071	0.403717	-0.073053427	0.1808294	-0.07305343	0.18082941
Stock Option Level	0.121833207	0.084938364	1.434372	0.15288	-0.045559607	0.289226	-0.04555961	0.289226021
Total Working Years	-0.028471292	0.019877602	-1.43233	0.153462	-0.0676452	0.0107026	-0.0676452	0.010702615
Training Times Last Year	-0.023684322	0.058022296	-0.408193	0.683527	-0.138032125	0.0906635	-0.13803213	0.090663482
Work Life Balance	-0.049626257	0.08927681	-0.55587	0.578862	-0.225569091	0.1263166	-0.22556909	0.126316577
Years At Company	0.070074927	0.028812772	2.4320787	0.015808	0.013291977	0.1268579	0.013291977	0.126857877
Years In Current Role	0.007748925	0.043790278	0.1769554	0.859706	-0.078551041	0.0940489	-0.07855104	0.09404889
Years Since Last Promotion	0.02579916	0.034231713	0.7536625	0.451854	-0.041663203	0.0932615	-0.0416632	0.093261522
Years With Current Manager	-0.063811836	0.040742793	-1.566212	0.11873	-0.144105951	0.0164823	-0.14410595	0.016482279

Table 7 Probability prediction for attrition

Predictive Analysis Parameters	Existing employees	
	Freq.	%
Years at company(0–5) yrs	614	50
Stock option (0)	477	39
Over time(Yes)	289	23
Years with current manager(0–1) yrs	243	20
Total working experience (0–5)	225	18

9 Conclusion

Researchers have observed that there is a strong relationship between the job satisfaction and attrition and failing of which it may lead to attrition. The employees will be satisfied with the organization when the company offers an interesting job, good policies, benefits, compensation. Hence, the management need to have an appropriate HR mechanism to deal with this.

The practice of overtime may be perceived as the source of additional income and a factor that may lead to physical stress. Further, it may exhaust the employees and may inversely affect the efficiency. From the data, it is ascertained that 54% of employees have left the organization due to overtime as one of the factors and it could be predicted from the analysis that 23% of employees may leave the organization.

Employees own a share in the company's success with stock option plans offered them a sense of belongingness. Holding of a stock automatically makes employees a contributor to the profit and loss of the company. The researchers have noticed that 65% of the employees who resigned were not owning stocks of the organization. This tendency may influence 39% of existing staff.

It is observed that the company recruits fresh people and offers them experience, skill, knowledge, and train them for their future. This staff expects higher salaries and growth opportunities. Based on the analysis it may be considered that employee between 0 to 5 years of experience may leave the organization. This demands better policy-making. The researchers observed that the role of supervisor/reporting authority is one of the factors that need to be considered.

Hence, as a result of this study, the researcher proposes to use predictive analysis techniques for better HR policy and strategy making.

References

1. Muscalu E, Serban A (2014) HR analytics for strategic human resource management. In: Proceeding of 8th international management conference
2. Momin, W.Y.M., Taruna, : HR analytics transforming human resource management. *Int J Appl Res* **9**(1), 2349–5869 (2015)
3. Sujeet, Mishra, Dev, Lama, Yogesh, Pal: Human resources predictive analytics (HRPA) For HR management in organizations. *Int J Sci Technol Res* **5**(5), 2277–8616 (2016)
4. Bindu, K.: A review of poor analytical skills. *Int J Innov Res Technol* **3**(1), 2349–6002 (2016)
5. Madhavi Lakshmi, P., Siva Pratap, P.: HR analytics- a strategic approach to HR effectiveness. *Int J Hum Resour Manag Res* **6**(3), 2249–7986 (2016)
6. Globcon Technologies Private Ltd (2016) Redefining HR through data analytics. Article published on www.analyticsindiamag.com
7. Rajbhar, A.K., Khan, T., Puskar, S.: A study on HR analytics transforming human resource management. *J Invest Manag* **6**(4), 2328–7721 (2017)
8. Shweta, T.: Analytics application in human resource context. *Int J Sci Eng Res* **6**(1), 2347–3878 (2018)

A Study of Business Performance Management in Special Reference to Automobile Industry



Gurinder Singh , Smiiti Kashyap , Kanika Singh Tomar 
and Vikas Garg 

Abstract In contemporary times, the automobile is one of the well-paid industries in the Indian market with an annual growth rate of 7.64% in the passenger-car market. The increasing disposable income of the people along with the ever-growing financial sector has led to this expendable growth. In accordance, there has been an increase in sales of passenger cars from 13.35% in –July 2018. This oligopoly market has fierce competition due to new entrants into the Indian markets. Thus, there emerges a need to grasp the knowledge to understand the ever-growing needs of the customers and dynamism in the technology-driven market. Every organization seeks to study consumer buying behavior and measure its business performance by analyzing customer perception toward the product. The understanding of customer’s perception is an ongoing process to survive the cut throat competition. Taking this into consideration, the study provides insights about several attributes which drive a consumer behavior toward buying a product of a brand. The study has used theories and exploratory research design along with analytical tools to identify the major attributes of consumer buying behavior toward sedan cars within Delhi/NCR among high-end consumers. This knowledge helps the car manufacturers in market segmentation which enables the organization to plan the market strategies toward consumer retention and product upgradation.

Keywords Business performance management · Consumer buying behavior · Market segmentation · Factor analysis · Brand reputation · Purchase service · Physical appearance · Automotive techniques

G. Singh · S. Kashyap · K. S. Tomar (✉) · V. Garg
Amity University, Noida, Uttar Pradesh 201308, India
e-mail: kstomar@gn.amity.edu

G. Singh
e-mail: gsingh@amity.edu

S. Kashyap
e-mail: skashyap@gn.amity.edu

V. Garg
e-mail: vgarg@gn.amity.edu

1 Introduction

In India, the automobile is an industry which plays a major role in the growth of the Indian economy. The industry encompasses namely, passenger cars, two-wheelers, three-wheelers, commercial vehicles, multi-utility vehicles, and its components. The share of the Indian motorcycle manufacturers is the largest in the world, followed by other two-wheelers and tractor manufacturers. The Indian Commercial Vehicle manufacturers are the fifth largest across the global market and fourth largest in Asia. India has strength in manufacturing low-cost, fuel-efficient cars of various automobile companies namely, Maruti Suzuki, Volkswagen, Toyota, Hyundai, and Nissan.

Looking at the ever-growing share of the automobile industry in the growth, the government of India focuses on Research & Development to provide the infrastructure and world-class automotive testing to the industry. The paper throws light on the several determinants of consumer buying behavior of passenger cars which can be utilized by the industry for market segmentation, advertisement strategies, product differentiation in terms of technology upgradation, safety and user friendliness, and sales maximization.

1.1 *Passenger Vehicles*

A passenger vehicle is a four-wheeler vehicle used as a passenger carriage in which up to nine people including a driver can be seated. India is leading in the car market with a 9% growth rate. Japan is the only market which could come closer to this growth rate. China which is the largest market of passenger vehicles had registered a decline in the growth rate by 2.59% and the US market shrunk by 10%. The paper studies the consumer behavior of sedan cars in India. A sedan car is a 3-box configuration passenger-car with A: B: C pillars along with compartments for other things such as the engine of the car, its passenger, and cargo. The best sedan cars in India are Maruti Dzire, Honda Amaze, Honda City, Hyundai Verna, Maruti Ciaz, Toyota Yaris, Toyota Etios, Hyundai Xcent, Tata Tigor, C Class 2018, Sonata, Civic, A6 2019, A Class Sedan, etc.

1.2 *Consumer Buying Behavior*

The decision-making of a consumer, during a purchase is necessary to be understood to capture the market and tap the potential gains. The understanding is required in terms of what the dynamics of consumer preferences are.

Post this understanding, the automobile industry segments and works on the marketing strategies to improve business performance. In theory, there is a six-stages process of the consumer demand behavior which comprises:

- Exploratory research to understand the problem
- Assimilating the information for the research
- Evaluating the solutions and alternatives
- Decision-making
- The final purchase
- Post-purchase experience.

The decision-making of the consumer buying behavior can be studied at a psychological level, physiological level, financial level, and spatial level.

2 Literature Review

In the automobile industry, the factors which impact the buying pattern of new cars and second-hand cars have been studied extensively. It has been observed that the factors affecting the buying behavior of second-hand cars are very different from first-hand cars [20].

The purpose of studying the consumer buying behavior for new car entrants is to fill the gaps between customer expectations and the present products. This helps to depict the total quality management of the cars to tap the existing and potential buyers [12].

The question for managers in the service industries regarding consumer loyalty has been studied based on online or offline consumers. They have depicted the correspondence among customer satisfaction and loyalist in an online and offline environment. It shows that the consumer satisfaction level among online and offline customers is equal, however, the loyalty of the customer is more toward online service provider [16].

A pattern to study the pricing and advertisement expenditure using time series data has been made in 3D Innovation Diffusion Model with empirical evidence of sales to understand the growth pattern of sales [3].

The consumers have a bend toward sporting cars and speed cars as they consider it as the good of distinction. There are several constructs to these concepts. The variables like R&D by the technical fraternity, disposable income, and liberalization in terms of foreign exchange, high-end models and integration in automobiles pave way to the dynamic consumer behavior in the automobile industry as observed today [17].

The car segments in terms of easy finance availability of rural and urban sectors have been studied to understand the bend of consumer based on brand loyalty, which helps the existing companies to survive when new entrants come in the market. All market segments within the Indian market have been considered such as

value-for-money, safety and driving comforts, and brand image impacts on sales of passenger cars [5].

The logistic regression modeling has also been used to depict the protectionism adopted by the government of India to help the domestic car manufacturers to grow in the competitive world. The domestic car manufacturers use product differentiations, innovative products, and quality and reliability as the attributes to capture the market share. The companies such as Honda, Toyota, General Motors, Ford, and Hyundai used these strategies prior to liberalization and foreign direct transformation in the Indian market [14].

The increase in disposable income at the rate of 25% per annum and the changes in the sociocultural environment are the factors affecting consumer behavior in terms of luxury cars. Now, even middle-class people like to purchase luxury cars. So, the study shows how luxury cars now occupy the place in the daily life of the people within the Indian Economy [18].

Automotive Technology and Human Factor Research related to the past, present, and future is primarily driven in terms of the features of car control, display for the driver, driver workspace, driver's conditions, etc. [1].

The aim of this paper is to find the significant factors which influence the consumer buying behavior of sedan cars. Further, the paper provides recommendations to the automobile industry to upgrade their products, market segmentation, and advertisement strategies to maximize the sales and hence, capture the market.

3 Objectives

1. To evaluate the measures to improve the business performance of the sedan car industry to compete in the Indian metropolitan markets.
2. To determine the agents affecting the consumer buying behavior of sedan cars in Delhi/NCR.
3. To segment the market of sedan cars with reference to the consumer preferences.

4 Data Collection

This paper aims to evaluate the measures to improve business performance in the automobile industry. The business performance management is studied based on different perspectives namely, learning-growth perspective, financial perspective, customer perspective, and internalized business process [2].

This paper focuses on the learning and growth aspect of the automobile industry. The market segmentation is significant for the existing firms and new entrants to compete in the market [7]. With the review of the literature and pilot surveys among

the consumers of Delhi/NCR (30 consumers for the pilot survey were used), we have used the following variables which impact the consumer buying behavior in the sedan car market to do the market segmentation:

The demographic variables under study are age, sex, household income, and occupation.

Other factors:

4.1 Brand Reputation

This attribute influences consumer buying behavior. We have studied brand reputation using the following variables:

Brand image, its name, consistent and affordable pricing for high-end customers, and resale value are important in the decision-making process of the consumer [9].

4.1.1 Brand Name

It is the name given by the company to the product or a range of products. A consumer may associate themselves with a brand name or may build trust toward a specific brand name.

4.1.2 Price

It is the manufacturer suggested retail price of the car which influences the purchase of the product by the consumers given their income levels and preferences. This is also known as “Sticker Price”.

4.1.3 Resale Value

It is a factor which is based on several attributes of the car which a customer keeps in mind while purchasing it. The customer likes to purchase a car with higher resale value.

4.2 Warranty

Sedan cars in India are sold by several companies such as Hyundai, Honda, Volkswagen, Maruti Suzuki, etc. The cars in the current era are more reliable and in cases of component failures, the companies provide for the replacement under the warranty period which is either in months/years or the kilometers driven from the date of delivery. A few companies even provide for extended warranty under the

name of “Forever yours” such as Maruti Suzuki. We have studied the warranty using the following variables:

4.2.1 Availability of Spare Parts

Availability of spare parts increases the longevity of the car and makes it economical.

4.2.2 Vehicle Availability

Availability of vehicle for use in the stipulated time of need.

4.3 Physical Appearance

The looks of a car play an important role in purchasing a car. We have studied physical appearance using the following variables:

4.3.1 Design and Exteriors

The design, shape, and surface are built by car manufacturers using upgraded technology which attracts the customer. The body of the car impacts the speed and safety of the car.

4.3.2 Comfort

It is a primary factor which influences consumer buying behavior. A comfortable vehicle with leg space, headroom, automatic starting are the definitive factors which a consumer investigates before buying the product.

4.4 Features

Apart from the looks, there are several other features which are kept into consideration while purchasing a car. We have used the following variables:

4.4.1 Interior Look

The well-defined body, neatly-cut lamp, installed lights, style, inside cabin, gearbox design, and several other interior designs are investigated by a consumer while purchasing a car.

4.4.2 Boot Space

The boot space ranges from 235 to 510 L in a sedan car which signifies how spacious a car is.

4.5 Technical Features

All the consumers look for an upgraded product along with user friendliness. We have used the following variables:

4.5.1 Technology

The sedan cars with high-tech features such as gadgets installed in the car, the automatic gear box, apple care play, compact hatch back, gas-electric hybrid technology, etc., are considered by the consumers prior to buying a car.

4.5.2 Engine

The cubic centimeters of the cylinder of the car engine are kept in mind by the consumer prior to purchase.

4.6 Value Addition

The additional services are provided by the car distribution channel to add value to the product. We have used the following variables:

4.6.1 Waiting Time

This is a time gap between the booking and delivery of the car.

4.6.2 Discount

The reduction in the price of the car makes the purchase attractive.

4.6.3 Safety

Several features in a car like airbag, crash rates, and platform, are important to a consumer and for auto-insurances.

4.6.4 Value for Money

If the utility received from the car is not in equilibrium with the pricing, the consumer tends to restrict the purchase.

4.7 Performance

The utility of the sedan car to a consumer can be studied using the following variables:

4.7.1 Mileage

The distance in kilometers covered by a car per liter of fuel is considered before purchase.

4.7.2 Fuel Economy

The fuel-efficiency of a car is a major performance parameter.

5 Research Methodology

This is an exploratory research where we try to identify the major factors which determine the consumer buying behavior of sedan cars in Delhi/NCR. For the study, we have used high-end consumers under the income bracket of 7 Lac and above (per annum). The primary data was collected using the survey method [8].

We have built a questionnaire under several heads as per the variables. The responses to the questionnaire were collected from the consumers who were the potential buyers of sedan cars.

We have divided the Delhi/NCR majorly into four regions namely, north, south, west, and east to represent the consumers (with income of 7 Lac and above annually) of sedan cars. The information regarding the consumers was collected using the data available at the showrooms. We have used the sample size of 306 consumers from all the four regions to truly represent the target population and performed personalized interviews to collect the information.

6 Data Analysis

To identify the significant factors which affect consumer buying behavior, we have collected the data using the questionnaire method. Every question was analyzed to understand consumer preferences using bar graphs, pie charts, tables, and statistical tools.

6.1 Demographic Distribution of Consumers

6.1.1 On the Basis of Gender

The data collected shows that out of the sample, 20% of the consumers of sedan cars are females and 80% of the consumers are males (refer to Fig. 1). This interpretation helps the automobile industry to focus on the male counterpart of the society while designing the marketing strategies and features of the product.

The data collected shows that out of the sample size, 20% of the consumers are females and 80% are males.

6.1.2 On the Basis of Age

During the data analysis, it was found that sedan cars are purchased by consumers of different age groups. The maximum percentage of users, i.e., 50% of the users lie

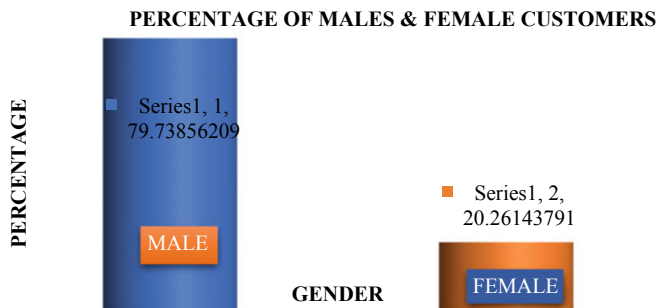


Fig. 1 Percentage of users of among males and females

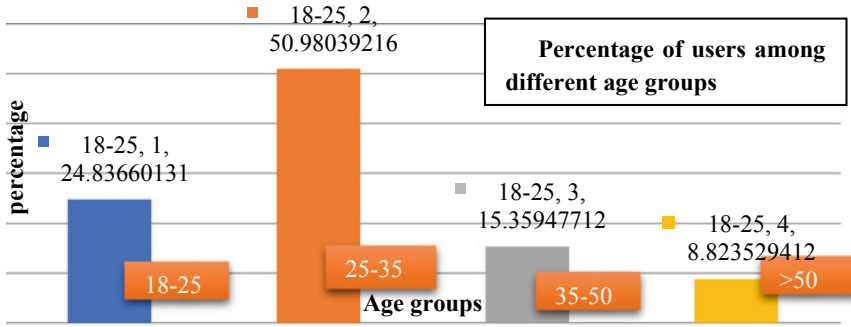


Fig. 2 Percentage of users among different age groups

in the age group of 25–35 years followed by the age group of 18–25 years constituting nearly 25% of the users (refer to Fig. 2). The rest 15% fall in the age group of 35–50 years and above. So, the companies should target the advertisements on the age group of 25–35 years to increase sales.

The consumers between 25–35 years are the people who demand 50% of the sedan cars in the market. The people between 35–50 years demand 15% of sedan cars in the market and 18–24 years demand 25% of the sedan cars in the market.

6.1.3 On the Basis of Occupation

The results show that the maximum users of sedan cars fall in the business class in terms of occupation constituting nearly 63%. The service class constitutes approximately 20% of the total users followed by students and others. (refer to Fig. 3). The automobile firms must focus on their marketing strategy on the business class.

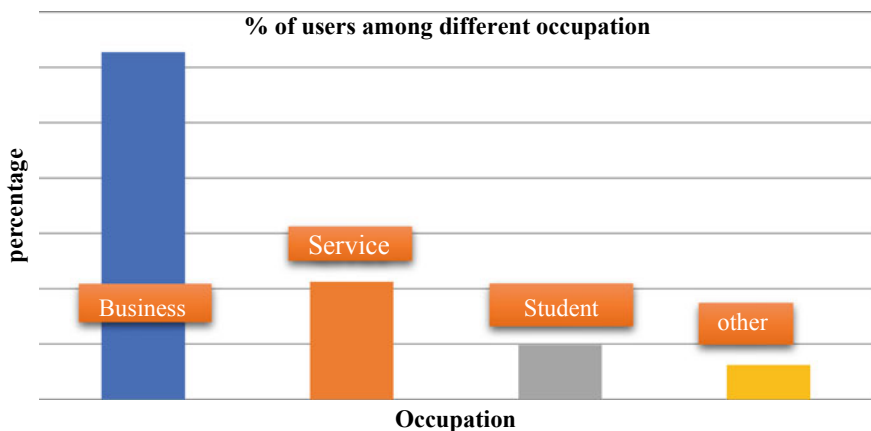


Fig. 3 Percentage of users among different occupation

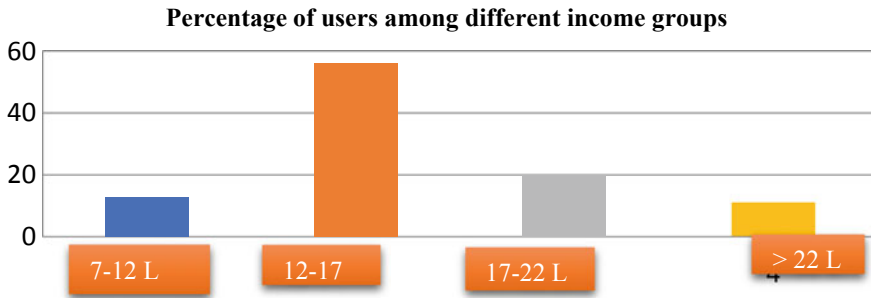


Fig. 4 Percentage of users among different income groups

Among the buyers, the maximum people are businessmen—approximately 63%. Nearly 20% are servicemen, 10% are students, and the rest 8% falls in the “others” category which includes self-employed people.

6.1.4 On the Basis of Income Groups

The data collected helps us to find that the maximum users lie in the income group of 12 to 17 Lacs constituting of up to 57%. 20% of the users fall in the income group of 17–22 Lacs and 12% of the users fall in the income group of 22 Lac and above and 11% fall in the income group of 7–12 Lacs (refer to Fig. 4). The companies’ pricing, products, and marketing strategies must focus on users falling under the age group of 12 to 17 lacs.

According to my study, 57% of the buyers are of the income group 12–17 lacs per annum and 12% people of income above 22 lacs buy sedan cars, as people of this group would rather go for sedan cars.

6.2 Motivational Factors of Consumers

6.2.1 On the Basis of Daily Drive

The results show that on an average 30% of the consumers in Delhi/NCR drive under 50 kms, 25% of the consumers drive up to 50–80 kms, 34% the consumers drive up to 80–100 kms and rest 11% of the consumers drive above 100 kms (refer to Fig. 5). As on an average, maximum consumers drive between 80–100 kms, they would prefer more economical cars on roads such as diesel engine, good mileage along with luxury and comfort.

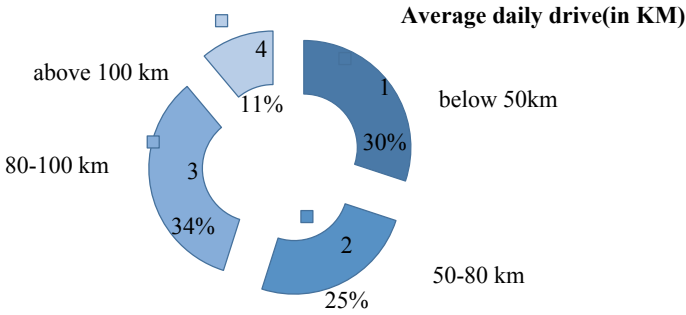


Fig. 5 The average daily drive of the consumers

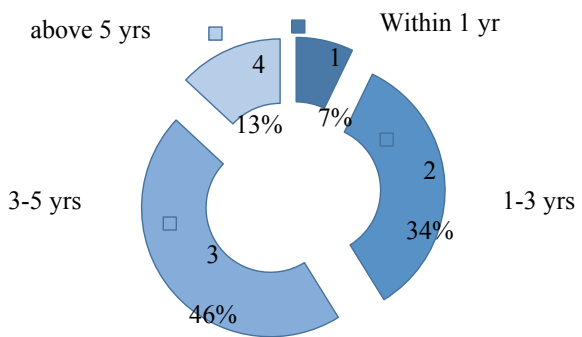
According to my survey, 34% of the buyers include people who drive 80–100 kms daily on an average and therefore they may be interested in a diesel engine sedan class car which would give them good mileage, luxury, and comfort.

6.2.2 On the Basis of Time Span Between Car Switch

The results show that on an average out of the total consumers of sedan cars in Delhi/NCR, 7% change their car within a time span of 1 year, 34% change it in 1 to 3 years, 46% change it in 3 to 5 years, and rest 13% change it in a span of above 5 years (refer to Fig. 6). The companies can use this data and intimate their consumers about the new releases within these time spans to retain the consumers.

According to my study, 46% people purchase/change their cars in 3–5 years. By keeping a track of such customers, who change their cars in every 3–5 years, the companies could send them invitations regarding their new releases and retain their customers.

Fig. 6 Time span in which the consumers change their cars



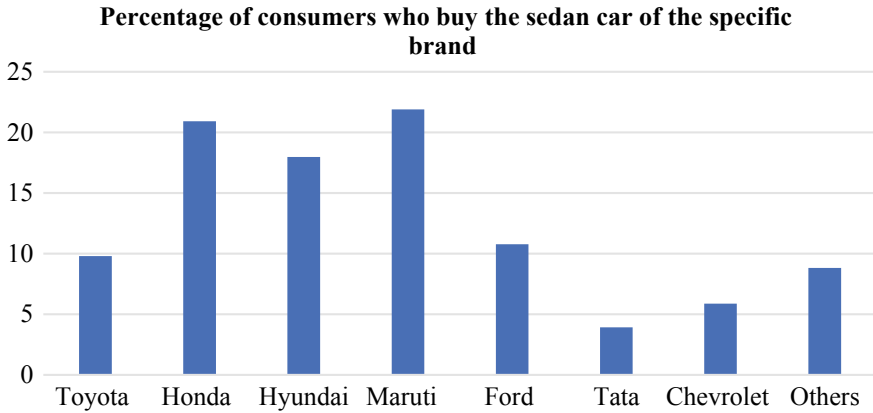


Fig. 7 The different brand sedan cars which consumers purchase

6.3 Preference of Consumers Toward Different Brands

The results show that in Delhi/NCR, 22% of the consumers buy Maruti sedan cars, 20% of the consumers buy Honda sedan cars, 18% of the consumers buy Hyundai sedan cars, 10% buy Ford, and the rest of the consumers prefer other brands like Toyota, Chevrolet, Tata, etc. (refer to Fig. 7). This shows the cut throat competition among the oligopolistic automobile industry.

According to the study, the consumers buy (cars) of other brands such as Honda, Hyundai, Maruti, and Ford which constitute 20, 18, 22, and 10%, respectively.

6.4 Consumer Preferences Based on Features of Sedan Cars

The study has used the 5-point Likert scale to analyze consumer preferences while buying a sedan car. The scale: 1 as Excellent, correspondingly 2 as Very-Good, followed by 3 as Satisfactory, 4 as Non-Satisfactory, and 5 as Poor. The study observes that when consumers purchase a sedan car, 20% is based on the comfort level of the car, 17% on the basis of engine strength, and 14% on the basis of luxury. Table 1 shows the weightage of different features while purchasing a sedan car.

Table 1 Frequency of consumer preference

Features	Frequency	Weightage (Percentage)
Comfort	64	20.91503
Luxury	43	14.05229
Engine	55	17.97386
Sunroof	36	11.76471
Cruise control	18	5.882353
Premium upholstery	24	7.843137
Push entry passive start (PEPS)	21	6.862745
Good mileage	39	12.7451
Others	6	1.960784

6.5 Factor Analysis

We have used the factor analysis technique using SPSS software. The factor analysis technique is used to extract the most significant variables from the list of several variables which influence the dependent variable. This tool calculates the maximum common variance of all variables to find the common score which determines the magnitude and direction of the impact of the independent variable on the dependent variable.

Factor analysis is a multivariate technique which is used to identify the underlying factors and then, eliminate the redundant variables [11].

We have calculated the eigenvalue which is used to consolidate the variance, characteristic roots, and latent values. Additionally, the scree plot of eigenvalue is used to find the significant variables of the study. All the variables on the scree plot before it becomes flat are taken into consideration [13]. An eigenvalue of either 1 or more than 1 is considered a good explanatory feature.

6.5.1 Descriptive Statistics

The data has been first analyzed to measure the central tendencies and dispersion using tools such as mean and standard deviation as depicted in Table 2. It is used to summarize the data set prior to applying factor analysis. The mean shows the influence of each factor on the consumer buying behavior of sedan cars. The weights which are closer to 0, hold no relevance while the larger the weights, larger is the influence on the dependent variable. The standard deviation shows the variation of these factors from the central tendencies. This data set is of total N (306) observations.

Table 2 Descriptive statistics

	Mean	Std. Deviation	Analysis N	Missing N
Technology	2.09	0.943	306	0
Brand_name	2.10	1.062	306	0
Price	3.67	1.133	306	0
Interior	2.32	0.659	306	0
Design_and_exterior_looks	2.42	0.630	300	6
Comfort_and_convenience	2.19	0.618	306	0
Safety_and_security	1.49	0.830	306	0
Engine	2.63	0.821	306	0
Mileage	1.84	0.733	306	0
Boot_space	1.51	1.031	306	0
Value_for_money	1.53	0.623	306	0
Warranty	3.18	0.747	306	0
After_sale_services	2.29	0.719	306	0
Resale_value	5.54	1.116	306	0
Availability_of_spare_parts	2.14	1.066	306	0
Discounts_or_special_offers	2.19	0.988	306	0
Fuel_economy	2.14	0.737	306	0
Vehical_availability	2.53	0.874	306	0
Waiting_time	2.63	0.762	306	0

6.5.2 KMO and Bartlett’s Test

The KMO and Bartlett’s Tests were performed to check the adequacy of the data set for factor analysis. The value of Kaiser test lying between 0.5 and above can be considered as middling to perform factor analysis and hence, acceptable. Additionally, Bartlett’s test is used to find how useful it is to apply the factor analysis on the data set. If the p value is less than 0.05, then it is considered suitable for further operations. The results in Table 3 show that the data set is suitable for factor analysis.

Table 3 KMO and Bartlett’s test

Kaiser–Meyer–Olkin measure of sampling adequacy.		0.585
Bartlett’s test of sphericity	Approx. Chi-square	1944.462
	df	171
	Sig.	0.000

Table 4 Communalities

	Initial	Extraction
Technology	1.000	0.619
Brand_name	1.000	0.592
Price	1.000	0.775
Interior	1.000	0.727
Design_and_exterior_looks	1.000	0.652
Comfort_and_convenience	1.000	0.773
Safety_and_security	1.000	0.635
Engine	1.000	0.649
Mileage	1.000	0.800
Boot_space	1.000	0.591
Value_for_money	1.000	0.695
Warranty	1.000	0.726
After_sale_services	1.000	0.720
Resale_value	1.000	0.671
Availability_of_spare_parts	1.000	0.688
Discounts_or_special_offers	1.000	0.701
Fuel_economy	1.000	0.772
Vehical_availability	1.000	0.666
Waiting_time	1.000	0.826

Extraction method: Principal component analysis

6.5.3 Communalities

We have calculated communalities to find the proportion of the variance of the variable that can be explained by each factor. The Initial values indicate the estimates of variance on the account of all factors; it is always 1 for principal component extraction. The extraction value is the estimate of the variance of the variable by all components. A high extraction value indicates that all the variables are represented well. The corresponding results are shown in Table 4.

6.5.4 Total Variance Explained

The purpose is to extract the variables which explain the maximum variance of the dependent variable. To extract the total variance explained by the factors, we have run the factor analysis. The number of factors used for factor analysis is 19. The initial eigenvalue depicts the variance; the total values show the variance in each factor in descending order; the percentage of variance shows the total variance accounted by each factor and the cumulative percentage shows the variance accounted by the current and all the preceding factors. The result in Table 5 shows that the first seven variables explain approximately 70% of the total variance.

Table 5 Total variance explained

Component	Initial eigenvalues		Extraction sums of squared loadings		Rotation sums of squared loadings	
	Total	% of Variance	Total	% of Variance	Total	% of Variance
1	2.884	15.176	2.884	15.176	2.446	12.875
2	2.707	14.247	2.707	14.247	2.156	11.346
3	2.093	11.016	2.093	11.016	2.009	10.573
4	1.553	8.172	1.553	8.172	1.926	10.136
5	1.418	7.462	1.418	7.462	1.661	8.743
6	1.332	7.011	1.332	7.011	1.653	8.700
7	1.292	6.802	1.292	6.802	1.428	7.514
8	0.889	4.679				
9	0.821	4.321				
10	0.678	3.566				
11	0.630	3.314				
12	0.551	2.897				
13	0.523	2.753				
14	0.433	2.281				
15	0.408	2.148				
16	0.279	1.470				
17	0.202	1.062				
18	0.181	0.951				
19	0.128	0.673				

Extraction method: Principal component analysis

The extraction sum of squared loadings calculates the value of retained factors using common variance. This shows that a total of 7 factors are retained from a total of 19 factors. The rotation sums of squared loading are used to maximize the variance of each of the 7 factors in order to redistribute the total variance [19].

6.5.5 Scree Plot

The scree plot is the graph where the eigenvalue is plotted against the number of factors used for the study. The factors from where the line becomes flat show the magnitude of variance becoming smaller and smaller. Therefore, the graph shown in Fig. 8 depicts that the maximum variance is explained by the first seven variables. [15].

To further validate the extraction of the seven variables from the 19 variables of the study, we have calculated the correlation and covariance of each factor with these seven factors. The results show the high correlation and the appropriateness of the variables extracted. The Component Matrix (Table 6), Rotated Component Matrix (Table 7), Component Transformation Matrix (Table 8), Component Score Coefficients (Table 9), and Component score covariance matrix (Table 10) are shown in the appendix for reference.

6.5.6 Factor Loadings

The factors retained after performing factor analysis are brand reputation; purchase service, physical appearance, features, technical features, value additions, and

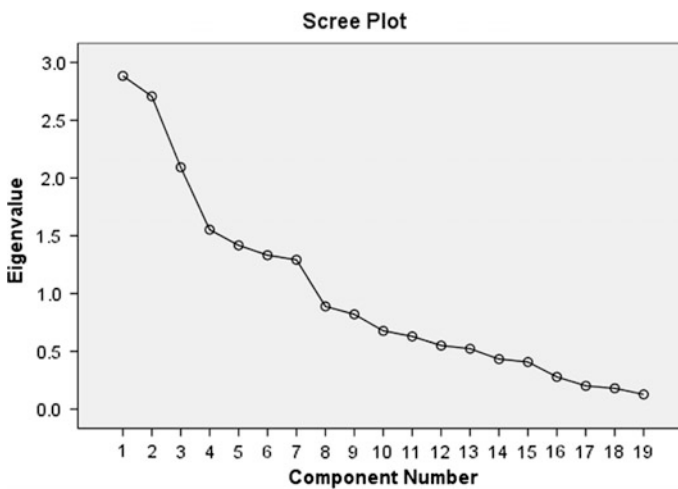


Fig. 8 Scree plot

Table 6 Component matrix (a)

	Component						
	1	2	3	4	5	6	7
Technology	-0.283	0.534	0.331	-0.056	0.286	-0.228	0.087
Brand_name	0.597	-0.234	-0.267	0.078	-0.128	-0.276	-0.104
Price	-0.077	0.383	0.181	0.561	-0.412	-0.137	0.294
Interior	0.342	-0.440	0.233	0.126	0.304	-0.227	0.450
Design_and_exterior_looks	0.639	-0.102	-0.298	0.200	0.252	-0.194	0.048
Comfort_and_convenience	0.574	-0.252	-0.274	-0.195	0.442	0.171	-0.207
Safety_and_security	0.159	0.470	0.251	0.055	0.460	0.259	-0.210
Engine	0.020	0.599	0.195	-0.399	0.086	-0.271	0.112
Mileage	0.497	0.715	-0.093	0.003	0.038	-0.157	0.088
Boot_space	-0.400	-0.216	0.071	0.276	0.066	0.476	0.269
Value_for_money	0.441	0.331	0.614	-0.005	-0.030	-0.103	0.045
Warranty	-0.275	0.100	-0.063	0.487	0.630	0.054	-0.017
After_sale_services	0.506	0.057	0.074	0.045	-0.186	0.477	0.437
Resale_value	-0.374	0.425	-0.485	0.161	0.211	0.126	0.170
Availability_of_spare_parts	0.376	-0.335	0.562	0.285	-0.039	0.073	-0.175
Discounts_or_special_offers	0.420	0.274	-0.498	0.361	-0.069	0.009	0.259
Fuel_economy	0.221	0.517	-0.415	-0.166	-0.262	0.364	-0.235
Vehicle_availability	0.326	0.219	0.426	0.163	-0.051	0.448	-0.316
Waiting_time	-0.120	0.096	-0.084	0.605	-0.144	-0.283	-0.573

Extraction method: Principal component analysis. 7 components extracted

performance. Several attributes are studied under these variables. The factor loadings of these are shown using the component matrices [10].

The major results from the component matrices are summarized from Tables 11, 12, 13, 14, 15, 16, and 17:

Factor 1 comprises brand reputation provided by the company. Here price has maximum factor loading of .863 so it becomes an important factor.

Factor 2 comprises purchase services provided by the company. Here warranty has maximum factor loading of .812, so it becomes an important factor.

Factor 3 comprises physical appearance provided by company. Here design and exterior looks have maximum factor loading of .768, hence it becomes an important factor.

Here, factor 4 comprises features. Here, interiors have maximum factor loading of .754 so it becomes an important factor.

Factor 5 comprises technical features of the car provided by the company. As we can see, engine has maximum factor loading of .783, so it is an important factor.

Factor 6 comprises value addition. Here, waiting time has maximum factor loading of .783, so it becomes an important factor.

Factor 7 comprises performance provided by the company. Here, fuel economy has maximum factor loading of .863 which makes it an important factor.

Table 7 Rotated component matrix (a)

	Component						
	1	2	3	4	5	6	7
Technology	-0.284	0.637	-0.014	-0.070	0.305	0.157	-0.095
Brand_name	0.652	-0.122	0.043	-0.079	-0.344	-0.113	-0.117
Price	0.863	0.093	0.103	0.016	0.069	0.073	-0.015
Interior	0.294	-0.034	0.057	0.754	0.009	-0.060	0.254
Design_and_exterior_looks	179	-0.005	0.048	-0.157	0.051	0.768	0.012
Comfort_and_convenience	0.477	-0.100	0.148	0.005	0.052	0.705	0.116
Safety_and_security	-0.003	0.389	0.025	0.177	0.465	-0.184	0.549
Engine	-0.099	0.783	-0.072	0.106	-0.067	0.023	0.072
Mileage	0.505	0.159	0.122	0.308	0.053	628	0.112
Boot_space	-0.352	-0.448	-0.004	0.689	0.388	0.172	0.280
Value_for_money	0.084	0.515	0.122	-0.163	-0.154	0.166	0.574
Warranty	0.019	-0.036	0.812	-0.157	0.022	-0.003	-0.202
After_sale_services	0.270	-0.095	0.315	0.089	-0.073	0.153	0.708
Resale_value	0.554	0.067	-0.444	0.369	0.014	0.157	0.040
Availability_of_spare_parts	0.064	-0.191	0.691	-0.383	-0.143	0.018	-0.058
Discounts_or_special_offers	0.184	-0.028	-0.115	0.246	0.161	0.277	0.700
Fuel_economy	0.203	0.095	0.046	0.060	0.863	-0.040	0.126
Veheical_availability	-0.034	0.027	0.775	0.234	0.064	0.007	0.073
Waiting_time	0.160	-0.169	0.166	0.146	0.132	0.305	0.783

Extraction method: Principal component analysis. rotation method: Varimax with Kaiser normalization a rotation converged in 10 iterations

Table 8 Component transformation matrix

Component	1	2	3	4	5	6	7
1	0.768	0.127	0.460	-0.029	-0.283	-0.196	0.250
2	0.040	0.712	0.065	0.556	0.269	0.324	0.021
3	-0.478	0.293	0.661	-0.458	-0.112	0.158	0.044
4	0.323	-0.375	0.262	-0.142	0.486	0.590	-0.286
5	0.086	0.243	-0.022	-0.334	0.713	-0.559	-0.016
6	-0.250	-0.426	0.372	0.453	0.287	-0.159	0.552
7	0.080	0.098	-0.374	-0.379	0.084	0.382	0.740

Extraction method: Principal component analysis. rotation method: Varimax with Kaiser normalization

Table 9 Component score coefficient matrix [6]

	Component						
	1	2	3	4	5	6	7
Technology	-0.089	0.316	-0.030	-0.125	0.146	0.027	-0.051
Brand_name	0.270	-0.033	-0.027	-0.036	-0.174	-0.007	-0.142
Price	0.079	-0.017	0.031	-0.047	-0.006	0.544	0.013
Interior	0.147	0.060	-0.059	-0.437	0.083	0.030	0.168
Design_and_exterior_looks	0.333	0.020	-0.033	-0.120	0.094	-0.064	-0.044
Comfort_and_convenience	0.153	-0.027	0.067	0.035	0.118	-0.420	0.025
Safety_and_security	-0.030	0.133	0.251	0.076	0.308	-0.188	-0.005
Engine	-0.050	0.392	-0.097	-0.029	-0.085	-0.035	0.035
Mileage	0.202	0.260	-0.002	0.074	0.019	0.076	0.031
Boot_space	-0.137	-0.252	0.054	-0.014	0.254	0.108	0.265
Value_for_money	0.002	0.225	0.230	-0.112	-0.079	0.092	0.038
Warranty	0.071	-0.023	0.031	-0.133	0.520	-0.049	-0.109
After_sale_services	0.055	-0.114	0.123	0.064	0.004	0.141	0.488
Resale_value	0.050	-0.003	-0.192	0.126	0.300	0.053	0.078
Availability_of_spare_parts	-0.001	-0.105	0.349	-0.136	-0.027	0.040	-0.080
Discounts_or_special_offers	0.315	-0.061	-0.094	0.071	0.110	0.207	0.115
Fuel_economy	0.028	-0.052	0.062	0.464	-0.069	-0.057	0.064
Veheical_availability	-0.080	-0.083	0.437	0.191	0.068	-0.029	0.014
Waiting_time	0.123	-0.115	0.147	0.089	0.045	0.165	-0.567

Extraction method: Principal component analysis. rotation method: Varimax with Kaiser normalization. Component scores

Table 10 Component score covariance matrix [4]

Component	1	2	3	4	5	6	7
1	1.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	1.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	1.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	1.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	1.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	1.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	1.000

Extraction method: Principal component analysis. rotation method: Varimax with Kaiser normalization. Component scores

Table 11 Factor 1 brand reputation

Attributes	Factor loading	Mean
Brand name	0.652	2.10
Price	0.863	3.67
Result value	0.554	0.554

Table 12 Factor 2 purchase service

Attributes	Factor loading	Mean
Warranty	0.812	3.18
Availability of spare parts	0.691	2.14
Vehicle availability	0.775	2.53

Table 13 Factor 3 physical appearance

Attributes	Factor loading	Mean
Design and exterior looks	0.768	2.42
Comfort	0.705	2.19

Table 14 Factor 4 features

Attributes	Factor loading	Mean
Interior look	0.754	2.32
Boot space	0.689	1.51

Table 15 Factor 5 technical features

Attributes	Factor loading	Mean
Technology	0.637	2.09
Engine	0.783	2.63

Table 16 Factor 6 value addition

Attributes	Factor loading	Mean
Waiting time	0.783	2.63
Discount	0.700	2.19
Safety	0.652	1.49
Value for money	0.863	1.53
After sales value	0.554	2.29

Table 17 Factor 7 performance

Attributes	Factor loading	Mean
Mileage	0.628	1.84
Fuel economy	0.863	2.14

7 Conclusion

The understanding of customer’s perception is an ongoing process to survive the cut throat competition. Taking this into consideration, the study provides insights about the several attributes which drive a customer’s behavior toward buying a product of a brand. The study has used theories and exploratory research design along with analytical tools to identify the major attributes of consumer buying behavior toward sedan cars within Delhi/NCR among high-end consumers. This

knowledge helps car manufacturers in market segmentation which enables the organization to plan the market strategies toward consumer retention and product upgradation.

We have the output of factor analysis for this problem—the un-rotated factor matrix, the final statistics comprising the communality for all 19 variables, and the eigenvalues of all factors having eigenvalues of more than 1. The seven factors are extracted using factor analysis namely, brand reputation, physical appearance, features, purchase services, technical features, value addition, and performance. These variables capture approximately 70% of the total variance information content out of the 19 variables under study.

Under the Brand Reputation, price is the most influential factor having factor loading of .863, followed by warranty under Purchase Services having factor loading of .812, design and exterior looks under physical appearance with factor loading of .768, interior look under features with factor loading of .754, engine strength under technical features with factor loading of .783, value for money under value addition with factor loading of .863, and fuel economy under performance with factor loading of .863.

The automobile companies must focus on their male consumers falling under the age group of 25–35 years who lie in the business class within an income group of 12–17 lacs. The marketing strategy, advertisement, and price strategy should be specially designed for this segment of the consumers based on good mileage, luxury, comfort, and engine strength and fuel economy.

The automobile companies should also pay attention to additional factor, i.e., the time span of car switch among the customers to retain them. The customers upgrade their cars usually between a period of 3–5 years by either purchasing a new car of their trusted brand or switching the brand. So, to retain the customers, it is important that the companies keep a focus on R&D to develop new upgraded cars and intimating their existing customers from time to time.

Acknowledgements This manuscript is supported by different individuals (students and faculty) associated with Amity University campuses: Ref. No. AUGN/Cert./VP(P) & R/2018/10/01. All procedures performed in the study involving human participation were in accordance with the ethical standards of the institution, and comparable ethical standards. Prior public disclosure approval to publish the study results have been taken by the university. The anonymity of the respondents of the questionnaire and data has been treated strictly confidential and not disclosed at any level.

Appendix

Tables 6, 7, 8, 9, 10.

References

1. Amy Van Looy, A.S.: Business process performance measurement: a structured literature review of indicators, measures and metrics. *5*(1) (2016)
2. Banerjee, S.: Study on consumer buying behavior during purchase of a second car. *J. Mark. Commun.* **6**(2) (2010)
3. Child, D.: *The Essentials of Factor Analysis*. Continuum International Publishing Group, New York (2006)
4. Field, A.P.: *Discovering Statistics Using SPSS for Windows: Advanced Techniques for the Beginner*. Sage, London (2000)
5. Field, D.A.: Chapter 14: Factor Analysis Using SPSS; *Discovering Statistics Using SPSS*, 2nd edn. Sage, London (2005)
6. Gabay, J.: *Brand Psychology: Consumer Perceptions, Corporate Reputations*. Kogan Page, London (2015)
7. Jadczaková, V.: Review of segmentation process. *ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS* 1215–1224 (2013)
8. Lorenzo-Seva, U.: How to report the percentage of explained common variance in exploratory research. *Tarragona* (2013)
9. Mathur, D.: Consumer buying behaviour of cars in India - a survey. In: *1st International Conference on New Frontiers in Engineering, Science & Technology*, New Delhi, India, pp. 468–473 (2018)
10. Menon, B.: A study on consumer behaviour of passenger car segments through logistic regression modelling 20–32 (2017)
11. Monga, N.: Car market and buying behavior- a study of consumer perception. *Int. J. Res. Manag. Econ. Commer.* **2**(2) (2012)
12. Motoyuki Akamatsu, P.A.: Automotive technology and human factors research: past, present and future. *Int. J. Veh. Technol.* **27** (2013)
13. Pearce, A.G.: A beginner's guide to factor analysis: focusing on exploratory factor analysis. *Tutor. Quant. Methods Psychol.* 79–94 (2013)
14. Raju, P.T.: Factors influencing consumer behaviour for buying luxury cars. *Int. J. Civ. Eng. Technol.* **292** (2018)
15. Sachdeva, N.: An innovation diffusion model for consumer durables with three parameters. *J. Manag. Anal.* **3**(3) (2016)
16. Seunggeun Lee, F.Z.: Convergence and prediction of principal component scores in high dimensional settings. *Ann. Stat.* 3605–3629 (2010)
17. Shende, V.: Analysis of research in consumer behavior of automobile passenger car customer. *Int. J. Sci. Res. Publ.* **4**(2) (2014)
18. Stebbins, R.A.: *Exploratory Research in the Social Sciences*. Sage, London (2001)
19. Venkatesh Shankar, A.K.: Customer satisfaction and loyalty in online and offline environments. *Int. J. Res. Mark.* 153–175 (2003)
20. Watkins, D.S.: *The matrix eigenvalue problem: GR and Krylov subspace methods*. SIAM (2007)

Secure Online Voting System Using Biometric and Blockchain



Dipti Pawade, Avani Sakhapara, Aishwarya Badgujar, Divya Adepu and Melvita Andrade

Abstract Elections play an important role in democracy. If the election process is not transparent, secure and tamper proof then the reliability and authenticity of the whole process is at stake. In this paper, we have discussed an online voting system which fulfills all the above system requirements. We have addressed the issue of user authentication through iris recognition. We have used One Time Password (OTP) to have an additional security check. We have also taken care that one valid user should not cast multiple votes. Use of Blockchain is another security measure implemented in order to provide decentralized, tamper proof storage of data related to users' biometric, personal details and votes casted by them. Thus we are not only focusing on user authenticity but also data security is also taken into consideration. The performance of the system has been tested for users from different age groups and different backgrounds and its inference is presented.

Keywords Online voting · Blockchain · Sidechain · Iris recognition

1 Introduction

In a democratic country like India, an election is considered as a vital process. Right from choosing the Prime Minister to the ward member, every election is carried out using a voting system. Earlier, the ballot paper voting system was used in the election process. Here people need to visit the appropriate polling booth and need to show the voter ID card to prove their authenticity. Casting of a vote is done by stamping on the ballot paper secretly and then the ballot paper is folded and deposited in the ballot box. After the voting is done, the ballot boxes are sealed and moved to the counting center under security and counting of votes is carried out manually. But there were some incidences of manipulation with ballot boxes. Also, this complete process took longer time for result declaration, required tremendous human resources and thus incurred high

D. Pawade (✉) · A. Sakhapara · A. Badgujar · D. Adepu · M. Andrade
Department of IT, K.J. Somaiya College of Engineering, Mumbai, India
e-mail: diptipawade@somaiya.edu

costing. Because of all these reasons, Electronic Voting Machine (EVM) is introduced in 1982 in 70-Parur Assembly Constituency of Kerala on trial basis [1]. At mass level it is used in 1998 assembly elections in Madhya Pradesh, Rajasthan and the NCT of Delhi. After that, EVM is being used in all the elections conducted by Election Commission of India. EVMs are under purview of Election Commission of India only and cannot be used for other elections like university senate elections, society election or elections carried out in various organizations. Thus they are still using the traditional ballot paper base election process.

In this paper we have addressed to the various issues related to the university senate election, where electoral roll of graduate teachers is made electronically but still ballot paper is used for vote casting. It has been observed that percentage of votes casted is very low as compared to the total number of eligible voters. One of the reasons for this reluctance is that for casting vote one needs to travel to the particular place. Most effective solution to this is like preparing electoral list and conducting election through e-portal. But due to security concerns, this option is ruled out. Hence we are proposing the concept of online voting system where voter identification is done using biometric iris recognition. As biometric trait is the basic parameter under consideration to verify the genuine voter, security of iris template database is also our concern. For that purpose we are storing all the details of voter including iris feature on Blockchain, so that it cannot be tampered. The system takes care of the issue related to multiple votes casted by single voter. Once the pre-set time of election conduction is over the portal gets locked and result statistics are automatically emailed to the election candidates and concerned authority.

Further outline of the paper is as follows: Sect. 2 consists of related work in the area of applications of online elections and research in iris recognition system. In Sect. 3 overview of Blockchain and related technologies is given. In Sect. 4, the design and implementation of the proposed online voting system is discussed. In Sect. 5 results are presented and in Sect. 6 conclusion and future work is stated.

2 Related Work

2.1 Literature Survey on E-Voting Application

Khasawneh et al. [2] discussed the implementation of biometric based secure e-voting system. For voter identification, initially they have used official voter ID cards which need to insert in magnetic card reader to fetch the details of the voters from local or central database. If record is found then the individual's fingerprints are scanned and matched against the pre-stored database. Thus only authenticated person is allowed to vote and once voting is done then flag is set so that same person cannot do the voting again. Main concern of this approach is that the authentication is solely dependent on central Authentication server. If that server is down, whole voting system will be affected.

Petcu et al. [3] describes a mobile e-system consisting of various phases. Firstly one needs to enroll themselves by visiting physically to the authentication center. Here they need to provide some valid identity proof and if person is eligible to vote then his/her mobile number, voice and fingerprint samples are collected and key token is given to the voter. In authentication phase, the user has to access the web portal and has to provide the ID card details and key token provided at the time of enrolment. The system will then send a SMS key on successful collection of three voice samples and fingerprint samples for both the hands of a voter. Then one needs to provide the unique code, key token and SMS key which is verified by the authority and then vote access is given to that voter. Now voter can use e-voting module to cast the vote securely and vote counting is also done automatically. Here author have used so many key values which need to be provided, which make the system complicated to use.

Sridharan [4] considered the smart card, voter identification number (VID) and fingerprints to authorize and uniquely identify the individual. The complete system has four modules viz. Franchise Excising Terminal to recognize smart card, Authentication Terminal, the Distributed Databases and Central Controlling Server. Use of distributed database makes the system scalable. But still biometric data security concern exists.

Agarwal et al. [5] proposed the AADHAR ID based online authentication system where voter first needs to provide AADHAR ID which is being verified by authority and then password is mailed to the particular person. One needs to login into online voting system to cast the vote and it is accepted only if fingerprint matching is done successfully. Author has not given the insight to how templates for fingerprint matching are collected and how the fingerprint server database is maintained.

Table 1 gives the overview of the different parameters used in various e-voting applications for voter authentication.

Table 1 Overview of available e-voting application

S. No.	Author	Parameter used for authentication
1	Khasawneh et al. [2]	Specially designed magnetic tape based Voter ID card, fingerprints
2	Petcu et al. [3]	Physical Identity proof, Voice sample, fingerprint of both hands, key token, 4 digit key, SMSkey
3	Sridharan [4]	Smart card, VID and fingerprints
4	Agarwal et al. [5]	AADHAR ID, Password, fingerprints

2.2 Literature Survey on Iris Recognition System

Garagad et al. [6] proposed iris identification system which is claimed to be invariant to distance and eye tiltation. They have Performed segmentation to take out the region of interest. It involves low intensity region recognition in input image, detecting iris and boundary inference point (edge detection) followed by selective image cropping followed by pupil localization and normalization. For feature extraction first cartesian plane data of an image is obtained through signature code generation and then it is converted to a binary iris signature code to optimize the memory utilization and speed up the processing. Finally for iris matching binary signature code of input iris image and template image are compared using XOR bitwise operation. The author [6] concluded that, this approach is simple to implement and effective way of iris authentication in case of scale and tilt variance.

Neda et al. [7] have suggested hybrid approach for iris recognition model based on a multi-layer perceptron NN (MLPNN) and particle swarm optimization (PSO) algorithms. Here, to separate the sclera and pupil, iris localization is carried out using circular Hough transform algorithm [8, 9]. Then normalization is carried out to convert image from the space of the Cartesian coordinates to polar coordinates using rubber-sheet model of Daugman followed by image segmentation [10]. 2D gabor kernel method is applied to the pre-processed image to extract the features. For pattern matching author have proposed a hybrid approach based on ANN-based MLP and PSO algorithm and called it as MLPNN-PSO algorithm. This hybrid approach achieves 95.36% accuracy rate which is much better when these approaches are used separately.

Muthana et al. [11] discussed two methods of iris feature extraction viz. the Fourier descriptors (FD) and Principle Component Analysis (PCA). Before feature extraction, iris segmentation is carried out using pupil and iris localization. The pupil localization consists of following steps: smoothing using non-linear median filter, binarization, and noise removal using morphological operation and finding the iris center. While the iris localization is done by applying Canny masks which is used to locate iris boundary and circular Hough transform which is used to get complete circular shape. Next step in pre-processing is normalization using rubber sheet model. Then for feature extraction FD and PCA is used and their results are classified using Manhattan, Euclidean and Cosine distance. According to their result discussion accuracy rate of FD with Manhattan distance is highest i.e. 96%.

Table 2 gives the overall summary for different iris recognition system based on various parameters.

Table 2 Summary of iris recognition system

Database	Preprocessing	Feature extraction	Classifier	Recognition rate	Peculiarity
High-quality UPOL iris images captured in constrained environment [6]	Segmentation	Gray signature code generation and binary signature code generation	Bit to bit comparisons of binary signature codes using XOR operation	99.64 (for Normal Test input images) 90.698 (for tilt & scale variation)	Approach is invariant to distance and tilt variations
(CASIA)-iris V3 and UCI [7]	Circular Hough transform algorithm, rubber-sheet model	2D Gabor kernel method	Hybrid MLPNN–PSO algorithm	95.36%	MLPNN and PSO algorithm gives better
CASIA v1 [11]	Pupil and Iris localisation, iris normalization	Fourier descriptors (FD) and Principle Component Analysis (PCA)	Manhattan, Euclidean and Cosine distance	96%	FD is better than PCA and FD with Manhattan gives highest accuracy

3 Overview of Blockchain Technology

Database is considered as a very important component when we are talking about full stack application development. Traditional databases follow the client-server architecture, where client/user has a provision to perform CRUD operations (Create, Read, Update, and Delete) and data is stored on a centralized server. As all clients have the rights to modify or update the data, there is a database administrator who authorizes the client and makes sure that the data is not being compromised. But here as data is maintained in centralized database, the whole system gets hampered if database server crashes. Also the load on one server is more thus the overall computation speed is low. To address this problem the concept of distributed database was introduced. A distributed database is considered as a single logical database where nodes are physically distributed across several geographic locations and communicate with each other over a network. This kind of database is normally suited for an organization, where all trust each other. Here in case if any node fails the system continues to work with other node.

Blockchain is said to be a decentralized database where data integrity and transparency is considered as parameter for public verifiability [12]. Opposite to traditional database, Blockchain allows only read and write operations to be performed on the data. In read operation user can read/query the data on Blockchain

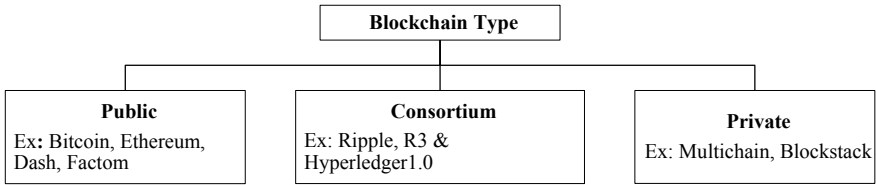


Fig. 1 Types of Blockchain

while write operation used to add the data on Blockchain. Whenever write operation is performed an additional block is appended in Blockchain. All previous data is available in block which cannot be altered. If it is being altered then it is added as new block and the original block data remains the same [13]. Thus it is considered as scam free, transparent system.

Figure 1 shows different types of Blockchain. Public Blockchain is also called as permissionless Blockchain, as there is no authority to sanction any Blockchain transaction. This means process of consensus building is public. Second type of Blockchain is consortium which is also known as public permissioned Blockchain. Here consensus process is monitored by a pre-selected group of nodes and for transaction verification minimum clearance is required. Private/private permissioned Blockchain is the third type of Blockchain where write operation rights are given to set of specific nodes while read rights can be public or restricted to specific nodes [14].

3.1 Sidechain

A sidechain is considered as separate blockchain that is attached to the blocks on the main blockchain. Main chain is also referred to as “Parent Blockchain” while sidechain is referred to as “childchain”. The assets can be interchanged between the main chain and the sidechain at a predetermined rate. The major advantage of using sidechain is that one can easily develop and deploy a quickly scalable Blockchain solutions that too at lower cost and under high level of security [15].

3.2 Ethereum Platform

For our application development we are using Ethereum public Blockchain platform which allows the development of decentralized application (dapp). Ethereum platform has two building blocks viz. Ethereum Virtual Machine (EVM) and peer-to-peer network protocol. In order to get consensus for transaction, each node

in network runs the EVM and processes the instruction. In simple words EVM is a run time environment which converts the smart contract into EVM bytecode and deploys it onto Blockchain for execution. Here smart contracts are the basic building blocks of a program which consist of contract instructions which are simply set of instructions to be executed upon certain event. We have written our smart contract in solidity language [15].

3.3 *Metamask*

Earlier the biggest issue with using dapps was, using normal browsers it was not possible to write to the Blockchain. There were special Blockchain browsers like Mist but using it was very tricky and thus the whole development process became tedious. The solution to this issue is provided in the form of Metamask which is an Ethereum web browser extension. Metamask basically acts like a Ethereum wallet which stores Ethereum related data like private keys, public addresses etc. MetaMask is considered as the easiest way to interact with dapps in a browser. It allows you to connect to an Ethereum network without running a full node on the browser's machine [16].

4 **Implementation Approach**

As shown in Fig. 2, system has three main modules running on client side,

- Voter enrollment module using which user needs to enroll themselves as a authorized voter
- Election authority interface through which the election process is initiated and election result calculation command is triggered.
- Election interface, which verifies the identity of the user as valid voter and provides interface to cast the vote.

In order to connect our web user interface with the Ethereum Blockchain we have used Metamask. The overall system implementation is carried out as follows:

4.1 *Working of Proposed System*

As per the AICTE rules, every institute affiliated to the university needs to send the details of their employees to the university. These details include information like name, AADHAR ID number (UID), Email ID, PAN number, address etc. So before university election, name and Email ID and UID of eligible candidates are collected

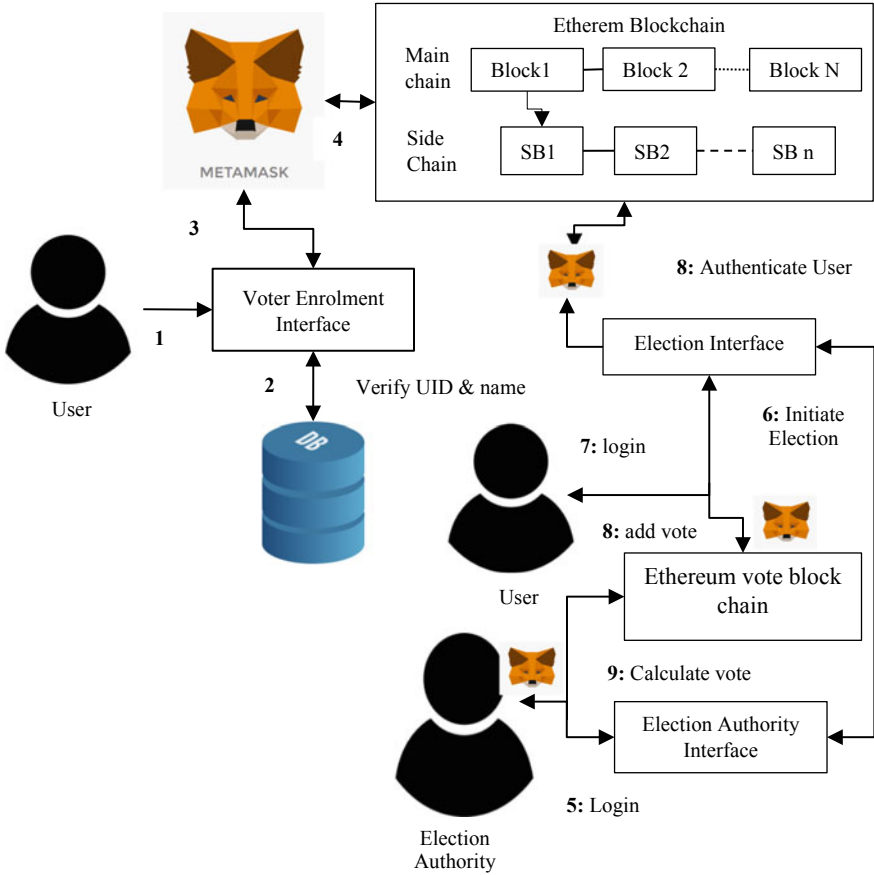


Fig. 2 Block diagram of proposed online voting system

from each institute and stored in database. A bot is designed to send auto-generated mail to notify eligible voters regarding voter enrollment deadline and guidelines. This bot will fetch the email IDs from the database and will send template mail to all specifying the link for registration and the description of the procedure to be followed.

Then the user needs to visit our election portal, for registration each user needs to provide basic information like name, address, valid mobile number and UID. This UID and name of the user is matched against the database for verification. If it gets matched then user has to choose unique username and password for them. This combination of username and password will be generated after successfully satisfying the required validation rules. For verification, link is sent on the respective Email Id. User needs to verify themselves by clicking on that link. Once verification is done, all the user data is pushed on Ethereum Blockchain through smart contract, then mining is carried out and genesis block is created [17].

The system will then ask the user to scan the iris. User can use webcam or front camera of mobile to take a picture of iris. The iris image is preprocessed and features are extracted. Preprocessing and feature extraction process is explained in Sect. 4.2. Once features are extracted, it is stored on associated sidechain along with other information like name, address, mobile number.

On the election day, the authorized person will instantiate the election portal. User have to login using the appropriate username and password which was generated during the enrollment phase. Then the user needs to provide his/her valid UID. System will then try to find the block on Blockchain with same UID. If the block is not found, then error message will be displayed otherwise system will scan the iris image through web-cam or front camera of mobile. Preprocessing and feature extraction is done using step in Sect. 4.2. These features are matched against the features stored on sidechain at the enrollment phase. If iris features do not match then the iris image is scanned again and matching is carried out. If match found, then One Time Password (OTP) is generated and sent to the mobile number stored on sidechain for that user. This OTP is valid for 40 s only. User needs to get verification done before that OTP expires. Because of some reason if OTP expires then the user can request for OTP again. Once OTP verification is done, e-ballot will appear on screen containing names and symbols of the candidates. In order to cast vote user needs to select the radio button in front of the particular candidate. Once voting is done flag is set for this user, so that same user should not cast the vote again. The vote casted by user is again pushed on Blockchain through smart contract and genesis block is created on voting Blockchain after mining is done.

The election authority person will disable the election portal once election period is over. On election authority console, calculate button is provided which will fetch all the votes from Blockchain, and calculate the number of votes gained by each candidate. Then the system will mail the result to the authorized members of election committee and the election candidates. Figure 3 depicts the flow of activities when user interacts with the election interface.

4.2 Iris Recognition System

Iris recognition system consists of following three steps:

- Preprocessing
- Feature Extraction
- Feature matching

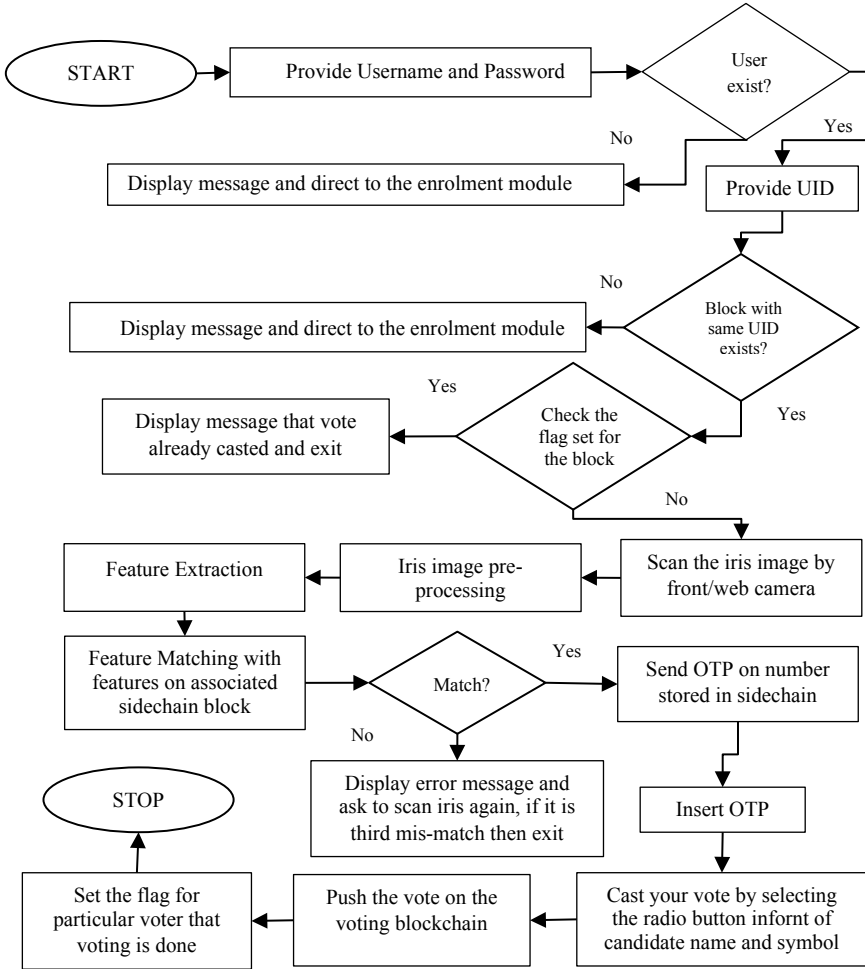


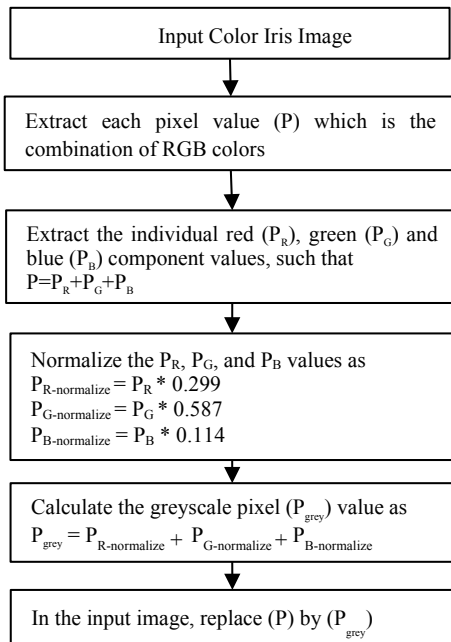
Fig. 3 Flowchart for online voting system for enrolled

4.2.1 Preprocessing

The first step is to convert the colored iris image to a greyscale image which contains only different shades of grey lying in the range of 0–255. This step is important because it reduces the unnecessary information in each pixel [18].

This grey color has the red, green and blue components in equal intensity in the RGB space. So henceforth, we only have one value in every pixel as compared to three values for every pixel of a colored image. Figure 4 depicts the process of color image conversion to greyscale image for any pixel at position (x, y) in colored image.

Fig. 4 Steps for color to greyscale Image conversion



The next step is to detect the edges of the iris and pupil. It works by detecting discontinuities in brightness. Sobel operator has been used for this purpose.

In sobel, we apply the derivative mask in horizontal and vertical direction to increase the edge intensity. The resultant is a faint boundary of the iris and pupil depending on the brightness of the image. It works by calculating the gradient of image intensity at each pixel within the image. It finds the rate of change of brightness.

The horizontal gradient G_x is the image that we get after applying the horizontal mask on the greyscale image (A) which is calculated as using the mask in Fig. 5a and the vertical gradient G_y is the image that we get after applying the vertical mask on the greyscale image (A) which is calculated by using mask in Fig. 5b. The masks contain positive and negative coefficients. This means the output image will also contain positive and negative values. The negative values are set to 0. The resultant image is the sum of the horizontal and vertical gradients.

(a)
$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * A$$

(b)
$$G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A$$

(c)

p4	p3	p2
p5	p	p1
p6	p7	p8

Fig. 5 a Sobel vertical mask, b Sobel horizontal mask, c CN mask

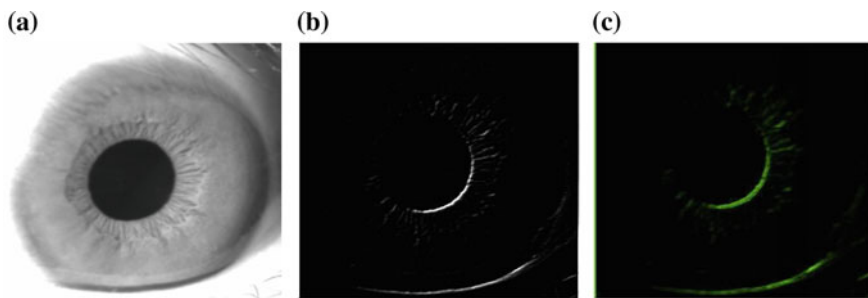


Fig. 6 a Sample input iris image, b after applying Gabor filter, c after applying CN mask

The next step is to use Gabor filter to analyze the texture of the iris. It basically does this by analyzing whether there is any specific frequency content in the iris in the localized region inside the iris circle. The image is convolved with a series of Gabor wavelets to perform a filtering operation. The buffered image is filtered and the resultant is produced as a rendered image [19, 20] (Fig. 6).

4.2.2 Feature Extraction

Minutiae representation is most widely used method of any biometric representation as it helps to determine uniqueness in an image. Minutiae or small details mark the regions of local discontinuity within a biometric image. For detecting minutiae points we have used Rutovitz's Cross Numbering technique.

This method extracts the minutiae points by identifying the bifurcation points (forking points) using a 3×3 mask on the local neighborhood pixels of the ridge pixel. These properties identify various properties (minutia) of iris like ridge endings, ridge bifurcations, etc. Figure 5c shows 3×3 Mask used for CN technique [21].

$$CN = \frac{1}{2} \sum_{i=1}^8 |P_i - P_{i+1}| \quad (1)$$

Using Eq. (1) we get iris feature vector. This output is then used for the matching algorithm.

4.2.3 Matching by Average Method

Step 1:

Here, each pixel is represented as 0xAARRGGBB (alpha, red, green, blue). By performing a bitwise-and with 0xFF, the individual RGB component is calculated as:

$$\begin{aligned} \text{red} &= (\text{rgbA} \gg 16) \& 0\text{ff} \\ \text{green} &= (\text{rgbA} \gg 8) \& 0\text{ff} \\ \text{blue} &= (\text{rgbA}) \& 0\text{ff} \end{aligned} \tag{2}$$

Step 2:

We calculate separate RGB values using Eq. (2) for both input and template iris images and calculate sum of difference between each of the values as

```
difference += Math.abs(redA - redB);
difference += Math.abs(greenA - greenB);
difference += Math.abs(blueA - blueB);
```

Step 3:

Then we calculate total number of pixels:

Number of pixels for each of the RGB component is calculated as given in Eq. (3) and then the summation of number of pixels for each RGB component is taken which gives the total number of pixels.

$$\text{Number of pixels for each RGB component} = \text{width} * \text{height} \tag{3}$$

Step 4:

For normalizing the value of different pixels and for accuracy we use average pixels per color component calculated as

```
average_different_pixels = difference/total_pixels
```

Step 5:

As there are 255 values of pixels(RGB) in total thus

```
Percentage = (average_different_pixels/255) * 100 * 10;
```

If percentage is greater than threshold value then it is iris mismatch otherwise it is considered as iris matched and the user is authenticated.

5 Results and Discussions

For the experimental purpose we have tested the system for 120 users. Now-a-days everyone has the basic knowledge of how to use the web based application. Still while using technology, comfort can be one of the key concerns in order to state the user friendliness of the application. Thus we have made 4 groups with 30 users in each group. Here users are the faculties of different age group and different background. The group details are as follows:

- Group A: contains users who are the faculties having age in between 27 and 42 years and belong to non-technical stream like Literature, Arts, Communication skill and Chemistry.
- Group B: contains users who are the faculties having age in between 27 and 42 years and belong to technical stream like Engineering, Polytechnic and Management.
- Group C: contains users who are the faculties having age in between 43 and 57 years and belong to non-technical stream like Literature, Arts, Communication skill and Chemistry.
- Group D: contains users who are the faculties having age in between 43 and 57 years and belong to technical stream like Engineering, Polytechnic and Management.

Table 3 shows the experimental results for users of different group based on age and their stream. For experimentation every group has used computer system and web camera of same configuration and same Internet speed. We have considered the standard environment with enough luminance and no tilt in iris image. Still it has been observed that there are few users whose iris matching is not done thus they

Table 3 Experimental results for users of different group based on age and their stream

Criteria	A	B	C	D
Average Time required for Enrollment (Sec)	4.38	3.89	5.99	4.46
Average Time required for Voting (Sec)	12.73	11.45	16.96	14.01
No. of user proceeding further with first OTP	24	25	19	25
No. of user not requesting OTP twice	02	02	03	00
No. of people not requesting OTP thrice or even more (max. 5 times)	00	00	01	01
No. of user for whom iris recognition is done in first attempt	16	21	10	18
No. of user for whom iris recognition is done in second attempt	08	06	07	03
No. of user for whom iris recognition is done in third attempt	02	00	06	05
No. of user for whom iris recognition is not done thus not casted the vote	04	03	07	04
No. of user casted the vote successfully	26	27	23	26

could not complete the voting process. We also found that technical expertise of the person and how well-versed the person is in dealing with web application and navigation has an impact on the total time required to complete the enrollment as well as voting process.

In second phase of experiment we have considered two groups viz. E and F each having 30 users. For group E, we asked the users to enroll during night with tilts and for group F enrollment is done during day time with tilts while voting for both the groups is carried out during day time. Here all the users are technically sound and between the age group of 27–48. For group E out of 30, only 19 users successfully casted the vote. For rest of the 11 users, biometric authentication failed. The reasons behind that is template iris features for this group is extracted and stored from image which is taken in night when luminance was poor. Thus the luminance plays an important role in the performance of system. For group F, out of 30, 23 people successfully did the iris authentication and casted vote. Thus even tilt in the iris image angle hampers the accuracy of the system.

Figure 7 gives the success and failure rate for all six groups. Same statistics can be considered for iris recognition accuracy and error rate of the system of the system. Thus the average accuracy of the system is 81.11% and average failure rate is 18.89%.

In Table 4, we have discussed the advantages of our proposed system over the existing system which are described in Sect. 2.1. In our system, we have completely eliminated the overhead of maintaining magnetic tape based voter ID cards or smart cards. This reduces the time as well as money required to complete the whole election process. Also there is no need for the voter to maintain these cards. We have also reduced the burden to generate and remember various security keys in order to use the system. Most of the researchers have used fingerprints for biometric

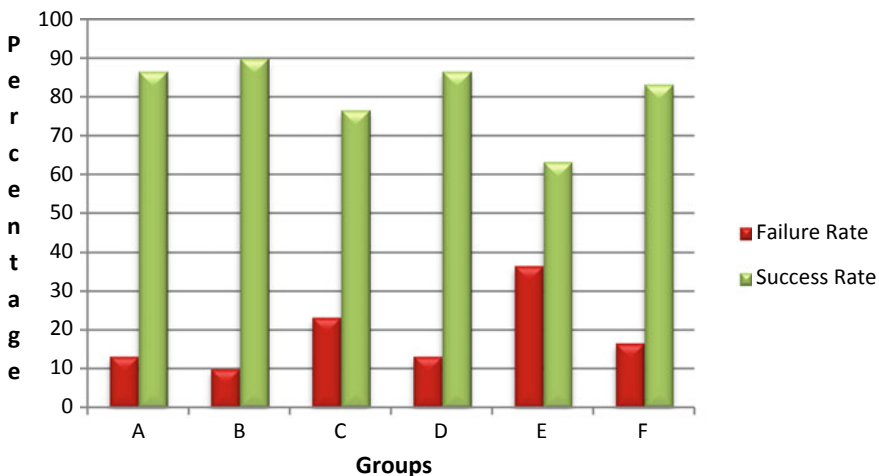


Fig. 7 Summary of success and failure rate for all five groups

Table 4 Advantage of proposed system over existing system

S. No.	Existing system discussed in Table 1	Advantage of proposed system over existing system
1	Khasawneh et al. [2]	No overhead of magnetic tape based Voter ID card, cost is low
2	Petcu et al. [3]	Easy to use as there is no need to generate and remember different keys
3	Sridharan [4]	Smart cards are not required to use the system
4	Agarwal et al. [5]	Biometric data storage is tamper proof and whole system is a decentralized, integrated and transparent

authentication in their system. Instead of fingerprints; we have used iris for biometric authentication as fingerprints are more vulnerable to attacks as compared to iris.

6 Conclusion and Future Work

This paper elaborates a technique of online voting empowered with biometric authentication and Blockchain Technology. By using iris recognition system, we make sure that only a verified person can cast the vote. Also as compared to the existing or proposed voting systems, our system does not store the captured iris image in the database. Instead, it extracts the features of the iris image and then removes the iris image from memory which makes the biometric data more secure. These extracted features along with all the other information of the user and the votes casted by them are stored on Blockchain. Thus system becomes transparent, decentralized and mishaps like information tampering or vote alternation can be easily detected. As now the average accuracy of the system is 81.11% and average failure rate is 18.89%. Accuracy rate is low because we have performed the experiment in low luminance and iris tilt variation. Thus capturing proper iris image with minimum noise is the prime requirement of the system. In future we are planning to improvise our iris recognition system, so that the overall system should work under non ideal condition without any constraint.

7 Acknowledgement

This research was carried out as an extension of final year dissertation titled “Fingerprint Authentication System Using Blockchain” for the degree of Bachelor in Information Technology of University of Mumbai, under the guidance of Prof. Dipti Pawade and Prof. Avani Sakhapara, at K.J. Somaiya College of Engineering, Mumbai, India. We express our gratitude to the Institute for its constant support and guidance. This research was carried out by adhering to the applicable ethical

standards and in accordance with the rules and regulations of the Institute. For this research, prior consent was taken from the participants to publish the results of research in public domain ensuring complete anonymity and confidentiality of the data.

References

1. http://eci.nic.in/eci_main1/evm.aspx
2. Khasawneh, K., Malkawi, M., Al-Jarrah, O., Barakat, L., Hayajneh, T.S., Ebaid, M.S.: A biometric-secure e-voting system for election processes. In: Proceeding of the 5th International Symposium on Mechatronics and its Applications (ISMA08), Amman, Jordan, May 27–29, 2008
3. Petcu, D., Stoichescu, D.A.: A hybrid mobile biometric-based e-voting system. In: 9th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, 7–9 May 2015, pp 37–42
4. Sridharan, S.: Implementation of authenticated and secure online voting system. In: 4th Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) 2013, Tiruchengode, India No. 6, July 2013. IEEE—31661
5. Agarwal, H., Pandey, G.N.: Online voting system for India based on AADHAAR ID. In: Eleventh International Conference on ICT and Knowledge Engineering, Bangkok, Thailand, 20–22 Nov 2013
6. Garagad, V.G., Iyer, N.C.: A novel technique of iris identification for biometric systems. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, India, 24–27 Sept 2014
7. Ahmadi, Neda, Akbarizadeh, Gholamreza: Hybrid robust iris recognition approach using iris image pre-processing, two-dimensional gabor features and multi-layer perceptron neural network/PSO. *IET Biom.* **7**(2), 153–162 (2018)
8. Wildes, R.P., Asmuth, J.C., Green, G.L., Hsu, S.C., Kolczynski, R.J., Matey, J.R., McBride, S.E.: A machine-vision system for iris recognition. *Mach. Vis. Appl.* **9**(1), 1–8 (1996)
9. Ahmadi, N., Akbarizadeh, G.: Iris recognition system based on canny and LoG edge detection methods. *J. Soft Comput. Decis. Support Syst.* **2**(4), 26–30 (2015)
10. Daugman, J.: High confidence recognition of persons by rapid video analysis of iris texture. In: *IET European Convention on Security and Detection*, pp. 244–251 (1995)
11. Hamd, Muthana H., Ahmed, Samah K.: Biometric system design for iris recognition using intelligent algorithms. *Int. J. Modern Educ. Comput. Sci. (IJMECS)* **10**(3), 9–16 (2018). <https://doi.org/10.5815/ijmeecs.2018.03.02>
12. David S., Harvey L., Vimi G., Alexander S., Stephen M., Tyler W.: What is blockchain? *Blockchain Enigma. Paradox. Opportunity – Deloitte*, 2016
13. Sloane B., Bhargav P.: *Blockchain basics: introduction to distributed ledgers* (2016). <https://www.ibm.com/developerworks/cloud/library/cl-blockchain-basics-intro-bluemix-trs/index.html>
14. Pawade, D., Jape, S., Balasubramanian, R., Kulkarni, M., Sakhapara, A.: Distributed ledger management for an organization using blockchains. *Int. J. Educ. Manag. Eng. (IJEME)* **8**(3), 1–13 (2018). <https://doi.org/10.5815/ijeme.2018.03.01>
15. <https://www.ethereum.org/>
16. <https://metamask.io/>
17. Pawade, D., Pawade, D., Sakhapara, A., Andrade, M., Badgujar, A., Adepu, D.: Implementation of fingerprint based authentication system using blockchain. In: *The International Conference on Soft Computing and Signal Processing (ICSCSP-2018)*, Hyderabad, 22–23 June 2018

18. Pawade, D., Chaudhari, P., Sonkambale, H.: Comparative study of different paper currency and coin currency recognition method. *Int. J. Comput. Appl.* **66**(23), 26–31 (2013)
19. Adhau, A.S., Shedge, D.: Iris recognition methods of a blinked-eye in non-ideal condition. In: *IEEE International conference on Information Processing*, pp. 75–79. ISBN: 4673-7758 (2015)
20. Puhan, N.B., Sudha, N., Xia, H., Jiang, X.: Iris recognition on edge maps. In: *IET Computer Vision*, 5th September 2007, <https://doi.org/10.1049/ietcvi:20080015>
21. <https://github.com/yoga1290/Fingerprint-Recognition>

An Approach: Applicability of Existing Heterogeneous Multicore Real-Time Task Scheduling in Commercially Available Heterogeneous Multicore Systems



Kalyan Baital and Amlan Chakrabarti

Abstract Interest in design and use of heterogeneous multicore architectures has been increased in recent years due to the fact that the energy optimization and parallelization in heterogeneous multicore architecture are better than that of homogeneous multicore architecture. In heterogeneous multicore architectures, cores have similar Instruction Set Architecture (ISA) but the characteristics of the cores are different with respect to power and performance. Hence, heterogeneous architecture provides new prospects for energy-efficient computation and parallelization. Heterogeneous systems, furnished with different types of cores provide the mechanism to take actions with respect to irregular communication patterns, energy efficiency, high parallelism, load balancing, and unexpected behaviors. However, designing such heterogeneous systems for the different platforms like cloud, Internet of Things (IoT), Smart Devices, and Embedded Systems is still challenging. This paper studies the commercially available heterogeneous multicore architectures and finds out an approach or method to apply the existing work on heterogeneous multicore real-time task scheduling model to commercially available heterogeneous multicore architecture to achieve the parallelism, load balancing, and maximum throughput. The paper shows that the approach can be applied very efficiently to some of the commercially available heterogeneous systems to establish a generic heterogeneous model for the platforms like cloud, Internet of Things (IoT), Smart Devices, Embedded Systems, and other application areas.

Keywords Big core · big.LITTLE · Heterogeneous · Multicore · Octa-core · Real-time task · Small core · Xeon Phi

K. Baital (✉)

National Institute of Electronics and Information Technology, Kolkata Centre, Kolkata, India
e-mail: kalyan_baital@yahoo.co.in

A. Chakrabarti

A. K. Choudhury School of Information Technology, University of Calcutta,
Kolkata, West Bengal, India
e-mail: amlanc@ieee.org

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_8

111

1 Introduction

Heterogeneous multicore systems are increasing day by day in the field of embedded systems as well as cloud environments to improve system throughput, parallelization, load balancing, and energy efficiency. Heterogeneous multicore system includes cores with various types or/and complexities to address throughput and productivity for different applications. Different types of heterogeneous multicore systems based on different application areas are commercially available in the market to offer a high degree of parallelism and energy efficiency. Thus, there is an increasing motivation to design a generic heterogeneous model for different application domains including the following technological domains which are in high demand in the market today:

- i. Cloud and IoT infrastructure which is high-performance and high-resource domain and offering a high degree of parallelism.
- ii. Smart devices and embedded domain which is low-performance and low-resource domain but offering great energy efficiency.

Therefore, our focus is to find an approach to establish a generic heterogeneous model which can be utilized as high-performance and high-resource domain (i.e., cloud and IoT). The same generic model can also be used as energy efficient and low resource domain (i.e., smart devices and embedded systems) along with other common application areas.

Heterogeneous multicore system has achieved significant importance in the area of High-Performance Computing (HPC) which offers a higher degree of parallelization, and hence it has enormous potential for future use in data centers and cloud platforms. In this regard, heterogeneous architectures such as Intel Xeon Phi coprocessor and Graphics Processing Unit (GPU) have significant roles in HPC systems.

Recently, heterogeneous architecture is being used in the mobile technology such as ARM big.LITTLE octa-core architecture. The usage pattern for smartphones and tablets varies drastically, for example, high-processing usage likes mobile gaming and web browsing and low-processing usage likes texting, e-mail, and audio. Octa-core heterogeneous architecture such as ARM big.LITTLE fulfills the demand of these different performance requirements by combining two different types of cores.

Hence, in our proposed approach, we apply different architectural parameters of Intel Xeon Phi coprocessor in our proposed generic model such a way that the generic model can additionally be used as cloud and IoT platform along with several common application areas. Further, we apply different architectural parameters of octa-core (ARM big.LITTLE) in the proposed model for using it as smart devices and embedded systems.

The following significant contributions are explained in this paper:

1. To study the architecture of some of the commercially available heterogeneous multicore architecture such as Intel Xeon–Xeon Phi platform and big.LITTLE mobile octa-core platform.
2. To find an approach to apply the existing heterogeneous multicore real-time task scheduling model as per work [1] into the commercially available heterogeneous multicore systems (Intel Xeon–Xeon Phi platform and big.LITTLE mobile octa-core platform) to achieve the parallelism, load balancing and maximum throughput.
3. To provide the performance comparisons between commercially available heterogeneous multicore systems (Intel Xeon–Xeon Phi platform and big.LITTLE mobile octa-core platform) and existing heterogeneous multicore real-time system [1] to highlight the achievements.

The organization of this paper is as follows: Sect. 2 explains the review of literature for commercially available heterogeneous multicore architecture, namely, Intel Xeon–Xeon Phi platform and big.LITTLE mobile octa-core platform and related works. In this section, it has been shown that (i) Intel Xeon–Xeon Phi system is a good candidate to be used in cloud and IoT infrastructure besides other application domains and (ii) octa-core system is a good candidate to be used in smart devices and embedded systems domain. Section 3 introduces an approach for applicability of the existing work [1] to commercially available heterogeneous multicore architecture such as Intel Xeon–Xeon Phi platform and big.LITTLE mobile octa-core platform. The section explains how the existing model [1] has been established as a generalized platform, where it can be used as a Xeon–Xeon Phi cluster in the cloud and IoT infrastructure as well as big.LITTLE mobile octa-core model in smart devices and embedded system domain. Section 4 summarizes our contribution.

2 Literature Review of Commercially Available Heterogeneous Multicore Systems and Related Works

Several heterogeneous multicore systems are available commercially in the market, namely, Intel Xeon processor–Intel Xeon Phi coprocessor multicore system platform, mobile octa-core processor, GPU multicore processor, etc. They are designed to be operated at specific types of application areas. Some of them are best suited for high-performance areas, some of them are best suited for graphics and image processing areas, and some of them are best suited for providing energy efficiency.

However, we study two heterogeneous multicore architectures, namely, (A) Intel Xeon processor–Intel Xeon Phi coprocessor platform and (B) Mobile octa-core processor due to the demand in current market scenarios as explained in Sect. 1. The architectures of both the systems are described below.

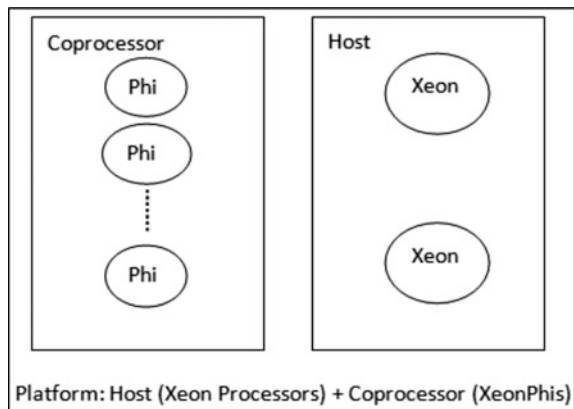
2.1 Intel Xeon Processor–Intel Xeon Phi Coprocessor Platform

In recent development of computer architecture, multicore is inclined toward many-core in high-performance computing applications and Intel Many Integrated Core (MIC) coprocessor which is known as Xeon Phi coprocessor is rising as single-chip many-core processor for high-performance computation [2]. The scaling capability of Intel Xeon processor-based system is increased with the association of Intel Xeon Phi coprocessor which offers additional power efficient scaling, vector support, memory bandwidth and maintains programmability and support associated with Intel Xeon processor [3]. Hence, Intel Xeon and Xeon Phi are generally integrated into a single platform and generally used for highly parallel applications which reach the scaling limits on Intel Xeon processors. A typical block diagram of the platform as per work [3] is shown as Fig. 1.

As per the existing work [3, 4, 5, 6, 7] the following architecture of the platform consisting of Intel Xeon and Intel Xeon Phi may be summarized:

- A platform cannot have only coprocessors.
- A typical platform consist of (i) 1–2 Intel Xeon processors (with a certain numbers of cores) with high frequency for the execution of highly serial tasks and (ii) 1–8 Intel Xeon Phi coprocessors per host with low frequency for execution of highly parallel tasks.
- Each coprocessor has around 60 cores (4 hardware thread on each core).
- All the cores composed of two levels of caches (L1 and L2).
- The Xeon Phi is attached with a host CPU through Peripheral Component Interconnect (PCI) express just like a Graphics Processing Unit (GPU).
- Both the Intel Xeon processor and Intel Xeon Phi coprocessor can be utilized by the specific application to achieve a higher degree of performance and parallelization. However, Xeon Phi coprocessor and Xeon CPU are different in performance characteristics with each other. The coprocessor may contribute

Fig. 1 A block diagram of platform consisting of Intel Xeon processor and Intel Xeon Phi coprocessor



considerably higher performance than the host processor for some applications, while the performance may be low for other applications.

- Preparing Intel Xeon Phi coprocessors should be done such a way that at first the performance of the Xeon Processor can be utilized fully for a given application. Additionally scaling of an application is very important to get the high performance and the application has to scale more than one hundred threads to achieve high parallelism.

Many research works are proposed to exploit the architecture of Xeon Phi to fit several applications. In paper [8], the authors presented new parallelization techniques that were used to search the biological sequence database in Xeon Phi-based architecture. The authors in paper [9] proposed a technique to enable the capabilities of Intel Xeon Phi for a virtual machine. In paper [10], the authors executed a well-known genomic aligner Burrows-Wheeler Aligner (BWA) in a heterogeneous system consisting of Xeon multicore processor and Xeon Phi manycore accelerator. More research work considering data analysis, performance, load balancing can be found in [11, 12, 13, 14].

From the above architecture on Intel Xeon–Xeon Phi system and all the existing research works based on the system, it may be concluded that Intel Xeon–Xeon Phi system is a good candidate to be used in cloud and IoT infrastructure besides other application domains for providing a high degree of parallelism and performance.

2.2 *Mobile Octa-Core Processor*

The functions of traditional desktop applications are steadily being taken over by mobile devices [15], and the end user uses it more closely than the Personal Computer (PC). Mobile devices have limited battery power and therefore, the thermal issue is more critical in mobile devices than the PC. More power is not required for a majority of tasks such as navigating through home screens, checking texts, and browsing the Web. However, power requirements become higher for tasks like HD video and gaming. Octa-core concept extends the battery life of modern handsets without compromising performance. Octa-core represents two separate operating setups of quad core integrated together on a single-chip for better energy efficiency. As per the work [16], it may be opinioned that ARM's big.LITTLE architecture is the basis of all octa-core mobile chips. As explained in the work [17], the architecture enables four (4) low-power Cortex-A7 cores (LITTLE cores) to operate with four (4) high-performance Cortex-A15 cores (big cores) and a simple block diagram as explained in the work [17] is shown as Fig. 2.

As per existing work [18, 19, 20] the following big.LITTLE architecture of ARM (mobile octa-core architecture) may be summarized:

- big.LITTLE octa-core architecture had been introduced to enhance the power optimization and performance simultaneously. big.LITTLE architecture consists

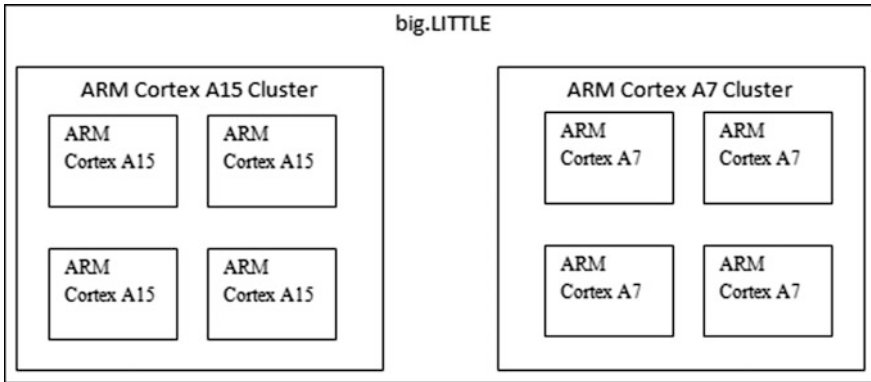


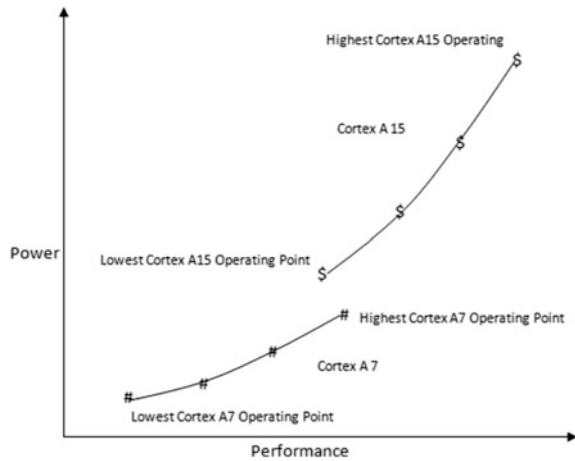
Fig. 2 Block diagram of big.LITTLE octa-core architecture of ARM

of performance-driven big core (but consumes more power) and power efficiency-driven LITTLE core (but delivers lower performance).

- Octa-core chips function as dual quad cores where tasks are balanced between them according to type. The low powered little set of cores is employed mostly while the faster big set of cores are used for advanced tasks. Different tasks have different performance and power requirements, and hence the tasks are required to be allocated to the correct processor according to their requirement.
- There are two major task scheduling techniques in big.LITTLE models as per below:
 - (1) **CPU Migration:** Here, each LITTLE core is paired with a big core and at any point of time, any single core in each pair is operational. The inactive core is powered off.
 - (2) **Global Task Scheduling:** The global scheduler is aware of the different capacities of big and LITTLE cores. The global scheduler keeps track of the performance requirement for each individual thread and decides which type of processor to utilize for each thread. Unused processors can be powered down. Global task scheduler is a flexible and popular model for all future development.
- Figure 3 shows power and performance and operating points of big (Cortex-A15) and LITTLE (Cortex-A7) processors [21]. As shown in Fig. 3, LITTLE core is more power efficiency than big core but performance is restricted than big core. Big core, on the other hand, provides high performance but with high power consumption. Altogether, big.LITTLE architecture in mobile environment can satisfy these two contradictory requirements—high performance with long battery life.

Many research works are found in the literature to use the big.LITTLE architecture in different applications. Paper [22] addressed the usage of low-power heterogeneous multicore such as ARM big.LITTLE architecture as micro servers

Fig. 3 Power and performance curve of ARM’s big.LITTLE (Cortex-A15, Cortex-A7) architecture (from [21])



which were used as a web search application. The authors in [23] introduced hipster which is a method to merge heuristic and reinforcement learning to deal with latency critical workload using the ARM big.LITTLE architecture. More research work on big.LITTLE architecture considering the parameters of energy and performance can be found in [24, 25].

From the above architecture on octa-core system (big.LITTLE architecture) and all the existing research works based on the system, it may be concluded that octa-core system is a good candidate to be used in smart devices and embedded systems domain for providing greater energy efficiency with low resources.

From all the existing work on Xeon Phi as well as big.LITTLE octa-core architectures, it can be concluded that in near future there is a need to have a generalized heterogeneous multicore platform for execution of different applications in Xeon–Xeon Phi platform in the cloud and IoT infrastructure besides other application domains as well as mobile big.LITTLE octa-core architecture in the smart devices and embedded systems domain.

3 Applicability of Existing Work of Heterogeneous Multicore Real-Time Task Scheduling into Commercially Available Heterogeneous Multicore System

The objective of our work is to analyze the existing heterogeneous multicore scheduling model [1] and to establish the model as a generalized platform where it can be used as a Xeon–Xeon Phi cluster as well as big.LITTLE mobile octa-core model. The heterogeneous architecture of work [1] illustrated a dynamic real-time task scheduling model and the model as depicted in the work is shown as Fig. 4.

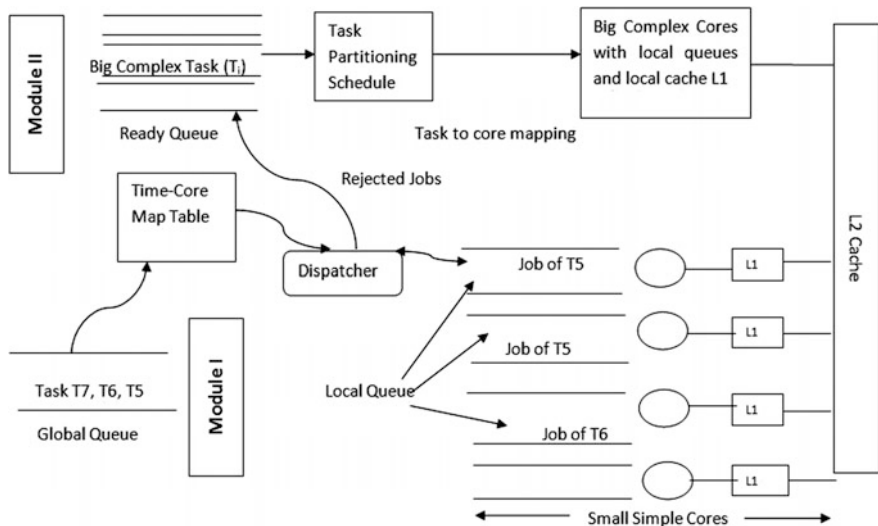


Fig. 4 Heterogeneous multicore real-time task scheduling model with different types of cores (from [1])

The heterogeneous architecture as explained in the work [1] may be summarized as per below:

- The architecture has two modules (module-I and module-II) along with a dispatcher to dispatch the matching or/and rejected job to/from particular core.
- The module-I is composed of low power small cores that are designed for the execution of low-power tasks to provide higher parallel system performance. The module-I is functional within a low-power range P (where $0 < P \leq P_l$ and $P_l = \text{maximum power to activate the small cores}$). The module has a global queue, where calculations of pseudo-release times of jobs (instances of low-power tasks) are completed and the time-core map table is searched with the calculated pseudo-release times. Consequently, run-time dispatcher forward the matching jobs to local queues of the corresponding cores. Subsequently, the jobs are assigned to cores from local queues meeting the deadline.
- The module-II is composed of high-power complex big cores that are incorporated for the execution of high-power tasks to provide higher serial system performance. The module-II is functional above a high power P (where $P_l < P \leq P_m$ and $P_m = \text{maximum power to activate the cores}$). The task instances (or jobs) of a big high-power task are stored in ready queue of particular big core. The rejected jobs from module-I are also accumulated in ready queues of module-II. Next task partitioning scheduling is used to forward all the task instances (or jobs) from ready queue to local queue and consequently into the corresponding big core.

- The model has C number of cores altogether, where $C = C_s + C_b$ ($C_s =$ Number of small cores; $C_b =$ Number of big cores). All the cores (small and big) consist of two levels of caches—L1 (own first level) and L2 (shared by all the cores). The work explained that low/high-power tasks are allocated to their corresponding small/big cores efficiently as per the requirements of serial and parallel performance.

We introduce an approach to apply this model [1] as Intel Xeon–XeonPhi cluster and mobile octa-core processors (big.LITTLE).

3.1 Applicability as Intel Xeon–Xeon Phi Multicore Platform

The high power big cores in module-II of work [1] are designed for highly serial tasks and can be treated Intel Xeon processors as the architecture of these cores are exactly similar to that of Xeon processors. The number of big cores of work [1] may be taken as n , i.e., $C_b = n$, where $n = i*j$; $i =$ number of cores per Intel Xeon processor and $j =$ number of Intel Xeon processor ($j \leq 2$).

Similarly, the low power small cores in module-I of work [1] are designed for highly parallel tasks and can be treated as Xeon Phi coprocessors as the architecture of these cores are exactly similar to that of Xeon Phi coprocessors. The number of small cores of work [1] may be taken as m , i.e., $C_s = m$, where $m = 60*k$; $k =$ number of coprocessors ($k \leq 8$) and number of cores per Intel Xeon Phi coprocessor = 60.

Complete architecture of the model can act as a platform consisting of Intel Xeon processor and Intel Xeon Phi coprocessor that can be used in cloud and IoT infrastructure besides other application domains.

In the work [1] it has been shown that the task allocation percentage of the model is approximately near to 100% at a fixed time period. The model further shows that the CPU utilization% is very high (between 95 and 100%) when the number of cores is beyond four (4). Further, we compared the average utilization% of the model [1] with the average performance in terms of utilization of existing work [3] related to Xeon and Xeon Phi cluster. For the comparison, we considered average system utilization% with respect to a different number of cores which we get by varying the number of cores of the model [1] and by varying the number of threads and subsequently calculating the number of cores of the work [3]. The number of cores for the work [3] has been calculated from the assumption that each Xeon Phi core has been assigned four numbers of threads and each Xeon core has been assigned two numbers of threads as explained in the work [3].

We plotted the comparison result as shown in Fig. 5. From Fig. 5, we found that performance in terms of utilization for Xeon processor reached maximum (nearly 100%) when the number of cores was beyond eight (8) and for Xeon phi coprocessor utilization% reached high when the number of cores was beyond thirty-two

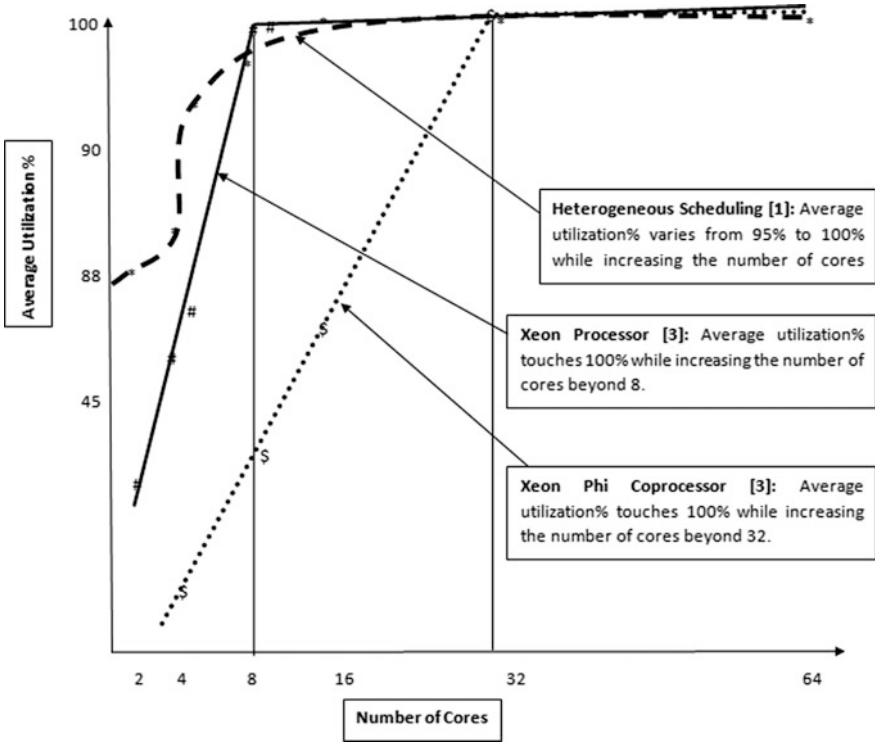


Fig. 5 Comparison of average system utilization: performance in terms of utilization for Xeon processor reached maximum (100%) when the number of cores was beyond eight and for Xeon phi coprocessor utilization% reached high when the number of cores was beyond 32. In the work [1], the average system utilization varied from 95% to 100% while increasing the number of cores beyond four. Hence, we can say that the model [1] can exactly fit into the system consisting of Intel Xeon processor and Intel Xeon Phi coprocessor

(32). Also, we found in the work [1], that the average system utilization varied from 95% to 100% while increasing the number of cores beyond four (4) only. Hence, we can say that the model [1] can exactly fit into the system consisting of Intel Xeon processor and Intel Xeon Phi coprocessor.

3.2 Applicability as Mobile Octa-Core Processor

The work of [1] exactly follows the architecture of ARM’s big.LITTLE mobile octa-core architecture where each type of core is operational between certain power levels to optimize power and performance as explained in Fig. 3. The number of big cores (module-II) as well as small cores (module-I) of work [1] may be taken as four,

i.e., $C_s = C_b = 4$, so that the complete architecture of the model becomes the mobile octa-core processors that can be used in smart devices and embedded systems.

The work [1] followed the global task scheduling model of ARM’s big.LITTLE architecture where a particular type of task including rejected task (small or big) was allocated to a particular type of core (small or big) through dispatcher maintaining constraints as explained previously.

The work [1] demonstrated that tasks of different types are allocated to cores of different types efficiently for optimization of power and performance maintaining very high task allocation percentage (100%) and CPU utilization percentage (varied from 95 to 100%). Further, we compared the job allocation% of the model [1] with the throughput in terms of job allocation% of existing work [26] related to mobile octa-core architecture (big.LITTLE). For comparison, we considered job allocation % with a fixed time period which we get by illustrating the variation in a number of jobs getting accommodated at a fixed time period.

We plotted the comparison result as shown in Fig. 6. From Fig. 6, we found that both the model accommodated 100% of the random task instances within a fixed time period. Hence, existing heterogeneous real-time task scheduling architecture [1] may be used as mobile octa-core architecture (big.LITTLE).

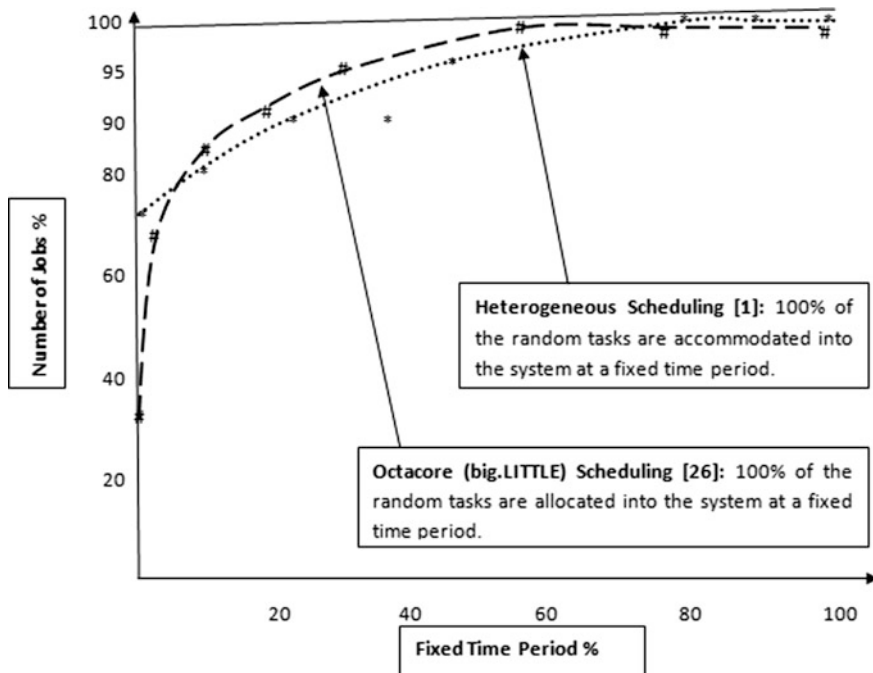


Fig. 6 Comparison of job allocation percentage between existing work [1] and mobile octa-core processor (big.LITTLE). The existing work explained that 100% of the random tasks are accommodated into the system at a fixed time period. As per work [26] 100% jobs are allocated in the ARM architecture of mobile octa-core

Therefore, from the above analysis considering the architectures of Xeon–Xeon Phi and big.LITTLE we can conclude that the existing heterogeneous multicore scheduling model [1] can be used as Intel Xeon–Xeon Phi platform in the cloud and IoT infrastructure besides other application domains as well as big.LITTLE mobile octa-core architecture in smart devices and embedded system domain maintaining the same power-performance benefit and parallelism.

4 Conclusion

This paper studies different parameters of commercially available heterogeneous multicore systems, namely, Intel Xeon–Xeon Phi cluster and mobile octa-core architecture. It also studies the scheduling algorithm for real-time tasks in existing heterogeneous multicore scheduling model [1]. It further finds out an approach or method to establish the existing heterogeneous system model [1] as a generic platform so that the generic model can be used as Intel Xeon–Xeon Phi processors and mobile octa-core processors as per requirements maintaining the same performance and parallelism.

The comparison results considering different parameters of existing model with Intel Xeon–Xeon Phi cluster and mobile octa-core architecture reveal that the existing model [1] can be applied as both the commercially available heterogeneous systems (Intel Xeon–Xeon Phi platform and mobile octa-core processors) for the use in cloud, IoT, smart devices, embedded systems, and other common application areas efficiently. Significant issues like resource sharing, energy enhancement, and dependency parameters will be incorporated in the existing model [1] to apply it as a more generic model.

References

1. Baital, K., Chakrabarti, A.: Dynamic scheduling of real-time tasks in heterogeneous multicore systems. *IEEE Embed. Syst. Lett.* **11**(1), 29–32 (2019)
2. Jha, S., He, B., Lu, M., Cheng, X., Huynh, H.P.: Improving main memory hash joins on Intel Xeon Phi processors: an experimental approach. *Proc. VLDB Endow.* **8**(6), 642–653 (2015)
3. Reinders, J.: An Overview of Programming for Intel® Xeon® processors and Intel® Xeon Phi™ Coprocessors, pp. 1–21. Intel Corporation, Santa Clara (2012)
4. Fang, J., Sips, H., Zhang, L., Xu, C., Varbanescu, A.L.: Test-Driving Intel Xeon Phi. In: *Proceedings of 5th ACM/SPEC International Conference on Performance Engineering*, NY, USA, 2014, pp. 137–148
5. Coviello, G., Cadambi, S., Chakradhar, S.: A coprocessor sharing-aware scheduler for Xeon Phi-based compute clusters. In: *Proceedings of IEEE 28th International Parallel and Distributed Processing Symposium*, Phoenix, AZ, 2014, pp. 337–346
6. Miyamoto, T., Ishizaka, K., Hosomi, T.: A dynamic offload scheduler for spatial multitasking on Intel Xeon Phi coprocessor. In: *Proceedings of SASIMI 2013*, pp. 261–266

7. Hirokawa, Y., Boku, T., Sato, S.A., Yabana, K.: Electron dynamics simulation with time-dependent density functional theory on large scale symmetric mode Xeon Phi cluster. In: Proceedings of IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Chicago, IL, 2016, pp. 1202–1211
8. Lan, H., Liu, W., Schmidt, B., Wang, B.: Accelerating large-scale biological database search on xeon phi-based neo-heterogeneous architectures. In: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, 2015, pp. 503–510
9. Gerangelos, S., Koziris, N.: vPHI: enabling xeon phi capabilities in virtual machines. In: Proceedings of IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Lake Buena Vista, FL, 2017, pp. 1333–1340
10. Chen, S., Senar, M.A.: Improving performance of genomic aligners on intel xeon phi-based architectures. In: Proceedings of IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Vancouver, BC, 2018, pp. 570–578
11. Xie, B., Liu, X., Zhan, J., Jia, Z., Zhu, Y., Wang, L., Zhang, L.: Characterizing data analytics workloads on Intel Xeon Phi. In: Proceedings of IEEE International Symposium on Workload Characterization, Atlanta, GA, 2015, pp. 114–115
12. Nookala, P., Dimitropoulos, S., Stough, K., Raicu, I.: Evaluating the support of MTC applications on Intel Xeon Phi many-core accelerators. In: Proceedings of IEEE International Conference on Cluster Computing, Chicago, IL, 2015, pp. 510–511
13. Pennycook, S.J., Hughes, C.J., Smelyanskiy, M., Jarvis, S.A.: Exploring SIMD for molecular dynamics, using Intel Xeon processors and Intel Xeon Phi coprocessors. In: Proceedings of IEEE 27th International Symposium on Parallel and Distributed Processing, Boston, MA, 2013, pp. 1085–1097
14. Misra, S., Pamnany, K., Aluru, S.: Parallel mutual information based construction of genome-scale networks on the Intel Xeon Phi™ coprocessor. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 5, pp. 1008–1020, 1 September–October 2015
15. Gao, C., Gutierrez, A., Rajan, M., Dreslinski, R.G., Mudge, T., Wu, C.: A study of mobile device utilization. In: Proceedings of IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Philadelphia, PA, 2015, pp. 225–234
16. Butko, A., Bruguier, F., Gamatie, A., Sassatelli, G., Novo, D., Torres, L., Robert, M.: Full-system simulation of big.LITTLE multicore architecture for performance and energy exploration. In: Proceedings of IEEE 10th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSOC), Lyon, France, 2016, pp. 201–208
17. Greenhalgh, P.: big.LITTLE Processing with ARM Cortex™-A15 & Cortex-A7, ARM White Paper, 2011, pp. 1–8
18. Jeff, B.: big.LITTLE technology moves towards fully heterogeneous global task scheduling, ARM White Paper, 2013, pp. 1–13
19. ARM White Paper: big.LITTLE Technology: The Future of Mobile, 2013, pp. 1–12
20. Kamdar, S., Kamdar, N.: big.LITTLE architecture: heterogeneous multicore processing. *Int. J. Comput. Appl.* **119**(1), 35–38 (2015)
21. Yu, K., Han, D., Youn, C., Hwang, S., Lee, J.: Power-aware task scheduling for big.LITTLE mobile processor. In: Proceedings of International SoC Design Conference (ISOC), Busan, 2013, pp. 208–212
22. Jain, S., Navale, H., Ogras, U., Garg, S.: Energy efficient scheduling for web search on heterogeneous microservers. In: Proceedings of IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Rome, 2015, pp. 177–182
23. Nishtala, R., Carpenter, P., Petrucci, V., Martorell, X.: Hipster: hybrid task manager for latency-critical cloud workloads. In: Proceedings of IEEE International Symposium on High Performance Computer Architecture (HPCA), Austin, TX, 2017, pp. 409–420
24. Holmgren, R.: Energy efficiency experiments on Samsung Exynos 5 heterogeneous multicore using OmpSs task based programming. MS Thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, pp. 1–55 (2015)

25. Yoo, S., Shim, Y., Lee, S., Lee, S., Kim, J.: A case for bad big.LITTLE switching: how to scale power-performance in SI-HMP. In: Proceedings of HotPower'15 ACM Proceedings of the Workshop on Power-Aware Computing and Systems, NY, USA, 2015, pp. 1–5
26. Villebonnet, V., Costa, G.D., Lefevre, L., Pierson, J., Stolf, P.: Towards generalizing “big.Little” for energy proportional HPC and cloud infrastructures. In: Proceedings of IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, NSW, 2014, pp. 703–710

Analyzing the Detectability of Harmful Postures for Patient with Hip Prosthesis Based on a Single Accelerometer in Mobile Phone



Kitti Naonueng, Opas Chutatape and Rong Phoophuangpairoj

Abstract This research studies the use of a single accelerometer inside a smart-phone as a sensor to detect those postures that may be risks for patients with hip surgery to dislocate their joints. Various postures were analyzed using Euclidean distances to determine the feasibility to detect eight postures that were harmful. With the mobile phone attached to the affected upper leg, it was found that there was one harmful posture that could not be detected due to its close similarity with a normal posture. Meanwhile, the other two harmful postures, although indistinguishable based on their measured data, were still detectable with the suitably selected threshold. The distance measure analysis is useful as an indicator as to which posture will be near to missing out in the detection process. This will form a guideline for further design of a practical and more robust detecting system.

Keywords Accelerometer · Euclidean distance · Hip prosthesis dislocation · Smartphone

1 Introduction

In a total hip replacement, there are basically four components that are used to replace the hip joint of the affected patient as shown in Fig. 1(left). The components are merged into an implant (center) that fits into the hip (right). With a sufficient

K. Naonueng

College of Engineering, Rangsit University, Pathum Thani, Thailand

e-mail: kittinn78@gmail.com

O. Chutatape

Department of Electrical Engineering, College of Engineering, Rangsit University, Pathum Thani, Thailand

e-mail: opas@rsu.ac.th

R. Phoophuangpairoj (✉)

Department of Computer Engineering, College of Engineering, Rangsit University, Pathum Thani, Thailand

e-mail: rong.p@rsu.ac.th

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_9

125

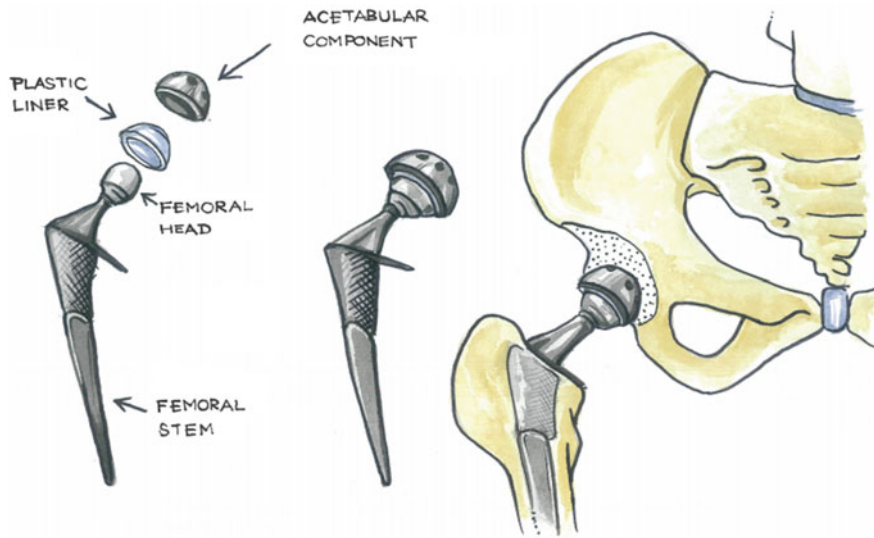


Fig. 1 Prosthetic components in total hip replacement

force from an inappropriate exercise or a repetitively unsuitable movement of the affected leg, the assembled joint can be dislocated.

For a patient who has undergone a hip replacement operation, during the recovery period which may last for a few months various postures and movements that are considered normal to ordinary people may be harmful to the person. These potentially harmful postures should be avoided to prevent possible dislocation of the joint that may lead to further complication and even require another surgery. Some of these forbidden postures or movements [1] are

- Cross the legs at the knees (during sitting, standing, or sleeping).
- Bring the knee up higher than the hip.
- Lean forward while sitting or as sitting down.
- Try to pick up something on the floor while sitting.
- Turn feet excessively inward or outward when bending down.
- Reach down to pull up blankets when lying in bed.
- Bend at the waist beyond 90°.

After being discharged from the hospital the person will need some help to get along with daily life at home. During this period a device that can help the person to stay more independently and safely will be most beneficial, and one such device, in this case, should be able to detect and give a warning against harmful movements or any postures that may cause a risk to the operated leg. Considering that the smartphone has already become an essential device for almost everybody's daily life and many applications are readily being built around its user-friendliness and myriad communication features, the next logical choice is to explore the

feasibility to incorporate these healthcare needs to it. Among various sensors embedded inside most smartphones are the most often used accelerometer in the form of a MEMS (Micro Electro Mechanical System) chip. The accelerometer is the sensor that is associated with the movement detection and therefore has applications in various core research areas such as gait analysis [2] human activity recognition [3–10], healthcare and fitness [11–17].

In this paper, the aim is to present the systematic approach to the collection of data and detection of improper postures based on a single accelerometer embedded in the smartphone, then to give the typical coordinate values of these harmful postures and of the relative distances that indicate the degree of detectability among themselves. This will simplify the remaining task of identifying the unrecommended postures and those that require further data collection to correctly identify them.

2 Related Work

Two major approaches have been used to detect and measure patterns of human motion in the three major research areas previously mentioned. One approach is based on the image processing techniques whereby video cameras were used to record 3D sequence of photographs of individual locomotion, which were then analyzed. Although the system is commercially available, highly accurate, and successfully implemented in a controlled environment, it is not publicly used and taken into everyday activities. Few main reasons are the relatively higher cost, the complexity of the procedures involved in the set-up, and the confinement to the controlled area to ensure proper operation. In another different approach, using accelerometers as movement and orientation sensors is more often the choice by most researchers. In recent research activities in the healthcare area, some new relevant initiatives have emerged. These are leg movement and hold detection for knee extension exercise [14], hip dislocation study [15], and upper arm posture and movement measurements [16]. According to the last report [16], an iPhone application was developed and validated against an optical tracking system (OTS). The small difference in comparison with the OTS indicated that the application which was based on sensors in a smartphone is a valid method. In another study, the method used to develop an Android-based mobile application which activated an alert signal and provided feedback in real time on the effectiveness of knee extension exercises was presented in [17]. Based on these reports it confirms that adopting a smartphone as a posture detecting device can give acceptable, valid and obviously more cost-effective results.

With a particular interest in the healthcare of the elderly, recent work on the detection of harmful postures leading to possible dislocation of the hip prosthesis was studied by using 3-axis accelerometer sensor present in any smartphone [15]. As far as the authors' knowledge, it was the first initiative to use a smartphone for the purpose of detecting (and further warning) the risky posture for hip-operated

patients. It was reported there that out of 14 postures under study, there were three harmful postures that could not be classified correctly by using simply one smartphone. To give an insight into the details of these postures, this paper presents further study and report on the new analysis of the problem that could offer a better implementation.

3 Materials and Method

3.1 Instruments

The Android-based Sensor Fusion software set developed by Linköping Universitet from Google Play Store was installed into Samsung Galaxy S6 smartphone and the application was used to record the sensor data. The smartphone was put into an extended armband and fastened to the thigh in the position as shown in Fig. 2.

For each typical posture selected as a reference, the 3-axis values of the accelerometer are recorded by the software. The deviation from this reference is measured using the Euclidean distance. To perform the detection analysis the application software was set to stream the data through the local area network to the program written in Microsoft Visual C# as shown in Fig. 3. This method is efficient and the data can be obtained in real time.

3.2 Proposed Method

For a smartphone testing, instead of using a traditional method that researchers have to develop an Android program, install and test it on a smartphone, this work used the publicly available software to read and send the sensor data and the PC-based program to study them. Changing a PC-based detection program is usually easier than changing the mobile program. It does not need to repeatedly load the written program to a smartphone installation and execution. Using this method an Android



Fig. 2 Position of the smartphone on the thigh

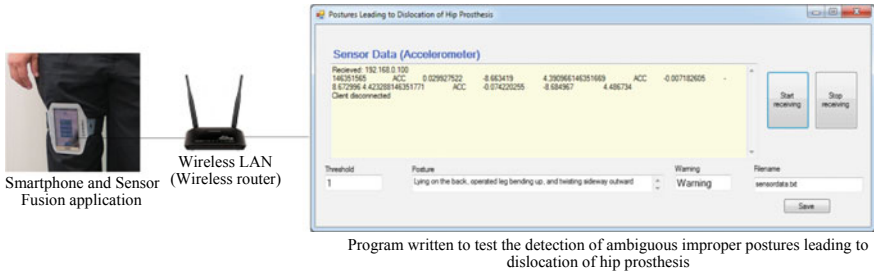


Fig. 3 Data collection of the test method

application to detect harmful postures can be developed after obtaining the desired test results.

The whole set-up worked as follows:

- (1) Receive the accelerometer data from the installed application using the TCP socket communication.
- (2) Extract the accelerometer x , y , and z values from the streaming data. Since the accelerometer values were received after the keyword “ACC”, the program checked the keyword and then captured the string values of x , y , and z , respectively.
- (3) Convert the string representation of the x , y , and z numbers to their equivalent real numbers.
- (4) Compute the Euclidean distance between the receiving x , y , and z values and the selected values of each harmful posture.
- (5) Detect harmful postures based on the Euclidean distance and a predefined threshold.

Steps 4–5 can be explained in details as follows:

Given that

P is the set of all detecting postures represented by

$$P = p_1, p_2, \dots, p_i \quad 1 \leq i \leq N \tag{1}$$

N = the number of detecting postures,

p_i is selected accelerometer x , y , and z (reference) values for the i th posture represented by

$$p_i = (\text{Ref}x_i, \text{Ref}y_i, \text{Ref}z_i), \quad 1 \leq i \leq N \tag{2}$$

O is an observation sequence (a sequence of accelerometer data) consisting of observations (o_t) at time 1 to time T .

$$O = (o_1 \ o_2 \ o_3 \ \dots \ o_T) \quad (3)$$

$$o_t = (x_t, y_t, z_t) \quad 1 \leq t \leq T \quad (4)$$

(x_t, y_t, z_t) are accelerometer x , y , and z values occurring at time t . The Euclidean distance between two accelerometer data p_i and o_t is given by $\text{dist}(p_i, o_t)$. The method to detect postures (DP*) at time t is described using the following equations:

$$\text{DP}^* = \arg \min_{p_i \in P} \text{dist}(p_i, o_t) \quad (5)$$

and

$$\text{dist}(p_{\text{DP}^*}, o_t) \leq \sigma \quad (6)$$

σ is a predefined detection threshold obtained from the experiments (e.g., 1).

By following the steps as indicated, manual intervention can be kept to a minimum and various results can be obtained for further analysis by using existing mathematical software packages.

4 Results

4.1 *Standing, Walking, and Sitting*

Figure 4 shows the accelerometer data x , y , z in the four scenarios, i.e., (1) straight standing, (2) walking, (3) sitting on a chair, and (4) sitting on a chair and slightly move the thigh up and down.

4.2 *Reaching Down to Collect Object on the Floor or to Tie Shoestrings*

Figure 5 shows the accelerometer data x , y , z when (1) sitting and bending up the leg to tie shoestrings, (2) sitting and stooping to collect an object on the floor or to tie shoestrings, and (3) stooping to collect object on the floor or to tie shoestrings with leg bending.

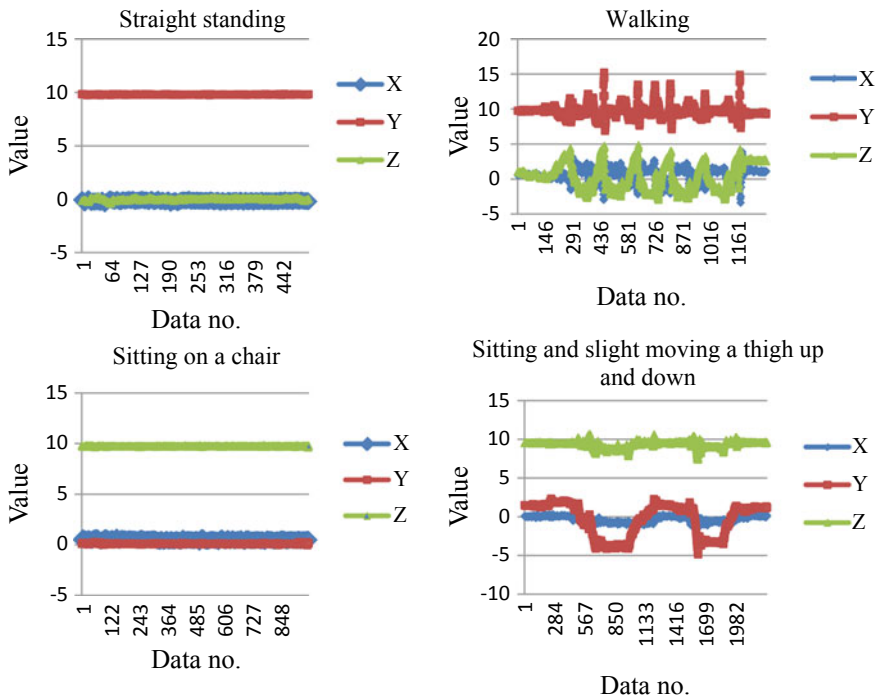


Fig. 4 Straight standing, walking, and sitting (with possibly a slight up-down thigh movement)

4.3 Lying on the Back

Figure 6 shows the accelerometer data x, y, z when (1) lying on the back, (2) lying on the back with the operated leg bent up and twisted sideways inward, and (3) lying on the back with the operated leg bent up and twisted sideways outward.

4.4 Twisting the Operated Leg Inward and Outward While Standing

Figure 7 shows the accelerometer data x, y, z when twisting an operated leg inward while standing, and outward while standing.

The next Fig. 8 shows the accelerometer data x, y, z when getting on a motorcycle.

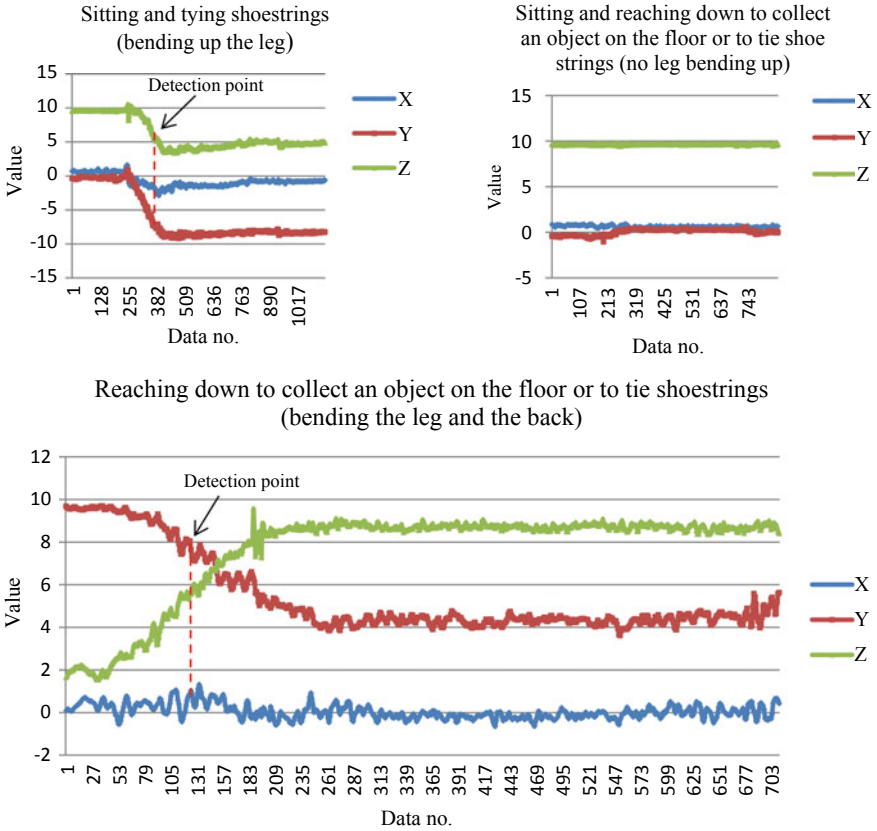


Fig. 5 Reaching down to collect object on the floor or to tie shoestrings

4.5 Accelerometer Outputs at Various Postures

The x - y - z values of the smartphone accelerometer at various different postures are shown in Table 1. In the last column, each posture is assigned either risk factor 0 (no risk), or 1 (with risk).

Euclidean distances between centroids are used to represent the relative differences among various postures as given in details in Table 2. Posture 2 (walking) and posture 3 (sitting) are two postures that have more different readings recorded. Notice that postures 1–4 are classified as “no risk” and postures 5–12 are classified as “at risk”.

By setting the threshold of the Euclidean distance at unity, it was seen that most postures, particularly those considered being harmful, could be separated. These are cells with corresponding distance values more than one and displayed as empty cells as shown in Table 3. On the other hand, posture 5 and 8 could not be

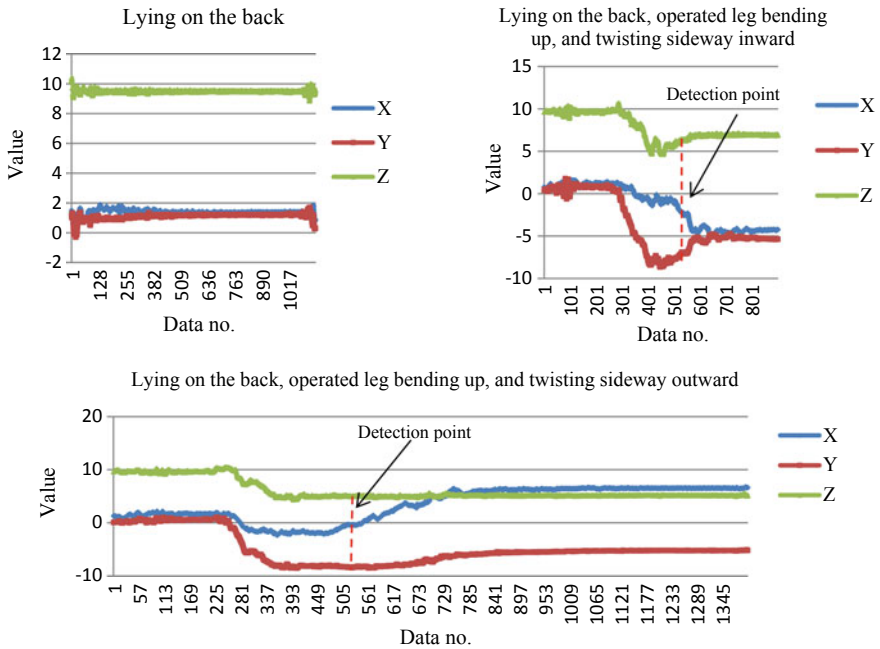


Fig. 6 Lying on the back posture

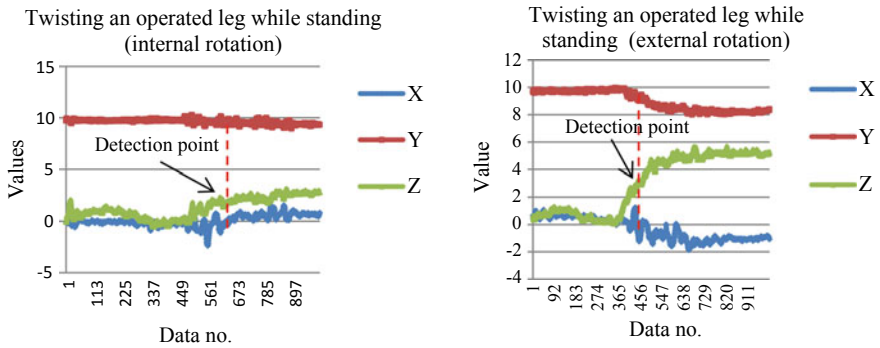


Fig. 7 Twisting an operated leg while standing (internal and external rotation)

distinguished from each other but they could be detected as harmful and give an alert signal since both were sufficiently at a far distance from the others. Posture 6 could not be recognized correctly because it was quite similar to posture 3 as far as the accelerometer values are concerned. These cells in the table are labeled with the letter “N” which indicates that the two postures indicated by row and column could

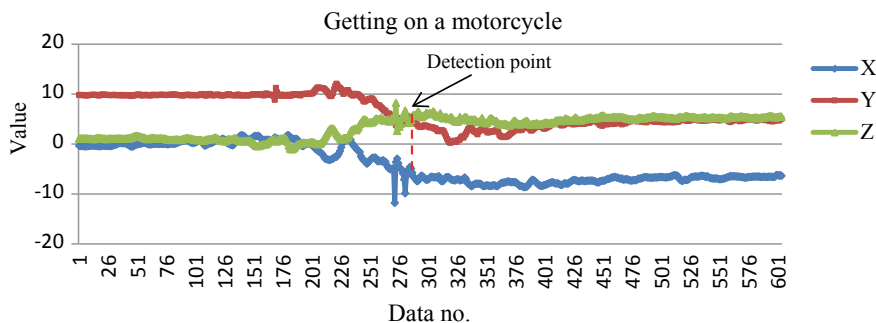


Fig. 8 Getting on a motorcycle

Table 1 x–y–z values of accelerometer at various postures

Posture	x	y	z	Risk factor
1. Straight standing	- 0.1545	9.802219	0.005719	0
2. Walking	0.501585	9.861717	0.52433	0
	0.659603	10.34175	0.802058	
	- 0.68235	10.52132	4.383783	
	1.740585	9.680955	- 1.14443	
	1.833959	9.56723	- 2.51152	
3. Sitting	- 2.23499	14.66808	2.592921	0
	0.485402	0.094524	9.734577	
	- 0.0826	1.447295	9.536105	
	- 0.90621	- 2.90297	8.707711	
4. Lying on the back	- 0.80325	- 3.95283	- 3.95283	0
	1.377871	1.099392	9.50181	
5. Sitting and tying shoestrings (bending up the operated leg)	- 1.99916	- 7.25922	5.436035	1
6. Sitting and reaching down to collect an object on the floor or to tie shoestrings (no leg bending up)	0.730254	0.020997	9.625291	1
7. Reaching down to collect an object on the floor or to tie shoestrings (bending the leg, and the back)	0.31364	8.137892	5.482722	1
8. Lying on the back, operated leg bending up, and twisting sideways inward	- 1.83994	- 7.34302	6.050148	1
9. Lying on the back, operated leg bending up, and twisting sideways outward	- 0.35674	- 8.43836	4.987122	1
10. Twisting an operated leg while standing (internal rotation)	- 0.20351	9.51336	2.000356	1
11. Twisting an operated leg while standing (external rotation)	- 0.50996	9.471462	2.958036	1
12. Getting on a motorcycle	- 5.89213	4.424485	5.469554	1

Table 2 Euclidean distance between postures

Posture	Euclidean distance																			
	1	2	3			4	5	6	7	8	9	10	11	12						
1	0.00	0.84	1.26	4.47	2.22	3.22	5.89	13.76	12.67	15.42	14.33	12.97	18.00	13.75	5.74	18.26	18.91	2.02	2.99	9.58
2	0.84	0.00	0.58	4.09	2.09	3.33	5.91	13.43	12.34	15.23	14.58	12.58	17.99	13.41	5.25	18.22	18.86	1.67	2.66	9.74
	1.26	0.58	0.00	3.83	2.32	3.60	5.51	13.60	12.49	15.50	15.14	12.71	18.39	13.58	5.19	18.62	19.27	1.69	2.60	9.99
	4.47	4.09	3.83	0.00	6.09	7.40	4.78	11.78	10.45	14.11	16.70	10.92	17.86	11.82	2.81	17.98	18.97	2.63	1.78	8.09
	2.22	2.09	2.32	6.09	0.00	1.38	7.39	14.55	13.61	16.20	14.15	13.68	18.55	14.50	6.95	18.83	19.24	3.70	4.68	11.39
	3.22	3.33	3.60	7.40	1.38	0.00	8.28	15.54	14.65	17.00	13.85	14.71	19.00	15.48	8.26	19.31	19.63	4.95	5.95	12.24
	5.89	5.91	5.51	4.78	7.39	8.28	0.00	16.46	15.09	18.65	19.79	15.65	22.11	16.52	7.58	22.28	23.31	5.57	5.49	11.25
3	13.76	13.43	13.60	11.78	14.55	15.54	16.46	0.00	1.48	3.46	14.33	1.36	8.87	0.28	9.10	8.62	9.80	12.21	11.61	8.81
	12.67	12.34	12.49	10.45	13.61	14.65	15.09	1.48	0.00	4.50	14.55	1.50	9.81	1.64	7.83	9.62	10.89	11.04	10.39	7.69
	15.42	15.23	15.50	14.11	16.20	17.00	18.65	3.46	4.50	0.00	12.70	4.68	5.56	3.47	11.57	5.26	6.69	14.13	13.65	9.44
	14.33	14.58	15.14	16.70	14.15	13.85	19.79	14.33	14.55	12.70	0.00	14.54	10.03	14.23	15.38	10.61	10.01	14.74	15.10	13.60
4	12.97	12.58	12.71	10.92	13.68	14.71	15.65	1.36	1.50	4.68	14.54	0.00	9.89	1.26	8.18	9.67	10.69	11.38	10.79	8.95
	18.00	17.99	18.39	17.86	18.55	19.00	22.11	8.87	9.81	5.56	10.03	9.89	0.00	8.83	15.57	0.64	2.07	17.22	16.98	12.32
	13.75	13.41	13.58	11.82	14.50	15.48	16.52	0.28	1.64	3.47	14.23	1.26	8.83	0.00	9.12	8.58	9.71	12.21	11.63	8.97
7	5.74	5.25	5.19	2.81	6.95	8.26	7.58	9.10	7.83	11.57	15.38	8.18	15.57	9.12	0.00	15.64	16.60	3.78	2.97	7.23
8	18.26	18.22	18.62	17.98	18.83	19.31	22.28	8.62	9.62	5.26	10.61	9.67	0.64	8.58	15.64	0.00	2.13	17.41	17.15	12.46
9	18.91	18.86	19.27	18.97	19.24	19.63	23.31	9.80	10.89	6.69	10.01	10.69	2.07	9.71	16.60	2.13	0.00	18.20	18.03	14.01
10	2.02	1.67	1.69	2.63	3.70	4.95	5.57	12.21	11.04	14.13	14.74	11.38	17.22	12.21	3.78	17.41	18.20	0.00	1.01	8.38
11	2.99	2.66	2.60	1.78	4.68	5.95	5.49	11.61	10.39	13.65	15.10	10.79	16.98	11.63	2.97	17.15	18.03	1.01	0.00	7.79
12	9.58	9.74	9.99	8.09	11.39	12.24	11.25	8.81	7.69	9.44	13.60	8.95	12.32	8.97	7.23	12.46	14.01	8.38	7.79	0.00

Italic represents distinctly the low Euclidean distance between posture 3 and posture 6 (0.28) and also posture 5 and posture 8 (0.64), which made the system difficult to distinguish the postures accurately

Table 3 Separability at unity threshold

Posture	1	2	3	4	5	6	7	8	9	10	11	12
1. Straight standing												
2. Walking												
3. Sitting						N						
4. Lying on the back												
5. Sitting and tying shoestrings (bending up the operated leg)								N				
6. Sitting and reaching down to collect an object on the floor or to tie shoestrings (no leg bending up)			N									
7. Reaching down to collect an object on the floor or to tie shoestrings (bending the leg and the back)												
8. Lying on the back, operated leg bending up, and twisting sideways inward					N							
9. Lying on the back, operated leg bending up, and twisting sideways outward												
10. Twisting an operated leg while standing (internal rotation)												
11. Twisting an operated leg while standing (external rotation)												
12. Getting on a motorcycle												

not be distinguished from each other at the given threshold. For a higher threshold, more of such postures that are indistinguishable will surface up.

The Euclidean distance-based algorithm to identify harmful posture may be initiated by periodically reading the 3-axis accelerometer outputs from the mobile phone, then calculate the Euclidean distance between the sample posture and the known, selected values of the harmful postures in the database. If the distance is less than a chosen threshold value, then the sample posture is likely to close to the harmful one and the warning signal should be given.

There however remain certain postures that cannot be identified as normal or harmful, and these require further data from additional sensors to be attached at the other major parts of the body. One possibility is to attach a second sensor on the chest. This will give the information about the movement of the upper part of the body which can differentiate posture 5 from 8, and posture 6 from 3. Nevertheless, considering that each individual may not move exactly in the same pattern with the same range of movement for the same posture, and the position of the attached sensor can vary, the threshold value is an individual-dependent parameter that will require custom adjustment.

5 Conclusion

The analysis on the detectability of various harmful postures for hip-operated patient based on the use of a single accelerometer in the smartphone has been presented. This is the first initiative and a contribution made to the research area in addition to the proposed method for other similar developments. It was found that at the selected threshold, six out of eight harmful postures, including the major ones, could be detected and so the warning signal could be generated by the smartphone. In the future, the method to distinguish posture 3 (sitting) from posture 6 (sitting and reaching down to collect an object on the floor or to tie shoestrings) ought to be developed. This ambiguous posture could be solved by implementing an additional sensor to detect the movement of the upper part of the body. However, adding more sensors on the body is not the favorable option from a user convenience point of view, and therefore using a smartphone alone to detect all or as many of these harmful postures probably through better utilization of sensors and intelligent time series data analysis should be further investigated.

Acknowledgements We would like to thank Dr. Nathee Ruangthong of Ban Mi Hospital, Lop Buri, Thailand, for many of his valuable suggestions. One was that no real patient was required to participate as it might cause injuries and some difficulties. With this recommendation and to follow the ethical guideline for clinical trial, the participant who was a normal person in this trial was fully explained of the procedure and the use of the data collected and was later given a proper form of consent to sign voluntarily.

References

1. Activities After Hip Replacement—OrthoInfo—AAOS: American Academy of Orthopaedic Surgeons. <https://orthoinfo.aaos.org/en/recovery/activities-after-hip-replacement/>. Accessed 07 Sep 2018
2. Muro-de-la-Herran, A., Garcia-Zapirain, B., Mendez-Zorrilla, A.: Gait analysis methods: an overview of wearable and non-wearable systems. *Highlighting Clin. Appl. Sens.* **14**(2), 3362–3394 (2014)

3. Noury, N., Fleury, A., Rumeau, P., Bourke, A. K., Rialle, V., Lundy, J.E.: Fall detection: principles and methods. In: Proceedings of 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBS), pp. 1663–1666 (2007)
4. Brezmes, T., Gorricho, J.-L., Cotrina, J.: Activity Recognition from Accelerometer Data on a Mobile Phone. Lecture Notes in Computer Science, vol. 5518, pp. 796–799. Springer, Berlin (2009)
5. Krishnan, N., Panchanathan, S.: Analysis of low resolution accelerometer data for continuous human activity recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3337–3340 (2008)
6. Long, X., Yin, B., Aarts, R.M.: Single accelerometer-based daily physical activity classification. In: Proceedings 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), pp. 6107–6110 (2009)
7. Mannini, A., Sabatini, A.M.: Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors* **10**(2), 1154–1175 (2010)
8. Huda, F.A., Tolle, H., Putra, K.P.: Human activity recognition using single accelerometer on smartphone put on user's head with head-mounted display. *Int. J. Adv. Soft Comput. Appl.* **9**(3), 239–249 (2017)
9. Mathie, M., Celler, B., Lovell, N., Coster, A.: Classification of basic daily movements using a triaxial accelerometer. *Med. Biol. Eng. Comput.* **42**, 679–687 (2004)
10. Kwapisz, J.R., Weiss, G.M., Moore, S.A.: Activity recognition using cell phone accelerometers. *ACM SIGKDD Explor. Newsl.* **12**(2), 74–82 (2010)
11. Putra, I.P.R.S., Brusey, J., Gaura, E., Vesilo, R.: An event-triggered machine learning approach for accelerometer-based fall detection. *Sensors* **18**(20), 1–18 (2017)
12. Gjoreski, H., Luštrek, M., Gams, M.: Accelerometer placement for posture recognition and fall detection. In: Proceedings of International Conference on Intelligent Environments, Nottingham, UK, pp. 47–54 (2011). <https://doi.org/10.1109/ie.2011.11>
13. Khawandi, S., Daya, B., Chauvet, P.: Automated monitoring system for fall detection in the elderly. *Int. J. Image Process.* **4**(5), 476–483 (2010)
14. Phoophuangpairoj, R.: Frame-based analysis of knee extension exercises using a smartphone accelerometer. In: Proceedings of 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) (2016)
15. Chutatape, O., Naonueng, K., Phoophuangpairoj, R.: Detection of dangerous postures leading to dislocation of hip prosthesis by a smartphone. In: Proceedings of 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (2017)
16. Yang, L., Grooten, W.J.A., Forsman, M.: An iPhone application for upper arm posture and movement measurements. *Appl. Ergon.* **65**, 492–500 (2017). <https://doi.org/10.1016/j.apergo.2017.02.012>
17. Phoophuangpairoj, R.: Developing an accelerometer-based mobile application to aid knee extension exercise. *Int. J. Adv. Soft Comput. Appl.* **10**(2), 132–147 (2018)

Software Development Process Evolution in Malaysian Companies



Rehan Akbar, Asif Riaz Khan and Kiran Adnan

Abstract GSD is a phenomenon mainly associated with the outsourcing of software development projects to some offshore company. Reduction in software development cost increased productivity and advantage of multisite development with respect to time are the main benefits that software development companies (SDCs) get from GSD. Besides benefits, a number of challenges associated with GSD are also observed. Consequently, the traditional processes to develop software are evolving and being replaced with a new set of processes which are lightweight and outcome-based. The process evolution has been deeply investigated in the context of companies mostly in Europe, Australia, USA and mainly other countries in those regions. In this regard, limited research has been carried out on Malaysian companies. The present research investigates the process evolution phenomenon in Malaysian companies. The current software development processes and the reasons for the evolution of software processes in Malaysian software companies have been identified. A qualitative approach using structured interviews has been followed for the collection of data collection and its analysis. The findings explain that software processes in most of the Malaysia companies are increasingly evolving or have been evolved. The companies are overwhelmingly adopting agile methods because of their support to GSD. Some of the companies are using ad hoc approaches for software development. The size of the company and project has been found as one of the main factors behind using ad hoc approaches. Mainly the small and medium-size companies and projects are involved in this practice.

Keywords Agile · Global software development · Malaysian companies · Software evolution · Software processes

R. Akbar (✉) · A. R. Khan · K. Adnan (✉)
Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman, 31900 Kampar, Perak, Malaysia
e-mail: rehan@utar.edu.my

K. Adnan
e-mail: kiranadnan@lutar.my

A. R. Khan
e-mail: asifriazkhan@gmail.com

1 Introduction

The Global Software Development (GSD) started in the late 1990's by mid-2000's it got an overwhelming response from the software companies [1]. GSD is a form of project outsourcing to offshore companies. GSD is carried out from different locations around the world that involved different teams for the development of software product [2, 3]. The availability of fast internet and latest communication tools are increasing the trend of developing software product from different locations [4]. GSD also developed the connections among different nations having different behaviors, social values, and different culture [5]. GSD helps the software companies to decrease the software development cost. Since, the GSD provides a number of benefits such as less development cost, development from different locations with respect to time, access to a large pool of experienced people and tax incentive [6–9]. Due to the GSD benefits, most of the SDCs have started the GSD to decrease the cost of software development and improve the software quality [10]. Moreover, GSD help to exchange the best expertise among teams [1, 8] and also provide access to skilled resources that would help the SDCs to increase their productivity and efficiency [1, 5, 8, 9]. In addition, the development of software with follow-the-sun (24/7) model also is an added advantage of GSD [8, 11].

Despite having advantages, a number of challenges are also associated with GSD [2, 5, 7, 12–16]. The main challenges being faced in GSD are related to communication and collaboration [5, 12, 13, 17–21], poor management [13, 22], project diversity and complexity,[12] and delayed response due to time zone difference [7, 12]. Although, the SDCs are using different communication applications like Skype, Instant Message (IM), Google chat, telephones, and emails [23, 24] to reduce these challenges. However, still, problems exist at certain levels and need to be addressed properly. To overcome the GSD challenges, many studies have proposed some strategies to meet the GSD challenges [6, 13, 25–28].

To deal with these challenges and to meet the market competition, software development processes have been evolving since the late 1990s and shifting towards new software development paradigms prominently agile methods [19]. The software project clients want to see the progress of the project on a weekly basis and also require to make necessary changes in requirements and deliverables anytime during the project. Consequently, to meet the client expectations the SDCs started following agile methods, as these approaches define procedures to provide an update to the client on a weekly basis [19, 29, 30]. According to Versionone survey [31] about 80% of the SDCs have been moved to the agile methodologies. The evolution in software development processes in companies operating in the UK, USA Australia, and other western countries have been investigated in different studies. However, limited work has been presented in the context of Malaysian companies. Most of the work presented on software development processes in Malaysian companies are very general and do not provide deep insight into process evolution and process paradigm shift. In the present research work, the aim is to check the trend of current software development processes in Malaysian companies

as a result of GSD. The evolution of software development processes, how the processes got affected and eventually changed in Malaysian companies, and reasons behind. The study also presents the expert views on the change in software development processes. In addition, some discussion on the software development practices in Malaysian companies has been made in the literature review section.

The section two of the paper presents the review of related literature while research methodology is presented in the third section. The results have been discussed in the fourth section and finally, conclusion has been summarized in section five followed by acknowledgment.

2 Related Work

A software process involves a set of procedures, tasks, and activities which are required to build a software [32]. The phenomenon to develop software from different locations is called GSD [19, 33]. Software development at different locations made the conventional software development process more difficult to implement. Consequently, most of the software development projects have been failed and over budget [34, 60, 61]. A number of factors behind the failure of GSD projects are misunderstanding of requirements, absence of face-to-face communication, absence of suitable software process, high coordination costs, lack of experience about distributed development, culture and languages problems, dependent modules in distributed development, and delay in responses due to time zone difference [4, 5, 7, 10, 12, 13, 32, 35, 36]. Numerous studies have suggested solutions and improvement to deal with these challenging factors [2, 6, 13, 25–28, 37–39].

The main purpose of using the software development process is delivering the project on time and within budget [32]. Therefore, software development companies need to follow a suitable software development process that helps to minimize these issues and deliver projects on time. The conventional approaches of software development have been deprecated because of less support to GSD and lightweight methodologies such as agile have appeared as popular practices by software development companies because of their support to frequent communication, short deliverables, and other disciplined practices [13, 18, 19, 26]. It is also observed that the application of agile methods in GSD environments has significantly reduced the failure rate of the projects [18, 26]. The existing studies describe that current software practices have been changed globally but are uncertain about Malaysian SDCs. As per our knowledge, there has not been carried out such study that investigates the current trends of software development in Malaysia as consequences of GSD. The research study presented in this paper accomplishes this research gap. Though a few studies have been conducted on Malaysian companies [40–46] but not in GSD perspective. Also, these studies insufficiently present the reasons behind the selection of the software development process as a result of GSD.

Almmani in a survey [41] investigated the difficulties in recent software processes improvement (SPI) practices in small and medium-size companies (SMEs) operating in Malaysia. The study has some limitations. Firstly, the study has been conducted through a questionnaire with small sample size, secondly, the target respondents are only those companies which are doing in-house development and are not involved in GSD. The study concludes that implementation of SPI is not very common in Malaysian SMEs because of inadequate knowledge, communication issues, and lack of resources. In addition, findings also show that many software development companies in Malaysia are using ad hoc approaches and agile methodologies. In 2005, Baharom [42] conducted a survey on software development practices in Malaysia. The survey has been conducted through a questionnaire and the findings of the survey explain that most of the software development companies are not following any software development process in their development environments. Moreover, software development companies have been facing problems in adopting standards and processes, software quality, over budget, and late delivery related issues.

In another research study [40], a survey using a questionnaire to observe the current practices of developing software in Malaysian companies has been conducted. The statistical analysis of the responses has been made to interpret the results. Findings show that the percentage usage of the methodology for managing the software project is still very low. Likewise, the use of the latest software development practices to develop software has also not been found satisfactory due to which the success rate of the projects is low. Another study conducted on “the adoption of software development methodologies among IT organization in Malaysia” has been found different from the present study in terms of (a) the study has been conducted on overall IT industry, not particularly on the companies involved in GSD, (b) the software development methodologies are defined and presented according to each phase of software development lifecycle (SDLC) such as requirements phase, design phase, development, and testing phase and (c) the focus of the study is too general [44]. Moreover, the results derived from these findings do not particularly list the names of software development processes rather it just generally discussed the software practices (without explicitly listing the software practices) being followed by Malaysian companies. However, it is found that few of the companies have developed their own processes while some of them are not following any software process. Furthermore, the author also proposed a framework with respect to each software development phase but sufficient discussion has not been made rather just discussed generally. The present study covers this gap and adequately investigates the most recent trends of software development practices in Malaysian companies as consequences of GSD and presents reason behind the process evolution.

3 Research Methodology

The research subject in present studies is software development companies operating in Malaysia which are practicing GSD in their software development environment. The qualitative research approach has been selected and data has been collected through structured interviews conducted face-to-face. Structured interviews are conducted to enhance the comparability of data [47]. Open-ended and pre-planned questions have been set for the interviews with the same replication in all the interviews [48]. The medium to conduct a structured interview could be face-to-face or online using the internet or via telephone. The interviews are divided into two sections such as (a) demographic data which is analyzed through statistical analysis and (b) open-ended structured questions which are analyzed qualitatively by using “general inductive approach” [49]. The general inductive approach is quite similar to grounded theory with the only difference is that it does not apply any coding technique for data analysis rather it analyzes the data more straight forward. [49] has discussed the general inductive approach in detail. The data from interviews have been managed and analyzed using qualitative data analysis software NVivo. Although, NVivo provides a platform to organize the qualitative data such as interviews and surveys at one place analysis of the data is totally dependent on the researcher that how he or she derived the result from it [50]. NVivo helps to reduce manual tasks and provide easy to manage, sort, and organize the qualitative data. The main features which NVivo provide are coding, querying text data, and representation of text with multiple charts and graphs in order to understand any particular phenomena. It is difficult to explain all the features of NVivo software, however, numbers of useful links are provided to understand the working of NVivo software in detail [51–57]. In NVivo, coding is performed in the ‘Nodes’ container. The information related to the particular question, theme or concept is stored under one node. In the present study, the answers of interviewees on current software development practices are stored in one node. It makes it easy to summarize the data and perform comparisons to derive the findings.

To conduct interviews for data collection, initially, 40 software development companies in Malaysia were selected. Only 25 software development companies agreed for the interviews out of which later 10 of companies refused to participate for interviews because of project deadlines and emergency meeting calls during the scheduled time. The privacy policies of the companies have been found as one of the big problems of low response rate as most of the companies also refused due to their company’s privacy policy. Therefore, a total of 15 interviews could be conducted. During face-to-face interviews, it is ensured that all the questions have been clearly understood by the respondents in order to get proper answers. In face-to-face interviews, the researcher has the opportunity to clarify the confusion about the questions, if arise any.

Very few questions have been set for the interview. Mainly the questions asked were about the current software development practices in their company along with the follow-up question about reasons. The interviews have been conducted with

expert professionals such as project manager (PM), technical manager (TM), team leads, senior software engineers, senior system analysts, and software developers working in the Malaysian software companies.

4 Results and Discussion

The interviews have been divided into two sections as the first section is about demographic data comprising of name, email, company location, and respondent experience, and the second section is comprised of open-ended questions related to the current software development practices along with follow-up question on reasons to select the current practices.

In the first section, the questions have been asked to know the participants' qualification, experience level, and company profile. As shown in Table 1, 40% of the interviewees are senior software engineers, 33% software engineers, and 20% are project manager/technical manager. Only 7% were system analysts.

Most of the interviewees are quite experienced professionals in the software development industry as shown in Table 2. 20% respondents have experienced between 10 and 15 years while similarly, 20% of the respondents have more than 15 years of experience. 13.3% interviewees have experience of 3–5 years and 5–10 years. However, 33.4% of them have 1–3 year(s) of experience in software development.

As shown in Table 3, the response rate of companies with respect to their location. Kuala Lumpur, Penang, and Selangor are the main states in Malaysia where the big software development companies are operating their business [46, 58]. The majority of the respondents of the present study also belong to companies in Kuala Lumpur, Penang, and Selangor as given in Table 3. The data shows that 40% of the interviewee companies are from Penang, 26.7% from Selangor, and 13.3% companies are from Kuala Lumpur. 20% of the total companies are from Ipoh.

The size of the company has been estimated based on the number of employees (staff strength) in that company. As shown in Table 4, most of the respondents, i.e., overall 90% are from small and medium-sized companies. According to [41, 42, 59],

Table 1 Employee designation

Employee designation	Interviewees	Percentage (%)
Project Manager/Technical Manager	3	20
Project Team Lead	0	0
Senior Software Engineer	6	40
Software Engineer	5	33.3
System Analyst	1	6.7
Total	15	100

Table 2 Experience level in software development

Experience level	Interviewees	Percentage (%)
1–3 year(s)	5	33.4
3–5 years	2	13.3
5–10 years	2	13.3
10–15 years	3	20
More than 15 years	3	20
Total	15	100

Table 3 Locations of respondent companies

Company location	Interviewees	Percentage (%)
Kuala Lumpur	2	13.3
Selangor	4	26.7
Penang	6	40
Ipoh	3	20
Total	15	100

the majority of the software companies in Malaysia are small and medium-sized. Table 4 shows that 40% of the software companies have staff less than 20, while 26.7% companies have staff between 20 and 50. 20% of the companies have 50–100 employees while only 13.3% companies are enterprise-level big companies and have more than 200 employees.

In the second section, questions about the effect of GSD, changes in software development processes and reasons behind it have been asked. The analysis of the qualitative data shows that GSD has brought a significant change in Malaysian SDCs and it has also affected the software development paradigms. It is evident from the results that the majority of the SDCs operating in Malaysia have replaced their existing processes with lightweight processes predominantly agile methods. 6 out of 15 respondents particularly stated that their companies are using agile methods in their software development environment to deal with the challenges and problems faced due to GSD. According to a project manager, software development companies strive to launch their products early in the market and agile method provides support in terms of “better visibility and faster development.” Another project manager added that the use of the agile method as “Managers can see the

Table 4 Company size w.r.t number of staff

Number of staff	Interviewees	Percentage (%)
Less than 20	6	40
Between 20 and 50	4	26.7
Between 50 and 100	3	20
Between 100 and 150	0	0
More than 200	2	13.3
Total	15	100

progress of projects and developers do the tasks with the given priority.” Software engineers also believe that agile methods are the best software development practices to handle the GSD related issues such that “as far as agile is concerned, it helps this company. There are many potential benefits and issues that can arise from GSD. The most frequently cited issue is a communication problem, so agile helps to sort out the communication problem.” A number of studies report that communication is the main issue in GSD [2, 4, 5, 12, 17, 20, 21]. The respondents of the current study also consider that communication is the main problem in GSD that can be overcome by following agile methods. Agile methods provide disciplined and best practices to develop the software product [19, 21, 26]. Among agile methods, Scrum is mostly being followed in the Malaysian SDCs because Scrum supports and defines processes for weekly updates and frequent communication among team members and clients [13, 18, 19, 26]. One of the software engineer respondents mentioned that “we use Scrum because we have daily meeting to discuss what we currently doing and possible problems we traced.” Furthermore, a senior software engineer also supported using the Scrum as “we’ve been using Scrum in-house, and then we tried to use it in some outsourcing projects and it really works in terms of managing project resources and tracking project status.” The analysis of the answers of the respondents shows that agile methods are suitable for the environments of software development companies and provide best practices to build the software as well as to manage the resources, tasks, and for better visibility of the project progress.

Subsequently, 3 out of 15 respondent companies have been found as following ad hoc approaches of software development. The small and medium-size of the companies has been found as the main reason for using ad hoc approaches in Malaysia [42]. Adding to this, one of the senior software engineers highlighted the reasons of using the ad hoc approaches in their software companies as “we don’t use any process because we are a small and medium-size software company. That’s why we don’t use any process.” Likewise, a project manager/technical manager also mentioned that “Because projects are not big usually 2 to 3 weeks long and this is the reason we don’t follow any process because the use of the process is unnecessary.”

Similarly, other studies conducted in Malaysia [40, 42–44] reports that most companies don’t follow any software process; similar trends also found in the present study. Surprisingly, Malaysian SDCs are also following the software product line (SPL) due to GSD as 4 interview respondent companies are following the SPL. The common reason reported by the respondents is “to reuse the software components and for better management.” According to [21], SPL and agile methods mostly share the common goals but achieved in different ways. However, few companies are still using traditional approaches of software development because the projects are at enterprise-level, as two of the participants are using the waterfall model and only one participant reported the ISO standards.

5 Conclusion

The research study presented in this paper successfully describes the trends of software development processes in Malaysian software development industry as well as the reasons from the experts. The findings show that GSD has brought a great evolution in Malaysian SDC and as a result software paradigm has been shifted. The findings also reveal that lightweight methodologies are the recent trends in the Malaysian SDCs. Agile and SPL methods are mostly being used as these provide fast development, better visibility of project progress, easy to manage resources and tasks as well as help to minimize the issues of GSD particularly related to communication and coordination. The small and medium-size of the companies as well as projects also urges them to follow ad hoc approaches. However, few of the respondent companies are still using traditional practices in their development environment as a result of GSD.

The results of the present study adequately explain the software development processes currently being used in Malaysian companies as a result of GSD. It also presents the views of IT experts in detail in order to know the reasons to adopt the current process and evolution in software development processes.

Acknowledgements This research project is supported under UTARRF; IPSR/RMC/UTARRF/2014-C1/R01, University Tunku Abdul Rahman, Malaysia. All procedures performed in the study involving human participants were in accordance with the ethical standards of the institution, and comparable ethical standards. Prior public disclosure approval to publish the results of the study have been taken from the university. The research study is led and supervised by Dr. Rehan Akbar as the sole member of the supervisory committee. The anonymity of the respondents of the questionnaire and data has been treated as strictly confidential, and not disclosed at any level.

References

1. Yu L., Mishra A.: Risk analysis of global software development and proposed solutions. *Automatika*. **51**(1), 89–98 (2010)
2. Niazi, M., et al.: Challenges of project management in global software development: a client–vendor analysis. *Inf. Softw. Technol.* **80**, 1–19 (2016)
3. Zafar, A., Ali, S., Shahzad, R.K.: Investigating integration challenges and solutions in global software development. In: *Frontiers of Information Technology (FIT)* (2011)
4. Alnuem, M.A., Ahmad, A., Khan, H.: Requirements understanding: a challenge in global software development, industrial surveys in Kingdom of Saudi Arabia, pp. 297–306 (2012)
5. Cho J.: Globalization and global software development. *Issues Inf. Syst.* **8**(2), 287–290 (2007)
6. Gomes, V., Marczak, S.: Problems? We all know we have them. Do we have solutions too? A literature review on problems and their solutions in global software development, pp. 154–158 (2012)
7. Javed, B., Minhas, S.S.: Process support for requirements engineering activities in global software development: a literature based evaluation (2010)
8. Conchúir, E.Ó., et al.: Exploring the assumed benefits of global software development (2006)
9. Nguyen, T., Wolf, T., Damian, D.: Global software development and delay: does distance still matter? pp. 45–54 (2008)

10. Khan, A.R., Akbar, R., Ten, D.W.H.: A study on Global Software Development (GSD) and software development processes in Malaysian software companies. *J. Telecommun. Electron. Comput. Eng.* **8**(2), 147–151 (2016)
11. Kroll, J., et al.: Handoffs management in follow-the-sun software projects: a case study. In: 2014 47th Hawaii International Conference on System Sciences (2014)
12. Holmstrom, H., et al.: Global software development challenges: a case study on temporal, geographical and socio-cultural distance (2006)
13. Hossain, E., Babar, M.A., Paik, H.-y.: Using scrum in global software development: a systematic literature review, pp. 175–184 (2009)
14. Gotel, O., et al.: Working across borders: overcoming culturally-based technology challenges in student global software development, pp. 33–40 (2008)
15. Mohagheghi, P.: Global software development: issues, solutions, challenges [cited 2016 10 Augst]. Available from: <http://www.idi.ntnu.no/grupper/su/publ/parastoo/gsd-presentation-slides.pdf> (2016)
16. Rao, N.M.: Challenges in execution of outsourcing contracts (2009)
17. Shah, Y.H., Raza, M., UIHaq, S.: Communication issues in GSD, p. 8 (2012)
18. Paasivaara, M., Lassenius, C.: Could global software development benefit from agile methods? p. 5 (2006)
19. Sriram, R., Mathew, S.K.: Global software development using agile methodologies: a review of literature (2012)
20. Kamaruddin, N.K., Arshad, N.H., Mohamed, A.: Chaos issues on communication in agile global software development (2012)
21. Ali Babar, M., Ihme, T., Pikkarainen, M.: An industrial case of exploiting product line architectures in agile software development (2009)
22. Mary, C.L., Joseph, W.R.: Effects of offshore outsourcing of information technology work on client project management. *Strat. Outsourcing: Int. J.* **2**(1), 4–26 (2009)
23. Kuusinen, K., Mikkonen, T., Pakarinen, S.: Agile user experience development in a large software organization: good expertise but limited impact (2012)
24. Sharp, J.H., Ryan, S.D.: Global agile team configuration. *J. Strat. Innov. Sustain.* **7**(1), 120–134 (2011)
25. Alzoubi, Y.I., Gill, A.Q.: Agile global software development communication challenges: a systematic review (2014)
26. Paasivaara, M., Durasiewicz, S., Lassenius, C.: Distributed agile development: using scrum in a large project, pp. 87–95 (2008)
27. Niazi, M., et al.: Establishing trust in offshore software outsourcing relationships: an exploratory study using a systematic literature review. *IET Softw.* **7**(5), 283–293 (2013)
28. Colomo-Palacios, R., et al.: Project managers in global software development teams: a study of the effects on productivity and performance. *Softw. Qual. J.* **22**(1), 3–19 (2014)
29. Guo, Y., Seaman, C.: A survey of software project managers on software process change, p. 7 (2008)
30. Martini, A., Pareto, L., Bosch, J.: Communication factors for speed and reuse in large-scale agile software development. In Proceedings of the 17th International Software Product Line Conference. ACM, Tokyo, Japan, pp. 42–51 (2013)
31. Versionone. State of agile survey. [cited 2016 25 Sep]. Available from: https://www.versionone.com/pdf/2011_State_of_Agile_Development_Survey_Results.pdf
32. Jain, R., Suman, U.: A systematic literature review on global software development life cycle. *SIGSOFT Softw. Eng. Notes* **40**(2), 1–14 (2015)
33. Lane, M.T., Agerfalk, P.J.: On the suitability of particular software development roles to global software development, pp. 3–12 (2008)
34. Betz, S., Makio, J., Stephan, R.: Offshoring of software development—methods and tools for risk management. In: International Conference on Global Software Engineering (ICGSE 2007) (2007)

35. Khan, A.A., Basri, S., Dominic, P.D.D.: A propose framework for requirement change management in global software development. In: 2012 International Conference on Computer & Information Science (ICCIS) (2012)
36. Niazi, M., et al.: GlobReq: a framework for improving requirements engineering in global software development projects: Preliminary results. In: 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012) (2012)
37. Lopez, A., Nicolas, J., Toval, A.: Risks and safeguards for the requirements engineering process in global software development. In: 2009 Fourth IEEE International Conference on Global Software Engineering (2009)
38. Silva, F.Q.B.d., et al.: Challenges and solutions in distributed software development project management: a systematic literature review. In: 2010 5th IEEE International Conference on Global Software Engineering (2010)
39. Khan, A.A., Keung, J.: Systematic review of success factors and barriers for software process improvement in global software development. *IET Softw.* **10**(5), 125–135 (2016)
40. Ahmad, M.A., Ubaidullah, N.H., Lakulu, M.: Current practices in monitoring software development process in Malaysia (2014)
41. Ali, R.Z.R.M., Ibrahim, S.: An integrated software process assessment for Malaysia's SME organizations (2011)
42. Almomani, M.A.T., et al.: Software development practices and problems in Malaysian small and medium software enterprises: a pilot study (2015)
43. Baharom, F., Deraman, A., Hamdan, A.R.: A survey on the current practices of software development process in Malaysia (2005)
44. Kamaroddin, J.H., et al.: The adoption of software development methodologies among IT organization in Malaysia (2012)
45. Mansor, Z., et al.: A survey on cost estimation process in Malaysia software industry (2012)
46. Mohamed, S.F.P., Baharom, F., Deraman, A.: An exploratory study on current software development practices in Malaysia focusing on agile based software development (2013)
47. Patton, M.Q.: *Qualitative Research & Evaluation Methods*. SAGE Publications, Thousand Oaks (2002)
48. Woods, M.: Interviewing for research and analysing qualitative data: an overview [cited 2016 6 June]. Available from: <http://owll.massey.ac.nz/pdf/interviewing-for-research-and-analysing-qualitative-data.pdf> (2016)
49. Thomas, D.R.: A general inductive approach for analyzing qualitative evaluation data. *Am. J. Eval.* **27**(2), 237–246 (2006)
50. Burnard, P., et al.: Analysing and presenting qualitative data. *BDJ* **204**(8), 429–432 (2008)
51. Dixon, H.: Qualitative data analysis using NVivo [cited 2016 20 June]. Available from: <http://www.slideshare.net/HelenDixon1/qualitative-data-analysis-using-n-vivo> (2016)
52. QSR International Pty Ltd. NVivo 11—getting started guide [cited 2016 1 July]. Available from: <http://download.qsrinternational.com/Document/NVivo11/11.3.0/en-US/NVivo11-Getting-Started-Guide-Starter-edition.pdf> (2016)
53. Lane Medical Library. Qualitative data analysis with NVIVO [cited 2016 1 June]. Available from: <https://www.youtube.com/watch?v=0YyVySrV2cM> (2016)
54. QSR International Learning how to explore and visualize your data with NVivo | NVivo Brown Bag Webinar [cited 2016 1 August]. Available from: <https://www.youtube.com/watch?v=Ez7PB6ZIA5I> (2016)
55. Rowe, D.: NVivo 10 coding [cited 2016 20 June]. Available from: <https://www.youtube.com/watch?v=4crQbeHKhtk> (2016)
56. Stanford University. Using NVivo for qualitative data analysis [cited 2016 17 March]. Available from: http://web.stanford.edu/group/ssds/cgi-bin/drupal/files/Guides/UsingNVivo9_0.pdf (2016)
57. Toolis, E.: Analyzing qualitative data using NVivo: an introduction [cited 2016 10 July]. Available from: <http://csass.ucsc.edu/images/NVivo.pdf> (2016)
58. Asnawi, A.L., Gravell, A.M., Wills, G.B.: An empirical study: understanding factors and barriers for implementing agile methods in Malaysia (2010)

59. Khaleel, Y., Sulaiman, R.: A system development methodology for erp system in SMEs of Malaysian manufacturing sectors. *J. Theor. Appl. Inf. Technol.* **47**(2), 504–513 (2013)
60. Hastie, S., Wojewoda, S.: Standish Group 2015 Chaos Report—Q&A with Jennifer Lynch [Online]. Available: <https://www.infoq.com/articles/standish-chaos-2015> (2015). Accessed 26 Sept 2016
61. THE STANDISH GROUP. The Standish Group Report Chaos [Online]. Available: <https://www.projectsmart.co.uk/white-papers/chaos-report.pdf> (2014). Accessed 12 April 2017

Automated Scheduling of Hostel Room Allocation Using Genetic Algorithm



Rayner Alfred and Hin Fuk Yu

Abstract Due to the rapid growth of the student population in tertiary institutions in many developing countries, hostel space has become one of the most important resources in university. Therefore, the decision of student selection and hostel room allocation is indeed a critical issue for university administration. This paper proposes a hierarchical heuristics approach to cope with hostel room allocation problem. The proposed approach involves selecting eligible students using rank based selection method and allocating selected students to the most suitable hostel room possible via the implementation of a genetic algorithm (GA). We also have examined the effects of using different weight associated with constraints on the performance of the GA. Results obtained from the experiments illustrate the feasibility of the suggested approach in solving the hostel room allocation problem.

Keywords Scheduling · Hostel space · Genetic algorithm

1 Introduction

A hostel management system is responsible for allocating students to available hostel spaces and managing students' resident and hostels information. The rapid rise in the student population of the tertiary institutions over the years in developing countries has become an inevitable challenge to university administration due to limited hostel spaces available in the university. Therefore, in order to cope with the issues above, optimization of the scheduling process for hostel room allocation is a must. Hostel room scheduling is the problem of allocating eligible students to hostel rooms while satisfying specified constraints and the scheduling process is

R. Alfred (✉) · H. F. Yu
Knowledge Technology Research Unit, Faculty of Computing and Informatics,
Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
e-mail: ralfred@ums.edu.my

H. F. Yu
e-mail: roronuozero@gmail.com

automated [2]. Automated scheduling is usually executed using a scheduling algorithm to generate a sequence of actions in order to achieve certain objectives. Some of the applications or problems which require automated scheduling are flow shop scheduling problem (FSSP) and job shop scheduling problem (JSSP) [1] including Hostel Space Allocation Problem [2]. Some examples of technique or method for dealing with scheduling problem are hybrid metaheuristic algorithm [3], relax and fix heuristics [4], and hierarchical heuristics approach [2]. The reason why these proposed techniques are advantageous is mainly because of the simplicity, flexibility, inexpensive computational cost, and derivation-free mechanism of heuristic or metaheuristic [5]. In this paper, a hierarchical heuristics approach will be proposed and implemented to optimize the scheduling process of hostel room allocation. Hierarchical heuristics approach is chosen because it is able to provide good results in dealing with multilevel allocation process of hostel space allocation problem [2].

The rest of the paper is organized as follow. Section 2 discusses related works with regard to implementation of GA-based heuristic in dealing with scheduling problems. Section 3 presents the design of the proposed hierarchical heuristics approach while the experimental setup is discussed in Sect. 4. This paper is concluded in Sect. 5.

2 Related Works

A large amount of research and study has been conducted in a variety of domains of scheduling problem such as JSSP [6], shelf space allocation problem (SSAP) [7], timetabling problem [8], knapsack problem [9], and bin-packing problem [10]. Yaqin et al. [6] studied on the job shop scheduling with multiobjectives-based on GA and according to the research, GA is relatively faster than other methods used for scheduling which are usually based on single point search as GA do multipoint searches on the whole population simultaneously. Their studies show that GA is feasible in optimizing NP-hard or combinatorial optimization problem. Castelli and Vanneschi [7] proposed a hybrid algorithm that combines GA and variable neighbor search (VNS) algorithm to increase the explorative ability of GA to prevent GA from being stuck in local optimum and results obtained from their experiments had shown the enhancement in explorative ability of GA. Yang et al. [9] introduced attribute reduction of rough sets into the crossover of GA to tackle multidimensional knapsack problem which is also known as the multi-constraints knapsack problem. The steps of the GA proposed by them are different than the steps in typical GA. If reduct is found in the genes of the parent chromosome, the particular genes in the reduct will be the points selected to crossover whereas in the event of no reduct is found, the single-point crossover is applied. Results obtained from their studies show that the likelihood of the proposed GA to reach a maximal solution is relatively higher than conventional GA. Other than that, Bennell et al. [10] proposed a multi-crossover genetic algorithm (MXGA) to solve

a new variant of the two-dimensional bin-packing problem where each rectangle is assigned a due date and each bin has a fixed processing time. Experiments conducted in their research shows that the MXGA performance better than the single crossover genetic algorithm in solving the bin-packing problem. One crucial similarity between the works reviewed above is such that GA is always a part of the proposed solution or method. Therefore, the motivation of our work is to implement GA in the proposed hierarchical heuristic approach in order to tackle the hostel room allocation problem. The hierarchical heuristics approach proposed in this work is different than the hierarchical heuristics approach proposed by Adewumi and Ali [2], in which instead of greedy-like heuristic, the rank based selection method is used in the selection stage. Moreover, the representation, as well as crossover and mutation mechanism of the GA, is different as well.

3 Hierarchical Heuristics Approach

This section explains the framework of the hierarchical heuristics approach. The main components of the approach, which are rank-based selection method and GA will be explained. The proposed hierarchical heuristics approach framework is shown in Fig. 1.

With respect to Fig. 1, necessary information of applicants, rooms and specified constraints will initially be provided to the framework. This information will be used during the selection stage; with the implementation of rank-based selection method, applicants will be ranked based on the weight they hold and higher ranker will be selected first. Weight hold by the applicant is determined via the degree in which the applicant satisfies the specified constraints and the order of priority of constraints will be shown in Sect. 4. A list of selected students or applicants will be produced at the end of the selection stage. During the allocation stage, the selected students will be allocated to the most suitable hostel room possible using GA-based on specified constraints. The constraints will be shown in Sect. 4.

3.1 Genetic Algorithm

A GA is a stochastic search algorithm which can be used to a variety of combinatorial optimization problems [11]. GA is also known as nature inspired algorithm for the reason that the steps in GA are based on the evolutionary process of biological organisms. GA evolves according to the principle “survival of the fittest”; this implies that offspring that carries best attributes genes will have higher chance to survive to the next generation. Generally, GA simulates the evolutionary process of biological organisms by generating individuals randomly that form up an initial population. Each individual is encoded into a string (chromosome) which represents a possible solution to a given problem. Fitness function is performed on

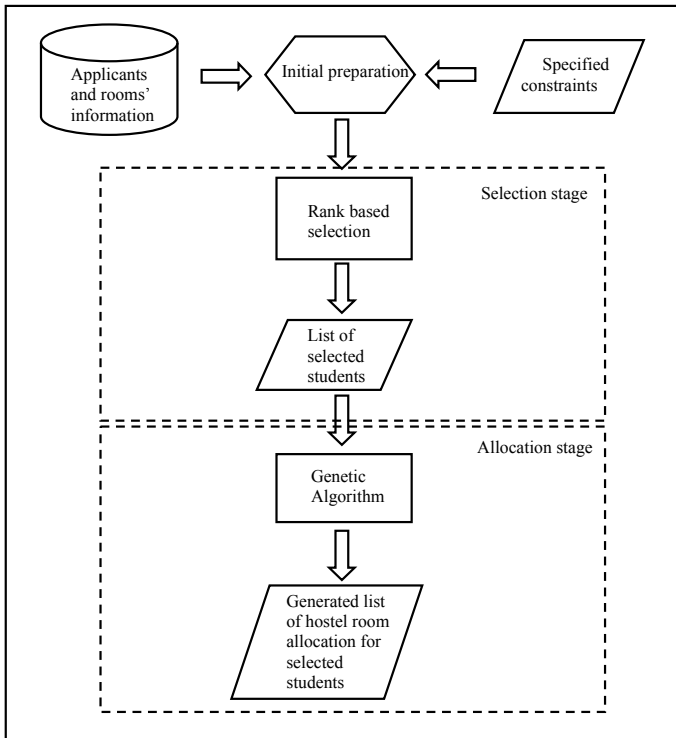


Fig. 1 Framework of hierarchical heuristics approach

each and every individual and each individual has its own fitness value. A probability of being selected during the selection process is assigned to each individual based on the individual's fitness value so as to ensure the characteristics of highly fit individuals or chromosomes are passed down to the next generation through a crossover procedure. The most general crossover method is called a one-point crossover; in which a single crossover point on both parents' chromosomes is selected first and then all information located beyond that single crossover point in either chromosome are swapped between these two parent chromosomes. The mutation is often applied to the offspring's chromosomes after crossover by altering some genes in the chromosomes, but with very low occurrence rate. In most of the case, less fit individuals will be replaced by the highly fit offspring and the new population will be reevaluated by the fitness function again. This sequence of processes will be repeated until a termination criterion is met.

3.2 Solution Encoding and Generic Operators

The encoding applied to represent candidate solutions for the GA used in the allocation stage as shown in Fig. 1 will be described in this section. The generic operators, which are crossover and mutation of the GA, will be described as well. Figure 2 illustrates how the candidate solutions are encoded. Based on Fig. 2, the integer values are the unique ID of selected students. The number of genes owned by each room varies with the types of room and as for this case, single room owns two genes while double room owns four genes. The length of the chromosome varies with the total number of rooms available. Figure 3 illustrates the process of crossover.

In Fig. 3, a self-crossover is used in order to ensure there is no repetition of integer values in the chromosome as the integer values are the unique ID of students. The crossover is done by randomly selects two portions with the same number of genes and swaps the genes of the selected portions in a random manner. The genes swapping phenomenon is controlled by a crossover rate of 0.45. On the other hand, the mutation process is almost the same as crossover process, in which the difference is that all the genes of the selected portions will be swapped completely in a random manner instead of controlled by a crossover rate of 0.45. The likelihood of the mutation process to happen is controlled by a mutation rate of 0.05.

3.3 Fitness Function

Fitness function, which is also known as evaluation function is used to evaluate individuals after initialization, crossover, and mutation steps to determine how fit the individuals are. In this work, the fitness of GA's individual will be evaluated based on the degree in which the constraints are satisfied. Each individual of the GA will have their own fitness value which is a set of real number in [0, 1]. A value of 0 indicates a complete violation of all given constraints while a value of 1 indicates no constraint is violated. The fitness of an individual, by referencing the fitness function in [2], is computed as:

$$f = \sum_{i=1}^n w_i u_i \in [0, 1] \tag{1}$$

where u_i in Eq. (1) is the utilization factor given as:

Fig. 2 Chromosome encoding of the GA in allocation stage

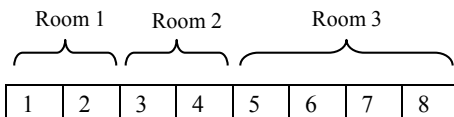
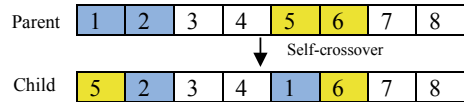


Fig. 3 Crossover process

$$u_i = \frac{1}{n} \quad (2)$$

w_i is the weight hold by gene i in a chromosome, n in Eqs. (1) and (2) is the total number of genes a chromosome has.

4 Experimental Setup

This section describes the experiments that are carried out in this work. The sample datasets for students (applicants) and rooms are shown in Tables 1, 2 and 3. The initial population of the GA will be randomly generated.

The main parameters for GA used are as follow 30 population size, 100 generations, crossover rate at 45% and mutation rate at 5%. The constraints applied to the selection and allocation process are as follow:

1. Selection:

- Hard constraints
 - All first year students must be selected (FY).
 - Student representative must be selected (SR).
 - Uniformed units must be selected (SU).
 - Handicapped students must be selected (HS).
- Soft constraints
 - Students with higher student development points are more likely to be selected (SD).
 - Students with higher CGPA are more likely to be selected (SC).

Table 1 Statistics of applicants per category

Category	Male	Female
First year	703	2635
Senior	297	1115
Student representative	4	8
Uniformed units	197	353
Handicapped	2	5
Food science and forestry	192	328
Others	605	3056

Table 2 Availability of rooms for male students

Type	Hostel				
	A/B	C/D	E	Angkasa	Usia
Single room capacity	52	52	52	52	
Double room capacity	148	148	148	148	0
Ground floor capacity	50	50	50	50	0

Table 3 Availability of rooms for female students

Type	Hostel				
	A/B	C/D	E	Angkasa	Usia
Single room capacity	140	140	140	140	140
Double room capacity	560	560	560	560	560
Ground floor capacity	150	150	150	150	150

2. Allocation:

- Hard constraints
 - Uniformed units must be allocated to hostel E (UH).
 - Handicapped students must be allocated to rooms situated on the ground floor (HG).
 - Single rooms are for senior students (SS).
- Soft constraints
 - Food Science (FS) and Forestry (FT) student should be allocated to hostel E (FE).
 - Students should be allocated to the hostel of their choices (SH).
 - First year students should be allocated to hostels inside the campus (SI).

Weights associated with selection and allocation constraints are shown in Tables 4 and 5. With respect to Table 4, weight associated to constraint FY is 0.51 and this is to make sure the sum of weights of other constraints is less than 0.51 as FY holds the highest priority in hard constraints. As for SR, SU, and HS, they have the same level of priority, in which fulfilling one or two or all of these constraints will give the student the same amount of weight, which is 0.25. This is to ensure the sum of weights for remaining soft selection constraints will not exceed 0.25. On the other hand, the four sets of weights’ distribution for allocation constraints shown in Table 5 are used to examine the effect of weight associated constraints on the performance of the GA; during allocation stage, the results obtained from the rank

based selection are processed by the GA using these four sets of weights' distribution and the results obtained are used as a basis for optimizing the weights' distribution for allocation constraints. The optimization of weights' distribution for allocation constraints is done by simulating the GA 30 times using the same set of results obtained from rank-based selection but with 30 different sets of weights' distribution for allocation constraints which are randomly generated.

The results obtained from the rank-based selection during selection stage are shown in Table 6 where it shows the fulfillment of hard constraints as soft constraints of selection, which are SD and SC cannot be quantified. However, with the implementation of rank-based selection and normalization of student development point and CGPA, SD, and SC will always be optimized. The optimized weights' distribution for allocation constraints is shown in Table 7 while the allocation results obtained from simulating the GA using the four sets of weights' distribution and optimized weights' distribution are shown in Table 7. Based on Table 8, the performance of the GA is measured in term of constraints satisfaction in percentage

Table 4 Weights' distribution for selection constraints

Type	Constraints	Weight
Hard	FY	0.51
	SR, SU or HS	0.25
Soft	SD	0.12
	SC	0.12

Table 5 Weights' distribution for allocation constraints

Type	Constraints	Set 1	Set 2	Set 3	Set 4
Hard	UH	0.22	0.25	0.27	0.30
	HG	0.22	0.25	0.27	0.30
	SS	0.22	0.25	0.27	0.30
Soft	FE	0.12	0.09	0.07	0.04
	SH	0.11	0.08	0.06	0.03
	SI	0.11	0.08	0.06	0.03

Table 6 Selection results obtained from rank-based selection method

Category	Selected		Not selected	
	Male	Female	Male	Female
First year	703	2635	0	0
Senior	97	865	200	250
Student representative	4	8	0	0
Uniformed units	197	353	0	0
Handicapped	2	5	0	0
Food science and forestry	103	207	89	121
Others	494	2927	111	129

Table 7 Optimized weights' distribution for allocation constraints

Type	Constraints	Weight
Hard	UH	0.231
	HG	0.243
	SS	0.216
Soft	FE	0.098
	SH	0.109
	SI	0.105

Table 8 Allocation results obtained Using GA with different weights' distribution

Type	Constraints	No of students	Results				
			Set 1	Set 2	Set 3	Set 4	Optimized weight
Hard	UH	550	462	505	509	512	506
	HG	7	4	5	5	6	6
	SS	962	714	801	832	825	822
Total		1519	1180 (77.68%)	1311 (86.31%)	1346 (88.61%)	1343 (88.41%)	1334 (87.82%)
Soft	FE	310	255	202	181	172	251
	SH	3338	2890	2458	2231	2091	2718
	SI	4300	3457	3321	2989	2786	3362
Total		7948	6602 (83.06%)	5981 (75.25%)	5401 (67.95%)	5049 (63.52%)	6331 (79.65%)

instead of average fitness value of GA's population. This is because fitness value is based on weights' distribution for allocation constraints and, therefore, it is not suitable for indicating the performance of the GA. The results obtained from both selection and allocation stages show the feasibility of the suggested approach in solving the hostel room allocation problem.

5 Conclusion

This paper proposed a hierarchical heuristics approach that includes the implementation of GA to cope with hostel room allocation problem. The effect of weights associated with constraints is examined by simulating GA with multiple sets of weights' distribution for allocation constraints. Optimized weights' distribution for allocation constraints is obtained during the simulation process. Results obtained from the experiments conducted show the feasibility of the proposed approach in dealing with the problem at hand. However, improvements can still be made in the

future to increase the overall performance of the proposed approach by tuning the GA components and parameters such as encoding method, crossover method, mutation method, crossover rate, mutation rate, population size, etc.

References

1. Zhang, X-F., Koshimura, M., Fujita, H., Hasegawa, R.: Combining PSO and local search to solve scheduling problems, pp. 347–354 (2011)
2. Adewumi, A.O., Ali, M.M.: A multi-level genetic algorithm for a multi-stage space allocation problem. *Math. Comput. Model.* **51**(1), 109–126 (2010)
3. Zhang, Q., Manier, H., Marie-Ange, M.: A hybrid metaheuristic algorithm for flexible job-shop scheduling problems with transportation constraints, pp. 441–448 (2012)
4. Deisemara, F., Reinaldo, M., Socorro, R.: Relax and fix heuristics to solve one-stage one-machine lot-scheduling models for small-scale soft drink plants. *Comput. Oper. Res.* **37**, 684–691 (2009)
5. Seyedali, M., Andrew, L., Sanaz, M.: Confidence measure: a novel metric for robust meta-heuristic optimization algorithms. *Inf. Sci.* **317**, 114–142 (2015)
6. Yaqin, Z., Beizhi, L., Lv, W.: Study on job-shop scheduling with multi-objectives based on genetic algorithms, vol. 10, pp. 10–294 (2010)
7. Castelli, M., Vanneschi, L.: Genetic algorithm with variable neighborhood search for the optimal allocation of goods in shop shelves. *Oper. Res. Lett.* **42**(5), 355–360 (2014)
8. Soria-Alcaraz, J., Carpio, M., Puga, H.: A new approach of design for the academic timetabling problem through genetic algorithms, pp. 96–101 (2010)
9. Yang, H., Wang, M., Chen, Y., Huang, Y., Kao, C.: Crossover based on rough sets—a case of multidimensional knapsack problem, pp. 2411–2415 (2010)
10. Bennell, J., Soon Lee, L., Potts, C.: A genetic algorithm for two-dimensional bin packing with due dates. *Int. J. Prod. Econ.* **145**(2), 547–560 (2013)
11. Reeves, C.R.: *Modern heuristic techniques for combinatorial problems*. Blackwell Scientific, Hoboken (1993)

Evaluation of ASTER TIR Data-Based Lithological Indices in Parts of Madhya Pradesh and Chhattisgarh State, India



Himanshu Govil, Subhanil Guha, Prabhat Diwan, Neetu Gill
and Anindita Dey

Abstract The present study was performed in some parts of Madhya Pradesh and Chhattisgarh State, India to compare the different quartz indices, feldspar indices and mafic indices according to Ninomiya (2005) and Guha (2016) using thermal infrared (TIR) bands (band 10, band 11, band 12, band 13, and band 14) of Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) data for detecting quartz, feldspar and mafic minerals. Results showed that these indices are equally useful for delineating quartz, feldspar or mafic minerals. It was noticed from the correlation coefficients that Guha's mafic index (GMI) and Ninomiya's mafic index (NMI) presented almost the same result. Guha's quartz index (GQI) was more powerful than Ninomiya's quartz index (NQI) in identifying quartz content in alkali granites and this GQI was also comparable with the Rockwall and Hofstra's quartz index (RHQI) in identifying quartz content in alkali granite.

Keywords ASTER · Feldspar · Mafic · Quartz · Thermal infrared

1 Introduction

The visible and near-infrared (VNIR) bands and the short wave infrared (SWIR) bands of satellite data are not so useful for detecting quartz and feldspar minerals [13] whereas thermal infrared (TIR) bands are useful for the delineation of the aforesaid minerals because of the vibration of Si–O bonds [7]. Lithological mapping is an important task of geological exploration. Earth observation techniques play an important role in lithological and mineralogical mapping. TIR region are

H. Govil · S. Guha (✉) · P. Diwan
Department of Applied Geology, National Institute of Technology Raipur, Raipur, India
e-mail: subhanilguha@gmail.com

N. Gill
Chhattisgarh Council of Science and Technology, Raipur, India

A. Dey
Department of Geography, Nazrul Balika Vidyalaya, Guma, West Bengal, India

less studied than VNIR or SWIR regions for mineral identification due to the less availability of satellite sensors in the TIR domain.

Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) satellite sensor has three VNIR bands (band 1–3), six SWIR bands (band 4–9), and five TIR bands (band 10–14). ASTER VNIR and SWIR bands were used extensively in the mapping of clay, carbonate, aluminium hydroxide, and iron hydroxide [15].

In later studies, a different combination of ASTER TIR bands were frequently used in differentiating various types of rock [4, 8–10].

Ninomiya et al. [9] applied some selected lithological parameters to propose a series of ASTER TIR band based lithological indices for quartz, mafic, feldspar, and carbonate rock detection [1, 4, 9, 12]. High ratio of alkali feldspar is present in alkali feldspar rich granite while in granodiorite and tonalite this ratio becomes lower [2]. Granitoids are also characterized by a low amount of mafic mineral which is found in high amount in tonalite.

The present study tried to detect quartz, feldspar, and mafic mineral in various types of rocks using ASTER TIR bands. A number of geoscientists examined the utility of ASTER TIR bands in the formulation of rocks and mineral indices to identify the basic mineral content in the lithological variation [3–5, 9]. The published geological map was used as the reference map to get a reliable interpretation of the final results. Different lithological indices were compared to each other for enhancing the sub-variants of granitoids.

2 Study Area and Geology

The parts of Madhya Pradesh and Chhattisgarh State of east-central India were selected as the study area (Fig. 1) of the present research work. This entire region is composed of felsic and mafic rocks in significant amount. Granitoid or quartz enriched felsic rock has a combination of quartz, mafic and feldspar minerals. Geological map published from the Geological Survey of India with 1:250,000 scale was used as the reference map for analyzing the derived lithological indices.

3 Materials and Methods

Thermal bands specification of ASTER data was provided in Table 1. Emissivity values of quartz and feldspar minerals were analyzed and it was observed that the band 12 reflects the lowest emissivity value and band 10 and band 13 reflect the higher emissivity values for quartz mineral. ASTER band 11 has a lower emissivity than band 10 for feldspar mineral (Fig. 2). ASTER band 13 indicates the lowest emissivity and band 12 and band 14 have higher emissivity for mafic rocks and

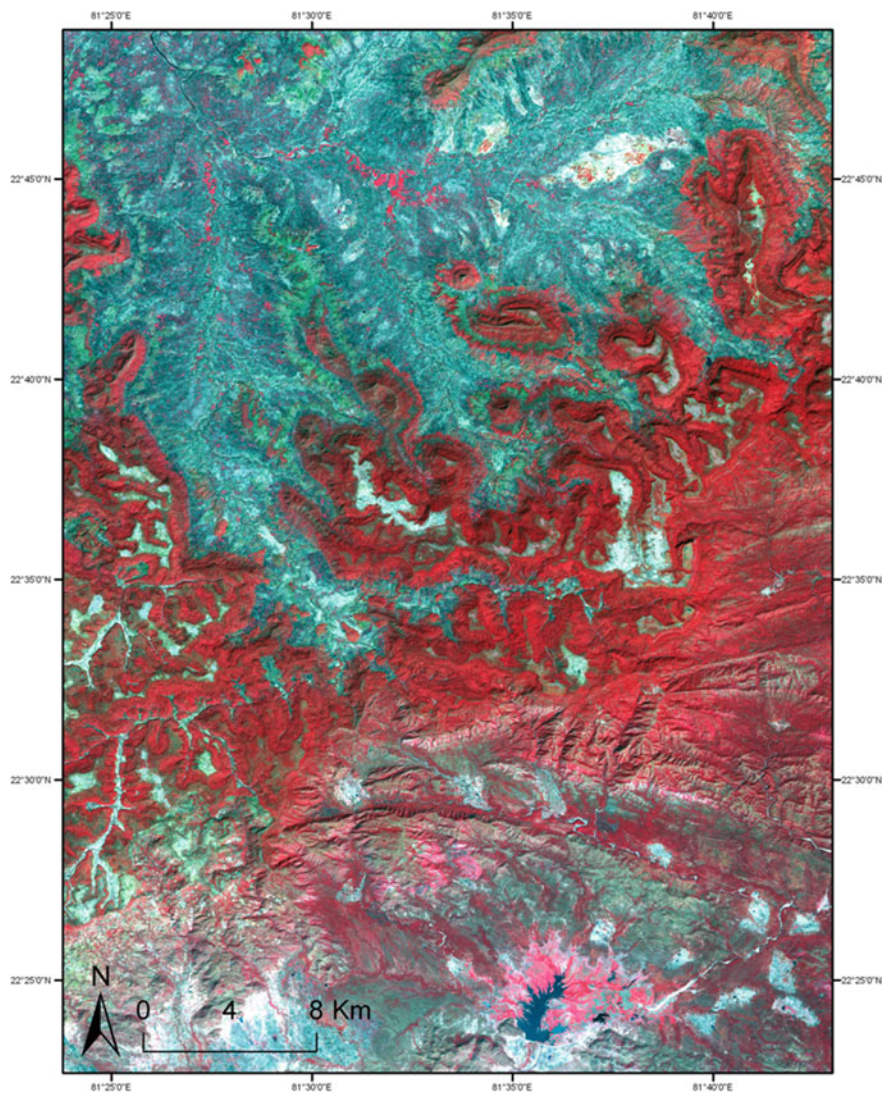


Fig. 1 Location of the study area

Table 1 General information of ASTER TIR bands

Data	Sensor	TIR bands	Wavelength (μm)	Spatial resolution (m)
ASTB070414051954	TIR	10	8.125–8.475	90
Date: 14 April 2007	TIR	11	8.475–8.825	90
	TIR	12	8.925–9.275	90
	TIR	13	10.25–10.95	90
	TIR	14	10.95–11.65	90

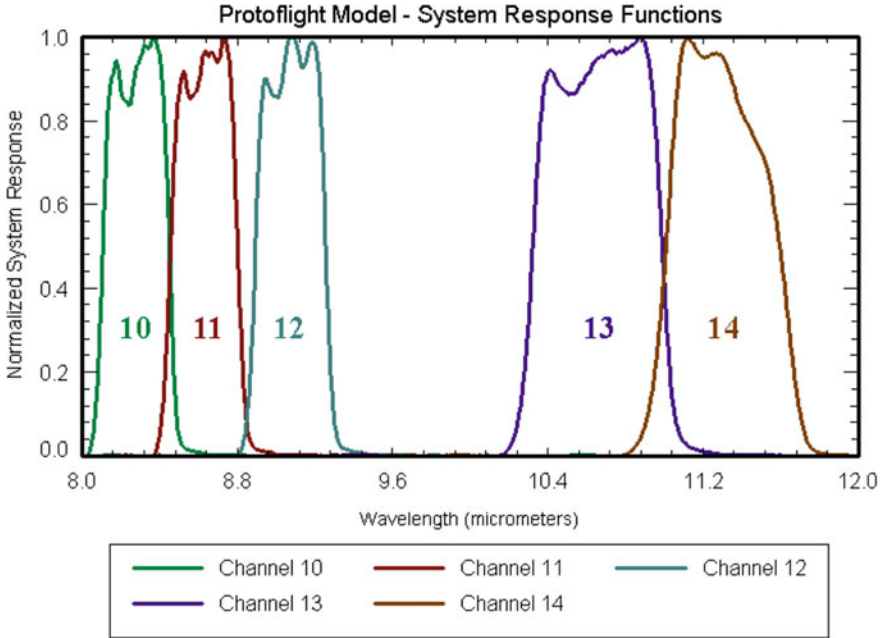


Fig. 2 Spectral ASTER TIR bands in the wavelength of the electromagnetic spectrum (Source NASA Jet Propulsion Laboratory)

minerals [14]. A significant reverse trend of emissivity values for TIR band 11 and band 12 was observed for quartz and feldspar minerals [14].

The lithological indices applied for detecting quartz, mafic and feldspar minerals were retrieved using TIR bands of ASTER sensor.

Guha's indices are described as follows [6]:

$$\text{Guha's quartz index (GQI)} = \frac{\text{band 10}}{\text{band 12}} \times \frac{\text{band 13}}{\text{band 12}} \quad (1)$$

$$\text{Guha's mafic index (GMI)} = \frac{\text{band 12}}{\text{band 13}} \times \frac{\text{band 14}}{\text{band 13}} \quad (2)$$

$$\text{Guha's feldspar index (GFI)} = \frac{\text{band 10}}{\text{band 11}} \times \frac{\text{band 12}}{\text{band 11}} \quad (3)$$

GQI, GMI and GFI were generally used to detect the rocks with high quartz content, high mafic minerals content, and high feldspar content, respectively. NQI and NMI were the two established lithological indices to detect quartz and mafic minerals, respectively [9]. Besides, Rockwall and Hofstra [11] proposed another index (RHQI) to detect quartz-rich rocks.

These indices are described as follows:

$$\text{Ninomiya's quartz index (NQI)} = \frac{\text{band 11} \times \text{band 11}}{\text{band 10} \times \text{band 12}} \quad (4)$$

$$\text{Ninomiya's mafic index (NMI)} = \frac{\text{band 12}}{\text{band 13}} \quad (5)$$

$$\text{Rockwall and Hofstra's quartz index (RHQI)} = \frac{\text{band 11}}{\text{band 10} + \text{band 12}} \times \frac{\text{band 13}}{\text{band 12}} \quad (6)$$

The present study compared GQI with respect to NQI in detecting quartz and GMI with respect to NMI in detecting mafic minerals. Applicability of RHQI and GQI was also examined.

4 Results

The study attempted to compare the lithological indices for identifying mafic, quartz and feldspar minerals. It was noticed that granite and quartz-rich granite were differentiated from mafic rich gneiss as yellow colour produced in the composite image of band 14 as red band, band 13 as green band and band 12 as blue band (Fig. 3a). Quartz particles were characterized with a red tint in the composite image of band 14 as red band, band 12 as green band and band 10 as blue band (Fig. 3c). NQI and NMI were the established rock bearing indices to distinguish quartz and mafic minerals in various types of lithological configurations. GQI provided a better result than NQI in the delineation of quartz content (Fig. 4a and b). Quartz enriched rocks were found as dark colour and bright colour in NQI and GQI images, respectively. This is because of the different emissivity values for quartz and feldspar in these TIR bands. GQI image was also compared to RHQI image. In RHQI image, the emittance value for quartz should be higher in band 11 [9]. Again, the emittance value for quartz was also high in band 13 compared to band 12 [6]. Thus, RHQI was used to delineate quartz content in granite and alkali granite (Fig. 4a and c). In GFI image, feldspar particles appeared as brighter in alkali granite (Fig. 4d). GMI and NMI built a strong correlation and thus these images were complementary to each other (Fig. 4e and f). It was found a relatively weak correlation (0.092) between GQI and NQI while GMI and NMI showed a strong (0.813) correlation (Table 2). However, GQI was comparable with RHQI in order to identify quartz enriched granitoids. GQI and RHQI built a very high correlation (0.808).

GFI and GQI had a very weak negative correlation (−0.090) because generally feldspar particle is increased slowly with the gradual decrease of quartz particle.

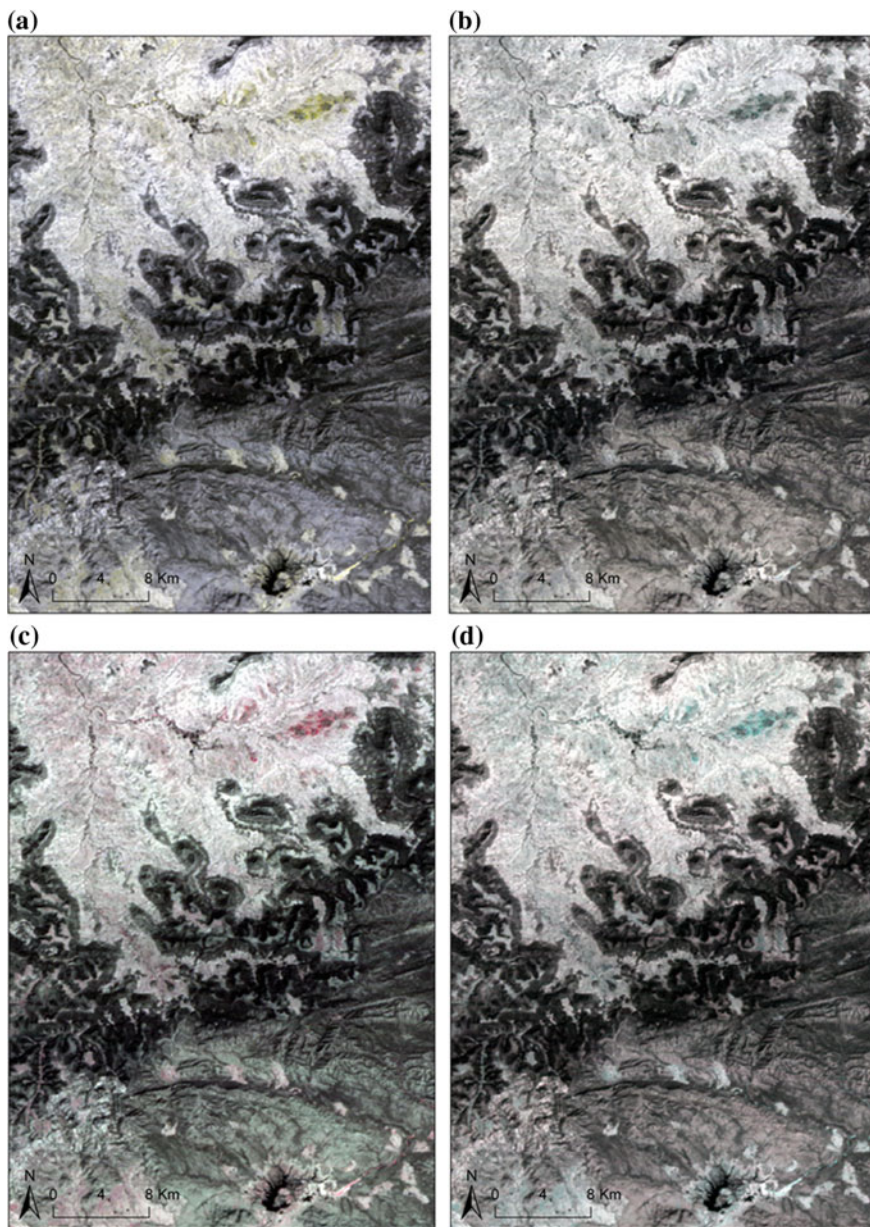


Fig. 3 ASTER TIR radiance composite images: **a** band 14 as red band, band 13 as green band, band 12 as blue band; **b** band 12 as red band, band 11 as green band, band 10 as blue band; **c** band 14 as red band, band 12 as green band, band 10 as blue band; **d** band 11 as red band, band 14 as green band, band 13 as blue band

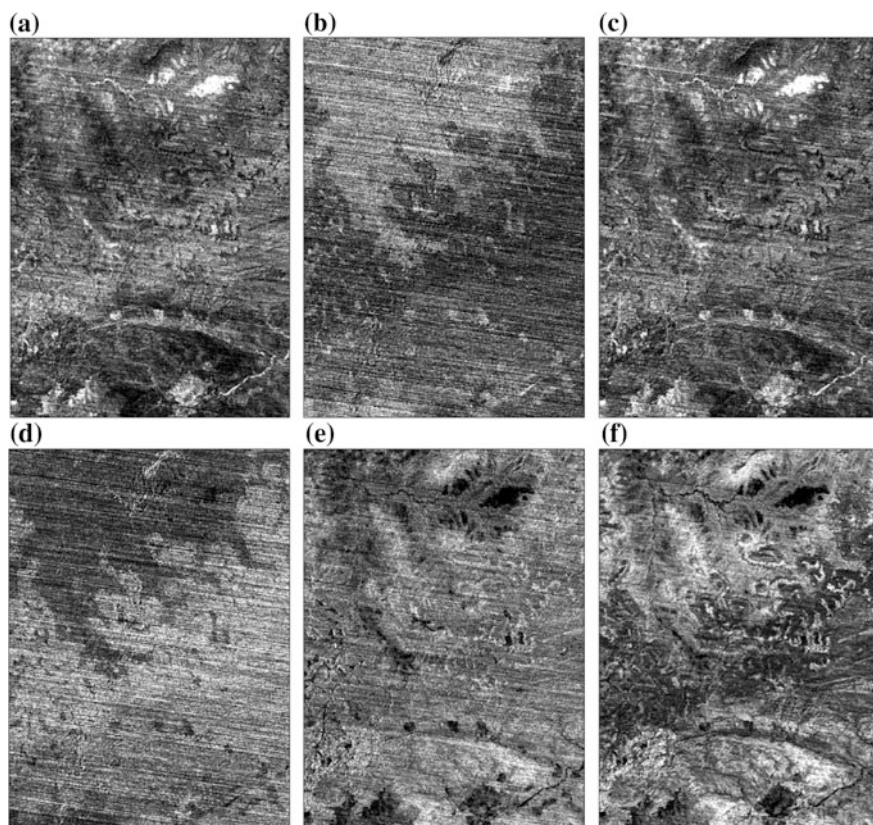


Fig. 4 a GQI image; b NQI image; c RHQI image; d GFI image; e GMI image; f NMI image

Table 2 Correlation Matrix of various lithological indices

	GQI	NQI	RHQI	GFI	GMI	NMI
GQI	1.00000	0.09179	0.80787	-0.09011	-0.80419	-0.88018
NQI	0.09179	1.00000	0.51516	-0.99992	-0.32162	-0.07198
RHQI	0.80787	0.51516	1.00000	-0.51341	-0.84916	-0.89145
GFI	-0.09011	-0.99992	-0.51341	1.00000	0.31993	0.06999
GMI	-0.80419	-0.32162	-0.84916	0.31993	1.00000	0.81306
NMI	-0.88018	-0.07198	-0.89145	0.06999	0.81306	1.00000

GMI and GFI were positively correlated with a regression value of 0.319 because feldspar particle is increased with the increase of mafic particle. GQI and GMI were negatively correlated with a very high regression value of -0.804 because quartz particle is increased rapidly with the decrease of mafic particle.

5 Discussions and Conclusions

The use of the lithological indices can be treated as a significant method for the delineation of quartz, mafic and feldspar minerals in granitoids. GQI was established as a more effective method compared to NQI (Fig. 4a and b). RHQI can retrieve quartz content almost as much as in GQI. GMI also provided a reliable comparison with NMI for delineation of mafic minerals (Fig. 4e and f). GMI and NMI were highly correlated (0.813) but the relationship became weak (0.091) between NQI and GQI (Table 2). GQI and QIRH also correlated with a high regression value (0.807) and hence can be used as important indices to determine quartz content in granite and alkali granite rocks. The results were significantly comparable to the published geological map. These indices can be evaluated in the different climatic and geologic environment to establish a more accurate conclusion.

Acknowledgements The authors are thankful to the United States Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA) Jet Propulsion Laboratory (JPL).

References

1. Bertoldi, L., Massironi, M., Visona, D., Carosi, R., Montomoli, C., Gubert, F., et al.: Mapping the Buraburi granite in the Himalaya of Western Nepal: remote sensing analysis in a collisional belt with vegetation cover and extreme variation of topography. *Remote Sens. Environ.* **115**(5), 1129–1144 (2011)
2. Bose, M.K.: *Igneous Petrology*. World Press, Kolkata (1997)
3. Chen, J., Wang, A.J.: The pilot study on petrochemistry components mapping with ASTER thermal infrared remote sensing data. *J. Remote Sens.* **11**, 601–608 (2007)
4. Ding, C., Liu, X., Liu, W., Liu, M., Li, Y.: Mafic and ultramafic and quartz-rich rock indices deduced from ASTER thermal infrared data using a linear approximation to the Planck function. *Ore Geol. Rev.* **60**, 161–173 (2014)
5. Ding, C., Xuqing, L., Xiangnan, L., Liting, Z.: Quartzose–mafic spectral feature space model: a methodology for extracting felsic rocks with ASTER thermal infrared radiance data. *Ore Geol. Rev.* **66**, 283–292 (2015)
6. Guha, A., Kumar, V.K.: New ASTER derived thermal indices to delineate mineralogy of different granitoids of an Archaean Craton and analysis of their potentials with reference to Ninomiya's indices for delineating quartz and mafic minerals of granitoids—an analysis in Dharwar Craton, India. *Ore Geol. Rev.* **74**, 76–87 (2016)
7. Kahle, A.B.: Thermal inertia imaging: a new geologic mapping tool. *Geophys. Res. Lett.* **3**, 419–421 (1976)
8. Kalinowski, A., Oliver, S.A.: ASTER Mineral Index Processing Manual. http://www.ga.gov.au/image_cache/GA7833.pdf (2004)
9. Ninomiya, Y., Fu, B., Cudahy, T.J.: Detecting lithology with advanced spaceborne thermal emission and reflection radiometer (ASTER) multispectral thermal infrared radiance-at-sensor data. *Remote Sens. Environ.* **99**, 127–139 (2005)
10. Rajendran, S., Nasir, S.: ASTER spectral sensitivity of carbonate rocks—study in Sultanate of Oman. *Adv. Space Res.* **53**, 656–673 (2014)

11. Rockwall, B.W., Hofstra, A.H.: Identification of quartz and carbonate minerals across northern Nevada using ASTER thermal infrared emissivity data—implications for geologic mapping and mineral resource investigations in well-studied and frontier areas. *Geosphere* **4**, 218–246 (2008)
12. Rowan, L.C., Mars, J.C., Simpson, C.J.: Lithologic mapping of the Mordor, NT, Australia ultramafic complex by using the advanced spaceborne thermal emission and reflection radiometer (ASTER). *Remote Sens. Environ.* **99**(1–2), 105–126 (2005)
13. Salisbury, J.W., Walter, L.S.: Thermal infrared (2.5–13.5 μm) spectroscopic remote sensing of igneous rock types on particulate planetary surfaces. *J. Geophys. Res.* **94**(B7), 9192–9202 (1989)
14. Son, Y.-S., Kang, M.-K., Yoon, W.-J.: Lithological and mineralogical survey of the Oyu Tolgoi region, South-eastern Gobi, Mongolia using ASTER reflectance and emissivity data. *Int. J. Appl. Earth Obs. Geoinf.* **26**, 205–216 (2014)
15. Zhang, X., Pazner, M., Duke, N.: Lithologic and mineral information extraction for gold exploration using ASTER data in the south Chocolate Mountains (California). *ISPRS J. Photogramm.* **62**, 271–282 (2007)

Analyzing Linear Relationships of LST with NDVI and MNDISI Using Various Resolution Levels of Landsat 8 OLI and TIRS Data



Himanshu Govil, Subhanil Guha, Prabhat Diwan, Neetu Gill and Anindita Dey

Abstract The present study used the Normalized Difference Vegetation Index (NDVI) and the Modified Normalized Difference Impervious Surface Index (MNDISI) to determine the linear relationship between Land Surface Temperature (LST) distribution and these remote sensing indices under various spatial resolutions. Four multi-date Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) images of parts of Chhattisgarh State of India were used from four different seasons (spring, summer, autumn and winter). The results indicate that LST established moderate to strong negative correlations with NDVI and weak negative to moderate positive correlations with MNDISI at various spatial resolutions (30–960 m). Generally, the coarser resolutions (840–960 m) possess stronger correlation coefficient values due to more homogeneity. The autumn or post-monsoon image represents the strongest correlation for LST–NDVI and LST–MNDISI at any resolution levels. The image of winter season reveals the best predictability of LST distribution with the known NDVI and MNDISI values.

Keywords Land surface temperature (LST) · Normalized difference vegetation index (NDVI) · Modified normalized difference impervious surface index (MNDISI) · Landsat · Spatial resolution

H. Govil · S. Guha (✉) · P. Diwan
Department of Applied Geology, National Institute of Technology Raipur, Raipur, India
e-mail: subhanilguha@gmail.com

N. Gill
Chhattisgarh Council of Science and Technology, Raipur, India

A. Dey
Department of Geography, Nazrul Balika Vidyalaya, Guma, West Bengal, India

1 Introduction

Generally, the spatial and temporal resolution of satellite sensor should follow the nature of surface configurations [24]. In any urban area, the high value of Land Surface Temperature (LST) is generated mostly in the Impervious Surface Area (ISA) with lower vegetation intensity [18]. Therefore, most of the thermal remote sensing methods used the Normalized Difference Vegetation Index (NDVI) [5] and ISA [1] as the significant indicators of LST, and a number of studies based on the relationships of LST with NDVI [6–8, 19], and impervious surface fraction [22], were used to explore high LST. Some scholars applied fractal techniques for analyzing LST–NDVI relationship [19]. An adjusted stratified stepwise regression method was proposed to sharpen the LST within a high-density urban land and the adjacent areas [25]. An improved error estimation method was developed considering the scale difference [4]. The relationships of LST with NDVI and Modified Normalized Difference Impervious Surface Index (MNDISI) were recently investigated for a single date image at different resolutions [11] in which the relationships became strongest at 30 m resolution. In the present paper, a detailed examination was performed by using four cloudless Landsat 8 OLI (Operational Land Imager) and TIRS (Thermal Infrared Sensor) data taken from the four different seasons (spring, summer, autumn and winter) to evaluate the relationships of LST with NDVI and MNDISI at 30–960 m spatial resolutions. The main aim of the present study was to assess the LST–NDVI and LST–MNDISI relationships generated in the satellite images of different seasons at various spatial resolution using Landsat 8 data.

2 Study Area and Data

Part of Chhattisgarh State including the capital city of Raipur and its surroundings was selected for the entire research work. The total study area extends between 20°59'24" N to 21°35'24" N and 81°22'48" E to 82°01'12" E covers a total geographical area of 4356 km² (Fig. 1). The study area has an elevation ranging from 147 m to 370 m. Mahanadi River flows along the east of Raipur city and the elevated southern part is covered by forest. The study area is considered as tropical wet (Savannah) type of climate. Spring season exists for only one and half months (February–March) characterized by pleasant weather. Summer or pre-monsoon months (March–June) are hot and dry and dust storms occur frequently. July–September months are considered as monsoon or rainy season. October and November months are often considered as the autumn or post-monsoon season, characterized by an excellent climatic condition with comparatively low temperature and moderate moisture content in the air. The presence of high density of green vegetation really adds an extra flavour in Chhattisgarh during autumn. Winter months (December–January) experience a cool and dry climate. The average annual

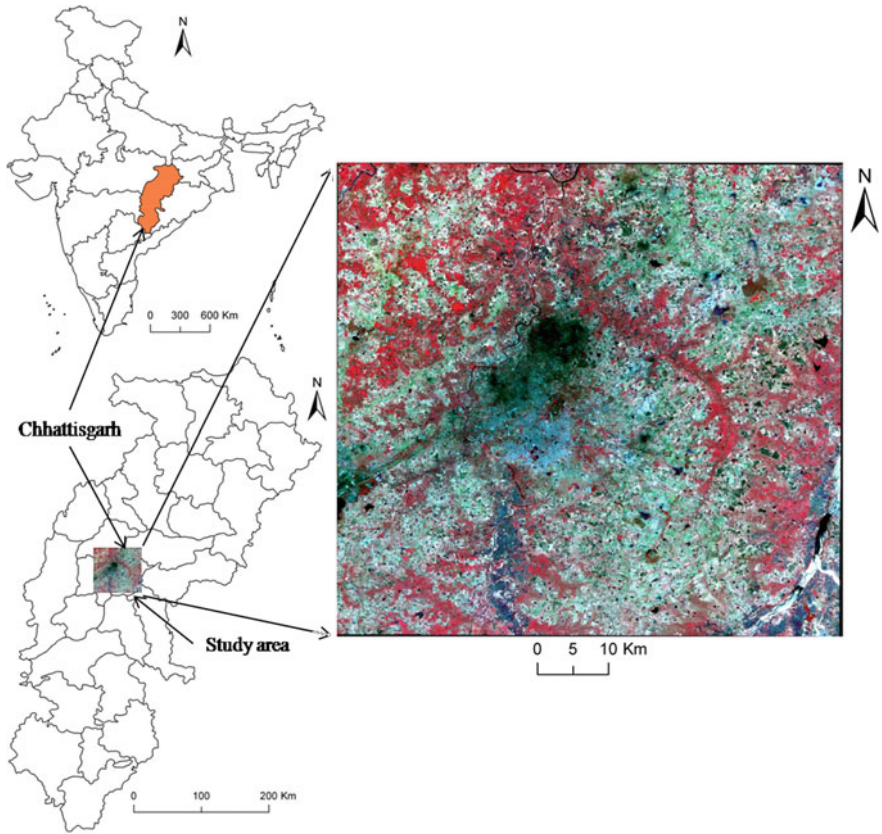


Fig. 1 Location of the study area

range of temperature is 20 °C–34 °C and average annual precipitation is 130 cm. An extensive bare land was observed throughout the study area including the capital city in the central part. The study area was also characterized by tropical mixed deciduous vegetation and red soil.

Four Landsat 8 OLI and TIRS data satellite images (Path/Row: 142/45) dated 3 February 2016, 23 April 2016, 16 October 2016 and 19 December 2016 were downloaded from the United States Geological Survey (USGS) Data Centre (Table 1) which were used as the representatives of the spring, summer or pre-monsoon, autumn or post-monsoon and winter seasons, prevails over the Chhattisgarh State of India, respectively. Landsat 8 TIRS dataset has two TIR bands (bands 10 and 11) in which band 11 has a larger calibration uncertainty. Thus, only TIR band 10 data was recommended in the present study. The TIR band 10 data was resampled to 30 m × 30 m pixel size of for all the satellite images.

Table 1 Specification of Landsat 8 OLI and TIRS satellite images

Data	Wavelength of Band 10 (μm)	Resolution of Band 10 (m)	Sun elevation ($^{\circ}$)	Sun azimuth ($^{\circ}$)	Cloud Cover (%)	Earth–Sun distance (astronomical unit–AU)	Path/row	Coordinated Universal Time (UTC)	Date of acquisition
Landsat 8	10.60–11.19	100×100	43.37	143.17	1.74	0.98	142/045	04:56:09	3 Feb 2016
Landsat 8	10.60–11.19	100×100	65.99	108.18	0.01	1.00	142/045	04:55:38	23 Apr 2016
Landsat 8	10.60–11.19	100×100	53.59	145.43	0.01	0.99	142/045	04:56:27	16 Oct 2016
Landsat 8	10.60–11.19	100×100	39.72	152.38	0.21	0.98	142/045	04:56:22	19 Dec 2016

Table 2 Various indices used for the extraction of Land Use/Land Cover (LU-LC) features

Indices	Formula	Significance in LU-LC Identification	References
NDVI	$(\text{NIR}-\text{Red})/(\text{NIR} + \text{Red})$	Green vegetation extraction	[13]
MNDISI	$[\text{TIR}-(\text{MNDWI} + \text{NIR} + \text{SWIR1})/3]/[\text{TIR} + (\text{MNDWI} + \text{NIR} + \text{SWIR1})/3]$	Impervious surface area extraction	[21]
MNDWI	$(\text{Green}-\text{SWIR1})/(\text{Green} + \text{SWIR1})$	Waterbody extraction	[23]

3 Methodology

3.1 Retrieving LST from Landsat 8 OLI Data

In this study, the mono-window algorithm was applied to retrieve LST from multi-temporal Landsat satellite image [6, 7, 14]. Ground emissivity, atmospheric transmittance and effective mean atmospheric temperature—these three parameters are needed to derive the LST using mono-window algorithm.

3.2 Extraction of NDVI and MNDISI

Various land surface biophysical parameters were applied to specify different types of land surface features [6, 7]. In this study, special emphasis was given on NDVI [9, 13] and MNDISI [10, 21] for determining the relationships with LST. NDVI was used as a vegetation index while MNDISI, based on modified normalized difference water index or MNDWI [23], acts as an index of impervious surface. These remote sensing indices were extracted by using the following formulas (Table 2).

4 Results and Discussion

4.1 Characteristics of LST, NDVI and MNDISI at Different Spatial Resolutions

There was a prominent variation of different time periods occurred in mean and Standard Deviation (STD) values of LST, NDVI and MNDISI under the different resolutions ranging from 30 to 960 m. The summer or pre-monsoon image had the maximum values of mean LST (44.26 °C) followed by spring image (27.37 °C), autumn or post-monsoon image (26.91 °C) and winter image (24.14 °C); and these

mean values were almost unaffected by the size of the pixel. But, Standard Deviation (STD) values of LST had been gradually decreasing with the increase of pixel size for each and every image. Again, the summer image had the maximum STD of LST (2.92 °C at 30 m resolution) followed by autumn image (2.09 °C at 30 m resolution), spring image (1.75 °C at 30 m resolution) and winter image (1.46 °C at 30 m resolution) and these STD values maintained a steady decreasing trend with the increase of pixel size, e.g., 2.24 °C, 1.64 °C, 1.34 °C and 1.02 °C, respectively for summer image, autumn image, spring image and winter image at 960 m resolution.

In the case of NDVI, the autumn image had the maximum value of mean NDVI (0.34) due to the abundance of green vegetation followed by winter image (0.14), summer image (0.14) and spring image (0.14) and these mean values remain almost unchanged at different spatial resolutions. But, STD values of NDVI was continuously decreasing as the spatial resolution becomes low and this phenomenon was noticed for the four multi-date images. The autumn image represented the maximum STD value of NDVI, both at 30 m resolution (0.10) and at 960 m resolution (0.07).

There was a very negligible variation in the multi-date images for mean MNDISI value at any particular spatial resolution. At 30 m resolution, the spring image (0.55) had the maximum values of mean MNDISI followed by winter image (0.54) because of the increase in imperviousness in soil due to lack of rainfall and vegetation. The summer image (0.53) had less value of mean MNDISI than spring image and winter image while it was least in the case of autumn image (0.52) because of the high percentage of soil moisture. STD values of MNDISI had a steady descending trend with the decrease of resolution or increase of pixel size and it was a common phenomenon for all the images. The winter image had the highest STD values of MNDISI at any resolution (0.05 at 30 m to 0.03 at 960 m resolution). But, for the rest of the three images, a dynamic effect was observed with the change of resolution. Up to 300 m resolution, the autumn image had greater STD values of MNDISI than summer and spring image while 360 m resolution onwards these STD values of MNDISI become lesser.

Figure 2 represented the spatial distribution pattern of LST, NDVI and MNDISI in the highest (30 m) resolution level for all the multi-date images. It was clear that in the spring image, the higher LST values generated over the western, central, northern and southern portions of the study area while these specific zones were simultaneously characterized by low NDVI and high MNDISI value. The figure also indicated that in the summer image, the lower LST values were found in the southern and the southeastern parts and the higher LST values were found in the northern portion of the study area. The NDVI values were normally inversely correlated with the LST values while MNDISI values were not very much significantly correlated to the LST values. In the autumn image, the central and western parts are characterized by high LST, low NDVI and high MNDISI values. The

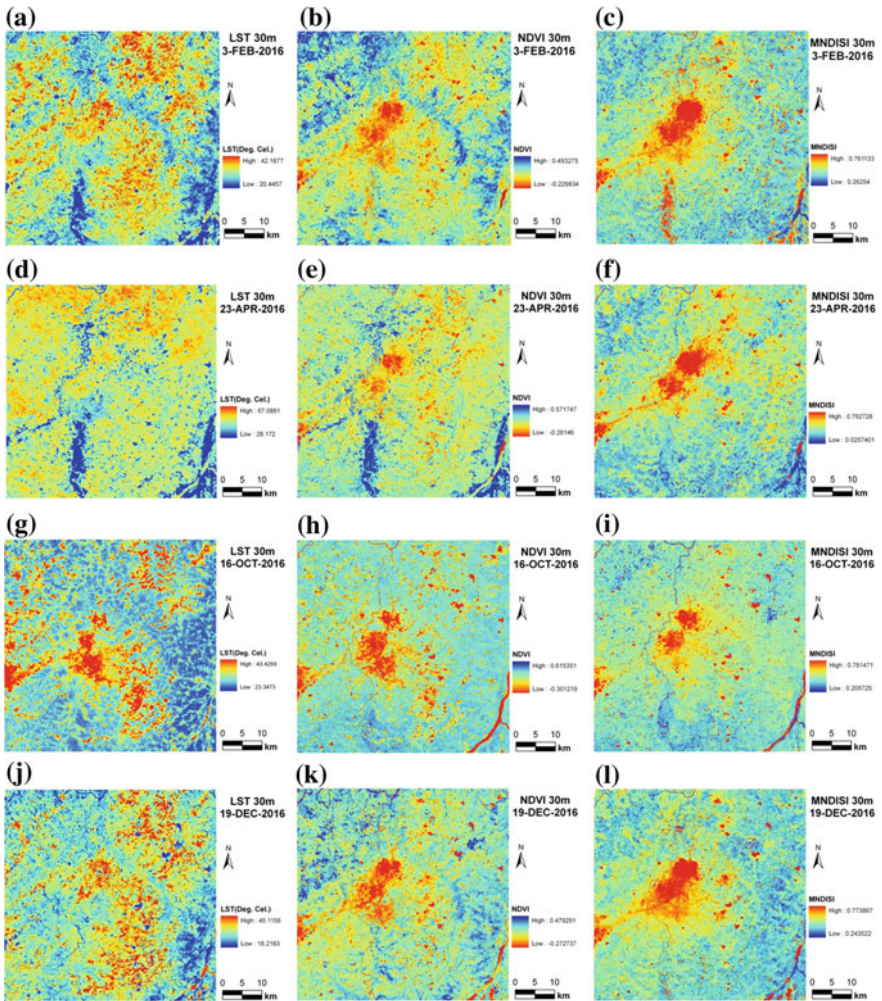


Fig. 2 Spatial distribution of LST, NDVI and MNDISI at 30 m resolution for spring image (a–c), summer image (d–f), autumn image (g–i) and winter image (j–l)

results remain almost similar at 30 m and 960 m resolution levels. In the winter image, no such evidence of significant changes was found due to the pixel size. The central part was characterized by high LST, low NDVI and high MNDISI values. Conversely, the northwest and the southeast portions had low LST, high NDVI and low MNDISI values.

4.2 Validation of Derived LST from Landsat 8 Data with Respect to MODIS Data

In the present study, MODIS data (MOD11A1 of 1 km resolution) was applied for the validation of LST retrieved from the aggregated Landsat 8 data of 960 m. Here, due to some missing pixels, MODIS data of exact corresponding date were not matched for LST validation process. MOD11A1 data (daily level 3 LST product) of the just preceding and the just following dates of Landsat 8 data were used in this study. The average MODIS LST values of the aforementioned two dates were taken for the final validation. No rainfall or strong winds occurred in between the acquisition of the Landsat 8 and MODIS imageries and almost similar weather conditions were noticed for the used Landsat 8 data and the corresponding MODIS data. Table 3 presented a strong correlation of retrieved LST from aggregated Landsat 8 data (960 m) and the corresponding MOD11A1 data (1 km) for multi-date satellite images.

4.3 Relationships Between LST–NDVI and LST–MNDISI Under Different Spatial Resolutions

The correlation coefficients between LST–NDVI and LST–MNDISI for various levels of spatial resolution (30–960 m) in four multi-date imageries were determined through pixel-by-pixel linear regression analysis. The 30 m LST, NDVI, and MNDISI were aggregated to various resolution levels (60–960 m with 60 m intervals). The Pearson’s correlation coefficients (significant at 0.001 level using one-tailed Student’s *t* test) between LST–NDVI and between LST–MNDISI under different resolution levels and different dates were clearly understood from Table 4. The LST values had developed a negative correlation with the NDVI values at all resolution levels for all images. The three highest negative correlations were found in between 840 and 960 m resolution levels and the three lowest negative correlations were found in between 30 and 120 m resolution levels for all multi-date images. A steady and gradual increasing trend of negativity was observed in LST–NDVI correlation analysis with the decrease in resolution levels, i.e., weaker correlations developed at the higher or finer resolutions while stronger correlations developed at the lower or coarser resolutions and it appeared for all images. The

Table 3 Validation of LST (°C) retrieved from Landsat 8 data (960 m resolution) with corresponding MODIS data

	3 February 2016	23 April 2016	16 October 2016	19 December 2016
Correlation coefficient	0.74	0.79	0.85	0.73

Table 4 Correlation coefficients between LST–NDVI and between LST–MNDISI

Pixel size (m)	Number of pixels	3-FEB-2016		23-APR-2016		16-OCT-2016		19-DEC-2016	
		Correlation coefficients		Correlation coefficients		Correlation coefficients		Correlation coefficients	
		LST and NDVI	LST and MNDISI	LST and NDVI	LST and MNDISI	LST and NDVI	LST and MNDISI	LST and NDVI	LST and MNDISI
30	4840000	-0.36	-0.12	-0.50	-0.02	-0.59	0.12	-0.16	-0.14
60	121000	-0.38	-0.10	-0.53	-0.01	-0.60	0.13	-0.18	-0.12
120	303520	-0.41	-0.09	-0.58	0.01	-0.64	0.16	-0.20	-0.11
180	134658	-0.43	-0.08	-0.59	0.03	-0.66	0.18	-0.22	-0.09
240	76158	-0.44	-0.07	-0.60	0.05	-0.68	0.20	-0.24	-0.08
300	48830	-0.45	-0.06	-0.61	0.07	-0.69	0.22	-0.26	-0.06
360	33850	-0.46	-0.05	-0.62	0.09	-0.70	0.24	-0.27	-0.04
420	24960	-0.46	-0.04	-0.63	0.10	-0.71	0.25	-0.28	-0.03
480	19041	-0.46	-0.03	-0.64	0.11	-0.72	0.27	-0.29	-0.02
540	15127	-0.47	-0.03	-0.64	0.12	-0.72	0.28	-0.30	-0.01
600	12319	-0.48	-0.02	-0.65	0.14	-0.73	0.29	-0.31	0.01
660	10200	-0.48	-0.02	-0.65	0.14	-0.73	0.30	-0.32	0.01
720	8464	-0.48	-0.01	-0.66	0.15	-0.74	0.31	-0.32	0.02
780	7225	-0.49	-0.01	-0.66	0.16	-0.74	0.32	-0.34	0.04
840	6241	-0.49	0.01	-0.67	0.16	-0.75	0.34	-0.35	0.05
900	5476	-0.49	0.01	-0.67	0.17	-0.75	0.34	-0.35	0.06
960	4761	-0.49	0.01	-0.67	0.17	-0.75	0.35	-0.35	0.06

strongest negative correlation (correlation coefficient values ranging from -0.59 to -0.75) was observed in the autumn image. The summer image also had a strong negative correlation (-0.50 to -0.67). A moderate negative correlation (-0.36 to -0.49) appeared in spring image while winter image experienced a low to moderate negative correlation (-0.16 to -0.35). The scenario became slightly different in LST–MNDISI correlation analysis (Table 4). The spring image (-0.12 to -0.01) revealed a very weak negative correlation between LST and MNDISI from 30 m to 780 m resolution while only the three coarser resolutions (840–960 m) had a very weak positive correlation. In the winter image (-0.14 to 0.06), up to 540 m spatial resolution, the correlation was very weak negative and it became very weak positive from 600 to 960 m resolutions. The summer image (-0.02 to 0.17) indicated very weak to weak positive correlation between LST and MNDISI except for 30 m and 60 m resolutions (very weak negative correlation). But, only the autumn image (0.12–0.35) holds a weak to moderate positive correlation between LST–MNDISI throughout the entire range of the resolution levels (30–960 m). The three highest positive correlations between LST and MNDISI were found between 840 m and 960 m resolution levels while the three lowest correlation values were found in between 30 and 120 m resolution levels for all the images. A constant increasing trend of positive correlation is observed between LST and MNDISI with the

decrease in resolution level and it occurs for each and every image (Table 4). Therefore, the correlations between LST–NDVI and between LST–MNDISI were best observed at 840–960 m resolution levels or at the coarser resolution levels due to the presence of more homogeneity in landscape features.

The stronger correlation between LST–NDVI and between LST–MNDISI at coarser resolution levels supports the relevance of linear correlation in determining LST with the known values of NDVI and MNDISI. The low density of vegetation and high density of impervious surface may together produce a high range of LST. The correlation graphs (significant at 0.05 levels) between LST–NDVI and between LST–MNDISI at various resolution levels and for all the images were presented in Fig. 3. Table 5 revealed the regression statistics of the correlation coefficients of LST–NDVI and LST–MNDISI relationship with respect to various resolution levels. In the correlation analysis between the correlation coefficients of LST–NDVI relationship and the various resolution levels, the values of coefficient of multiple

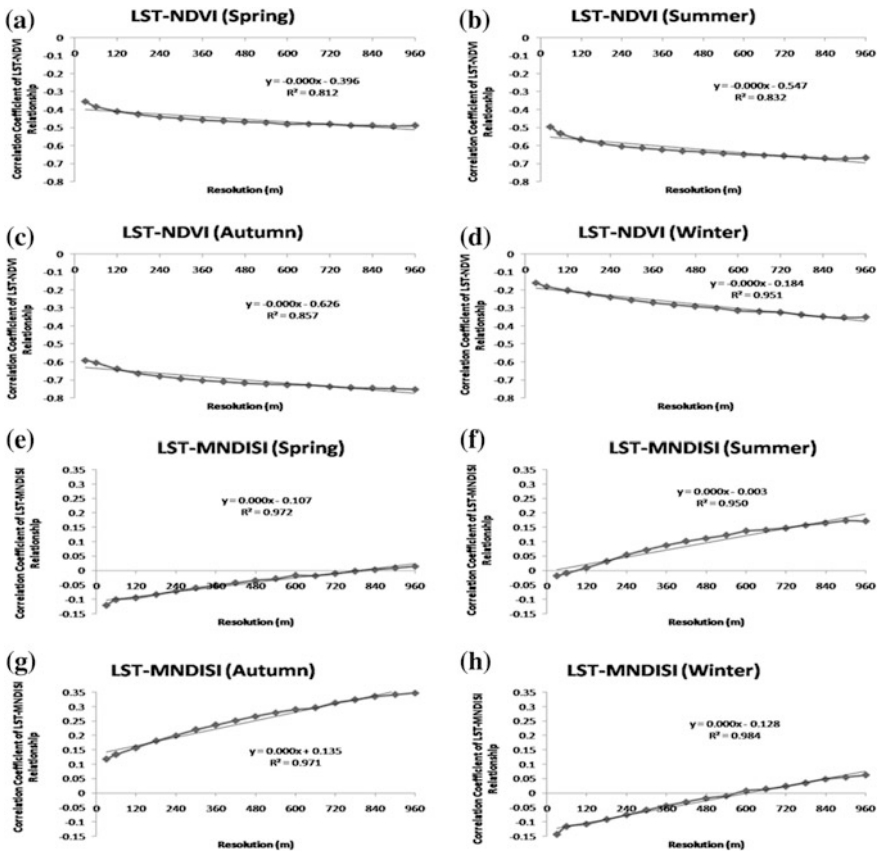


Fig. 3 Seasonal variation on LST–NDVI (a–d) and LST–MNDISI (e–h) correlation coefficients for various spatial resolutions

Table 5 Regression statistics of correlation coefficients of LST–NDVI and LST–MNDISI relationships with respect to different resolution levels for all images

	LST–NDVI correlation coefficients with various resolution levels (30–960 m)				LST–MNDISI correlation coefficients with various resolution levels (30–960 m)			
	Spring image	Summer image	Autumn image	Winter image	Spring image	Summer image	Autumn image	Winter image
Intercept	−0.40	−0.55	−0.63	−0.18	−0.11	−0.01	0.14	−0.13
Slope	−0.01	−0.01	−0.01	−0.01	0.01	0.01	0.01	0.01
Multiple R	−0.90	−0.91	−0.93	−0.98	0.99	0.98	0.99	0.99
R-squared	0.81	0.83	0.86	0.95	0.97	0.95	0.97	0.98
Adjusted R-squared	0.79	0.81	0.84	0.94	0.97	0.94	0.97	0.98
Standard error	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01

determination or R-squared values are 0.81, 0.83, 0.86 and 0.95 for the spring image, summer image, autumn image and winter image, respectively while in the correlation analysis between the correlation coefficients of LST–MNDISI relationship and the various resolution levels, these R-squared values became 0.97, 0.95, 0.97 and 0.98 for the spring image, summer image, autumn image and winter image, respectively (Table 5). These R-squared values implicit a very strong correlation between the correlation coefficient values and the resolution levels for all the four multi-date satellite data. Hence, it can be concluded that 81.21%, 83.26%, 85.78% and 95.14% of the variance in the correlation coefficient values of LST–NDVI were predictable from the resolution levels for the images of the spring, summer, autumn and winter, respectively and 97.22%, 95.07%, 97.11% and 98.40% of the variance in the correlation coefficient values of LST–MNDISI were predictable from the resolution levels for the images of the spring, summer, autumn and winter, respectively. On the basis of multiple R, R-squared and adjusted R-squared values (Table 5), the winter image estimated the strongest correlation and may be considered as the best image under any possible resolution levels for the predictability of LST distribution with the known NDVI and MNDISI values. The overall results indicated that the NDVI and the MNDISI may be accepted as influential factors for LST distribution.

5 Conclusion

In this study, LST, NDVI and MNDISI were computed with 30–960 m resolutions for four multi-date imageries by using Landsat 8 OLI and TIRS data in parts of Chhattisgarh State, India. The linear relationships of LST–NDVI and LST–MNDISI varied with various resolution levels but yielded the higher correlation coefficient values around the coarser resolution (840–960 m) levels for all the four

images due to more homogeneity. For each data, the LST values tended to be negatively (weak to strong) correlated with the NDVI values and showed weak negative to moderate positive correlation with the MNDISI values. Autumn image (16 October 2016) reflected the strongest correlation followed by summer image while spring and winter images experienced weaker correlation for both the relationships. Moreover, a very strong negative correlation was also estimated between LST–NDVI correlation coefficients and various resolution levels. But, LST–MNDISI correlation coefficients and various resolution levels generated a very strong positive correlation. These strong relationships were sustained for all four different imageries. Amongst the four imageries, the winter image possessed the strongest correlation and ensured the best prediction level of LST distribution when NDVI and MNDISI values were known. NDVI and MNDISI may also be considered as significant variables for the prediction of LST distribution.

The study further recommends that these results may be investigated by the inclusion of other satellite imageries of low or high resolution and other LU-LC indices. Any nonlinear or nonparametric regression models may also be evaluated. A new study area may be selected with more homogeneous or heterogeneous landscape features. An important task of the present study was to observe the variation of correlation analysis for LST–NDVI and LST–MNDISI relationships in multi-date imageries. Hence, areas having extreme climatic conditions (hot, cold, wet or dry) may also be recommended for the future examination.

Acknowledgements The authors are indebted to the United States Geological Survey (USGS).

Disclosure Statement No potential conflict of interest was reported by the authors.

References

1. Arnold, C.L., Gibbons, C.J.: Impervious surface coverage—the emergence of a key environmental indicator. *J. Am. Plan. Assoc.* **62**(2), 243–258 (1996). <https://doi.org/10.1080/01944369608975688>
2. Carlson, T.N., Ripley, D.A.: On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sens. Environ.* **62**, 241–252 (1997)
3. Chen, X.L., Zhao, H.M., Li, P.X., Yi, Z.Y.: Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. *Remote Sens. Environ.* **104**(2), 133–146 (2006)
4. Chen, X., Li, W., Chen, J., Zhan, W., Rao, Y.: A simple error estimation method for linear-regression-based thermal sharpening techniques with the consideration of scale difference. *Geo-spat. Inf. Sci.* **17**(1), 54–59 (2014). <https://doi.org/10.1080/10095020.2014.889546>
5. Goward, S.N., Xue, Y.K., Czajkowski, K.P.: Evaluating land surface moisture conditions from the remotely sensed temperature/vegetation index measurements: an exploration with the simplified simple biosphere model. *Remote Sens. Environ.* **79**, 225–242 (2002). [https://doi.org/10.1016/S0034-4257\(01\)00275-9](https://doi.org/10.1016/S0034-4257(01)00275-9)

6. Guha, S., Govil, H., Dey, A., Gill, N.: Analytical study of land surface temperature with NDVI and NDBI using Landsat 8 OLI and TIRS data in Florence and Naples city, Italy. *Eur. J. Remote Sens.* **51**(1), 667–678 (2018). <https://doi.org/10.1080/22797254.2018.1474494>
7. Guha, S., Govil, H., Mukherjee, S.: Dynamic analysis and ecological evaluation of urban heat islands in Raipur city, India. *J. Appl. Remote Sens.* **11**(3), 036020 (2017). <https://doi.org/10.1117/1.JRS.11.036020>
8. Gutman, G., Ignatov, A.: The derivation of the green vegetation fraction from NOAA/AVHRR data for use in numerical weather prediction models. *Int. J. Remote Sens.* **19**(8), 1533–1543 (1998). <https://doi.org/10.1080/014311698215333>
9. Ke, Y.H., Im, J., Lee, J., Gong, H.L., Ryu, Y.: Characteristics of landsat 8 oli-derived NDVI by comparison with multiple satellite sensors and in-situ observations. *Remote Sens. Environ.* **164**, 298–313 (2015). <https://doi.org/10.1016/j.rse.2015.04.004>
10. Liu, C., Shao, Z., Chen, M., Luo, H.: MNDISI: a multi-source composition index for impervious surface area estimation at the individual city scale. *Remote Sens. Lett.* **4**(8), 803–812 (2013). <https://doi.org/10.1080/2150704X.2013.798710>
11. Mao, W., Wang, X., Cai, J., Zhu, M.: Multidimensional histogram-based information capacity analysis of urban heat island effect using Landsat 8 data. *Remote Sens. Lett.* **7**(10), 925–934 (2016). <https://doi.org/10.1080/2150704X.2016.1182656>
12. Markham, B.L., Barker, J.K.: Spectral characteristics of the LANDSAT thematic mapper sensors. *Int. J. Remote Sens.* **6**(5), 697–716 (1985)
13. Purevdorj, T.S., Tateishi, R., Ishiyama, T., Honda, Y.: Relationships between percent vegetation cover and vegetation indices. *Int. J. Remote Sens.* **19**, 3519–3535 (1998)
14. Qin, Z., Karnieli, A., Barliner, P.: A mono-window algorithm for retrieving land surface temperature from landsat TM data and its application to the Israel-Egypt border region. *Int. J. Remote Sens.* **22**(18), 3719–3746 (2001). <https://doi.org/10.1080/01431160010006971>
15. Sobrino, J.A., Raissouni, N., Li, Z.: A comparative study of land surface emissivity retrieval from NOAA data. *Remote Sens. Environ.* **75**(2), 256–266 (2001)
16. Sobrino, J.A., Jimenez-Munoz, J.C., Paolini, L.: Land surface temperature retrieval from Landsat TM5. *Remote Sens. Environ.* **9**, 434–440 (2004). <https://doi.org/10.1016/j.rse.2004.02.003>
17. Sun, Q., Tan, J., Xu, Y.: An ERDAS image processing method for retrieving LST and describing urban heat evolution: a case study in the Pearl River Delta Region in South China. *Environ. Earth Sci.* **59**, 1047–1055 (2010)
18. Voogt, J.A., Oke, T.R.: Thermal remote sensing of urban climates. *Remote Sens. Environ.* **86**, 370–384 (2003). [https://doi.org/10.1016/S0034-4257\(03\)00079-8](https://doi.org/10.1016/S0034-4257(03)00079-8)
19. Weng, Q.H., Lu, D.S., Schubring, J.: Estimation of land surface temperature-vegetation abundance relationship for urban heat island studies. *Remote Sens. Environ.* **89**, 467–483 (2004). <https://doi.org/10.1016/j.rse.2003.11.005>
20. Wukelic, G.E., Gibbons, D.E., Martucci, L.M., Foote, H.P.: Radiometric calibration of Landsat Thematic Mapper thermal band. *Remote Sens. Environ.* **28**, 339–347 (1989)
21. Xu, H.Q.: A new remote sensing index for fastly extracting impervious surface information. *Geomat. Inf. Sci. Wuhan Univ.* **11**, 1150–1153 (2008). <https://doi.org/10.13203/j.whugis2008.11.024>
22. Xu, H.Q., Lin, D.F., Tang, F.: The impact of impervious surface development on land surface temperature in a subtropical city: Xiamen, China. *Int. J. Climatol.* **33**(11), 1873–1883 (2013). <https://doi.org/10.1002/joc.3554>
23. Xu, H.X.: A study on information extraction of water body with the modified normalized difference water index (MNDWI). *J. Remote Sens.* **9**, 589–595 (2005)

24. Zhang, H.K., Huang, B., Zhang, M., Cao, K., Yu, L.: A generalization of spatial and temporal fusion methods for remotely sensed surface parameters. *Int. J. Remote Sens.* **36**(17), 4411–4445 (2015). <https://doi.org/10.1080/01431161.2015.1083633>
25. Zhu, S., Guan, H., Millington, A.C., Zhang, G.: Disaggregation of land surface temperature over a heterogeneous urban and surrounding suburban area: a case study in Shanghai, China. *Int. J. Remote Sens.* **34**(5), 1707–1723 (2013). <https://doi.org/10.1080/01431161.2012.725957>

Automatic Robot Processing Using Speech Recognition System



S. Elavarasi and G. Suseendran

Abstract Nowadays, speech recognition is becoming a more useful technology in computer applications. Many interactive speech-aware applications exist in the field. In order to use this kind of easy way of communication technique into the computer field, speech recognition technique has to be evolved. The computer has to be programmed to accept the voice input and then process it to provide the required output, using various speech recognition software. Speech recognition is the process of converting speech signal to a sequence of words using appropriate algorithm. This provides an alternative and efficient way for the people who are not well educated or not having sufficient computer knowledge to access the systems and where typing becomes difficult. This speech recognition technique also reduces the manpower to accept and process the commands. In our research work, we have to implement this speech recognition technique in customer care center, where many queries have to be processed every day. Some of the queries are repeated often and the responses also seem to be the same. In such cases, we have to propose a methodology to automate the query-processing activities using this speech recognition technique. The ways of how to automate the system and how to process the queries automatically are explained in our methodology with suitable algorithm.

Keywords Feature extraction · Speech modeling · Speech automation · Robot processing · Automate query processing · Interactive speech-aware applications

S. Elavarasi (✉) · G. Suseendran

Department of Information and Technology, School of Computing Sciences, Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, India
e-mail: elavarasi_msc@yahoo.co.in

G. Suseendran

e-mail: suseendar_1234@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_14

185

1 Introduction

In this decade, people want to make all the actions to be carried out in a much easier and faster way. In order to accomplish this, most of the things have to be modernized and computerized to perform the action efficiently. As the new trends emerge in each and every field, things have to be changed to accommodate the trend. In such a way, many new technologies have emerged in the computer field also. One such technology is *speech recognition* or *voice recognition*.

Speech recognition is a technology that has emerged in order to make easier the process of providing the requirements to the system by the user through speech rather than text. That is, speech recognition accepts the user's spoken words as input and then translates the words to appropriate text or commands to accomplish the task. It is also called as a software to accept and understand dictation to undertake the commands. While accepting the voice input, the system also gathers the information such as gender, expression or emotion and in some cases, the identity of the speaker. Thus, speech recognition converts the speech to text, and so it is also termed as *automatic speech recognition* or *ASR* or *speech to text*. This is also a potential for human-computer interaction.

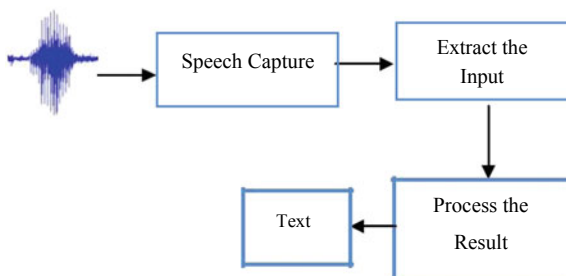
Speech recognition is mostly used in many areas, since it is more useful for those who are not well educated or not having sufficient computer knowledge. The essential requirement for the user is to have sufficient language skills and proper pronunciation efficiency such that the user is able to process the commands. The major research challenges in this area are noise, speaker variation, language variation and vocabulary (Fig. 1).

While designing the speech recognition system, the researcher requires providing special attention toward the challenges:

- Speech module and illustration
- Speech pre-processing
- Feature extraction
- Database maintenance
- Performance evaluation.

In recent years, the enhancement and effective progress of speech recognition technology have been implemented in many areas to simplify the process of user requirements and to improve the performance. Some of the applications are:

Fig. 1 Speech to text conversion



- Aerospace
- Automatic subtitling
- Hands-free computing
- Robotics
- Video games and so on.

In our research work, we have to implement the speech recognition technology in *customer care service* area to improve the customer satisfaction and experience some pre-defined operations with reduced manual efficiency. The major premise is to promote better customer service and to improve customer relationships.

Customer care service is the act of taking care of customer requirements by offering and delivering high-quality service in a short period of time. In order to provide such kind of response, more manual power has to be empowered. To avoid this dilemma, we have to provide suitable methodology to accomplish the process of responding the customer requirement through speech recognition technology automatically. This paper describes the development of an efficient speech recognition system over customer care service using various technologies.

2 Related Works

Gupta et al. discussed that with development in the requirements for installed figuring and the interest for rising implanted stages, it is necessary that the discourse acknowledgment frameworks (SRS) are accessible on them as well. PDAs and other handheld gadgets are ending up increasingly ground-breaking and reasonable also. It has turned out to be conceivable to run interactive media on these gadgets. In that paper, different methods about discourse acknowledgment framework were discussed. Likewise, it displayed the rundown of procedure with their properties of feature extraction and feature coordinating. Through this audit paper, it is discovered that MFCC is generally used to include extraction and VQ is better over DTW [1].

Singh et al. expressed that using fake neural systems (ANNs), numerical models of the low-level circuits in the human mind, to enhance discourse acknowledgment execution, through a model known as the ANN-hidden Markov model (ANNHMM) have indicated guarantee for expansive vocabulary discourse acknowledgment frameworks. Accomplishing higher recognition precision, low word mistake rate, creating discourse corpus relying on the idea of dialect and tending to the issues of wellsprings of fluctuation through methodologies like missing data techniques and convolutive non-negative matrix factorization are the significant contemplations for building up an effective ASR. In this paper, an exertion has been gained to feature the ground made so far for ASRs of various dialects and the mechanical viewpoint of programmed discourse acknowledgment in nations like China, Russian, Portuguese, Spain, Saudi Arab, Vietnam, Japan, UK, Sri Lanka, Philippines, Algeria and India [2].

Arora et al. endeavored to portray a writing audit of automatic speech recognition. It talked about past years' propels gained in order to give ground that has been

refined here of research. One of the vital difficulties for scientists is ASR precision. The speech acknowledgment system centers around troubles with ASR, essential building squares of discourse preparing, highlight extraction, discourse acknowledgment and execution assessment. The primary target of the survey paper is to expose the advancement made in ASRs of various dialects and the mechanical perspective of ASR in various nations and to thoroughly analyze the procedures used in different phases of speech acknowledgment and recognize the examined subject in this testing field. They are not displaying comprehensive depictions of frameworks or numerical definitions, yet rather, they are introducing unmistakable and novel highlights of chosen frameworks and their relative benefits and demerits [3].

Rashmi conveyed a review of various calculations that can be used in discourse acknowledgment in light of the points of interest and disservices. Likewise, it helps in picking the better calculation in light of the examination done [4].

Gamit et al. depicted that speech acknowledgment is the procedure of consequently perceiving the talked expressions of individual in view of the data content in discourse flag. This paper presents a concise overview on automatic speech recognition and examines the different characterization procedures that have been refined in this wide region of discourse handling. The target of this survey paper is to outline a portion of the notable strategies that are broadly used in a few phases of discourse acknowledgment system [5].

Shaikh Naziya et al. examined about speech advances that are limitlessly used and have boundless employments. These advances empower machines to react accurately depending on human voices, and give helpful and profitable administrations. The paper gave a diagram of the discourse acknowledgment process, its fundamental model, its application and approaches, and furthermore, examined near investigation of various methodologies that are used for discourse acknowledgment framework. The paper additionally gives a review of various strategies of discourse acknowledgment framework and furthermore demonstrates the rundown portion of the notable techniques used in different phases of discourse acknowledgment system [6].

Navneet et al. introduced the programmed discourse acknowledgment framework and examined the significant subjects and advances made in the previous 60 long stretches of research, in order to give a mechanical point of view and energy about the principal advance that has been proficient in this vital region of discourse correspondence. Following quite a while of innovative work, the exactness of programmed discourse acknowledgment stays as one of the imperative research challenges. The outline of speech recognition framework requires watchful considerations to the accompanying: definition of different sorts of discourse classes, discourse acknowledgment process, ASR configuration issues and discourse acknowledgment strategies. The target of this audit paper is to outline and look at a portion of the notable techniques used in diverse periods of talk affirmation system and perceive an investigation on subject and applications which are at the front line of this stimulating and testing field [7].

Lawrence et al. looked into some of the significant features in the innovative work of programmed discourse acknowledgment amid the most recent couple of decades to give a mechanical point of view and a valuation for the principal advance that has been made in this imperative region of data and correspondence technology [8].

Prabhakar et al. gave a description of major inventive perspective and valuation for the vital progression of talk affirmation; gave graph technique made in each period of talk affirmation, besides condense; took a gander at changed talk affirmation systems; and perceived an investigation on subjects and applications which are at the front line of this stimulating and testing field [9].

Bhavneet Kaur gave a concise presentation of SRS portraying how the innovation functions, and after that, talks about the general engineering of SRS, strategies, benefits of using this framework, major mechanical point of view and valuation for the principal advancement of discourse acknowledgment. It gives a way to deal with the acknowledgment of discourse flag using recurrence ghashly data with Mel recurrence for the change of discourse that includes portrayal in a HMM-based acknowledgment approach, and furthermore, gives a review of strategies created in each phase of discourse acknowledgment alongside the momentum and explores future options on the same. This paper depicts the real difficulties for SRS framework which have been went over by clients' criticism and different examinations, which must be settled as quickly as time permits for better execution outcome [10].

Swati Atame et al. discussed that one of the technologies used in these fields is automatic singer identification, which is used to recognize from features of the audio signal, the singer of the song or who is the one who is singing, and the genuine singer [5]. This same area of singer identification and recognition can be used in the bioinformatics whereby the voice of the singer is used to gain access to a particular singer. The system would be very much useful to singers of the song who are actual singers and also to the common man who can store their speech and can later use this input signal to access the system. For the professional singers, there are many possibilities that the original singers' voice is get mimicked which may in this situation lead to pirated copies of the voice of the singer which are then sold into the market [11].

Hori et al. exhibited strategies for discourse-to-content and speech-to-discourse programmed synopsis in light of discourse unit extraction and connection. For the previous case, a two organized outline strategies comprising essential sentence extraction and word-based sentence compaction are examined. Sentence and word units which augment the weighted total of etymological probability, measure of data, certainty measure and linguistic probability of linked units are separated from the discourse acknowledgment results and connected for star ducing outlines. For the last case, sentences, words and between-filler units are researched as units to be removed from unique discourse.

Renals et al. portrayed the advancement of a framework to translate and outline voice messages. The consequences of the exploration introduced in this paper are two overlays. Initially, a cross-breed connectionist way to deal with the voicemail interpretation undertaking demonstrates that focused execution can be accomplished using a setting autonomous framework with less parameters than those in view of blends of Gaussian probabilities. Second, a successful and vigorous blend of factual with earlier learning hotspots for term weighting is used to remove data from the decoder's yield keeping in mind the end goal to convey rundowns to the message beneficiaries by means of a GSM short message service (SMS) entryway [12].

3 Proposed Methodology

3.1 Proposed Method

Speech recognition has been the most explored point since mid 1960s and is a standout among the most famous and dynamic territory of research. In speech recognition, there exists lot of fields for research such as:

- Speech recognition
- Speaker recognition
- Speech conversion
- Feature extraction
- Noise reduction.

Speech recognition is used in many applications since it does not need any syntax, procedure or coding to access the commands. It just simply accepts the spoken words as input and then translates it into appropriate text to perform the task. Thus speech recognition technology has unlimited users with exciting range of tasks. The main theme of the speech recognition technology is to “*listen*”, “*identify*”, “*understand*” and “*respond*” to the spoken information. It has the potential to act like an interface between the humans and computer, that is, human–computer interface.

In our research work, we choose this speech recognition technique to be implemented in a customer care organization in order to provide better performance in a short period of time. In customer care organization, they have the task of providing response for the queries submitted by the user. The queries may be of different types such as account details, server problem, server restart, password reset, account lock, account balance, account transaction and last transaction.

In some situations, the customer may submit the queries continuously, for which the executive provides response for it. Some of the queries may be repeated often that the executive must be processed every time. In this case, it requires huge manual power with 24×7 services and high cost.

In order to overcome these circumstances, we have to propose a methodology to automate the process of providing responses for the queries with some pre-defined activities. This has to be carried out by programming the customer care server with some set of instructions with appropriate keywords in the database. The methodology can be executed as follows:

When the customer calls for sending query, the server accepts the query as voice input and records it. Upon receiving the input, the server converts the voice input into text and then starts to search in the database for the keyword found in the text. As a result of this database search, the appropriate instructions for the query have been found from the database, and the response to it is sent to the customer. Thus, the query has been automatically processed by the speech recognition technology like ROBOT processing, and hence our methodology can also be termed as automatic ROBOT processing (ARP) (Fig. 2).

In case the customer needs to change his/her password means, it comes under the category “*Password Reset*”. The customer first calls the customer care service and asks for password reset through voice command as: “*I Need Password Reset*”. This command has been recorded by the server and then processed to convert into text. Then the server starts to search from the database for the keyword “*Password Reset*”. In database, there exist several instructions for the password reset category that are to be programmed by the programmer to process the request in an appropriate manner.

In that database, if there exists as “*Please tell your Username*” and “*Please tell your Date of Birth*” for “password reset” category, then the server gets the instructions and converts it into speech and then transfers to the customer as to submit their username and date of birth through speech. Upon receiving the details from the customer, the server starts converting the details into text from speech and then validates the details. Only if the information provided by the customer matches with the database, the server allows the customer to reset their password. Otherwise, he/she has to be denied (Fig. 3).

Thus, the small pre-determined activities in the customer care service have been automated with the speech recognition technique to simplify the process and to reduce the manual power. It also reduces the cost of processing the pre-determined

Fig. 2 Processing the customer query

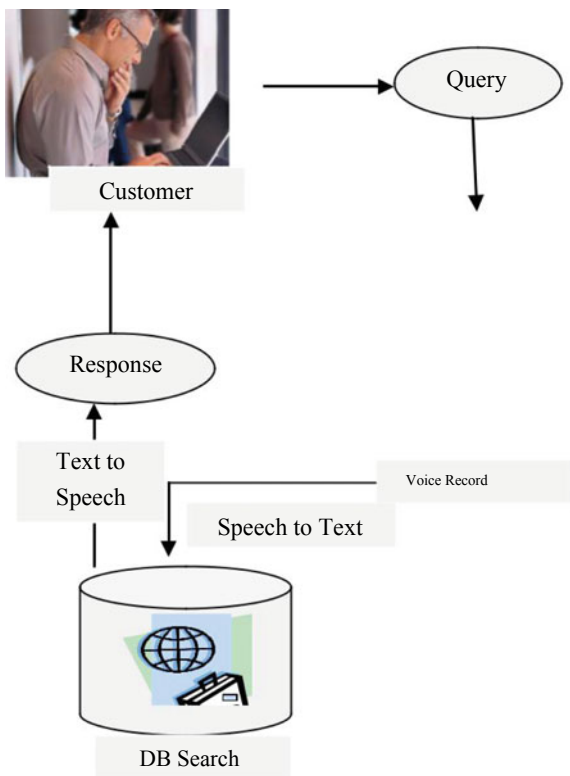
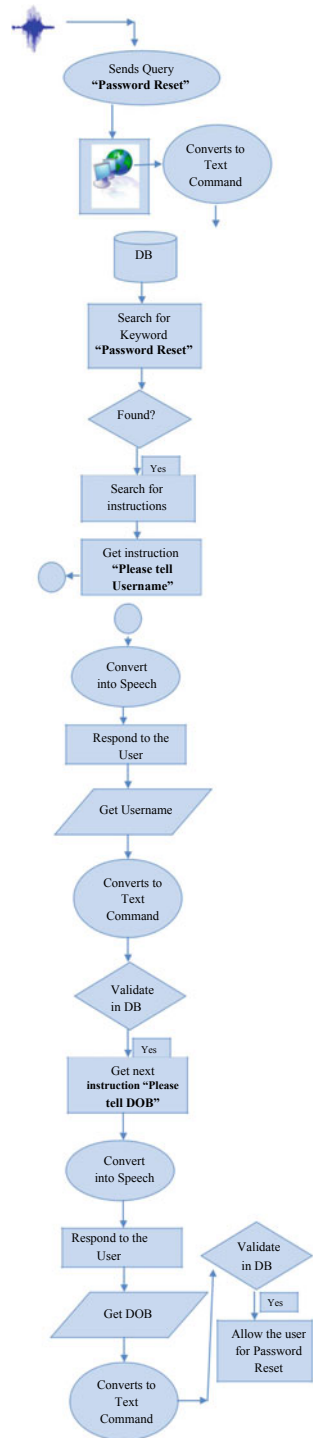


Fig. 3 Processing query for password reset



repetitive query processing. This kind of automatic processing the customer queries resembles ROBOT-type processing and thus we termed our methodology as ***automatic robot processing (ARP)***.

Along these lines, our proposed strategy acknowledges the voice information and follows the procedures to process the client inquiries effectively. The proposed strategy comprises a calculation to depict the stream of the procedure.

3.2 Algorithm-The Flow of Automatic Robot Processing (ARP)

B. Algorithm- The Flow of Automatic Robot Processing (ARP)

Begin

Load the server with Voice Recognition Software

Load Database with instructions for Query Processing

If customer enters then

do

Get the voice Command

Translate the Voice Command to Text Command

Search for relevant keywords

Get appropriate instructions from the database

Translate the instructions into speech commands

Responds the customer with the speech command

until the query processing ends

End If

End

4 Experimental Setup and Performance Metrics

We have to evaluate the performance of our proposed methodology in order to prove that our ARP query processing performs better than the existing techniques. This has to be carried out by taking some set of queries, and those queries have to be processed both manually and through our proposed ARP algorithm. The performance ratio is analyzed and the result has to be tabulated in order to prove that our ARP algorithm performs much better than the existing technique. While the

Fig. 4 Comparison chart

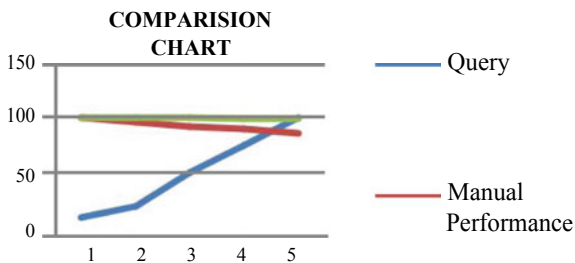


Table 1 Comparison results

Query	Manual performance	ARP performance
10	100	100
20	96	100
50	92	99.5
75	90	99
100	86	98

queries have to be processed manually, as the number of queries becomes increased, the ratio of performance seems low. However, when the queries have to be processed through our proposed algorithm, the performance becomes invariable. The results are tabulated and the comparison chart is shown below. With these results, it becomes clear that our ARP algorithm performs better and the queries are processed successfully (Fig. 4; Table 1).

5 Conclusions

Speech recognition is one of the most assimilating areas of artificial intelligence. This speech recognition technique helps the people who were uneducated or not having sufficient English knowledge or computer knowledge. After undertaking deep research, it has been proved that speech recognition has made a role for handling the speech-oriented activities and has been seen in many walks of life. In this computer world, various techniques are discussed about speech recognition system. This speech recognition is one of the most challenging problems to deal with. We have attempted in this paper to provide a solution to automate some pre-defined activities in the customer care organization. We also provide suitable algorithm to process this robot processing. In this case, it seems easy to access the queries in customer care center without any manual resources.

In future, our approach must be improved with much propelled route in order to handle more operations automatically with reduced manpower.

References

1. Gupta, S., Pathak, A., Saraf, A.: A study on speech recognition system: a literature review. *Int. J. Sci. Eng. Technol. Res. IJSETR* **3**(8), 2193–2194 (2014)
2. Ghai, W., Singh, N.: Literature review on automatic speech recognition. *Int. J. Comput. Appl.* **41**(8), 44–46 (2012)
3. Arora, S.J., Singh, R.P.: Automatic speech recognition: a review. *Int. J. Comput. Appl.* **60**(9), 37–41 (2012)
4. Rashmi, C.R.: Review of algorithms and applications in speech recognition system. *Int. J. Comput. Sci. Inf. Technol.* **5**(5), 5258–5261 (2014)
5. Gamit, M.R., Dhameliya, K., Bhatt, N.S.: Classification techniques for speech recognition: a review. *Int. J. Emerg. Technol. Adv. Eng.* **5**(2), 59–61 (2015)
6. Shaikh Naziya, S., Deshmukh, R.R.: Speech recognition system—a review. *IOSR J. Comput. Eng.* **18**(4), 3–8 (2016)
7. Kaur, I., Kaur, N., Ummat, A., Kaur, J., Kaur, N.: Automatic speech recognition: a review. *Int. J. Comput. Sci. Technol.* **7**(4), 44–46 (2016)
8. Juang, B.H., Rabiner, L.R.: Automatic speech recognition—a brief history of the technology development. *Int. J. Biomed. Eng. Technol.* **3**(2), 4–17 (2005)
9. Prabhakar, O.P., Sahu, N.K.: A survey on: voice command recognition technique. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **3**(5), 578–584 (2013)
10. Kaur, B.: Major challenges of voice command recognition technique. *Int. J. Sci. Eng. Res.* **5**(8), 216–224 (2014)
11. Atame, S., Shanthi Therese, S., Gedam, M.: A survey on: continuous voice recognition techniques. *Int. J. Emerg. Trends Technol. Comput. Sci.* **4**(3), 38–40 (2015)
12. Koumpis, K., Renals, S.: Transcription and summarization of voicemail speech. In: *Proceedings of ICSLP* (2000)

Banking and FinTech (Financial Technology) Embraced with IoT Device



G. Suseendran, E. Chandrasekaran, D. Akila and A. Sasi Kumar

Abstract In recent years the traditional financial industries have motivated for a new technology of financial technology (FinTech) which embraced with internet of things (IoT). The requirements of FinTech and IoT need to be integrated into new business environment. Several companies are affected because of the financial-level investments. So, there is a need to improve the next level of the business. FinTech can introduce a new service of tools and products for the emergent businesses through the internet of services which provide ideas linked in internet. Nowadays, increasing number of companies uses the IoT and creates new added values. The administrators of existing money-related organization in the direct society are dreadful by means of budgetary innovation. The social innovation is accomplished by new innovation. To make a powerful business plan and action, the FinTech and IoT are combined in order to create new innovative ideas based on the requirements.

Keywords FinTech · IoT · Innovation · Business model · Cyber security · Banking

G. Suseendran (✉) · D. Akila · A. Sasi Kumar
Department of Information Technology, School of Computing Sciences,
Vels Institute of Science Technology & Advanced Studies (VISTAS), Chennai, India
e-mail: suseendar_1234@yahoo.co.in

D. Akila
e-mail: akiindia@yahoo.com

A. Sasi Kumar
e-mail: askmca@yahoo.com

E. Chandrasekaran
Department of Mathematics, Veltech Dr. RR & Dr. SR University, Chennai, India
e-mail: e_chandrasekaran@yahoo.com

1 Introduction

The current generation is surrounded by smart phones, which connect devices working from home, public places and office, colleges, schools and everywhere. The consumers use this opportunity to work smarter with the help of internet of things (IoT). Banking, finance and insurance companies are easily embraced with IoT devices. The banking and financial sectors create a new way of collecting the valuable information about the customer through IoT sensor devices using smart phones. FinTech is similar to electric vehicle innovation which works with capital valuation, trading, investment and asset valuation, and provides new improving accounting systems when compared to existing financial technology. Financial engineers handle hundreds of millions of datasets. It is not easy for them to maintain all the data. So FinTech is adapting to a larger field of the IoT and will transform all the customer information to cloud using IoT devices. Most of the banking and financial sectors can rapidly improve in e-commerce by connecting cloud business through the profitable offerings of business tactics, where the cloud is stored with a lot of customers' data. In recent years, all the banks are communicating with the customer through smart phones, social networks and any new sensors, in financial technology to create some new industries [1].

1.1 *IoT-Aided Banking Services*

The internet of things is a big approach in financial services. The IoT connected to the concerned bank through the internet will send and receive data that are stored in cloud. Figure 1 explains the details of the customer, bank and IoT transactions [2].

Around the world, billions of devices are connected with each other. These devices share the information on the cloud with the permission of the bank and allow the entire customers to view the account details and provide access power while using smart devices. It is the easiest way of communicating with the customer and also conveying the personal information through messages and alerts awaiting works.

1.2 *Benefits of IoT in Banking Services*

The most important benefit of IoT in banking services is providing the credit and debit cards for easy access of the services of the banks. Also, the bank can analyze the usage of ATMs in the specific area to increase and decrease the installations of ATMs. While using IoT device, all the customer information are stored in the devices. So, the bank uses this opportunity to help in identifying the customer's business needs, like supplier, retailer and distributors [3]. Figure 2 shows the details of the bank providing valuable services to a customer.

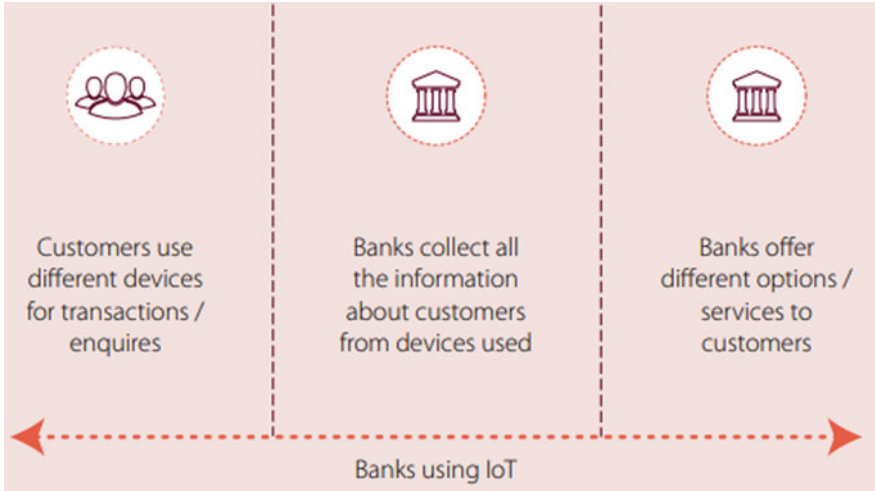
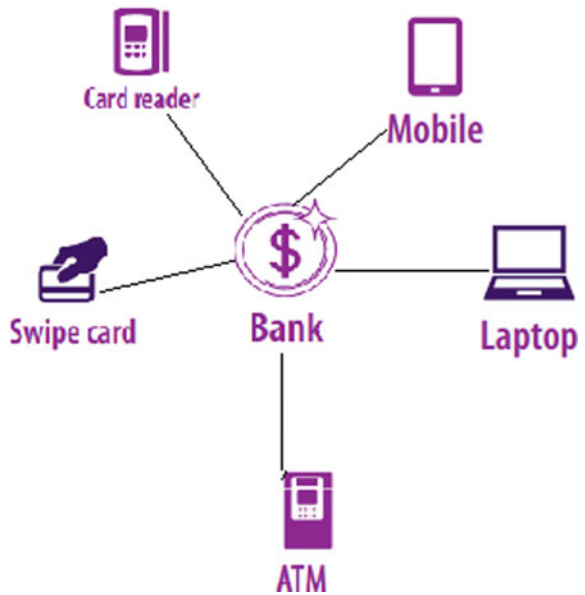


Fig. 1 How to connect customer, bank and IoT devices

Fig. 2 Bank providing valuable services



2 Cyber Security in the Banking and FinTech on IoT

The banking and financial industry is the main target of cyber attacks. In this network world, personal data are easily accessed by the attacker. So, cybercrime is feared all around the world today. We will protect the cyber security in the banking and FinTech services industries. In Fig 3 IoT connects device through the internet and then connects to the FinTech, cloud, machine learning and industrial productivity.

Cyber Attack

Cyber attack is a careful mistreatment of a computer system, technology and networks. The hacker uses malicious code or software to alter the system and secrete code that can compromise the data to cybercrimes, such as health care documents, banking accounts details and hacking lock of the system.

As we are increasing the use of IoT connecting devices in the banking sector, the risk of cyber attacks also increases. The IoT connecting device communicates, analyzes and presents some new ways for technology. It is not only the data but other kinds of sensitive information are also shared through the IoT. Hence the risks are exponentially high [4].

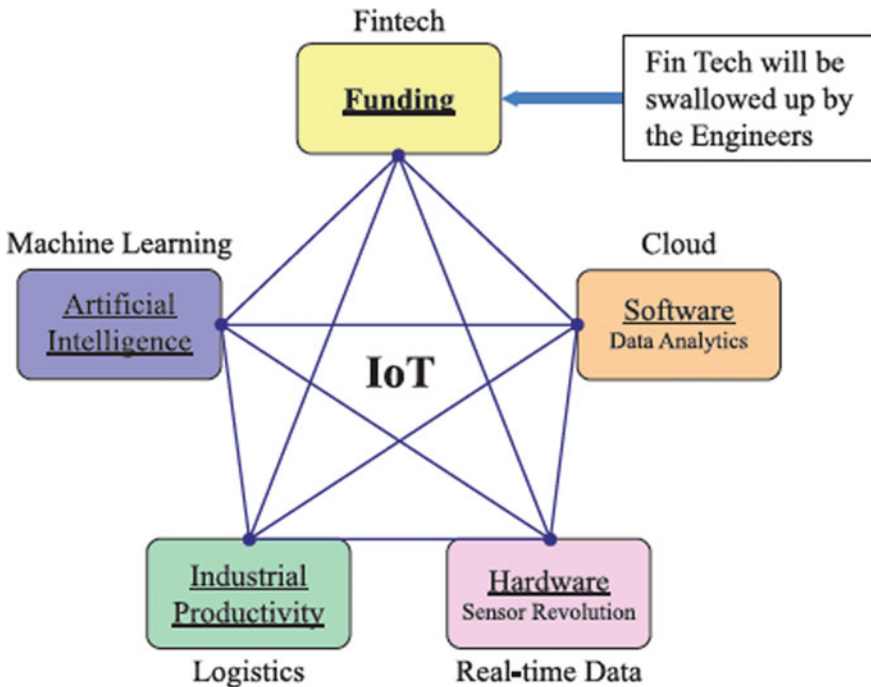


Fig. 3 Internet of things

Unencrypted Data

The data breach happens due to improper encryption, and the stolen data have immediate access.

Unprotected Third-Party Services

The internet services are an extremely worldwide connector; therefore, the cyber attacker can easily access the data of the targeted user, because third-party services are unprotected.

Unsecured Mobile Banking

In recent years, mobile banking users have increased. Using this opportunity, the mobile hacker accesses the data due to small computation time. For securing the data in mobile banking, a cryptography method of encryption and decryption is used.

A Constantly Changing Treat Landscape

In the past few years, cyber threat landscape has changed the financial services. Cybercriminals must change the low-value payment into high-value payment. So, a number of breaches affect the financial sector.

False-Positives

Anti-money laundering (AML) monitoring system is false-positive. The issue will point out the fake activity accessing time and calculated with the help of an analyst.

The Big Breach

The huge volume of financial data is increasing the risk on customers' security from hackers and cybercrime as it occurs at the night time. So beware of the breach in banking sectors.

These are major cyber security threats in banking and FinTech [5].

3 Challenges of Banking and Financial Institutions

Mainly four challenges are faced by banks and financial industries which respond consumer expectations, heavy competition on financial companies, regulatory pressure and not making enough money [6, 7].

i. **Not making enough money**

Many of the banks run on unprofitability because the financial industry is still not making enough return on investment.

ii. **Consumer expectation**

Many bank employees are feeling at most pressure because the expectations of the consumer are high. To maintain the standard they need to work hard which leads to the pressure.

iii. **Increased competition from financial technology companies**

FinTech companies is using the start up business to provide financial services to technologies. FinTech is a big challenge because of using a traditional banking system. They can change the modification and technical operation quickly, and another backup process is handled.

iv. **Regulatory pressure**

The bank requirements are continuously changing and because of that banks invest a lot of amount in some other business. So, the system processes to keep up with the higher goal requirements.

4 New Tools and Product for FinTech Sector

Traditional financial sectors are lacking because of some redundant files and were forced to move with the new tools and new products. With the advantages from new technologies, which are requirements, design and modeling, investments and delivery models are merged with IoT through the internet facilities. Table 1 lists out the new tools and products that are developed and developing in the financial sector. In 2018, 15 top Indian financial markets had radical transformation by technology and innovation in India. The FinTech sector in India with USD 1.2 billion in the year 2016 is expected to touch USD 2.4 billion in the year 2020. **MobiKwik**—Indian digital wallet company, **Capital Float**—digital financial company serving businesses, **Bank Bazaar**—online marketplace for bank loans, credit cards and insurance policies, **Incred**—web-based financial services, **PolicyBazaar**—online insurance aggregator, **Fino Payments Bank**—providing a

Table 1 New tools and product of FinTech company [9]

S. No.	FinTech sector	New tool	New products
1	Banking	Improved loan risk Monitoring Emp Debt Profit Analysis of SMS lending Financing failing applications	Record keeping with sensors Factoring and leasing Trade finance and energy finance Security for accounts Goolglization of accounts
2	Wealth management	Real-time IoT data for stocks IoT as main source for ideas IoT replaces bond derivatives	Telematics as a metric for start ups Link health monitor to wealth management Data to profit customer
3	Insurance	Insuring in high-risk areas Sensors data for smart payload Unbiased vehicle data	Pricing assets in risk-prone areas Accurate pricing product liability Weather detection reduces claims
4	Capital market	Leveraging IoT for crowd investing Block chain IPOs	New commodity data streams New banks for public access to capital

technology solutions for institutions like banks, governments and insurance companies, **CCAvenue**—popular payment gateway, **Razorpay**—a product suite that manages the entire payments lifecycle for all business are the topmost FinTech companies in India [8].

5 Use Cases of IoT—Digital Future

In banking, IoT is interconnected with connecting devices; the system that provides services does machine-to-machine (M2M) communications and is connected to lot of protocols, domains and applications. The IoT has impacted the traditional financial process such as trade financial, payments, personal financial management (PFM) and insurance [10]. There are many use cases that can be implemented in banking in the period ranging from small to long term (Fig. 4).

5.1 Account Management on Things

Biometrics (voice/touch) can make the accounts’ access anywhere simpler through the digital channel. Using the new technology called Wet Ink, the customer can sign in remotely through any touch screen gadget and can be marked promptly onto physical paper with Wet Ink.

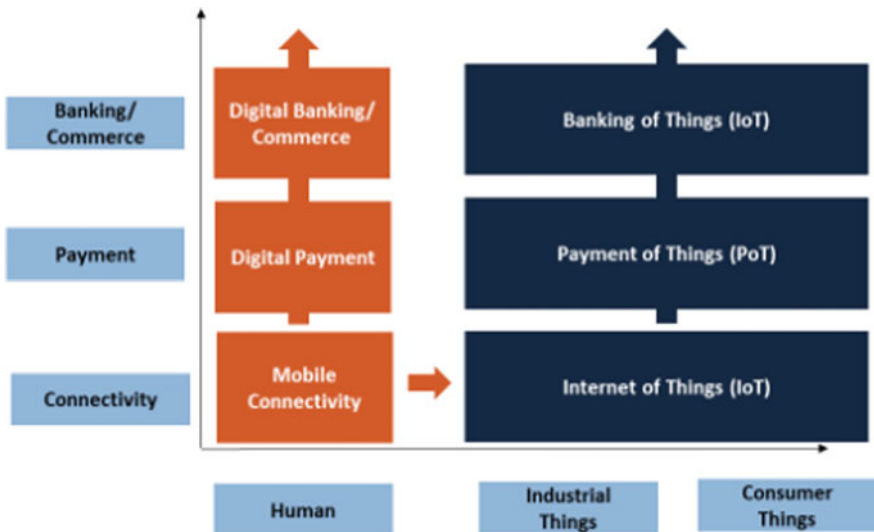


Fig. 4 IoT enabling banking of things in digital future

5.2 Leasing Finance Automation

New daily leasing models have enabled the digital assets worldwide and are effectively turning the traditional products and its services. For delineations, the rented resources could be bolted or debilitated remotely by the bank.

5.3 Smart Collaterals

An IoT device can empower the money-related banks to all the more likely command over client hypothecated resources, for example, observing their wellbeing, autos and home. Financial offerings such as manufacturing machinery, cars, building home loans as collaterals are provided in short term and can be done in a digital way automatically. For representations, in the event of EMIs are not paid, the motor can be impaired and the nature of security can be observed continuously.

5.4 Automated Payment Through Internet of Things

The traditional banking of payment transaction is automated and integrated of services. IoT can raise some conditions on security concerns and digital security in payments. So, the customer can do the payment transaction automatically like one bank to other bank or one bank to other companies.

5.5 Risk Mitigation in Trade Finance

High estimation of products are utilize the RFID monetary space. With the help of IoT the shipments including the delicate merchandise are monitored for example, restorative atoms. These executions of the hazard are relief and more educated choices at banks for including exchange back.

5.6 Wallet of Things

Wallet is associated with each device, where more devices have become digital and smart and that all banks have automated payments through IoT. For example, upkeep administrations utilizing wallet automatically can be stopping the payment transactions.

6 International Global Fintech Benchmark Report 2017

An international global FinTech benchmark report led an online review on FinTech chiefs from monetary establishments around the globe. An in-depth interview was conducted among senior executives from leading FinTech companies (Fig. 5).

FinTech technologies will come to emerging with Amazon, Google and business-based platforms by the next three years. From overview reacting investigation and huge information be arranged anticipated that would a large portion of consideration. 76% of back up plans, 65% of banks and 58% of advantage administration organizations are positioned the information examination as one of the best most FinTech advancements. Moreover, application programming interface advancements and mechanical autonomy and robo-guides are positioned in the high range of the world wide. Figure 6 shows the rising of FinTech with more enthusiasm in the upcoming next three years [7].

Insurance companies are more intent towards IoT because all valuable information and pricing are provided to the customer. The IoT connecting with the insurer person and his information from risk partners. On the off chance that any

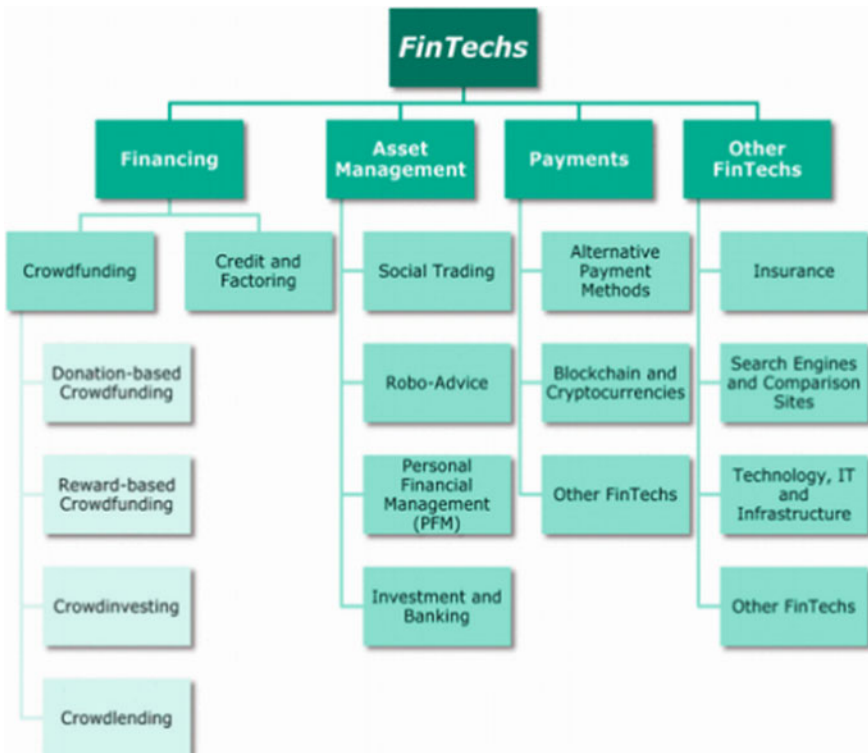


Fig. 5 Segments and elements of FinTech

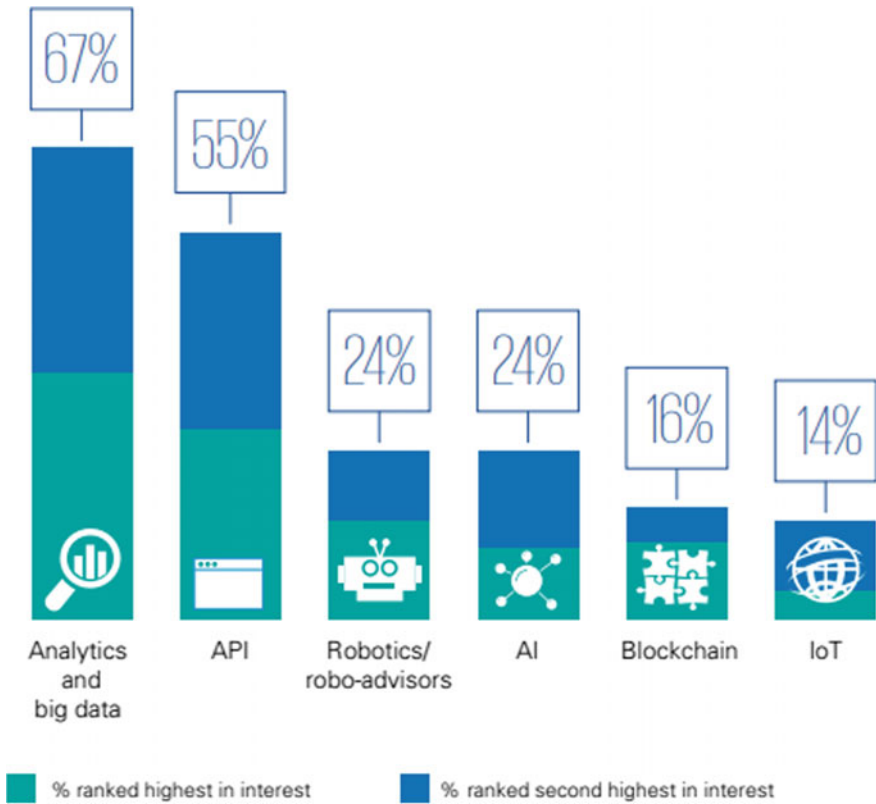


Fig. 6 Emerging FinTech in next three years in the field

mishaps to giving security after the occasion as a case. 36% resource administration organizations additionally positioned man-made consciousness as a best of advancements of intrigue. Safety net providers have more grounded enthusiasm for advances identified with IoT at that point banks or resource administration organizations.

In the course of recent months, the square chain has essential measurement of consideration. It was anticipating the high positioned organizations in the benefit administration. But, it was just 35% positioned zone intrigue.

7 Analysis and Discussion

For this part of work all information and data are collected from various research articles and magazines, and these data were discussed and investigated in this section. We divided the analyses into two categories: the first part is the status of

FinTech block chain and crypto currencies. EU, USA and India have how alternative payment methods and investment are to be demonstrated. Banking is providing positive characteristic as well as negative characteristic using FinTech global financial service sector elements [11].

7.1 FinTech Current Status and Positive Characteristics

FinTech is an imperative piece of open monetary administration segment. It gives information on monetary and keeping money of conventional individuals. The prospect getting diverse sorts of money related administrations. The fiasco and the wretchedness in all the three locales, for instance, the EU, USA and India, did negative effects and impacts on the advancement and interruption in the monetary administration sector [12].

The new FinTech companies are using block chain innovation that has given the opportunity and benefits for more money transaction related administrations. The term digital revolution was reported by Accenture in 2015.

According to Table 2, the Luxembourg is in the top position in digitalization index of 1.00 and USA has 0.92 in digitalization index, which means this is closing to reach the full potential value, and EU also has the average range digitalization value of 0.62, which means the EU member countries are positioned with different values. But India is lagging in digitalization world as it has the value 0.29 and positioned 83rd in the digitalization index.

The FinTech companies bid these services at a lower level price compared to traditional financial service sectors because fully automated technology accessing the operation and process work completely or partially. The motivation or main concept of FinTech service is to reduce the human errors and increase banking business transaction and accessing. In this system automatically applicable by India. India will reached the highest digitalization index ranking position when compared to EU and USA ranking level.

7.2 Fintech Current Status and Negative Characteristics

They are many positive characteristics identified by the people, which are block chain and crypto currencies, an alternative payment system and FinTech technology and banking solutions. Nevertheless, the treats related to the FinTech essentials are really negative effects in the FinTech financial service sectors. The negative elements are due to affecting the FinTech operations and that related process failed incompletely. That's why India is not able to develop it, while the EU and USA quickly increased the work regularly.

Table 2 BBVA digitalization index 2015 [13]

S. No.	Country	Index	S. No.	Country	Index
1	Luxembourg	1.00	44	Kazakhstan	0.47
2	United Kingdom	0.97	45	South Africa	0.47
3	Hong Kong SAR	0.95	46	Slovakia	0.46
4	United States	0.92	47	Mauritius	0.46
5	Netherlands	0.90	48	Colombia	0.45
6	Japan	0.88	49	Russian Federation	0.45
7	Singapore	0.87	50	Italy	0.44
8	Norway	0.86	51	Azerbaijan	0.44
9	Finland	0.85	52	Poland	0.43
10	Sweden	0.84	53	Romania	0.43
11	Switzerland	0.82	54	Croatia	0.43
12	Iceland	0.82	55	Montenegro	0.42
13	Canada	0.81	56	Kuwait	0.41
14	New Zealand	0.80	57	Mexico	0.41
15	Australia	0.79	58	Greece	0.40
16	Germany	0.78	59	Armenia	0.40
17	Denmark	0.77	60	Georgia	0.40
18	Korea, Rep.	0.76	61	Panama	0.40
19	Estonia	0.76	62	Macedonia FYR	0.39
20	France	0.76	63	China	0.38
21	Austria	0.73	64	Thailand	0.38
22	United Arab Emirates	0.71	65	Morocco	0.37
23	Belgium	0.69	66	Philippines	0.35
24	Ireland	0.68	67	Sri Lanka	0.34
25	Island	0.68	68	Egypt	0.33
26	Bahrain	0.65	69	Indonesia	0.33
27	Lithuania	0.65	70	Bulgaria	0.33
28	Maita	0.64	71	Moldova	0.33
29	Malaysia	0.63	72	Tunisia	0.33
30	Spain	0.62	73	Argentina	0.32
31	Qatar	0.61	74	Kenya	0.32
32	Saudi Arabia	0.59	75	Peru	0.32
33	Portugal	0.59	76	El Salvador	0.31
34	Chile	0.58	77	Serbia	0.31
35	Latvia	0.55	78	Dominican Rep	0.31
36	Czech Republic	0.52	79	Vietnam	0.31
37	Oman	0.51	80	Honduras	0.30
38	Turkey	0.50	81	India	0.29
39	Costa Rica	0.49	82	Albania	0.28
40	Jordan	0.48	83	Albania	0.24

(continued)

Table 2 (continued)

S. No.	Country	Index	S. No.	Country	Index
41	Cyprus	0.48	84	Senegal	0.24
42	Hungary	0.48	85	Guatemala	0.22
43	Uruguay	0.47	86	Ukraine	0.21

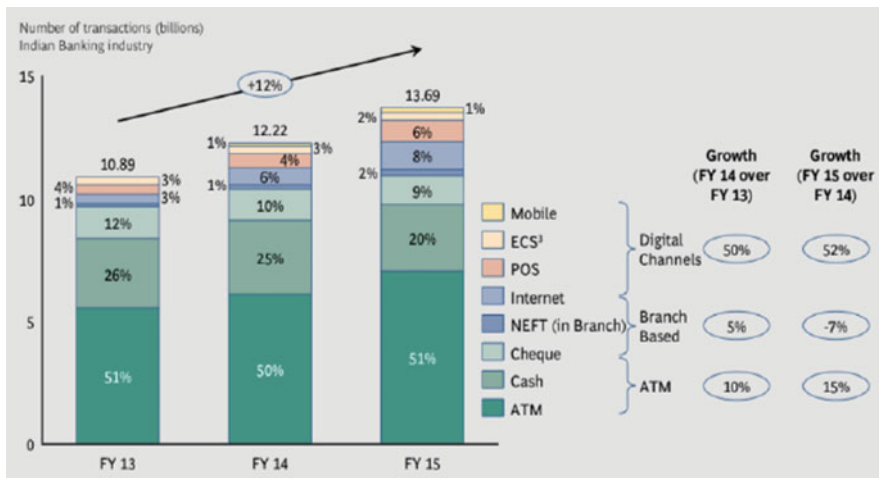


Fig. 7 Digitalization transaction of India in 2013–2015

According to the data, Fig. 7 displays the digitalization transaction in India’s growth; however, the Billion Eco System (BCG) research in India that surveyed the ATM mobile transaction and withdrawal transaction at ATM, NEFT-based transaction in mobile, mobile internet and traditional transaction of cash and cheque and all our day-to-day our life usage showed 12% increase in the year 2013–2014 and also in 2014–2015; so, year by year as we increase FinTech operations, banking financial services also increased [14].

This is very a good example of block chain and crypto currencies method. The customer will need to safety for payment transaction because the hackers follow to your regular work. This cyber security and data privacy also comes from the USA and EU. The regularly EU, USA and India using the FinTech methods and block chain Methods for money transaction [15].

8 Conclusions

In this study the overall functions of financial advanced technologies, challenges of banking and financial industries, cyber security of banking while connecting IoT through the internet to store the all information's on cloud and benefits of the FinTech while connecting IoT device have been described. In addition, the use cases of FinTech on IoT in digital society in futures are mentioned in detailed. All banks, financial industry, insurance companies are moved to automated technology that reduces the workloads and easy to mingle the customer activities. From authors' point of view, the FinTech embraced with IoT in future will boom the society into the next generation.

References


1. Rega, F.G.: The bank of the future, the future of banking—an empirical analysis of European banks. SSRN Electron. J. (2017). <https://doi.org/10.2139/ssrn.3071742>
2. Infosys Limited: IOT-Enabled Banking Services. <https://www.infosys.com/industries/financial-services/white-papers/Documents/IoT-enabled-banking.pdf>
3. Nakashima, T.: Creating credit by making use of mobility with FinTech and IoT. IATSS Res. **42**(2), 61–66 (2018)
4. Us, A., Us, C., Account, M.Y.: Cybersecurity in the Banking and Financial Services Sector—IoT Threats, Potential Solutions and Blockchain (2018). <https://www.stoodnt.com/blog/cybersecurity-in-banking-financial-services/>
5. Cuipa, E., Ramani, S., Shetty, N., Smart, C.: Financing the Internet of Things: An Early Glimpse of the Potential M-RCBG Associate Working Paper Series, No. 85 (2018)
6. Schubert, J.: 4 Top Challenges Facing The Banking Industry Right Now (2015). <https://www.digitalistmag.com/industries/banking/2015/08/27/4-top-challenges-facing-banking-industry-right-now-03352186>
7. Foes, O.R.: Fintech and banks, regulation, and the real sector. Eur. Econ. Banks Regul. Real Sect. **2**, 1–162 (2017)
8. S. in Join: 15 Top Fintech Companies in India 2018 (2018). <https://www.whizsky.com/2018/04/20-top-fintech-companies-in-india-2018/>
9. Schulte, P., Liu, G., Alerts, E., Journals, I.I.: FinTech is merging with IoT and AI to challenge banks: how entrenched interests can prepare. Inst. Investig. J. **20**(3), 41–57 (2018)
10. Story, C., Talk, I., Study, C.: Accelerating your journey to #truly digital banking (2017). file:///C:/Users/Nathiya/Downloads/FinacleConnect-Apr-Sep-2017-Vol09-Issue34.pdf
11. Pejkovska, M.: Potential negative effects of FinTech on the financial services sector Examples from the European Union, India and the United States of America (2018). <https://www.theseus.fi/handle/10024/148416>
12. Fenwick, M., Mccahery, J.A., Vermeulen, E.P.M., Fenwick, M., Mccahery, J.A.: FinTech and the Financing of Entrepreneurs : From Crowdfunding to Marketplace Lending FinTech and the Financing of Entrepreneurs (2017). https://ecgi.global/sites/default/files/working_papers/documents/fenwick-mccahery-vermeulen.pdf
13. Accenture 2015: The Future of FinTech and Banking (2015). https://www.accenture.com/t20150707T195228Z_w_us-en/_acnmedia/Accenture/Conversion-Assets/DoCom/Documents/Global/PDF/Dualpub_11/Accenture-Future-Fintech-Banking.pdf#zoom=50

14. Shah, A., Vibha, K., Prateek, R., Abhishek, A.: Digital Payments 2020: The making of a \$500 billion ecosystem in India. [ebook] Boston Consulting Group and Google. http://image-src.bcg.com/BCG-GoogleDigitalPayments2020-July2016_tcm21-39221.pdf
15. Pulse, K., Share, F.R.: Global FinTech market sees quiet Q1'17 as M&A slows, VC funding holds steady: KPMG Pulse of FinTech Report (2018). <https://home.kpmg.com/in/en/home/media/press-releases/2017/05/kpmg-pulse-fintech-report.html>

Big Data Management

GRNN++: A Parallel and Distributed Version of GRNN Under Apache Spark for Big Data Regression



Sk. Kamaruddin  and Vadlamani Ravi 

Abstract Among the neural network architectures for prediction, multi-layer perceptron (MLP), radial basis function (RBF), wavelet neural network (WNN), general regression neural network (GRNN), and group method of data handling (GMDH) are popular. Out of these architectures, GRNN is preferable because it involves single-pass learning and produces reasonably good results. Although GRNN involves single-pass learning, it cannot handle big datasets because a pattern layer is required to store all the cluster centers after clustering all the samples. Therefore, this paper proposes a hybrid architecture, GRNN++, which makes GRNN scalable for big data by invoking a parallel distributed version of K-means++, namely, K-means||, in the pattern layer of GRNN. The whole architecture is implemented in the distributed parallel computational architecture of Apache Spark with HDFS. The performance of the GRNN++ was measured on gas sensor dataset which has 613 MB of data under a ten-fold cross-validation setup. The proposed GRNN++ produces very low mean squared error (MSE). It is worthwhile to mention that the primary motivation of this article is to present a distributed and parallel version of the traditional GRNN.

Keywords Apache Spark · Big data regression · GRNN · HDFS · K-means++ · K-means||

Sk. Kamaruddin · V. Ravi (✉)

Centre of Excellence in Analytics, Institute for Development and Research in Banking Technology, Castle Hills Road No. 1, Masab Tank, Hyderabad 500057, India

e-mail: padmarav@gmail.com

Sk. Kamaruddin

e-mail: skkamaruddin@gmail.com

Sk. Kamaruddin

SCIS, University of Hyderabad, Hyderabad 500046, India

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_16

1 Introduction

The term prediction is used for regression as well as classification. When the target variable is discrete the associated task belongs to classification and when the target variable is continuous the associated task is a regression. The regression or forecasting model takes the past data to predict the target variable. There are different neural network architectures used for regression, namely, MLP, RBF, WNN, GRNN, and GMDH.

MLP has been used for regression in several research works, namely, Kusakunniran et al. [1] presented a gait recognition system based on motion regression using MLP; Agirre-Basurko et al. [2] suggested MLP to forecast gas levels; Gaudart et al. [3] performed a regression for epidemiological data using MLP.

Similarly, RBF also has been used for regression, namely, Mignon and Jurie [4] proposed a face reconstruction algorithm, based on RBF regression in eigenspace; Hannan et al. [5] presented heart disease diagnosis using RBF; Taki et al. [6] used RBF for energy consumption prediction for wheat production.

WNN is used for regression models for forecasting reservoir inflow [7]. Vinaykumar et al. [8] implemented WNN for the estimation of cost of software development; Chauhan et al. [9] used DE-trained WNN for bankruptcy prediction in banks. Rajkiran and Ravi [10] used WNN for predicting reliability of software.

In addition, GMDH is used for regression, namely, Astakhov and Galitsky [11] presented tool life prediction in gundrilling; Elattar et al. [12] used GMDH for short-term load forecasting; Srinivasan [13] used GMDH for energy demand prediction; Ravisankar and Ravi [14] implemented GMDH for bankruptcy prediction in banks; Mohanty et al. [15] used GMDH for prediction of software reliability; Reddy and Ravi [16] proposed kernel GMDH for regression.

The ANN with its vast variants of architectures have been implemented in different application domains, namely, wireless networks [17], robot manipulator control [18], wind energy systems [19], cancer prediction in healthcare [20], crop production [21], and business [22].

Among these different architectural variants of ANN, the GRNN architecture is preferable as it uses a single pass for learning, and being non-parametric estimation produces a good result. In this article, we have proposed a hybrid version of GRNN for prediction. The motivation and contribution for the proposed method are introduced in Sect. 2.

The paper is structured in the following way: Motivation and contribution are presented in Sect. 2. A literature review is presented in Sect. 3. The proposed methodology is described in Sect. 4. The system setup for conducting experiments is discussed in Sect. 5, and the information about the dataset used in the experiment is described in Sect. 6. The results and discussion are presented in Sect. 7. Section 8 concludes the paper with future directions.

2 Motivation and Contribution

We have presented in the introduction section the different ANN architectures for regression. Among them, GRNN is an optimal choice for exploration being a one-pass trainer and non-parametric estimator. According to Specht [23], if a large dataset is involved, we should have a process of clustering to have neurons in the pattern layer. But, clustering of massive data is not efficient enough through any standard clustering approach. Even if we opt for the parallel version of K-means, it is not efficient enough with respect to optimal clustering. Again, we find there is no published work involving prediction of large-scale dataset using the GRNN. These shortcomings and adversaries are the strong cause of the motivation for the current research work.

In this proposed work GRNN is used for prediction of the large-scale dataset. Here, we have used a clustering approach using K-means|| [24] which is a parallel version of K-means++ [25] to reduce the pattern layer neurons. It is a variant of K-means which has optimal initial seed selection approach. The K-means|| is embedded into the architecture of GRNN for carrying out clustering before the pattern layer and implemented in a parallel distributed computing framework of Apache Spark which is henceforth called as GRNN++. The GRNN++ has shown very high accuracy in the form of mean squared error. The hybrid architecture has proved the implementation of the best features of two worlds, that is, of clustering and regression.

3 Literature Review

There are several variants of K-means proposed in the literature by the research community. Also, a large effort has been dedicated toward parallelization of K-means algorithm. Zhao et al. [26] have propounded a MapReduce framework for parallel K-means. Liao et al. [27] have presented an improved version of parallel K-means using MapReduce and a method for selecting the initial centroids. The K-means++ [25] is optimized by the process of selection of initial cluster centers, but it has a major drawback as being sequential. This limitation restricts its application to the massive dataset. It has to perform k-passes over the dataset to figure out k initial cluster centers. This drawback is addressed by K-means|| [24]. Apart from K-means, evolving clustering method is another clustering approach which has been parallelized by Kamaruddin and Ravi [28].

The regression has always been the fascinating area of machine learning, and different approaches have been presented by the researchers. One of the efficient neural network architecture for regression is GRNN [23] which has the advantage of non-parametric estimation along with one-pass of training. These advantages enable the GRNN to be faster with better accuracy. There are several contributions by the researchers in different applications areas using GRNN, namely, Leung et al.

[29] have forecasted exchange rate using GRNN; Kayaer and Yildirim [30] implemented GRNN for medical diagnosis of diabetes; Li et al. [31] implemented GRNN for image quality assessment; Li et al. [32] implemented GRNN with fruit fly optimization algorithm for power load distribution prediction.

But the primary disadvantage of GRNN is that all the training data should be stored in the pattern layer, which requires more computational time. Thus, results in an inefficient architecture for the large-scale dataset.

Some researchers have presented hybrid architectures, and those are generally found to be more efficient. Ravi and Krishna [33] implemented generalized regression with auto-associative neural network (GRAANN) for imputing missing values. Similarly, the hybrid architecture of auto-associative extreme learning machine and multiple linear regression proposed by Tejasviram et al. [34] has shown better prediction capability. Kamaruddin and Ravi [35] presented a hybridization of particle swarm optimization and AANN for one-class classification and had applied it to credit card fraud detection which had demonstrated good performance.

From the study of the related literature, we found that GRNN has not been implemented in a parallel distributed environment and is not able to handle large-scale dataset. So the current research work addresses the issue.

4 Introduction to Proposed Methodology: GRNN++

GRNN is not able to handle large-sized dataset without efficient clustering. It will raise memory issues and increase the execution time. These drawbacks are due to the incapability of the standard clustering approach to handle large-sized dataset. In the proposed GRNN++ method this drawback has been addressed. The GRNN++ is implemented with Apache Spark and HDFS. It has two components, namely, (i) K-means|| clustering, and (ii) GRNN for prediction. The two components are introduced in the following two subsections.

4.1 *K-Means++: An Overview*

The K-means clustering algorithm has an objective of minimal intra-cluster variance; that is, the sum of the squared distance of the points belonging to a cluster from their cluster center should be minimum. The K-means chooses random initial “K” cluster centers and updates the cluster centers with iterations until there is no further change in the cluster center, and thus finds reasonable solutions quickly. However, it suffers from at least two major drawbacks: (1) the algorithm has the worst-case execution time which is super-polynomial of input size; (2) the formed clusters may not be optimal with respect to the objective function.

The K-means++ [25] has addressed the deficiency of optimal clustering by the introduction of a method for cluster center initialization before execution of the iterations of K-means algorithm. The K-means++ clustering algorithm using an innovative method for initial cluster center selection or seed selection was proposed by Arthur and Vassilvitskii [25]. The algorithm has the following steps:

K-Means++ algorithm:

Let D be the dataset of p points in R^d and $x \in D$. Let $d(x)$ be the smallest distance from any given data point to its nearest cluster center which has been already selected.

1. Select first center C_1 uniformly at random from D .
2. For each data point $x \in D$, compute $d(x)$ and take a new center C_i , with probability $\frac{d(x)^2}{\sum_{x \in D} d(x)^2}$.
3. Repeat Step 2 until k centers have been selected altogether.
4. Proceed with the iterations of the K-means.

4.2 *K-Means||: An Overview*

The K-means|| is the parallel implementation of K-means++. Say, there are n number of samples, and k initial cluster centers are needed. The K-means++ performs k number of passes to sample one initial cluster center in each pass. But, K-means|| samples first center uniformly at random. Then, the next centers will be chosen non-uniformly with a given probability which is stochastically biased by the already chosen centers. This happens in a parallel way across all partitions. Thus, the so-obtained $O(k \log n)$ points are finally re-clustered into k initial centers for the standard K-means iterations.

4.3 *GRNN: An Overview*

GRNN [23] is a feed forward neural network having roots in statistics. GRNN represents an advanced architecture in the neural networks that implements non-parametric regression with one-pass training.

The topology of GRNN is depicted in Fig. 1 and involves four layers of neurons, namely, input, pattern, summation, and the output.

The pattern layer comprises training neurons. The test sample is fed to the input layer. The distance, d_i , between the training samples presents as a neuron in the pattern layer, and the data point from the test set used for regression is used to figure out how well each training neuron in pattern layer can represent the feature space of the test sample, X . This probability density is calculated with Gaussian activation function. Thus, the summation of the product of target value and the result of activation function

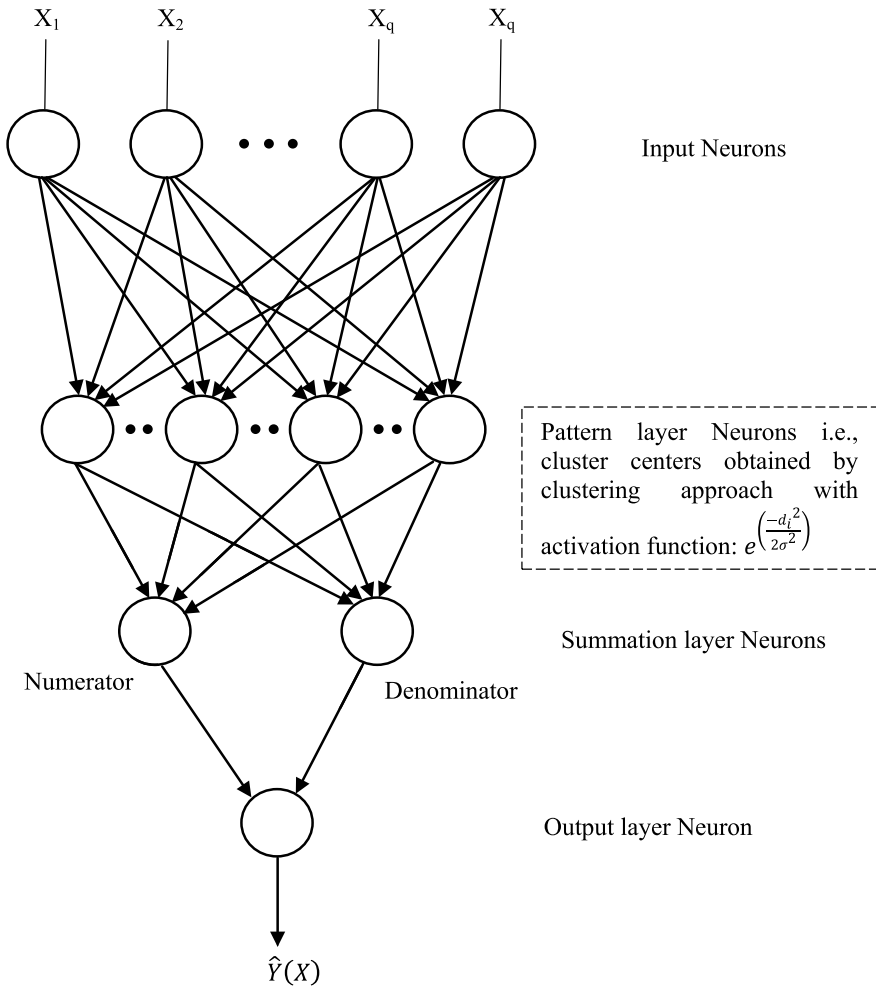


Fig. 1 Architecture of GRNN

for each neuron that is, $\sum_{i=1}^n Y_i * e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}$ forms the numerator term in the summation layer and the summation of the term $e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}$ that is, $\sum_{i=1}^n e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}$ forms the denominator term in the summation layer. Then, in the output layer, the prediction is calculated as

$$\hat{Y}(X) = \frac{\sum_{i=1}^n Y_i * e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}}{\sum_{i=1}^n e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}}. \text{ Here } d_i^2 = (X - X_i)^T * (X - X_i). X \text{ is a test sample and } X_i \text{ is the}$$

neuron present in the pattern layer. For big datasets, the equations involved in

summation and output layer change to accommodate the information about the cluster centers. For further details please see the training algorithm presented in Sect. 4.4.

4.4 GRNN++: The Proposed Method

The training data is clustered with K-means||. Thus, the cluster centers represent the data samples belonging to the cluster most effectively. Then the Dunn-like Index (a modified Dunn Index), a cluster validity index, is used for measuring the cluster validity.

The Dunn Index, proposed by Dunn [36], presents the approach for measuring the cluster quality so that the best clusters were selected based on the high compactness of clusters and high separation among them, that is, compact and well-separated clusters. The compactness with good separation is found by the ratio of minimum intra-cluster distance to the maximum inter-cluster distance. The Dunn Index can be presented as follows:

$$Dunn\ Index(D) = \min_{1 \leq i \leq M} \left\{ \min_{i+1 \leq j \leq M} \left(\frac{dist(c_i, c_j)}{\max_{1 \leq l \leq M} (diam(c_l))} \right) \right\}. \quad (1)$$

where M is the total number of clusters under consideration, $dist(c_i, c_j)$ is the inter-cluster distance and $diam(c_l)$ is the maximum intra-cluster distance or the largest diameter among all the clusters.

As the Dunn Index performs a calculation of maximum intra-cluster distance, that is, the diameter of a cluster under consideration, it is not suitable for large-scale dataset as it will involve large computational overhead. Also, it is affected by noisy data and outliers. So, we have used a Dunn-like Index [37] for cluster validity measurement. In this approach, the diameter, that is, the denominator of (1) is calculated by finding the radius using the distance from each sample present in the cluster to its cluster center, that is, $d(x, \bar{c}_j)$, where $x \in c_j$ cluster and \bar{c}_j is its center. This Dunn-like Index is not affected by noise or outlier present in the data. The Dunn-like Index can be presented as:

$$Dunn - like\ Index(DL) = \min_{1 \leq i \leq M} \left\{ \min_{i+1 \leq j \leq M} \left(\frac{dist(c_i, c_j)}{\max_{1 \leq l \leq M} \left(2 * \frac{1}{|c_l|} \sum_{x \in c_l} d(x, \bar{c}_l) \right)} \right) \right\}. \quad (2)$$

The cluster centers form the clusters generated by the K-means|| algorithm and validated by Dunn-like Index. They now represent the pattern nodes of GRNN++.

Then, each test sample is passed on to the neurons present in the pattern layer with a Gaussian probability density function resulting in higher value if the neuron has a strong similarity to the test sample. Now, the summation layer represents the numerator, denominator and the output layer presents the prediction as described in the overview of the GRNN section (refer to Sect. 4.3).

Training Algorithm for the proposed GRNN++

- Step 1:** Perform data normalization with the range [0,1].
- Step 2:** Divide the data to perform 10-Fold cross-validation.
- Step 3:** Take the training set and cluster them with K-Means|| with the $k = 2$ to 10.
- Step 4:** Compute Dunn-like Index by varying the value of k and find the k which produces optimal clustering.
- Step 5:** Consider the cluster centers represented by the optimal k as the neurons in the pattern layer of GRNN++ and pass the training or test data, as the case may be, to the pattern layer nodes with a Gaussian transfer function.
- Step 6:** The product of A_i , i.e., the sum of the output values of the samples present in the i^{th} cluster and the result of activation function for i^{th} neuron in the pattern layer $e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}$ i.e., $\sum_{i=1}^n A_i * e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}$ forms the numerator term in the summation layer (say *Num*).
- Step 7:** The product of B_i , i.e., count of the samples present in the i^{th} cluster and the result of activation function for i^{th} neuron in the pattern layer $e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}$ i.e., $\sum_{i=1}^n B_i * e^{\left(\frac{-d_i^2}{2\sigma^2}\right)}$ forms the denominator term in the summation layer (say *Denom*).
- Step 8:** In the output layer, the prediction is computed as $\hat{Y}(x) = \text{Num} / \text{Denom}$.
- Step 9:** The accuracy of prediction is measured by computing MSE.
- Step 10:** Steps 3 to 9 are repeated for the 10 folds of data.
- Step 11:** Average MSE for 10-FCV is computed and reported.

The architecture of GRNN++ is the same as the traditional GRNN except for the presence of the clusters (obtained through K-means||) occupying the pattern layer in place of the patterns themselves (see Fig. 1). The third and fourth layer of GRNN++ are identical to that of GRNN.

5 Experimental Setup

The system configuration for carrying out experimental work consists of standalone Spark cluster 2.2.0 as the computational framework on top of HDFS which is the distributed storage system. The program development environment used is Apache

Table 1 System configuration description

Central processing unit	Intel® Core™ i7-6700 CPU @ 3.40 GHz
No. of cores	4 Physical cores or 8 logical cores
Primary memory	32 GB
OS	Ubuntu 16.04 LTS

Table 2 Details of resource allocation in the cluster and coding environment

Allocated resource details	Node type	
	Master	Worker
Memory allocated for driver process	14 GB	–
Count of worker daemons	2	4
Memory allocated to worker daemon	7 GB	7 GB
Number of executors per node	2	4
Memory allocated to each executor	7 GB	7 GB
Cores allocated per executor	2	2
Memory allocated to all executors	14 GB	28 GB
Total memory utilized (out of 32 GB)	28 GB	28 GB
Framework for computation	Apache Spark 2.2.0	
Distributed storage system	HDFS (Hadoop 2.7.3)	
Coding interface	Apache Zeppelin 0.7.3	
Programming language	Scala 2.11.8	

Zeppelin 0.7.3 with Scala 2.11.8 as the coding environment. The Spark cluster consists of a system as a master node and seven systems as worker nodes. The driver program resides in the master node along with two instances of worker daemons. Each of the worker nodes executes four instances of worker daemons. All the seven systems have a similar configuration (refer to Table 1). In a standalone cluster system, each worker daemon has one executor. The details of resources allocated to worker daemon or executor, driver process, and programming environment details are presented in Table 2.

The GRNN++ was implemented with the above cluster configuration and programming environment. We found the best execution time by adjustment of the amount of memory allocation to the executors along with the optimum data partitions.

Table 3 Attribute details of ethylene–CO gas sensor array dataset

Attribute name	Attribute details
Time (s)	Time measured in seconds for the sensor readings
CO conc. (ppm)	The concentration of CO measured in ppm
Ethylene conc. (ppm)	The concentration of ethylene measured in ppm
Sensor readings (16 channels)	Readings from 16 chemical sensors forming the data comprising 16 columns

6 Dataset Description

The current work has analyzed the data from an array of gas sensors involving different gas mixtures. The dataset [38] was collected from UCI ML repository [39]. The dataset contains the readings gathered from 16 sensors exposed to varying concentrations of gas mixtures. The original dataset contains readings of two different mixtures of gas in air, namely, ethylene with methane, and ethylene with CO. We have analyzed only the ethylene and CO mixture in air dataset for prediction of ethylene and CO concentration in air. The sensor readings were the data from 16-sensor array signals gathered uninterruptedly for about 12 h.

The ethylene and CO mixture in air dataset, which is analyzed in the current work, presents 19 features. Here, the first feature represents the time of sensor reading; the second feature represents CO concentration in ppm; and the third feature represents the ethylene concentration in ppm. The next 16 features represent the sensor readings from 16 chemical sensors (see Table 3). The dataset contains 4.2 million samples. The dataset is of 643 MB.

For our regression work, we have dropped the first feature being the time series values. Out of the second and third feature, one is selected as the dependent variable. We have selected the features from the fourth to the last as independent variables for our experimentation.

7 Results and Discussion

The research work conducted comprises the analysis of the dataset mentioned above with the proposed approach. The results of the carried out work cannot be compared with any other technique as we could not find any work carried out in the same domain and using such hybrid architectural techniques.

The experiment was carried out with ten-fold cross-validation (10-FCV) of the min–max normalized dataset. The K-means|| was carried out with k value ranging from 2 to 10 and then the Dunn-like Index was taken into account for cluster validity for finding out the optimal cluster centers. Then, with the selected cluster centers GRNN++ was carried out with sigma (σ) value ranging from 0.001 to 1.0

Table 4 Mean MSE for 10-FCV

	Ethylene concentration prediction from ethylene to CO dataset	CO concentration prediction from ethylene to CO dataset
Gaussian activation function	0.075	0.076

with an increment of 0.001. The error is calculated as MSE. Table 4 presents the average MSE of 10-FCV.

The activation function used in GRNN++ is the Gaussian function which captures the features of a test data and correspondingly presents the prediction value.

8 Conclusion and Future Directions

The GRNN++, a parallel and distributed version of GRNN, is proposed in Apache Spark, where HDFS takes care of the distributed data storage. The GRNN++ is implemented in Scala programming language. The GRNN++ could achieve big data regression in a single pass and by yielding a very low mean squared error of 10-FCV. The hallmark of the proposed method is the invocation of the parallel distributed K-means++ aka K-means|| into the architecture of the traditional GRNN. GRNN being the non-parametric estimation provides better accuracy.

The innovation of the proposed method resulted in a drastic reduction in the count of neurons in the pattern layer of GRNN. This feature has made the system overcome the major shortcoming of GRNN which is computational overhead.

The effectiveness of GRNN++ is tested on a static dataset, but it can be modified to suit online streaming data mining by invoking parallel evolving clustering method (PECM) [28] in place of the K-means||. Thus, it can be extended to handle real-time or quasi-real-time prediction from streaming data. Further, the logistic activation function can also be considered in place of the Gaussian in the pattern layer.

References

1. Kusakunniran, W., Wu, Q., Zhang, J., Li, H.: Multi-view gait recognition based on motion regression using multilayer perceptron. In: 2010 20th International Conference on Pattern Recognition, pp 2186–2189. IEEE, Istanbul (2010)
2. Agirre-Basarco, E., Ibarra-Berastegi, G., Madariaga, I.: Regression and multilayer perceptron-based models to forecast hourly O₃ and NO₂ levels in the Bilbao area. *Environ. Model Softw.* **21**, 430–446 (2006)

3. Gaudart, J., Giusiano, B., Huiart, L.: Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data. *Comput. Stat. Data Anal.* **44**, 547–570 (2004)
4. Mignon, A., Jurie, F.: Reconstructing faces from their signatures using RBF regression. In: *Proceedings of the British Machine Vision Conference 2013*, pp 103.1–103.11. British Machine Vision Association, Bristol (2013)
5. Hannan, S.A., Manza, R.R., Ramteke, R.J.: Generalized regression neural network and radial basis function for heart disease diagnosis. *Int. J. Comput. Appl.* **7**, 7–13 (2010)
6. Taki, M., Rohani, A., Soheili-Fard, F., Abdeshahi, A.: Assessment of energy consumption and modeling of output energy for wheat production by neural network (MLP and RBF) and Gaussian process regression (GPR) models. *J. Clean. Prod.* **172**, 3028–3041 (2018)
7. Budu, K.: Comparison of wavelet-based ANN and regression models for reservoir inflow forecasting. *J. Hydrol. Eng.* **19**, 1385–1400 (2014)
8. Vinaykumar, K., Ravi, V., Carr, M., Rajkiran, N.: Software development cost estimation using wavelet neural networks. *J. Syst. Softw.* **81**, 1853–1867 (2008)
9. Chauhan, N., Ravi, V., Karthik Chandra, D.: Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. *Expert Syst. Appl.* **36**, 7659–7665 (2009)
10. Rajkiran, N., Ravi, V.: Software reliability prediction using wavelet neural networks. In: *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, pp 195–199. IEEE, Sivakasi (2007)
11. Astakhov, V.P., Galitsky, V.V.: Tool life testing in gundrilling: an application of the group method of data handling (GMDH). *Int. J. Mach. Tools Manuf* **45**, 509–517 (2005)
12. Elattar, E.E., Goulermas, J.Y., Wu, Q.H.: Generalized locally weighted GMDH for short term load forecasting. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **42**, 345–356 (2012)
13. Srinivasan, D.: Energy demand prediction using GMDH networks. *Neurocomputing* **72**, 625–629 (2008)
14. Ravisankar, P., Ravi, V.: Financial distress prediction in banks using group method of data handling neural network, counter propagation neural network and fuzzy ARTMAP. *Knowl. Based Syst.* **23**, 823–831 (2010)
15. Mohanty, R., Ravi, V., Patra, M.R.: Software reliability prediction using group method of data handling. In: Sakai, H., Chakraborty, M.K., Hassanien, A.E., Ślęzak, D., Zhu, W. (eds.) *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. RSFDGrC 2009*, pp 344–351. Springer, Berlin (2009)
16. Reddy, K.N., Ravi, V.: Kernel group method of data handling: application to regression problems. In: Panigrahi, B.K., Das, S., Suganthan, P.N., Nanda, P.K. (eds.) *Swarm, Evolutionary, and Memetic Computing. SEMCCO 2012*, pp 74–81. Springer, Berlin (2012)
17. Ahad, N., Qadir, J., Ahsan, N.: Neural networks in wireless networks: techniques, applications and guidelines. *J. Netw. Comput. Appl.* **68**, 1–27 (2016)
18. Jin, L., Li, S., Yu, J., He, J.: Robot manipulator control using neural networks: A survey. *Neurocomputing* **285**, 23–34 (2018)
19. Marugán, A.P., Márquez, F.P.G., Perez, J.M.P., Ruiz-Hernández, D.: A survey of artificial neural network in wind energy systems. *Appl. Energy* **228**, 1822–1836 (2018)
20. Agrawal, S., Agrawal, J.: Neural network techniques for cancer prediction: a survey. *Proc. Comput. Sci.* **60**, 769–774 (2015)
21. Khoshroo, A., Emrouznejad, A., Ghaffarizadeh, A., Kasraei, M., Omid, M.: Sensitivity analysis of energy inputs in crop production using artificial neural networks. *J. Clean. Prod.* **197**(Part 1), 992–998 (2018)
22. Tkáč, M., Verner, R.: Artificial neural networks in business: two decades of research. *Appl. Soft Comput.* **38**, 788–804 (2016)
23. Specht, D.F.: A general regression neural network. *IEEE Trans. Neural Netw.* **2**, 568–576 (1991)
24. Bahmani, B., Moseley, B., Vattani, A., Kumar, R., Vassilvitskii, S.: Scalable K-means++. *Proc. VLDB Endow.* **5**, 622–633 (2012)

25. Arthur, D., Vassilvitskii, S.: k-means ++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp 1027–1035 (2007)
26. Zhao, W., Ma, H., He, Q.: Parallel K-means clustering based on MapReduce. In: Jaatun, M. G., Zhao, G., Rong, C. (eds.) Cloud Computing, pp. 674–679. Springer, Berlin (2009)
27. Liao, Q., Yang, F., Zhao, J.: An improved parallel K-means clustering algorithm with MapReduce. In: 2013 15th IEEE International Conference on Communication Technology, pp 764–768. IEEE (2013)
28. Kamaruddin, S., Ravi, V., Mayank, P.: Parallel evolving clustering method for big data analytics using apache spark: applications to banking and physics. In: Reddy, P., Sureka, A., Chakravarthy, S., Bhalla, S. (eds.) Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, pp. 278–292. Springer, Cham (2017)
29. Leung, M.T., Chen, A.-S., Daouk, H.: Forecasting exchange rates using general regression neural networks. *Comput. Oper. Res.* **27**, 1093–1110 (2000)
30. Kayaer, K., Yildirim, T.: Medical diagnosis on Pima Indian diabetes using general regression neural networks. In: Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), pp 181–184 (2003)
31. Li, C., Bovik, A.C., Wu, X.: Blind image quality assessment using a general regression neural network. *IEEE Trans. Neural Netw.* **22**, 793–799 (2011)
32. Li, H., Guo, S., Li, C., Sun, J.: A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. *Knowl. Based Syst.* **37**, 378–387 (2013)
33. Ravi, V., Krishna, M.: A new online data imputation method based on general regression auto associative neural network. *Neurocomputing* **138**, 106–113 (2014)
34. Tejasviram, V., Solanki, H., Ravi, V., Kamaruddin, S.: Auto associative extreme learning machine based non-linear principal component regression for big data applications. In: 2015 Tenth International Conference on Digital Information Management (ICDIM), pp 223–228. IEEE, Jeju (2015)
35. Kamaruddin, S., Ravi, V.: Credit card fraud detection using big data analytics: use of PSOANN based one-class classification. In: Proceedings of the International Conference on Informatics and Analytics—ICIA-16, pp 1–8. ACM Press, Pondicherry (2016)
36. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.* **3**, 32–57 (1973)
37. Bezdek, J.C., Pal, N.R.: Some new indices of cluster validity. *IEEE Trans. Syst. Man Cybern. B Cybern.* **28**, 301–315 (1998)
38. Fonollosa, J., Sheik, S., Huerta, R., Marco, S.: Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring. *Sensors Actuators B Chem.* **215**, 618–629 (2015)
39. Gas sensor array under dynamic gas mixtures Data Set, <https://archive.ics.uci.edu/ml/datasets/Gas+sensor+array+under+dynamic+gas+mixtures>

An Entropy-Based Technique for Conferences Ranking



Fiaz Majeed and Rana Azhar Ul Haq

Abstract Ranking of conferences is carried out to guide the researchers so as to publish their work in top-level venues. Existing works are needed to be improved in regard to accuracy. In this paper, we have proposed a ranking technique of conference papers named as influence language model (InLM). It uses entropy to calculate score for the purpose of ranking. It identifies level of the paper as well as level of the conference based on entropy score. Experiments have been performed on bigger dataset. The results reflect better position in comparison to the existing systems.

Keywords Conference ranking · Language model · Influence language model · Entropy · Scientific research

1 Introduction

The ranking of the conferences is an important area of research to assess quality of research work. Such a ranking provides recommendations to the researchers to publish their valuable work in reputed conferences. Several measures have been developed to rank the conferences [3].

This paper solves the problem of ranking the conference papers so that relevant papers of top ranked conferences are placed at first places in the retrieved results. The quality of paper is measured based on its frequency of citations. The main purpose of this research is the semantic ranking of conference papers using influ-

F. Majeed (✉)

Department of Information Technology, Faculty of Computing and Information Technology,
University of Gujrat, Gujrat, Pakistan
e-mail: fiaz.majeed@uog.edu.pk

R. A. U. Haq

Department of Computer Science, Faculty of Computing and Information Technology,
University of Gujrat, Gujrat, Pakistan
e-mail: azhar4u@live.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_17

229

ence language model (InLM). In conference paper ranking, we utilize search and ranking technique that use query string to rank the conference papers. To evaluate the InLM technique, a dataset containing 32,000 papers and 50 user queries is built. We have performed comparative analysis of InLM with the existing methods.

The motivation of this research work is that quality researchers need to search the quality conference papers. As conferences have no standardized impact factor system similar to journals, this research work provides them quality conference papers (which is measured based on citations of the particular paper). The relevant documents are then ranked based on calculation of entropy that uses citations of the paper.

The key contributions of this paper are summarized as follows:

- We have proposed a conference papers ranking technique named as InLM.
- We have presented a ranking algorithm based on entropy to rank the conference papers.

The paper has been organized as follows: The review of related works is comprehensively described in Sect. 2. Later, the proposed solution is presented in Sect. 3, and experimental evaluation is provided in Sect. 4. Finally, conclusion and future work is given in Sect. 5.

2 Related Works

Existing methods on conferences ranking have achieved less accuracy in comparison to the journals ranking. A review of those methods is provided to justify further improvement in this domain: the ranking of authors and conferences based on citations is performed by adding time factor. It is the extension of PagRank for yearly conferences ranking. The DBLP dataset is used for experiments and top 10 authors and conferences have been provided [7]. In another work, conferences are ranked based on publication of their extended papers in top journals. For this purpose, acknowledgment information of finance journals is taken and analyzed. The dataset comprised footnotes of 3000 journal articles and 9000 conference articles. Based on this, 47 conferences have been ranked [4].

Currently, two citation analysis systems only exist for journals ranking, which are CiteSeer and Institute for Scientific Information (ISI). The ranking system for the conferences is equally important. Thus, Scientific Collection Evaluator by using Advanced Scoring (SCEAS) system is developed that considers age of the conference. Using DBLP dataset, a web-based application is built [6].

In computer science (CS) domain, few ranking methods exist. Thus, ranking of CS conferences is performed using RsIT technique. It uses a set of conferences, relatedness measure and baseline method to rank the top CS conferences. The Microsoft Academic dataset is used for experiments. The comparison is carried out with the existing CCF rating list. This method provides better results in comparison

to the CCF list [2]. An analysis of the criteria of existing measures for CS conferences is performed. The conference ranking is compared using bibliometric indicators and acceptance rate. The experiments are carried out on three datasets: RankX, Perfil-CC and ERA2010 [3]. The health of conferences related to the software engineering is checked by defining a suite of metrics. The comparison of conferences by parameters like author, PC candidates and scientific prestige is used. Based on the results, 11 top conferences are selected for the interval of 10 years [8]. A technique to rank the CS conferences is proposed with the extension of existing metrics. It generates self-organizing maps for the purpose of ranking. This method is evaluated in correlation to ERA using Microsoft Academic Search dataset including 3442 conferences [1].

The extension of h-index is presented for the journals and conferences ranking. For this purpose, yearly h-index and normalized yearly h-index have been proposed. The DBLP dataset is used with 2024 conferences [5].

In the existence of above conferences ranking methods, no system could be used as standard similar to the ISI. This is due to the fact that current methods do not achieve the required level of accuracy. It is still required to work on achieving the accuracy. This work is build based on this idea.

3 Proposed Solution

The architecture of the InLM is shown in Fig. 1. The input query is processed on the pre-processed dataset. Further, the retrieved results are ranked based on the proposed InLM-based ranking algorithm.

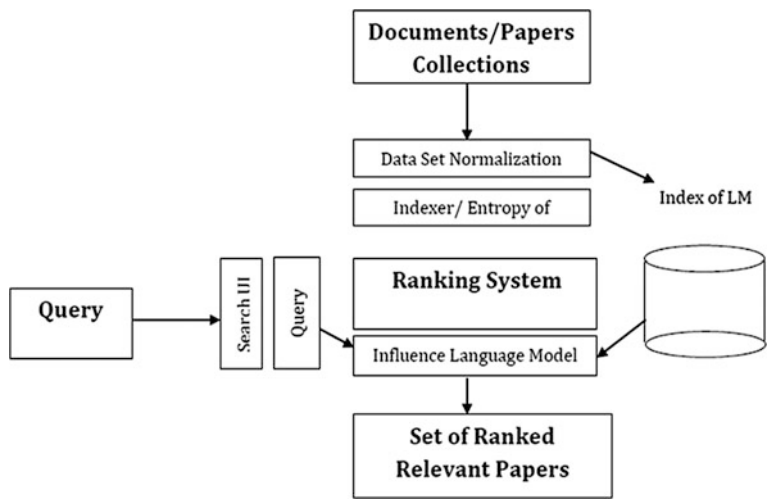


Fig. 1 InLM ranking system

3.1 Entropy Score

As shown in Eq. (1), the entropy of the conferences can be calculated as:

$$\text{Ent}(v) = - \sum_{i=1}^m w_i \log_2(w_i) \quad (1)$$

where w_i has the probability w_i at venue v . Individually, all publications of each conference can be used for computation of the probability for input query.

The entropy of papers is calculated as given in Eq. (2):

$$\text{Ent}(p) = \sum_{i=1}^m w_i \log_2(w_i) \quad (2)$$

In a paper (p), w_i provides the probability of word $_i$.

3.2 Calculation of Citations

In comparison to the DBLP, it was tricky to develop a crawler for the Google Scholar. The procedure was to input the query in the search box of Google Scholar and then find out the relevant paper from retrieved results after pressing the search button. Google Scholar provides the citation count for each paper. This process was required to be automated. Initially, the papers were searched manually for analysis. The URL of the relevant results was changed programmatically in such a way that half part of URL was updated to route the citations in our database. The search page of Google Scholar was explored by using the PHP's html document object model for required data, thus put all the citations in Wamp Server and updated the database. For the improved language model, citations of each ranked paper were measured.

3.3 Influence Language Model

This method measures the worth of a paper, whether it is published in low- or high-level conference. The entropy of the conference is calculated to check the level of a conference. The publication of a paper in certain level of conference is proportional to the number of citations. For instance, a paper published in high-level conference will have increased citations, whereas low citations will be received for the paper published in a low-level conference. The InLM increases the score for high-level conference having low entropy, whereas for the low-level conference with high entropy, the method decreases the score.

Two thresholds of entropy 1.8 and 1.85 have been defined and taken as limiting values. The conference or venue is considered as high level in case the entropy value remains below 1.8. Further, in case the entropy value reaches above 1.85, the venue is considered as low level. In influence language model, the probability of a query over document $P(q|d)$ and the resulted value with the entropy result is multiplied. The entropy of the high-level venue is multiplied to increase the score as high-level venue gain smaller value of entropy then multiply the resulted entropy value with $P(q|d)$ (Eq. 3). The entropy is first divided with 3 for decreasing its value for low-level venues and then $P(q|d)$ is multiplied with the resulted entropy value (Eq. 4). We simply multiply $P(q|d)$ with entropy for the venues whose value of entropy is between the range 1.8 and 1.85. The proposed composite model does not consider venues influence.

$$\begin{aligned} &\text{If ent} < 1.8 \\ &\text{ent_value} = \text{ent} \times 3 \end{aligned} \quad (3)$$

$$\begin{aligned} &\text{Elseif ent} > 1.85 \\ &\text{ent_value} = \text{ent}/3 \end{aligned} \quad (4)$$

The values of language model are multiplied with entropy values. So the main Eq. (5) of InLM is as follows:

$$\text{InLM} = \text{LM} \times \text{ent_value} \quad (5)$$

In Eq. (5), the entropy value of each conference paper is multiplied with each paper of language model result.

It is necessary to prepare the data before input to the machine learning algorithm, so the pre-processing procedure is given in Algorithm 1.

Algorithm 1

1. Extract all the titles from html pages.
 - Read the html file in java.
 - `Patternpt2 = Pattern.compile("b>.*?</b");`
2. Remove stop words from extracted data.
 - Give input of the raw file.
 - By using the string tokenizer, string is matched. The stop words are removed if those are matched.
 - Then line-by-line matching of each token is performed.
 - `StringTokenizer stnz = new StringTokenizer("“*+<>,&-:;()[]$#@!\\^” ~ ^? ”);`
 - By using the above code, each token is matched one by one.

3. Conversion in lower case

- The search becomes much easier by converting these letters in lowercase. Lowercase function is used for this purpose.
- `Token = token.toLowerCase ();`.

4. Frequency Removal

In the research paper titles, there are many words which are used frequently by many researchers. If a word is matched more than thrice, then it will be removed from the data.

5. Normalize the whole data.

3.4 Application of the InLM

For instance, three papers say P1, P2 and P3 are chosen from the dataset. According to the existing language models, all have the same score because of having equal $p(q|d)$. Let according to previous models,

$$P(q|d) \text{ for } P1 = 0.85$$

$$P(q|d) \text{ for } P2 = 0.85$$

$$P(q|d) \text{ for } P3 = 0.85$$

These papers should not have the same scores as they contain the same length of query and published in different conferences, as well as have same number of query terms.

The venue's entropy can be calculated as:

Let

doc 1 entropy = 1.87 (P1 belongs to low-level conference)

doc 2 entropy = 1.69 (P2 belongs to high-level conference)

doc 3 entropy = 1.95 (P3 belongs to low-level conference)

The following result is achieved when the results of previous language models are divided with entropies of conference.

Doc 1 = $0.85 * (1.87/3) = 0.52$ (P1 is also low-level conference)

Doc 2 = $0.85 * (1.69/3) = 4.30$ (P2 is of high-level conference)

Doc 3 = $0.85 * (1.95/3) = 0.55$ (P3 is of low-level conference)

According to the values, the level of conference is being decided.

4 Experimental Evaluations

The experimental evaluation is performed on big dataset. In this respect, the details of the dataset are given below.

4.1 Dataset

Conferences data:

- Data of 100 conferences
- Around 32,000 papers
- Collection of all the title of publications

Query set:

We have developed a query set for experimental purpose. Few of the queries are given below:

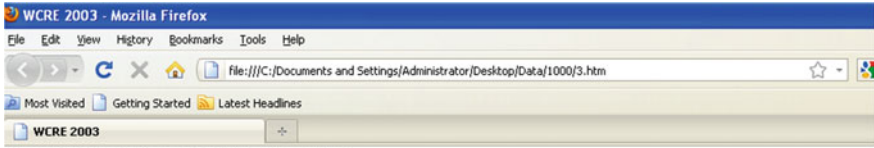
- Machine language
- Data mining
- Formal languages
- Group theory
- Information retrieval
- Computer networks
- Semantic web
- Analysis of algorithm
- Natural language processing
- Databases

These queries are used to find the best conference papers from the conferences data. In other words, quality papers of the authors in the relevant field are ranked by the proposed InLM. The dataset is made up of different CS conferences and more than 32,000 research papers of these conferences. The searching technique used for this purpose is content-based ranking. Specifically, there are 100 CS conferences that are extracted from the DBLP. The research papers are taken in the duration of 2003–2007 (5 years).

The support vector machine (SVM) algorithm has been used for the classification of the data and the results are measured using precision, recall and F-measure.

4.2 Pre-processing

Initially, the whole data are pre-processed. In Fig. 2, raw data in html pages are shown, which contain conference information, that is, author's name, title of the paper and name of the conference where the paper has been published.



Low-Level Reverse Engineering

- [Noah Snavely, Saunmya K. Debray, Gregory R. Andrews](#): Unscheduling, Unpredication, Unspeculation: Reverse *Engineering +Itanium Executables. 4-13 [Electronic Edition](#) (link) [BBIX](#)
- [Lori Vinciguerra, Linda M. Wills, Nidhi Keirival, Paul Martino, Ralph L. Vinciguerra](#): An Experimentation Framework for Evaluating Disassembly and Decompilation Tools for C++ and Java. 14-23 [Electronic Edition](#) (link) [BBIX](#)
- [Lewis B. Baumstark Jr., Murat Guler, Linda M. Wills](#): Extracting an Explicitly Data-Parallel Representation of Image-Processing Programs. 24-35 [Electronic Edition](#) (link) [BBIX](#)

Software Architecture

- [Rainer Koschke, Daniel Simon](#): Hierarchical Reflexion Models. 36-45 [Electronic Edition](#) (link) [BBIX](#)
- [Christoph Stoermer, Liam O'Brien, Chris Verhoef](#): Moving Towards Quality Attribute Driven Software Architecture Reconstruction. 46-56 [Electronic Edition](#) (link) [BBIX](#)
- [Lionel C. Briand, Yvan Labiche, Y. Miaou](#): Towards the Reverse Engineering of UML Sequence Diagrams. 57-66 [Electronic Edition](#) (link) [BBIX](#)
- [Minmin Han, Christine Hofmeister, Robert L. Nord](#): Reconstructing Software Architecture for J2EE Web Applications. 67-79 [Electronic Edition](#) (link) [BBIX](#)

Fig. 2 html pages including raw data

4.3 Results

In this section, we produce the ranking result of our language model and some key terms are used for showing the results of both language models. These terms include Name (title of research paper), Cit (citations), LM (language model) and InLM. The results about one of the query from the query set are given in Table 1.

These are the average results of given base queries using the InLM, including the citations and entropy of the papers for particular input query strings.

In Table 2, the comparison of two models has been taken under consideration, one is previously implemented LM for average of 10 queries and showed the value

Table 1 Top 2 conferences against “data mining” query

S#	Name	Cit	InLM	Name	Cit	LM
1	Region restricted clustering geographic data mining	6	1.875	Data mining track editorial	0	0.675
2	Measuring effectiveness agile methodologies data mining knowledge discovery	4	1.627	Optimization of language data mining	0	0.607

Table 2 Average citations results of top 10 queries

Models	Average citations for 10 queries
LM (VSM)	184.92
InLM	457.36

Fig. 3 Graphical representation

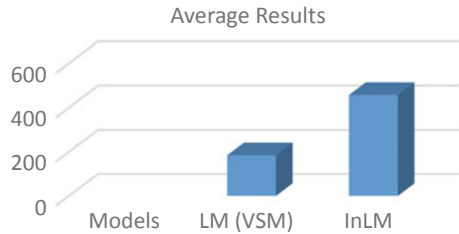
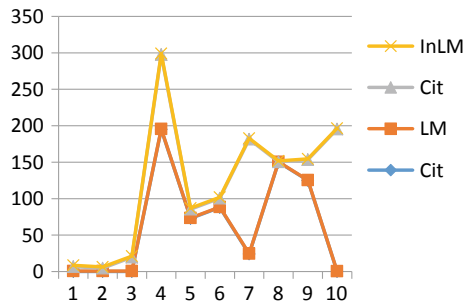


Fig. 4 Comparison of results for data mining query



of 184.92 after setting parameters. On the contrary, the proposed model InLM with the same parameters has given a value of 457.36 with more accuracy and precision for average value of 10 queries.

Figure 3 shows the models along the horizontal axis and number of citations along the vertical axis. The graph clearly mentions that the model InLM has better average results, including entropy results, in more than 400 citations than the LM which is less than 200 citations.

In Fig. 4, the result of InLM is better than LM for “data mining” query.

In Fig. 5, graph depicts score comparison of LM and our proposed model InLM, which shows much better results than LM and has a better ranking for the conference papers.

In Fig. 6, line chart shows ranking result of LM and Fig. 7 shows the result for the InLM.

It has been clearly seen that InLM clearly shows better ranking including entropy of the paper for query. In addition, classification results are shown in Table 3.

Fig. 5 Comparison of language models results for data mining query

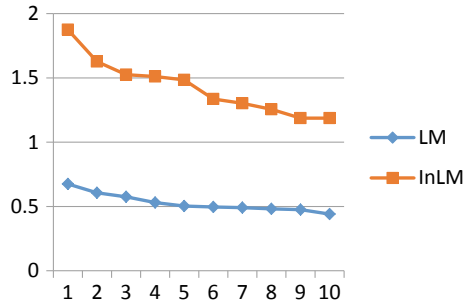


Fig. 6 Language model results of ranked conference papers

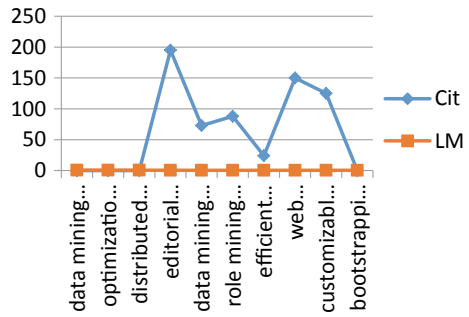


Fig. 7 Language model results of ranked conference papers

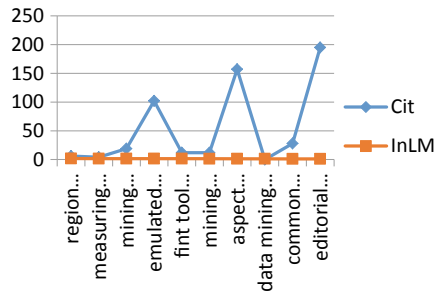


Table 3 Classification results

Precision	Recall	Measure-F
82%	79%	81%

5 Conclusions

In this work, the ranking of the conferences is performed using the entropy measure. The InLM system computes entropy of the conferences to generate a ranking list. Further, a comparative analysis is performed with LM. The InLM shows better results in comparison to LM. Furthermore, time factor and names of authors' papers can also be incorporated to improve the ranking.

References

1. Da Silva Almendra, V., Enăchescu, D., Enăchescu, C.: Ranking computer science conferences using self-organizing maps with dynamic node splitting. *Scientometrics* **102**(1), 267–283 (2015)
2. Effendy, S., Yap, R.H.: Investigations on rating computer sciences conferences: an experiment with the Microsoft Academic Graph dataset. In: *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 425–430 (2016)
3. Küngas, P., Karus, S., Vakulenko, S., Dumas, M., Parra, C., Casati, F.: Reverse-engineering conference rankings: What does it take to make a reputable conference? *Scientometrics* **96**(2), 651–665 (2013)
4. Reinartz, S.J., Urban, D.: Finance conference quality and publication success: A conference ranking. *J. Empir. Finance* **42**, 155–174 (2017)
5. Sidiropoulos, A., Katsaros, D., Manolopoulos, Y.: Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics* **72**(2), 253–280 (2007)
6. Sidiropoulos, A., Manolopoulos, Y.: A new perspective to automatically rank scientific conferences using digital libraries. *Inf. Process. Manag.* **41**(2), 289–312 (2005)
7. Singh, A.P., Shubhankar K., Pudi, V.: An efficient algorithm for ranking research papers based on citation network. In: *Proceedings of the 2011 3rd Conference on Data Mining and Optimization (DMO)* (2016)
8. Vasilescu, B., Serebrenik, A., Mens, T., van den Brand, M.G., Pek, E.: How healthy are software engineering conferences? *Sci. Comput. Program.* **89**, 251–272 (2014)

MapReduce mRMR: Random Forests-Based Email Spam Classification in Distributed Environment



V. Sri Vinitha and D. Karthika Renuka

Abstract The furthestmost standard message transfer system used on the internet for communication is email. These days spam is a serious concern that causes major problems in today's internet. Spam emails are uninhibited messages that are sent to a large number of beneficiaries arbitrarily. Owing to an overgrowing rise in reputation, the number of unsolicited data has also increased promptly and has led to many security concerns. Although the sufficient number of spam filtering techniques exists, nowadays spammers start discovering innovative practices to escape data that are filtered using the spam filters. Spammers use this communication source for spreading the malware in the name of an executable file. These spam emails waste user's system memory, computing power, and bandwidth of the network. Spam emails have been initiated to progressively damage the integrity of email and destroy the online experience. The research revealed that if the classification algorithms are used with feature selection then that will return the exact results than the standard classification. In this paper, feature selection is done through minimum redundancy and maximum relevance (mRMR) and the classification is done by means of Random Forests in the MapReduce environment. The performance is compared using various measures, namely sensitivity, correctness, and accuracy with the Random Forests in the distributed environment using Spambase dataset.

Keywords Email spam · Minimum redundancy maximum relevance (mRMR) · Random Forests · Spambase dataset

V. Sri Vinitha (✉)

Department of IT, Bannari Amman Institute of Technology, Sathyamangalam 638401, India

e-mail: srivinithavellingiri@gmail.com

D. Karthika Renuka

Department of IT, PSG College of Technology, Coimbatore 641004, India

e-mail: karthirenu@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_18

241

1 Introduction

Nowadays, the internet has been mostly used for the purpose of exchanging information. Email has started rising on the internet because of its simplicity, cost-effectiveness, and commercial tool for exchanging information and business communication, and so on [1]. It is exposed to many threats because of its easiness. One among them is email hazard which becomes a big trouble over the internet. Spam email is sent to an enormous amount of users since it charges very less amount for transferring data between the different users. When the user receives the wide range of spam emails then it becomes extremely tough to predict whether the received emails are spam or ham, as it consumes more bandwidth on the internet, wasting user's time and memory. Gradually, spam emails are directed through "zombie systems", or through worm-contaminated PCs around the world. Advanced worms permit spammers to access the user's PC indirectly for malicious purposes. At present, there are some important work-related emails in the spam folder. Globally, the amount of users using email is increased from over 88.6 million in 2014 to over 200.3 million in 2018. Today, various types of spam email exist, and certain things include making money through advertisement and so on. [2]. The sender of a spam email looks like real mail that is not judged by some of the existing filtering techniques. The unwanted spam emails getting into user's inbox should be avoided and to increase the accuracy of email classifiers on the user side, spam filtering techniques are used.

2 Related Works

Nowadays, most people have the habit of using the internet for browsing data, communication purpose, business promotion, and an innovation, providing business services and sharing knowledge with others. Henceforth, the internet has started developing promptly. One of the common threats is spam. Spam message is sent to massive number of recipients for the purpose of advertising, marketing products, job offers, and so on with large incomes for spammers. It also wastes user's time and occupies more memory. These days most of the emails received are spam. Therefore, Bassiouni et al. [3] proposed a technique for classifying spam emails with better performance using ten classification algorithms. The ten different classifiers such as logistic regression, decision tree, Bayes net, support vector machine, naïve Bayes, k-nearest neighbors, Random Forests, artificial neural network, random tree, and radial basis function are used for classification. The ten classification algorithm provides accuracy such as 92.4, 90.3, 89.8, 91.8, 89.8, 90.7, 95.4, 92.4, 91.5, and 82.6. The experimental result shows that the system achieves a correctness of about 95.45% by using Random Forest's approach for Spambase UCI datasets.

Email is the most active and fastest method of correspondence to trade data over the web. It is so popular for its cost-effectiveness, easiness, and fastest mode of communication. Since there is an expansion in the number of users using the diverse social sites, there is a massive increase in unwanted messages. Spam can cause the severe financial disaster, loss of data transmission, and wastage of memory. Spam emails have been created to progressively damage the integrity of email and destroy online experience. The problem of spam mail getting into user's inbox is resolved by using spam filtration approach. In this paper, Kaur et al. [4] suggested the N-gram technique for feature selection and improved multilayer perceptron for spam classification. The result comparison is performed on emails collected from Enron dataset shows that the proposed N-gram-based K-enhanced MLP demonstrated higher performance than existing MLP along with low error rate. Future work uses optimization techniques and some other feature selection techniques for selecting the best features for numerical data, text data, and image corpus.

Currently, the internet has turned out to be an essential one in everyone's life. With this increased use, the number of email users is also enlarged nowadays. It is predicted that 90% of the emails sent are spam. This leads to a problem initiated by spontaneous bulk email stated as spam. It is expected that 90% of emails are sent every day with the intention of advertising any product or spreading any kind of malware to the user PCs without any notification. Spam email is sending unwanted communications to different email users. This spam email causes severe harms for internet users such as the delay in sending legitimate messages, wastage of network bandwidth, storage, and computation power. To avoid this problem, various spam filtering techniques exist. Vijayasekara et al. [5] propose machine learning algorithm naïve Bayesian with three-layer structure for classifying spam in bulk emails. The feature selection method is used to select the best attributes and categorize spam emails using naïve Bayesian algorithm which helps to increase the correctness of the system.

To select the best attributes from the high-dimensional data, it uses feature selection techniques. The methodologies used for feature selection should be efficient to manipulate multivariate historical data. To elucidate this problem, Radovic et al. [6] recommend the significant and repeated data with some modification for the assessment process. The features can be discrete as well as continuous. F-statistics can be considered for significant data and Pearson correlation or mutual information for assessing repeated data. The recommended technique is measured using gene expression temporal datasets and provides high performance compared to other systems. Forthcoming work will be applied to various algorithms with feature selection approach that chooses the slightest repetition and more importance.

At present, electronic mail has become the popular mode of communication on the internet which is used to exchange information to a large number of beneficiaries. It is used for getting and transferring data among different users. Because of its fastest delivery, less time, and low cost for sending emails to make people use it for advertising products, sending fraudulent mails and phishing mails resulted in

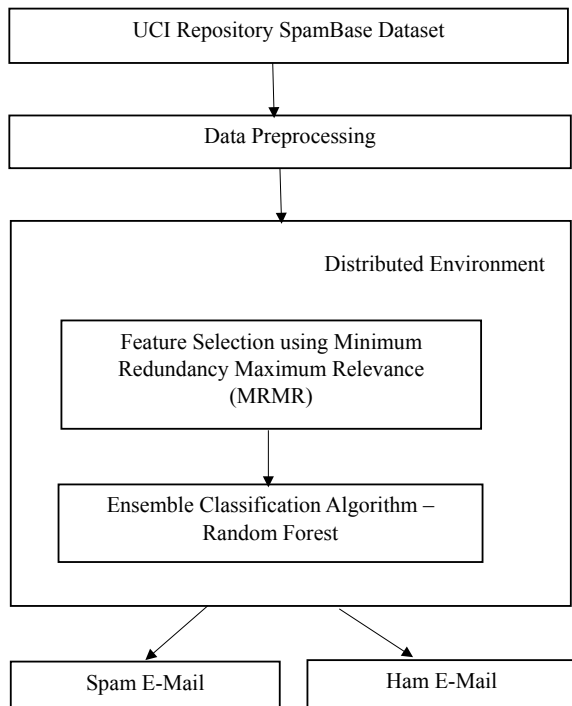
unwanted spam emails. Spam is the unsolicited, unwanted mail that sends messages to a large number of recipients. It also wastes user’s time, transmission capacity, and user’s system space. Spam filtering approaches categorize the spam email based on the information it contains and sends the spam mails into spam folder. Easwaramoorthy et al. [7] proposed the system consisting of the training and testing phase. In the preliminary phase, the hybrid Firefly GSO algorithm takes the input mails to choose the best attributes from the huge dataset. Then the selected features are classified using three classification algorithms, namely decision tree, neural networks, and naïve Bayes algorithm. Then the result is evaluated using UCI repository Spambase dataset through various metrics.

3 Proposed Methodology

The process of the proposed email spam classification technique as in Fig. 1 can be described in the succeeding phases:

- Prepare the UCI repository Spambase dataset for the proposed system
- In the MapReduce framework, do the feature selection and classify the spam emails

Fig. 1 Framework of the proposed spam email classification



- Feature selection is done by using minimum redundancy and maximum relevance technique to select the best attributes from Spambase dataset
- Then the selected attributes are classified by using Random Forest classification algorithm
- Finally, compare the proposed framework with Random Forest classification in a distributed environment.

3.1 Dataset

In this paper, UCI repository Spambase dataset has been taken for email spam classification. Totally, there are 4601 emails in the Spambase. In this 39.4%, that is, 1813 emails are spam and remaining 60.6%, that is, 2788 are ham emails [8]. There are 58 features in UCI repository Spambase dataset out of which 57 attributes are continuous and one with class label (d). The description of the attributes is given in Table 1 as.

3.2 Data Preprocessing

In real world, Spambase dataset contains some missing values, and this missing data may lead to incorrect or even misleading statistics. Data preprocessing procedure is used to remove noise, inconsistent, and incomplete data in order to obtain a quality result. It is the initial step which makes sure that data is prepared to be analyzed [9]. There are various preprocessing steps available, and here data normalization is done in order the make the dataset suitable for further processing.

Data discretization: Discretization states the method of converting continuous attributes into discretized variables. It transforms data from numeric into nominal data type. The generic representation of Spambase dataset is given by:

$$\text{Spambase Dataset} = d_k; \quad 1 \leq k \leq p \quad (1)$$

Table 1 Description of Spambase dataset

Attribute number	Type of attribute
1–48	word_freq_WORD
49–54	char_freq_CHAR
55	capital_run_length_average
56	capital_run_length_longest
57	capital_run_length_total
58	Class attribute

The Spambase dataset is given to the discretization function to transfer the input data into discretized one. Discretization converts the data into specific range. The highest and smallest value of every attribute is identified and the I interval is followed by taking the proportion between the differed value and the I value.

- For example, at-first, the deviation is calculated for every k value

$$\text{Dev}(k) = \frac{\text{Max}(d_k) - \text{Min}(d_k)}{2} \quad (2)$$

- After calculating deviation values for each row, values are converted to the following condition:

$$\left. \begin{array}{l} 0, \text{ input} < 1(\text{Dev}(k)) \\ 1, \text{ input} < 2(\text{Dev}(k)) \\ 2, \text{ input} < 3(\text{Dev}(k)) \\ 4, \text{ input} < 4(\text{Dev}(k)) \end{array} \right\} \quad (3)$$

Then, the value which comes under particular range is substituted by interval value in order to convert the input data into discretized data. After discretization function, the training dataset is converted to above conversion Eq. (3) and then each value is converted into 8421 binary conversions as discretized format.

3.3 *MapReduce*

MapReduce is a framework of Hadoop system designed to examine large dimensions of files in a parallel mode on commodity machines. Its objective is to get used to the available computational resources and process data in a massively scalable way in two processing phases, that is, map and reduce phase.

The entire process of MapReduce can be illustrated using the following notation [10]:

$$\text{Map:}(k1, v1) \rightarrow \text{list} \{k2, v2\} \quad (4)$$

$$\text{Shuffle \& Sort: list}\{k2, v2\} \rightarrow (k2, \text{list}\{v2\}) \quad (5)$$

$$\text{Reduce:}(k2, \text{list}\{v2\}) \rightarrow \text{list}\{k3, v3\} \quad (6)$$

In the Map step, Spambase dataset after preprocessing into discretized data is loaded once into a mapper function as in Eq. (4). Then the feature selection is done by using minimum redundancy maximum relevance and the features are delivered by means of a list of key-value objects. The output of mapper will be the sequence of new objects with maximum relevance features without any repetition.

In the Reduce step, the input of a reducer is the intermediate object as in Eq. (5), which shares the same features that will be collected together. The features are classified by means of Random Forest algorithm and the output is obtained by the majority voting of the classification trees that have been formed. The final output will be produced as in Eq. (6).

3.3.1 Minimum Redundancy Maximum Relevance (mRMR)

The process used for obtaining the subset of relevant attributes is a feature selection process, that is, the dimensionality reduction method [11]. Its objective is to attain a better result with the minimum number of features. The filter method used in this paper selects the feature with minimal repetition and maximal importance. mRMR depends on two metrics: the least repetition among the attributes that have chosen and the greatest significance among candidate features and target class. In the Mapper step, the Spambase dataset D is given as input with N samples (4601 instances) and M features (57 attributes) $X = \{x_1, x_2, x_3, \dots, x_m\}$ and the target classification variable c . The feature selection technique finds the subset of m attributes, R^m from the M -dimensional data, R^M .

Maximal relevance is selecting the features from the Spambase dataset with the maximum significance to the target class c . Either correlation or mutual information may be the measure used to define dependency of variables. Max-relevance is an average of all collective information values among individual attribute x_i in the Spambase dataset and class c as in Eq. (7).

$$\max D(S, c), D = \frac{1}{|S|} \sum_{x_i \in S} I(X_i, c) \quad (7)$$

The subsets that are identified by the maximum relevance often contain relevant repeated data. mRMR challenges to provide the solution to this problem by eliminating those repeated subsets. The minimal redundancy property is included to choose mutually exclusive features from the Spambase dataset as in Eq. (8).

$$\min R(s) = \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i, X_j) \quad (8)$$

Minimum redundancy maximum relevance is combined to select the features that have the high association with the criterion without any repeated data that can be denoted as in Eq. (9). This gives the subset of features S^* selected using mRMR, which is the output of Map step.

$$S^* = \arg \max_{S \subseteq F} [\max D(S, c) - \min R(s)] \quad (9)$$

mRMR Algorithm

```

INPUT: Data, Features
// Data – Spambase Dataset
// Features - 57 features in Spambase Dataset
OUTPUT: SubsetofFeatures // Set of features selected from Spambase dataset.
for features  $f_i$  in SpambaseDataset do
    rel = mutualInfomation( $f_i$ , classlabel);
    redundant = 0;
    for features  $f_j$  in SpambaseDataset do
        redundant += mutualInformation( $f_i$ ,  $f_j$ );
    end for
    mRMRData[ $f_i$ ] = rel - redundant;
end for
SubsetofFeatures = sort(mRMRData). take(Features);

```

The algorithm for minimum redundancy maximum relevance is given above [11]. In the mRMR system, the core problem is associated with the calculation of the mutual information among two components: either exists as a couple of input features or feature by means of the class. This task is figured among every single pair of features; some of the pairs of features are still unrelated for the final outcome, which is insignificant in high-dimensional problems.

3.3.2 Random Forests

Random Forest is an ensemble learning technique that works best for high-dimensional data and was developed by Leo Breiman and Adel for classification and regression not pruned trees. The bagging technique uses a decision tree for selecting arbitrary features to improve classification performance by combining the results of different classifiers. It uses randomized node optimization. Based on the arbitrary choice of data and variables, the progress of the trees can be determined.

The selected feature S^* from mapper phase is given as input to the reducer phase. With the selected feature, it builds an m number of decision trees and forms a forest. Then the new object which is necessary to be classified is placed under each of the

decision trees for classification. The mail is categorized as ham or junk mail based on each trees decision by voting as in Fig. 2. The prediction of the spam emails is attained by majority voting of the decision trees that have been formed, which will be the results of the reducer phase.

Every tree is developed as follows [12]:

1. Spambase dataset after feature selection contains S^* features; the training set will be sample M cases from the subset of features S^* .
2. Choose the best split between the M features.
3. Develop each one of the trees to the biggest possible level without pruning.
4. The spam emails are estimated by grouping the predictions of m trees by majority voting technique.

Certain features of the Random Forest algorithm are given as follows:

1. It provides better accuracy and also lighter compared to other ensemble algorithms.
2. It does not overfit the data because it does not have any insignificant data.
3. It is sturdy to noise and outliers and can also handle uneven datasets.

Pseudocode for Random Forests Algorithm

Begin RandomForests

Input: N:number of nodes

F: Spambase Dataset with the subset of features S^*

T: Total amount of trees to be created

Output: MV : the class label spam or ham with majority voting

While ending condition is incorrect do

The underneath phases are used to build the tree from the training data:

- 1) Select the features f arbitrarily from F ; where $f \ll F$
- 2) For node t , compute the best split point between the f features
- 3) Based on the best split, divide the node into two child nodes
- 4) Iterate the steps 1, 2 and 2 up to n amount of nodes has been reached

Construct the forest by iterating steps from 1–4 for T periods

End While

Display all the created trees

The new sample is applied to all the created trees beginning from the initial node

Allocate the new object to the external node with the resultant class label.

Aggregate the votes of all the nodes in the trees.

Display the final output MV, with the highest vote.

End RandomForests

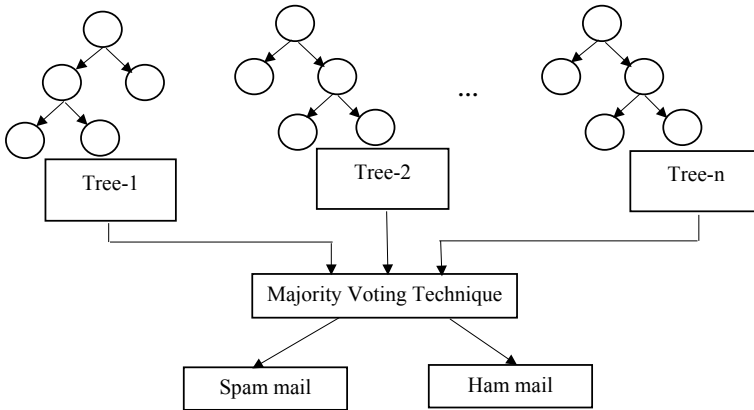


Fig. 2 Random Forests algorithm

4 Performance Analysis

Table 2 shows the performance comparison for Spambase dataset using Random Forests in the distributed environment, and MapReduce minimum relevance maximum redundancy (mRMR) feature selection with Random Forests algorithm. The different assessment metrics have been used for investigating the performance.

Table 2 Comparison of precision, recall and accuracy

Model	Precision	Recall	Accuracy
Random Forests in distributed environment	0.837624	0.846754	0.840721
MapReduce mRMR feature selection with Random Forests	0.852746	0.864657	0.860824

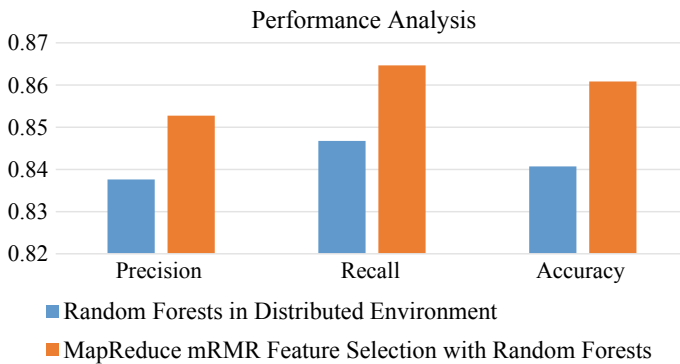


Fig. 3 Performance analysis

By analyzing the plotted graph in Fig. 3, the performance of the proposed email spam classification is significantly improved by selecting the best features using minimum redundancy maximum relevance (mRMR) in MapReduce environment, and the classification is done using the Random Forest.

5 Evaluation Metrics

The efficiency of the recommended methodology for discovering spam emails has been estimated by means of some metrics such as precision, recall, and accuracy. The assessment metrics can be defined as [13]:

Precision

It is the fraction of spam emails that are correctly categorized as spam from the given input emails. Precision illustrates the exact correctness and can be represented using Eq. (10) as:

$$\text{Precision} = \frac{\text{Spam emails correctly classified}}{\text{Spam emails correctly classified} + \text{Ham emails wrongly classified}} \quad (10)$$

Recall

It is the fraction of spam emails that are categorized as spam from the given input emails. Recall illustrates the completeness and is represented using Eq. (11) as:

$$\text{Recall} = \frac{\text{Spam emails correctly classified}}{\text{Spam emails correctly classified} + \text{Spam emails wrongly classified}} \quad (11)$$

Accuracy

It is determined as the percentage of addition of all true positives and true negatives to the entire amount of emails. Accuracy can be denoted by means of Eq. (12) as:

$$\text{Accuracy} = \frac{\text{Spam emails correctly classified} + \text{Ham emails correctly classified}}{\text{Total emails}} \quad (12)$$

6 Conclusion

Email is an efficient, fastest, and cost-effective communication system in today's world. In the email, spam is the common problem, which sends data to the large number of recipients for distributing malware, advertising products, phishing, and so on. Spam emails take excess time in deleting junk mails, surplus disk space, and network bandwidth. Therefore, the filter is needed with great precision to filter the unsolicited emails. To filter the spam emails, here feature selection is done using MapReduce minimum redundancy maximum relevance (mRMR) to select the best features for classification. The selected features are classified by using the Random Forests algorithm. In the Random Forests, by means of voting technique, it classifies the junk mails and ham emails. The comparison indicates that the proposed email spam classification gives better accuracy than classification using Random Forests in the distributed environment.

Acknowledgements The authors sincerely thank the University Grants Commission (UGC), Hyderabad for granting funds to carry out this work.

References

1. Zhang, Y., He, J., Xu, J.: A new anti-spam model based on e-mail address concealment technique. *J. Nat. Sci.* **23**(1), 79–83 (2018)
2. Khalaf, O.I., Abdulsahib, G.M., Salman, A.D.: Handling dimensionality reduction in spam e-mail classification. *J. Adv. Res. Dyn. Control Syst.* **10**(1), 691–697 (2018)
3. Bassiouni, M., Ali, M., El-Dahshan, E.A.: Ham and spam e-mails classification using machine learning techniques. *J. Appl. Secur. Res.* **13**(3), 315–331 (2018)
4. Kaur, J., Priyanka: Feature selection based efficient machine learning technique for email spam prediction. *Int. J. Eng. Appl. Sci. Technol.* **2**(12), 13–19 (2018)
5. Vijayasekaran, G., Rosi, S.: Spam and email detection in big data platform using naive bayesian classifier. *Int. J. Comput. Sci. Mob. Comput. (IJCSMC)* **7**(4), 53–58 (2018)
6. Radovic, M., Ghalwash, M., Filipovic, N., Obradovic, Z.: Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinform.* **18**(8), 1–14 (2017)
7. Easwaramoorthy, S., Thamburasa, S., Aravind, K., Bhushan, S.B., Rajadurai, H.: Heterogeneous classifier model for e-mail spam classification using FSO feature selection method. In: *International Conference on Inventive Computation Technologies (ICICT)*, pp. 1–6 (2017)
8. Awad, M., Foqaha, M.: Email spam classification using hybrid approach of Rbfneural network and particle swarm optimization. *Int. J. Netw. Secur. Appl. (IJNSA)* **8**(4), 1–12 (2016)
9. Sri Vinitha, V., Karthika Renuka, D., Bharathi, A.: E-mail spam classification using machine learning in distributed environment. *J. Comput. Theor. Nanosci.* **15**(5), 1688–1694 (2018)
10. Nesi, P., Pantaleo, G., Sanesi, G.: A hadoop based platform for natural language processing of web pages and documents. *J. Vis. Lang. Comput.* **31**, 130–138 (2015)
11. Ramirez-Gallego, S., Lastra, I., Martinez-Rego, D., Bolon-Canedo, V., Benitez, J.M., Herrera, F., Alonso-Betanzos, A.: Fast-mRMR: fast minimum redundancy maximum relevance algorithm for high-dimensional big data. *Int. J. Intell. Syst.* **00**, 1–19 (2016)

12. Ozarkar, P., Patwardhan, M.: Efficient spam classification by appropriate feature selection. *Glob. J. Comput. Sci. Technol. Softw. Data Eng.* **13**(5), 49–57 (2013)
13. Vaishnavi, N., Thiyagarajan, K.: A study on prediction of malicious program using classification based approaches. *Int. J. Comput. Sci. Mob. Comput. IJCSMC* **7**(5), 38–46 (2018)

The Impact of Sustainable Development Report Disclosure on Tax Planning in Thailand



Sathaya Thanjunpong and Thatphong Awirothananon

Abstract This paper aims to examine the impact of sustainable development report disclosure (hereafter called SDRD) on the tax planning (hereafter called TP) of listed companies, which exclude the financial sector, in the Stock Exchange of Thailand. The data of this paper is based only on the year 2016 and the sample size consists of 337 companies from seven industries. The questionnaires from Global Report Initiative are used for evaluation of SDRD. The TP is also measured by the ratio of total tax expenses to total assets. Overall, this paper finds that the SDRD had a statistically negative effect on the TP. This indicates that companies with good SDRD practices could have a higher level of TP. Regarding control variables, financial leverage and capital intensity had a statistically positive effect on the TP, while profitability and family control had a statistically negative effect on the TP. This paper further divides the sample into family and non-family companies to examine whether there is any different effect of SDRD on the TP. The results further indicated that the relationship between SDRD and TP was significantly negative for the family companies. The relationship was, thus, weak and insignificant for the non-family companies. The results are useful to market regulators since they could make some decisions about adjusting some rules and regulations or give some incentives to encourage the Thai listed companies to perform better sustainable development practices.

Keywords Listed companies · Sustainable development report disclosure · Tax planning · Thailand

This paper is our original contributions and has not been submitted or accepted for publication anywhere else. All remaining errors are ours.

S. Thanjunpong · T. Awirothananon (✉)
Faculty of Business Administration, Maejo University, Sansai, Chiang Mai 50290, Thailand
e-mail: thatphong@hotmail.com

S. Thanjunpong
e-mail: sathayat@gmail.com

1 Introduction

Corporate social responsibility (hereafter called CSR), as a consequence of corporate governance (hereafter called CG), in Thailand particularly seems to have become more important. The Corporate Governance Center is established in the year 2002 by the Stock Exchange of Thailand (hereafter called SET) to support the listed companies in developing their CG system by complying with the international assessment regarding the improvement of CG [8]. As a result, the voluntary CG rating was subsequently introduced in the year 2008. The SET was set up [31] to continue the long-term sustainable growth of the capital market in the year 2007 and launched the “Thailand Sustainability Investment” scheme in the fourth quarter of 2015 by applying three criteria, which are economic, social, and governance criteria, respectively. The Thai listed companies and related organisations consequently become more involved with CSR. This is supported by the additional sections on the disclosure reports and the annual report; however, it is not officially required to disclose the CSR section. There is also a significant increase in the number of activities and projects of listed companies. Sustainable development report disclosure (hereafter called SDRD) is also a report published by companies about the practice of measuring, disclosing, and being accountable for their performance from their everyday activities, while they try to operate their business towards the sustainable development. This report could provide balanced and reasonable representations of sustainability performance, including both positive and negative contributions. It could also be considered as non-financial reporting, or triple bottom line reporting, or CSR reporting. The quality of SDRD is expected to increase the overall value creation, which is the increase in the performance of organisation.

The tax planning (hereafter called TP) is also along the continuum depending upon how aggressive the activity is in reducing total tax expenses. The tax avoidance (hereafter called TA) is some activities that reduce companies’ cash-effective tax rate within the legal framework of tax regulations over a long period of time [12, 29]. These activities take some advantages of the loopholes in the tax regulations in order to reduce companies’ tax liabilities [19, 34]. Companies, for example, could choose some advantageous methods in the financial reporting and could reduce their tax liabilities; consequently, the TA for this paper is a part of TP, although in some prior studies, including Chen et al. [6], Lestari and Wardhani [24], Zhang et al. [39], and Ying et al. [38], the effective tax rate is popularly used as the TP measurement. This paper, therefore, has developed a new measurement of TP. It is measured by the ratio of total tax expenses to total assets. This measurement is popularly used as the TP measure in Thailand [35, 36]. Under the assumption of corporation with the same profitability we should have the same total tax expenses at any condition, as well as an earning management of the statement of financial position is more difficult than that of the income statement. This leads to

the conclusion that companies with CSR should have the total tax expenses as well. The companies with higher SDRD practices would, therefore, have higher total tax expenses, which indicate the lower TP. The SDRD quality is also expected to improve the overall value creation, which is the reduction in TP.

Many previous papers show that SDRD has a significant impact on the increasing TP [5, 9, 22, 28]. In developed countries, the authors of [22] investigated a legitimacy theory during the 2001–2006 period in the Australian corporations by comparing the CSR disclosures of tax aggressive corporations with those of non-tax aggressive corporations. Their results show that CSR has a statistically and positively significant effect on the TA. Davis et al. [9] further investigated the relationship between CSR and corporate tax payment in the US-listed companies during the year 2002–2011. Their results show that CSR has a significantly positive effect on the TP. Preuss and Preuss [28] also found that CSR and corporate tax payments act as substitutes in the European public companies. In the developing region, Chen [5] found that the CSR disclosure significantly strengthens the possibility of TP in the Chinese-listed companies from 2008 to 2014. However, many prior studies show that SDRD has a significant impact on the decreasing TP. In the developed countries, Hoi et al. [18] found a statistically negative relationship between irresponsible CSR and TA in the US public companies during the 2003–2009 period. Lanis and Richardson [23] further showed that companies with the higher level of CSR disclosure could have the lower level of corporate TA for the 2008–2009 financial year. The CSR disclosure, therefore, has a significant negative effect on the TA in the Australian listed companies. In the Asia region, Sari and Tjen [30] recently revealed that the CSR disclosure has a significantly negative effect toward the TA in the Indonesian-listed companies during the 2009–2012 period.

Some earlier papers, including Laguir et al. [20], Nadiyah et al. [27], and [33], further showed that SDRD is not significant or cannot summarise the relationship between CSR and TA. In the developed region, Laguir et al. [20] revealed that the relationship between various CSR dimensions and TA is either positive or negative in the French companies during the 2003–2011 period. In the developing countries, Nadiyah et al. [27] further found the relationship between CSR disclosure and corporate TA is either positive or negative in Malaysia in the year 2014. Susanti [33] also investigated the effect of CSR on the TA for manufacturing listed companies in the Indonesia Stock Exchange in the year 2012–2014. These results show that CSR does not have any influence on the TA. On the other hand, SDRD is not associated with the TA. Recent empirical studies find a mixed evidence of the impact of SDRD on the TP. This paper, therefore, aims to search the benefit of SDRD implementation to the TP.

2 Research Methodology

The population of this paper are listed companies, which are excluded financial sectors (58 companies), in the Stock Exchange of Thailand (hereafter called SET) in the year 2016. The purposive sampling method is used for sample selection. The selection is done in accordance with two criteria as follows:

- (a) Companies that publish financial statements, annual report, and Form 56-1 in their website and from the SETSMART database with completed period of 2016.
- (b) Companies that have the sustainable development report during the study period and their necessary data of TP is available.

According to the above two criteria, the sample consists of 337 companies in this paper. These companies are categorised into seven industries according to SET categorisation. It comprises (1) 35 agro and food industry companies, (2) 28 consumer products companies, (3) 64 industrial companies, (4) 74 property and construction companies, (5) 24 resources companies, (6) 81 services companies, and (7) 31 technology companies, respectively.

The dependent variable is the TP, which is calculated as the total tax expenses for the year divided by the total assets. The most important independent variable is SDRD, which is measured by the Global Report Initiative (hereafter called GRI) index score. This index score is calculated from 18 separate criteria to quantify the overall GRI principles from the Global Sustainability Standards Board [14]. The scorecard criteria span three dimensions of the GRI principles, which are economic dimension that has five criteria, environmental dimension that has nine criteria, and social dimension that has four criteria, respectively, as shown in Table 1.

This paper acquires SDRD index scores from Form 56-1, SETSMART database, and annual reports for the year 2016. These index scores are also adjusted to consider the subtleties of Thai laws and regulations. The inclusion of each attribute is scored on a binary basis as “yes” (included) or “no” (not included). Each “yes” answer is equal to the value of one point, while “no” is equal to the value of zero point. This paper also includes some control variables that have been shown to have a statistically significant impact on the TP [36]. This paper uses firm size (hereafter called SIZE), which is measured by the natural logarithm of total assets. The financial leverage variable (hereafter called LEV) is also included in this paper. The LEV variable, which is a proxy of the differences in the financial structure of companies, is computed as the percentage of total debt to total assets. This paper also includes the return on equity (hereafter called ROE) by dividing the net profit to total equity, as an indicator of companies’ financial performance. In addition, this paper uses the capital intensity variable (hereafter called CAP) by dividing the property, plant, and equipment to total assets, as a proxy of companies’ working capital management.

Table 1 SDRD index score

Dimensions	Details
Economy dimension	1. The results on economic growth as net profit of companies were shown
	2. The results on a growth of community economy as nearby community incomes created partly by companies were demonstrated
	3. The results on a partly growth of national economy as corporate income tax and others related to business operation were exhibited
	4. The results on a circular flow of capital among those diverse stockholders in different business companies were presented
	5. The results on corporate information related to policy and campaign against corruption inside their business organisations were proven
Environment dimension	1. Companies that affected human life and exiting organism were demonstrated
	2. Companies that affected ecosystem and quality of water, soil, and air were exhibited, respectively
	3. Companies that affected our natural resources were shown
	4. Companies that affected input factors, namely, energy, and water were demonstrated, respectively
	5. Companies that followed relevant laws and environmental regulations were provided
	6. A disclosure of costs of environmental disposal
	7. Companies that dealt with waste and garbage disposal were provided
	8. Companies that are concerned with their productions and environmental-friendly logistics were demonstrated
	9. Companies that handled with diversified activities like campaign, public relation events, publications, and donation against environmental damage were exhibited, respectively
Society dimension	1. Companies that are concerned with labour forces and employment were demonstrated
	2. Companies that are concerned with human rights were provided
	3. Companies that are concerned with their public accountability were shown
	4. Companies that are concerned with their productions were described

Source Improved from the Global Sustainability Standards (14)

There is a large number of Thai-listed companies that are still family-owned companies, where the family members being the owners have substantial control over both ownership and management [7, 32]. This paper, therefore, uses family control (hereafter called FAM) to separate between family and non-family companies as some previous papers, including Tantiyavarong [35], Chen et al. [4], Landry et al. [21], and Landry et al. [36], show that family companies have a significant impact on the TP. In the developed countries, Chen et al. [4] and Landry

et al. [21] revealed that the relationship between family companies are less TA than their non-family companies because family companies being more concerned with the reputation damage and potential penalty. Bauweraerts and Vandernoot [1] and Martinez and Ramalho [25], however, found that the family companies are more TA than non-family. Recently, Gaaya et al. [13] revealed that the family ownership has a positive effect on corporate TA in Tunisian-listed companies during 2008 and 2013.

In the Asian region, Tantiyavarong [35] found that family companies have positive effects on the TP during the 2001–2007 period. In addition, Thanjunpong and Bangmek [36] further showed that the relationship between family companies is more tax planning than their non-family companies in Thailand during the 2012–2014. Masripah et al. [26], however, found that the entrenchment effect of controlling shareholder has negative effect on the TA in manufacturing industries on the Indonesia Stock Exchange companies during the 2010–2013 period. Furthermore, Bousaidi and Hamed [3] found that the concentration ownership exhibits a negative association with the effective tax rate, which indicates there is a higher TP in Tunisian-listed companies during the 2006–2012 period. Therefore, the FAM variable, which is a dummy variable, has a value of one for the family companies and zero otherwise for non-family companies. The family companies are defined as when a stake of 25% or above of companies' outstanding shares belongs to the major family shareholders.

3 Research Methods

This paper uses the multiple regression analysis technique to determine the relationship between SDRD and TP. According to the objective of this paper, the multiple regression on the full sample model will be utilised as follows:

$$\text{Model 1 : } TP_i = \beta_0 + \beta_1 \text{SDRD}_i + \beta_2 \text{SIZE}_i + \beta_3 \text{LEV}_i + \beta_4 \text{ROE}_i + \beta_5 \text{CAP}_i + \beta_6 \text{FAM}_i + \sum_{j=7}^{12} \beta_j \text{IND}_j$$

In addition, this paper separates the full data to two sub-groups (family and non-family companies) to examine whether there is any different effect of SDRD on the TP among the family companies (Model 2.1) and the non-family companies (Model 2.2) as follows:

$$\text{Model 2 : } TP_i = \beta_0 + \beta_1 \text{SDRD}_i + \beta_2 \text{SIZE}_i + \beta_3 \text{LEV}_i + \beta_4 \text{ROE}_i + \beta_5 \text{CAP}_i + \sum_{j=6}^{11} \beta_j \text{IND}_j$$

4 Results and Discussion

The descriptive statistics for the full and separate sub-samples for family and non-family companies in the SET for the year 2016 are reported in Table 2. It also presents the mean values and the *t*-statistics that test for the mean differences of all variables between family and non-family companies. The results show that the average value of TP for the full sample was 0.011%, which was a low level (representing the TP was a high level), while the values for both family and non-family companies were 0.011 and 0.009%, respectively. The results also reveal that the TP value shows statistically significant difference between family and non-family companies at the 0.10 level. The results show that the family companies had a lower TP than the non-family companies. This implies that the family companies had higher level of TP than the non-family companies.

The values of descriptive statistics in Table 2 show that the mean value of SDRD for the full sample was 11.580. It also shows that the mean value of SDRD for the family companies was 11.730, which is higher when compared with the non-family companies, 11.110. However, the *t*-statistics for the mean difference of SDRD between family and non-family companies was not statistically significant. This indicates that both family and non-family companies had the same level of SDRD implementation. Additionally, the average value of SIZE for the full sample was 22.606, while the mean values for both family and non-family companies were 22.700 and 22.330, respectively. The results indicated that the mean value of SIZE for the family companies was greater than that for the non-family companies, since the *t*-statistics of the mean differences were statistically significant at the 0.05 level. Moreover, the average value of LEV for the full sample was 0.438. The mean value for the family companies, which was 0.437, was also higher than that for the non-family companies (0.442). However, the results indicated that the non-family companies had the same LEV value on the average as the family companies, since

Table 2 Descriptive statistics

Variables	Full sample (<i>n</i> = 337)		Family companies (<i>n</i> = 254)		Non-family companies (<i>n</i> = 83)		<i>t</i> -statistics of mean difference
	Mean	SD	Mean	SD	Mean	SD	
TP	0.011	0.013	0.011	0.014	0.009	0.010	1.865*
SDRD	11.580	3.048	11.730	3.118	11.110	2.789	1.612
SIZE	22.606	1.513	22.700	1.578	22.330	1.263	2.154**
LEV	0.438	0.252	0.437	0.258	0.442	0.237	- 0.168
ROE	0.070	0.776	0.077	0.256	0.046	1.505	0.189
CAP	0.340	0.237	0.343	0.235	0.330	0.244	0.422
FAM	0.754	0.431					

Note One, two, and three asterisks indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively

the *t*-statistics for the mean differences were not statistically significant. Furthermore, the mean value of ROE for the full sample was 0.070, while the values for both family and non-family companies were 0.077 and 0.046, respectively. Furthermore, the average value of CAP for the full sample was 0.034, while the mean values for both family and non-family companies were 0.343 and 0.330, respectively. Nevertheless, both family and non-family companies had the same level of CAP as the *t*-statistics for the mean difference was not statistically significant. The mean value of FAM for the full sample was 0.754.

The Pearson's product-moment correlation, as shown in Table 3, is computed to examine the correlation among all variables of this paper. The correlation between SDRD and TP, for example, was -0.121 , which was statistically and negatively significant. It indicated that higher SDRD practices could lower the TP, representing the TP of companies is at a high level. The correlations among explanatory variables were between -0.025 and 0.369 . For instance, the correlations between CAP and other explanatory variables were not statistically significant. The correlations of FAM and ROE variables with others were also not statistically significant. Overall, the correlation among independent variables was also moderately low (below ± 0.7), indicating that there is no any multicollinearity problem [17]. This paper further considers whether data has the normal distribution or not. According to Berenson et al.'s [2] statement, the sampling distribution becomes almost normal, regardless of the shape of population since the sample size is large enough (which is greater than 30 observations). The data of this paper, therefore, has the normal distribution.

As previously mentioned, this paper uses the multiple regression analysis technique for analysing data and tests for the heteroscedasticity problem. According to the White's [37] test, all models, including the full sample and both family and non-family companies, did not have the problem of heteroscedasticity. The problem of autocorrelation also does not exist in all the models, as the Durbin-Watson [10, 11] statistics are between 1.50 and 2.50 [16]. The values of variance inflation factor (hereafter called VIF) in Table 4 for all models are further less than 10, indicating that the multicollinearity does not exist [15]. Overall, the results in Table 4 show

Table 3 Pearson correlation matrix

Variables	TP	SDRD	Size	LEV	ROE	CAP	FAM
TP	1.000	-0.121^{**}	0.026	0.197^{***}	-0.128^{**}	0.060	-0.101
SDRD		1.000	0.369^{***}	0.034	0.079	0.043	0.088
SIZE			1.000	0.335^{***}	0.068	-0.015	0.105
LEV				1.000	-0.005	-0.025	-0.009
ROE					1.000	-0.018	0.017
CAP						1.000	0.023
FAM							1.000

Note One, two, and three asterisks indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively

Table 4 Effects of sustainable development report disclosure and family control on the tax planning

Independent variables	Full sample (n = 337)			Family companies (n = 254)			Non-family companies (n = 83)		
	β	t-statistics	VIF	β	t-statistics	VIF	β	t-statistics	VIF
Intercept	- 0.013	- 1.025		- 0.041	- 3.151***		0.048	2.416**	
SDRD	- 0.000	- 1.992**	1.204	- 0.000	- 1.657*	1.339	- 0.000	- 0.652	1.203
SIZE	0.000	0.260	1.501	0.002	2.494**	1.718	- 0.003	- 3.235***	1.266
LEV	0.010	3.418***	1.172	0.003	0.984	1.279	0.016	3.347***	1.239
ROE	- 0.002	- 2.183**	1.013	- 0.022	- 7.055***	1.129	- 0.000	- 0.460	1.067
CAP	0.006	1.910*	1.197	0.007	2.151**	1.176	0.003	0.590	1.352
FAM	- 0.003	- 1.855*	1.063						
IND1	- 0.001	- 0.395	2.012	- 0.003	- 0.974	2.085	0.006	1.103	1.888
IND2	0.003	1.012	1.855	0.002	0.607	2.032	- 0.001	- 0.159	1.256
IND3	- 0.002	- 0.870	2.642	- 0.002	- 0.711	2.702	0.001	0.179	2.509
IND4	0.000	0.160	2.694	- 0.002	- 0.555	2.506	0.006	1.572	3.446
IND5	0.000	0.010	1.732	- 0.003	- 0.744	1.747	0.003	0.635	1.745
IND6	- 0.002	- 0.955	2.824	- 0.006	- 1.930*	2.705	0.001	0.263	3.383
Durbin-Watson	1.813			2.080			2.122		
F-statistics	2.964***			6.904***			2.031**		
R ²	0.099			0.239			0.239		
Adjusted R ²	0.066			0.204			0.121		

Note One, two, and three asterisks indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively

that the coefficient for SDRD and TP was statistically significant at the 0.05 level. It indicated that companies with good SDRD practices could decrease the TP, representing the TP of companies is at a high level. These findings are consistent with prior studies, including Hoi et al. [18], Sari and Tjen [30], and Chen [5]. When dividing the data into family and non-family companies, the result indicated that there was no significant relationship between SDRD and TP for the non-family companies. On the other hand, the relationship between SDRD and TP for the family companies was negatively and statistically significant at the 0.10 level. This finding is consistent with previous papers, including Chen et al. [4], Landry et al. [21], and Thanjunpong and Bangmek [36]. Moreover, sustainable development is related to higher disclosure and transparency, higher accounting quality, and audit quality. Therefore, high SDRD companies might be able to decrease their TP, resulting in the TP of companies at a high level.

For the control variables, the relationship between SIZE and TP was not significant for full sample, as presented in Table 4. When dividing full sample into two sub-groups (family and non-family companies), the results show that the relationship between SIZE and TP was positively and statistically significant at the 0.05 level for the family companies, while the relationship was negatively and statistically significant at the 0.01 level for the non-family companies. Similarly, the effects of LEV on the TP for the full sample and the non-family companies were positively and statistically significant at the 0.01 level. Moreover, the effects of ROE on the TP for the full sample and the family companies were negatively significant at the 0.01 level. In addition, the effects of CAP on the TP for the full sample and the family companies were positively and statistically significant, while the effect of CAP on the TP for the non-family companies was not statistically significant. Finally, the relationship between FAM and TP was negatively statistically significant at the 0.10 level. This relationship is consistent with prior studies, including Chen et al. [4], Landry et al. [21], and Thanjunpong and Bangmek [36]. The relationship between IND and TP, however, was not statistically significant for the full sample.

This paper further does an additional analysis by excluding the industry variable, which is a dummy variable, for a robustness investigation. All models are also tested for the problems of autocorrelation and multicollinearity. The results in Table 5 show that the Durbin–Watson [10, 11] statistics are between 1.50 and 2.50, resulting that the autocorrelation problem does not exist [16]. Each variable of all models also has the VIF value less than 10, indicating that the problem of multicollinearity does not exist [15]. The results in Table 5 show that the coefficient for SDRD and TP was negatively and statistically significant at the 0.05 level. It indicates that the TP could be reduced, when companies have a good SDRD implementation. This implies that companies have a high level of TP. These findings are consistent with previous papers, including Hoi et al. [18], Sari and Tjen [30], and Chen [5]. Once dividing the full sample into family and non-family companies, the result indicated that relationship between SDRD and TP for the non-family companies was not statistically significant. However, the relationship between SDRD and TP for the family companies was negatively and statistically

Table 5 Effects of sustainable development report disclosure and family control on the tax planning (robustness analysis)

Independent variables	Full sample (n = 337)		Family companies (n = 254)		Non-family companies (n = 83)				
	β	t-statistics	VIF	β	t-statistics	VIF	β	t-statistics	VIF
Intercept	- 0.013	- 1.269		- 0.040	- 3.339***		0.035	1.863**	
SDRD	- 0.001	- 2.141**	1.178	- 0.001	- 2.044**	1.295	- 0.000	- 0.471	1.135
SIZE	0.000	0.418	1.329	0.001	2.502**	1.480	- 0.002	- 2.614**	1.115
LEV	0.010	3.146***	1.141	0.003	0.894	1.242	0.017	3.594***	1.183
ROE	- 0.002	- 2.188**	1.009	- 0.021	- 6.821***	1.084	- 0.000	- 0.524	1.011
CAP	0.004	1.328	1.004	0.005	1.663*	1.005	0.000	0.028	1.047
FAM	- 0.003			- 1.713*			1.016		
Durbin-Watson	1.780			2.005			2.016		
F-statistics	4.883***			13.303***			3.412***		
R ²	0.082			0.211			0.181		
Adjusted R ²	0.065			0.196			0.128		

Note One, two, and three asterisks indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively

significant at the 0.05 level. These results were similar to the above results, which were reported in Table 4, and prior studies. Consequently, sustainable development is associated with higher disclosure and transparency, higher accounting quality, and audit quality. As a result, high SDRD companies might be able to decrease their TP, representing the TP of companies is at a high level.

5 Conclusion

This paper investigates the effect of SDRD on the TP of listed companies, excluding financial sectors, in the SET. The sample size consists of 337 companies, based only on the data for the year 2016. The questionnaires of GRI are used for evaluation of SDRD. The TP is also measured by the ratio of total tax expenses to total assets. Overall, this paper finds that the SDRD had a significant and negative effect on the TP. This indicates that companies could have a higher level of TP, when they have higher SDRD implementation. Regarding some control variables, the financial leverage and capital intensity variables had a positive effect on the TP. However, the profitability and family control variables had a significantly negative impact on the TP. In addition, the relationship between SDRD and TP was significantly negative for the family companies. The relationship was, however, weakly insignificant for the non-family companies. As a result, the results are useful to tax officers for making a decision of initial audit. They, for instance, should pay attention on companies that have low SDRD implementation. Moreover, the results are useful to the SET and Securities and Exchange Commission (hereafter called SEC). They could further make some decisions about adjusting some rules and regulations or give some incentives to encourage the Thai-listed companies to perform better SDRD practices. For example, market regulators, including SET and SEC, should focus on family companies and encourage them to improve their SDRD implementation by giving them some incentives, and the results indicate that the relationship between SDRD and TP for the family companies is negatively significant.

References

1. Bauweraerts, J., Vandernoot, J.: Are family firms more tax aggressive than non-family firms? Empirical evidence from Belgium. *Int. J. Manag.* **30**(4), 235–243 (2013)
2. Berenson, M.L., Levine, D.M., Krehbiel, T.C.: *Basic Business Statistics: Concept and Application*, 12th edn. Pearson Education, New Jersey (2012)
3. Boussaidi, A., Hamed, M.S.: The impact of governance mechanisms on tax aggressiveness: empirical evidence from Tunisian context. *J. Asian Bus. Strategy* **5**(1), 1–12 (2015)
4. Chen, S., Chen, X., Cheng, Q., Shevlin, T.: Are family firms more tax aggressive than non-family firms? *J. Financ. Econ.* **95**(1), 41–61 (2010)

5. Chen, X.: Corporate social responsibility disclosure, political connection and tax aggressiveness: evidence from China's capital markets. *Open J. Bus. Manag.* **6**(1), 151–164 (2018)
6. Chen, X., Hu, N., Wang, X., Tang, X.: Tax avoidance and firm value: evidence from China. *Nankai Bus. Rev. Int.* **5**(1), 25–42 (2014)
7. Chienwittayakun, J., Mankin, D.: Strategic management planning process (SMPP) as an organization development intervention (ODI) to align values, goals and objectives and improve employee teamwork, engagement and performance: a case study of a family-owned business in Thailand. *ABAC J. Vis. Action Outcome* **2**(1), 74–92 (2015)
8. Corporate Governance Center: Good governance assessment of listed companies. The Stock Exchange of Thailand, Bangkok (2003)
9. Davis, A.K., Guenther, D.A., Krull, L.K., Williams, B.M.: Do socially responsible firms pay more taxes? *Account. Rev.* **91**(1), 47–68 (2016)
10. Durbin, J., Watson, G.S.: Testing for serial correlation in least squares regression: I. *Biometrika* **37**(3/4), 409–428 (1950)
11. Durbin, J., Watson, G.S.: Testing for serial correlation in least squares regression. II. *Biometrika* **38**(1/2), 159–177 (1951)
12. Dyreng, S.D., Hanlon, M., Maydew, E.L.: Long-run corporate tax avoidance. *Account. Rev.* **83**(1), 61–82 (2008)
13. Gaaya, S., Lakhal, N., Lakhal, F.: Does family ownership reduce corporate tax avoidance? The moderating effect of audit quality. *Manag. Audit. J.* **32**(7), 731–744 (2017)
14. Board, Global Sustainability Standards: G4 Sustainability Reporting Guidelines. Global Reporting Initiative, Amsterdam (2013)
15. Greene, W.H.: *Econometric Analysis*, 8th edn. Pearson, New York (2018)
16. Gujarati, D.N.: *Basic Econometrics*, 4th edn. McGraw-Hill, New York (2004)
17. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: *Multivariate Data Analysis*, 7th edn. Pearson Education, New Jersey (2010)
18. Hoi, C.K., Wu, Q., Zhang, H.: Is corporate social responsibility (CSR) associated with tax avoidance? Evidence from irresponsible CSR activities. *Account. Rev.* **88**(6), 2025–2059 (2013)
19. Hope, O.-K., Ma, M., Thomas, W.B.: Tax avoidance and geographic earnings disclosure. *J. Account. Econ.* **56**(2–3), 170–189 (2013)
20. Laguir, I., Staglianò, R., Elbaz, J.: Does corporate social responsibility affect corporate tax aggressiveness? *J. Clean. Prod.* **107**, 662–675 (2015)
21. Landry, S., Deslandes, M., Fortin, A.: Tax aggressiveness, corporate social responsibility, and ownership structure. *J. Acc. Ethics Public Policy* **14**(3), 611–645 (2013)
22. Lanis, R., Richardson, G.: Corporate social responsibility and tax aggressiveness: a test of legitimacy theory. *Acc. Audit. Account. J.* **26**(1), 75–100 (2012)
23. Lanis, R., Richardson, G.: Corporate social responsibility and tax aggressiveness: an empirical analysis. *J. Account. Public Policy* **31**(1), 86–108 (2012)
24. Lestari, N., Wardhani, R.: The effect of the tax planning to firm value with moderating board diversity. *Int. J. Econ. Financ. Issues* **5**(1S), 10–11 (2015)
25. Martinez, A.L., Ramalho, G.C.: Family firms and tax aggressiveness in Brazil. *Int. Bus. Res.* **7**(3), 129–136 (2014). <https://doi.org/10.5539/ibr.v7n3p129>
26. Masripah, M., Diyanty, V., Fitriasar, D.: Controlling shareholder and tax avoidance: family ownership and corporate governance. *Int. Res. J. Bus. Stud.* **8**(3), 167–180 (2015)
27. Nadiyah, A.H., Syakirah, W.N.W.K., Mastora, Y., Rohayu, Y., Rozainun, A.A.: Corporate social responsibility (CSR) disclosure and its impacts towards corporate tax aggressiveness. *J. Appl. Environ. Biol. Sci.* **7**(5S), 10–15 (2017)
28. Preuss, A., Preuss, B.: Corporate tax payments and corporate social responsibility: complements or substitutes? Empirical evidence from Europe. *Bus. Econ. J.* **8**(4), 1–8 (2017)
29. Sandmo, A.: The theory of tax evasion: a retrospective view. *Natl. Tax J.* **58**(4), 643–663 (2005)
30. Sari, D., Tjen, C.: Corporate social responsibility disclosure, environmental performance, and tax aggressiveness. *Int. Res. J. Bus. Stud.* **9**(2), 93–104 (2016)

31. Social Responsibility Center: Sustainable Development Meaning. Wanida Printing Limited Partnership, Nonthaburi (2014)
32. Suehiro, A., Wailersak, N.: Family business in Thailand its management, governance, and future challenges. *ASEAN Econ. Bull.* **21**(1), 81–93 (2004)
33. Susanti, M.: Corporate social responsibility, size and tax avoidance. *Int. J. Econ. Perspect.* **11**(1), 1639–1650 (2017)
34. Tang, T., Firth, M.: Can book–tax differences capture earnings management and tax Management? Empirical evidence from China. *Int. J. Acc.* **46**(2), 175–204 (2011)
35. Tantiyavarong, T.: A Study of the Determinants of Tax Planning and the Association between Tax Planning and Firm Value: An Empirical Evidence of Thailand. Chulalongkorn University, Bangkok (2009)
36. Thanjunpong, S., Bangmek, R.: The influence of board of directors, audit committee and ownership structure impact on tax planning: an empirical evidence of Thailand. *J. Acc. Prof.* **13**(37), 29–44 (2017)
37. White, H.: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**(4), 817–838 (1980)
38. Ying, T., Wright, B., Huang, W.: Ownership structure and tax aggressiveness of Chinese listed companies. *Int. J. Acc. Inf. Manag.* **25**(3), 313–332 (2017)
39. Zhang, C., Cheong, K.C., Rasiah, R.: Corporate tax avoidance and performance: evidence from China’s listed companies. *Inst. Econ.* **8**(3), 61–83 (2016)

Clustering and Labeling Auction Fraud Data



Ahmad Alzahrani and Samira Sadaoui

Abstract Although shill bidding is a common fraud in online auctions, it is however very tough to detect because there is no obvious evidence of it happening. There are limited studies on SB classification because training data are difficult to produce. In this study, we build a high-quality labeled shill bidding dataset based on recently scraped auctions from eBay. Labeling shill bidding instances with multidimensional features is a tedious task but critical for developing efficient classification models. For this purpose, we introduce a new approach to effectively label shill bidding data with the help of the robust hierarchical clustering technique CURE. As illustrated in the experiments, our approach returns remarkable results.

Keywords Auction fraud · Shill bidding · Hierarchical clustering · CURE · Silhouette · Data labeling

1 Introduction

In the last three decades, we witnessed a significant increase in exchanging goods and services over the Web. According to the World Trade Organization, the worldwide merchandise during the period 1995–2015 was over 18 billion.¹ Online auctions are a very profitable e-commerce application. For instance, in 2017, eBay claimed that the net revenue attained 9.7 billion US dollars, and the number of active users hits 170 million.² Regardless of their popularity, e-auctions remain very vulnerable to cybercrimes. The high anonymity of users, low fees of auction services, and flexibility

¹https://www.wto.org/english/res_e/statis_e/its2015_e/its2015_e.pdf.

²<https://www.statista.com>.

A. Alzahrani (✉) · S. Sadaoui
Computer Science Department, University of Regina, Regina, SK, Canada
e-mail: alzah234@uregina.ca

S. Sadaoui
e-mail: sadaouis@uregina.ca

of bidding make auctions a great incubator for fraudulent activities. The Internet Crime Complain Center announced that auction fraud is one of the top cybercrimes [2]. As an example, the complaints about auction fraud in only three states, California, Florida, and New York, reached 7,448 in 2016 [2]. Malicious moneymakers can commit three types of fraud, which are pre-auction fraud, such as auctioning of black market merchandise, in-auction fraud that occurs during the bidding time, such as Shill Bidding (SB), and post-auction fraud, such as fees stacking. Our primary focus is on the SB fraud whose goal is to increase the profits of sellers by placing many bids through fake accounts and colluding with other users. SB does not leave any obvious evidence, unlike the two other auction fraud. Indeed, buyers are not even aware that they have been overcharged.

Identifying relevant SB strategies, determining robust SB metrics, crawling and preprocessing commercial auction data, and evaluating the SB metrics against the extracted data make the study of SB fraud very challenging as demonstrated in our previous technical paper [1]. In addition, labeling SB instances with multidimensional features is a critical phase for the classification models. In the literature, labeling training data is usually done manually by the domain experts, which is quite a laborious task and prone to errors. Due to the lack of labeled SB training datasets, the prime contribution of this paper is to produce high-quality labeled SB data based on commercial auction transactions that we extracted from eBay and preprocessed [1]. As illustrated in Fig. 1, we introduce a new approach to effectively label SB data with the help of data clustering. First, we split the SB dataset into several subsets according to the different bidding durations of the extracted auctions. Second, we efficiently partition each SB subset into clusters of users with similar bidding behavior. Last, we apply a systematic labeling method to each cluster to classify bidders into normal or suspicious.

Hierarchical clustering is significantly preferable over partitioning clustering because it provides clusters with a higher quality [10]. This type of data clustering has been utilized successfully in numerous fraud studies [8, 10]. In fact, we employ the Clustering Using REpresentatives (CURE) technique to produce the best differentiation between normal and suspicious activities. CURE [11] has proved over the years to be a highly efficient clustering method in terms of eliminating outliers and producing high-quality clusters, especially for large-scale training datasets. The

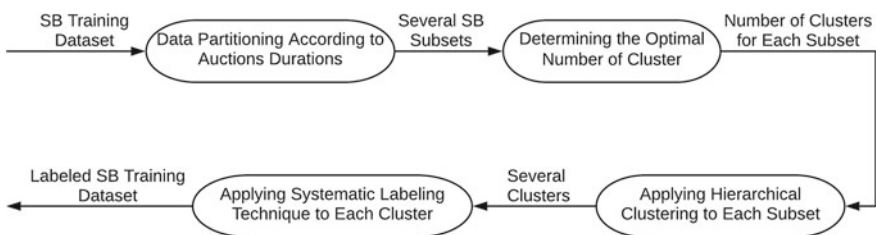


Fig. 1 The labeling process of shill bidding training data

labeled SB dataset that we produced can be utilized by the state-of-the-art supervised classification methods. Furthermore, the accuracy of new predictive models can also be tested using our SB training dataset.

The rest of the paper is organized as follows. Section 2 discusses related work on data clustering in the context of online auctions. Section 3 describes the SB patterns used in this research as well as the SB data produced from the commercial website eBay. Section 4 explains how to apply the hierarchical clustering CURE according to different bidding durations. Section 5 exposes a new approach to label the clusters of bidders. Finally, Sect. 6 summarizes the results of this study and highlights important research directions as well.

2 Related Work

Many researchers utilized data clustering to examine auctions from different angles, such as studying the dynamics of auction prices and bidder behavior. For instance, [13] introduced an approach to model and analyze the price formation as well as its dynamics to characterize the heterogeneity of the price formation process. The proposed functional objects represent the price process by accommodating the structure format of bidding data on eBay. Then, the curve clustering is used to partition auctions by grouping similar price profiles. Finally, differential equations are used to specify the price of each group.

Another work [7] measured the similarity of bidder behavior using specific attributes, such as bidder feedback rating, average increment difference, and number of bids. Then, a centroid-based hierarchical clustering approach is presented to group similar bidders. Each produced cluster is then labeled manually according to the overall bidder behavior in that cluster.

The study [14] suggested an SB detection model utilizing k-mean clustering technique. The latter groups similar buyers in one class to differentiate between general buyers and shill bidders. There are four features that represent a buyer: “how long the buyer has been in the auction,” “the times of buyer bids,” “the average response time of the buyer,” and “the absolute average discrepancy of buyer bids.” Based on these features, k-mean classifies bidders into one of the two clusters: general buyers and shill bidders. However, we believe the second and third features do not really reflect SB since a buyer might be very interested in winning the auction. Besides, there are stronger patterns that highly identify SB. Since SB behavior is somehow similar to real bidding [5], there is a possibility that some SB samples fell in the normal class.

In [4], the authors proposed a two-step clustering model to recognize bidding strategies. The hierarchical clustering is the first step to produce a dendrogram and an agglomeration schedule table to find the best number of clusters. The next step employs k-mean clustering to provide more details about the bidders’ strategies in each cluster. The experiment was operated on an outdated data (2003) collected

from Taobao.com. According to the agglomeration coefficients, the optimal number of clusters is three: “early bidding strategy”, “snipe bidding strategy,” and the third one groups bidders that enter the auction early and remain for a long time. The first cluster has shown low values for the given features, which indicates that the bidders’ strategy is to enter and exit auctions early and participate infrequently. The second cluster has displayed high values for some features and low values for others. This illustrates that bidders enter and exit auctions late and rarely participate. The final cluster has administered the bidders’ strategy where bidders enter early and stay for a long time and highly participate in the auctions.

Zhang et al. [20] proposed a model based on hedonic regression and Fuzzy Logic Expert System (FLES) to analyze bidder behavior. The hedonic regression is used to select key variables that are passed to FLES to produce a knowledge base about the relationships between variables, like auction characteristics. Since the examined data have no relational information, k-mean is employed to obtain the minimum squared-error clusters. So, each training sample is classified to low, medium, or high membership degree. The issue here is that the study is based on an outdated dataset (2004). Also, there is a potential for fabricating feedback ratings conducted between shill bidders and fraudulent bidder rings [7].

More recently, [12] applied k-mean clustering to categorize bidders’ habits. The observations are obtained by the k-mean and then passed to the Baum–Welch algorithm and Hidden Markov model. Three main clusters were suggested according to the values of the given features, which are low, medium, and high cluster values. A bidder habit with values beyond these clusters values is considered as a fraudster. The experiment showed that only two simple features were given to identify the clusters: the number of auctions that a bidder participated in and the number of submitted bids by that bidder. Thus, if more features were considered to define the clusters, then the samples’ distribution on each cluster might be changed. As a result, this may influence the outcomes of the detection model. Also, the clustering is based on a dataset that is not adequately described, and only ten samples were used for explaining the results.

Lastly, [10] applied a hierarchical clustering to group users with similar bidding behavior. The centroid linkage is used as the similarity measure. The described SB patterns were computed for all the bidders of each of the generated clusters. Then, the authors introduced a semi-automatic approach to label each cluster according to the general behavior of users in that cluster and the weights of the fraud patterns.

3 Shill Bidding Overview

SB is a well-known auction fraud, and yet it is the most difficult to detect since it behaves similarly to normal bidding [5, 8]. The aim of this fraud is to increase the price or desirability of the auctioned product through imitation accounts and collusion

with other users. In other words, shill bidders do not tend to win the auction but to increase the revenue of the seller. SB leads buyers to overpay for the items, and for high priced items, buyers will lose a substantial amount of money. As mentioned in [5], excessive SB could lead to market failure. Thus, e-auctions may lose their credibility [5]. In fact, several sellers and their accomplices have been prosecuted due to SB activities, including

- In 2007, a jewelery seller was accused of conducting SB fraud on eBay and had to pay \$400,000 for a settlement. Also, he and his employees were prevented from engaging in any online auctioning activities for 4 years.³
- In 2010, a seller faced a £50,000 fine after being found outbidding himself on eBay. He claimed that “eBay let me open up the second account and I gave all my personal details and home address to do so.”⁴
- In 2012, the online auction Trade Me had to pay \$70,000 for each victim after the investigation discovered SB fraud conducted by a motor vehicle trader in Auckland. The fraud was carried out for 1 year and caused a significant loss for the victims. Trade Me blocked this trader from using their site, and referred the case to the Commerce Commission for a further investigation.⁵
- In 2014, a lawsuit was filed against Auction.com by VRG in California claiming that the website allowed SB. The bid of \$5.4 million should have secured the property as the plaintiff declared, and yet the winning price was 2 million more. Auction.com was accused of helping the property loan holder, which is not fair for genuine bidders. The California state passed a law on July 1, 2015, which requests the property auctioneers to reveal bids they submit on a seller’s behalf.⁶ The spokeswoman for the California Association of Realtors said “*To the best of our knowledge, we are the only state to pass this sort of legislation, even though we believe shill bidding to be prevalent all over the country.*”

By examining throughly the literature on the SB strategies [6, 8, 17], we compiled in Table 1, the most relevant SB patterns. Each pattern, which is a training feature, represents a unique aspect of the bidding behavior in auctions. The feature uniqueness will lead to an improved predictive performance.

³<https://www.nytimes.com/2007/06/09/business/09auction.html>.

⁴<http://www.dailymail.co.uk/news/article-1267410/Ebay-seller-faces-fine-bidding-items-raise-prices.html>.

⁵<https://www.trademe.co.nz/trust-safety/2012/9/29/shill-bidding>.

⁶<https://nypost.com/2014/12/25/lawsuit-targets-googles-auction-com>.

Table 1 SB patterns and their characteristics

Name	Definition	Category	Source	Weight
Bidder tendency (BT)	Engages exclusively with few sellers instead of a diversified lot	Bidder	User history	0.5
Bidding ratio (BR)	Participates more frequently to raise the auction price	Bid	Bidding period	0.7
Successive outbidding (SO)	Successively outbids himself even though he is the current winner	Bid	Bidding period	0.7
Last bidding (LB)	Becomes inactive at the last stage to avoid winning	Bid	Last bidding stage	0.5
Early bidding (EB)	Tends to bid pretty early in the auction to get users attention	Bid	Early bidding stage	0.3
Winning ratio (WR)	Participates a lot in many auctions but rarely wins any auctions	Bidder	Bidder history	0.7
Auction bids (AB)	Tends to have a much higher number of bids than the average of bids in auctions selling the same product	Auction	Auction history	0.3
Auction starting price (ASP)	Offers a small starting price to attract genuine bidders	Auction	Auction history	0.3

4 Production of Shill Bidding Data from Online Auctions

4.1 Auction Data Extraction and Preprocessing

To obtain a reliable SB training dataset, it must be built from actual auction data. Nevertheless, producing high-quality auction data is itself a burdensome operation due to the difficulty of collecting data from auction sites on one hand, and the challenging task of preprocessing the raw data on the other hand. The latter consumes a significant time and effort, around 60–80% of the entire workload [15]. In our technical study [1], we employed the professional scraper Octoparse⁷ to collect a large number of auctions for one of the most popular products on eBay. The extracted dataset contains all the information related to auctions, bids and bidders. We crawled completed auctions of the iPhone 7 for 3 months (March to June 2017). We chose iPhone 7 because it may have attracted malicious moneymakers due to the following facts:

- Its auctions attracted a large number of bids and bidders.
- It has a good price range with the average of \$578.64 (US currency). Indeed, there is a direct relationship between SB fraud and the auction price [5].
- The bidding duration varies between 1 (20.57%), 3 (23.2%), 5 (16.23%), 7 (38.3%), and 10 (1.7%) days. In long duration, a dishonest bidder may easily mimic usual bidding behavior [5]. However, as claimed in [3], fraudulent sellers may receive a positive rating in short duration. Thus, we considered both durations.

⁷<https://www.octoparse.com>.

Table 2 Preprocessed auctions of iPhone 7

No. of auctions	807
No. of records	15145
No. of bidder IDs	1054
No. of seller IDs	647
Avg. winning price	\$578.64
Avg. bidding duration	7
No. of attributes	12

Table 2 presents the statistics after preprocessing the scraped auction data. This operation was very time consuming as it required several manual operations [1]: (1) removing redundant and inconsistent records, and also records with missing bidder IDs; (2) merging several attributes into a single one; (3) converting the format of several attributes into a proper one; (4) assigning an ID to the auctions. For example, the two attributes date and time in each auction are converted into seconds; as an example, 1-day and 10-day durations are converted into 86,400 and 864,000 seconds, respectively.

4.2 SB Data Production

The algorithms to measure the SB patterns are presented in [1, 17]. Each metric is scaled to the range of [0, 1]; a high value indicates a suspicious bidding behavior. We evaluated each metric against each bidder in each of the 807 auctions [1]. Therefore, we obtained a SB training dataset with a total of 6321 samples. A sample is described as a vector of 10 elements: the eight SB features along with Auction, and Bidder identification numbers. Once labeled, an instance will denote the conduct (normal or suspicious) of a bidder in a certain auction.

5 Hierarchical Clustering of Shill Bidding Data

Since the produced SB data are not labeled, data clustering, an unsupervised learning method, can be utilized to facilitate the labeling operation. Clustering is the process of isolating instances into K groups w.r.t. their similarities. The clustering techniques fall into one of the following categories: (1) Partitioning-based, such as K-medoids and K-means; (2) Hierarchical-based, such as BIRCH, GRIDCLUST and CURE; (3) Density-based, such as DBSCAN and DBCLASD; (4) Grid-based, such as STING and CLIQUE. In our work, we select agglomerative (bottom-top) hierarchical clustering where instances are arranged in the form of a tree structure using a proximity matrix.

5.1 CURE Overview

Among the hierarchical clustering methods, we choose CURE because it is highly performant in handling large-scale multidimensional datasets, determines non-spherical shapes of the clusters, and efficiently eliminates outliers [11]. Random sampling and partitioning techniques are utilized to handle the large-scale problem and to speed up the clustering operation. Each instance is first considered as an individual cluster, and then the cluster with the closest distance/similarity is merged into it in order to form a new cluster [11]. Two novel strategies have been introduced in CURE:

- **Representative Points (RPs)**, which are selected data points that define the cluster boundary. Instead of using a centroid, clusters are identified by a fixed number of RPs that are well dispersed. Clusters with the closest RPs are merged into one cluster. The multiplicity of RPs allows CURE to obtain arbitrary clustering shapes.
- **Constant shrinking factor (α)**, which is utilized to shrink the distance of RPs toward the centroid of the cluster. This factor reduces noise and outliers.

The worst-case computational complexity of CURE is estimated to $O(N^2 \log N)$, which is high when N is large (N is the number of instances) [18]. Since the SB data clustering is an offline operation, so the running time is not an issue. The only disadvantage of CURE is that the two parameters RP and α have to be set up by users.

To run the experiments, we utilize the Anaconda Navigator environment for running Python 3, and incorporate the CURE program developed by Freddy Stein and Zach Levonian. CURE code (in Python) is available at GitHub.com.⁸

5.2 SB Data Preparation

Since the bidding duration is used as a denominator for the two patterns Early Bidding and Last Bidding, the large gap between different durations greatly affects the computation results. The computed value of the fraud pattern for 10 days is far smaller than for 1 day. So, before applying CURE, we first partition the SB dataset into five subsets according to the five durations (1, 3, 5, 7, and 10 days) as presented in Table 3.

5.3 Optimal Number of Clusters

It is always difficult to decide about the optimal number of clusters of a training dataset. Besides, normal and skill bidder behavior is somehow similar. Thus, deter-

⁸<https://github.com/levoniaz/python-cure-implementation/blob/master/cure.py>.

Table 3 SB dataset partitioning according to bidding durations

Subset	1 Day	3 Days	5 Days	7 Days	10 Days
No. of auctions	166	187	131	309	14
No. of Instances	1289	1408	1060	2427	137

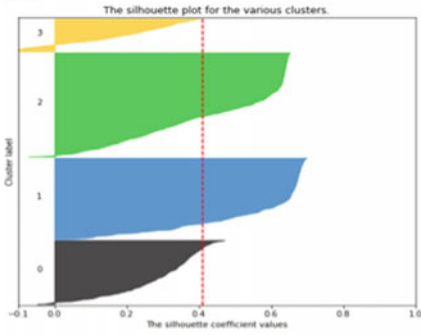
mining the best number of clusters is an essential step to achieve a better interpretation for classifying similar SB instances [19]. There are several methods to address this problem, such as Elbow, Dendrogram, the Rule of Thumb and Silhouette. In our study, we employ the Silhouette method where each group is represented as a silhouette based on the separation between instances and the cluster's tightness [16]. The construction of a silhouette requires the clustering technique to generate the partitions and also to collect all approximates between instances [16]. K-mean clustering algorithm has been successfully utilized for this task due to its simplicity and effectiveness [19]. Consequently, we apply K-mean to estimate the number of clusters for each of the five SB subsets.

Next, for each subset, we examine the silhouette scores of 19 clusters (2–20 clusters), and choose the best number based on the best silhouette score. We have noticed that once the peak (denoting the optimal number) of the silhouette score is reached on a certain number of clusters, the silhouette score gradually decreases with the increasing of the number of clusters. In Fig. 2, we give an example for the 7-day bidding duration for which the optimal number of clusters is eight since the highest silhouette score (0.4669) is obtained on that number. In Table 4, we expose the best number of clusters for each of the five SB subsets. The total number of produced clusters is 29.

5.4 Cluster Generation

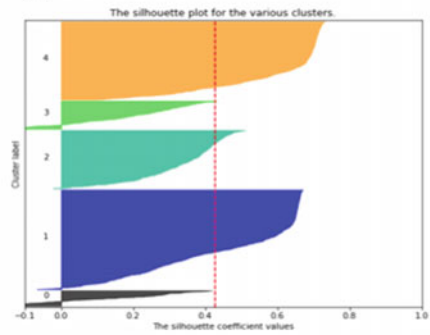
CURE has three parameters that need to be setup: representative points (RPs), shrinking factor (α) and an optimal number of clusters. Based on the results of silhouette, we have obtained the optimal number of clusters for each of the five SB subsets. The two parameters RPs and α are defined by selecting the configuration that provides the best instance distribution among the specified clusters. Thus, CURE is applied with different values of RPs and α starting from the default values (5 for RPs and 0.1 for α). We thoroughly conducted trial-and-error experiments for all the clusters (29 clusters in total) of the five SB subsets to determine the best values of the parameters (Table 4). The best parameters' values are selected based on the best distribution of a subset population between the defined clusters. As an example, in Table 5, we present the results for the eight clusters that we have generated previously for the 7-day bidding duration subset. The best value configuration is shown in bold. Each

(a) 4 clusters



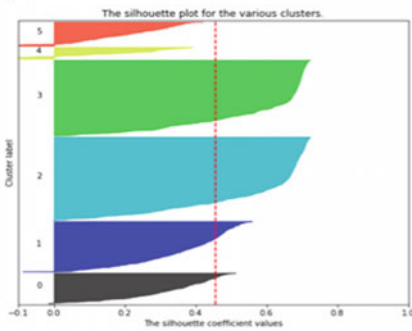
Silhouette score is 0.4104

(b) 5 clusters



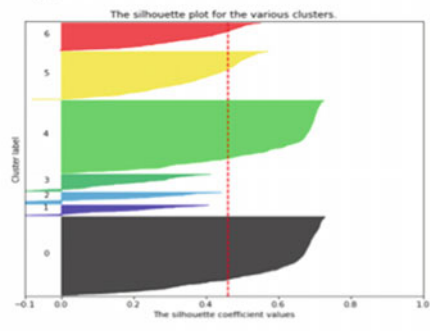
Silhouette score is 0.4282

(c) 6 clusters



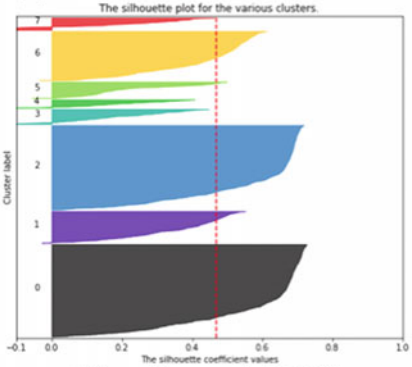
Silhouette score is 0.4571

(d) 7 clusters



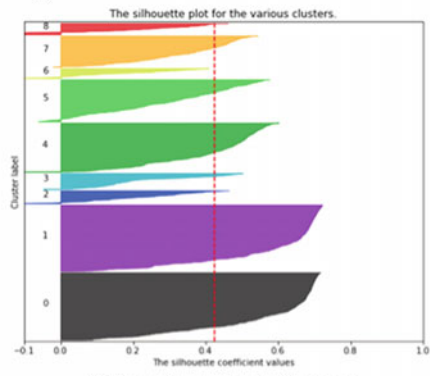
Silhouette score is 0.4608

(e) 8 clusters



Silhouette score is 0.4669

(f) 9 clusters



Silhouette score is 0.4250

Fig. 2 Optimal number of clusters for 7-day bidding duration. Silhouette score is examined 19 times. We show the top Silhouette scores

Table 4 Optimal number of clusters and optimal CURE parameters

SB subset	No. of samples	No. of clusters	Silhouette score	RPs	α
1 Day	1289	7	0.4597	5	0.05
3 Days	1408	7	0.4672	5	0.01
5 Days	1060	5	0.4758	5	0.05
7 Days	2427	8	0.4669	10	0.001
10 Days	137	2	0.5549	5	0.1

Table 5 CURE clustering for 7-day bidding duration (8 clusters)

RP	α	Cl. # 1	Cl. # 2	Cl. # 3	Cl. # 4	Cl. # 5	Cl. # 6	Cl. # 7	Cl. # 8
5	0.1	136	1438	1	2	657	190	2	1
5	0.05	657	328	2	1408	1	28	2	1
5	0.01	1438	640	1	1	17	1	1	328
5	0.001	21	166	2	1	2	25	2209	1
10	0.1	2	1	657	1410	22	8	137	190
10	0.05	2	133	2	1	2	31	2066	190
10	0.01	1	1	1	1	135	31	2067	190
10	0.001	189	654	1410	3	137	31	1	2

cluster consists of users with similar bidding behavior. As we can see, the clusters 4, 7, and 8 have very few bidders, which are most probably outliers i.e., suspicious.

6 Labeling Shill Bidding Data

In Algorithm 1, we show the steps to label the bidders of a given cluster. A cluster belongs to a certain SB subset. As shown in Table 6, for each subset, we first compute the mean and Standard Deviation (STD) of each fraud pattern for all the instances in that subset. Then, we compute the average of the means (Avg. Means) and average of the STDs (Avg. STDs) of all the patterns for that subset. We consider the value of $(Avg. Means + \frac{1}{2} Avg. STDs)$ since it produces the best decision line that separates between normal and suspicious instances as depicted in Fig. 3. Then, we calculate the average of the means of all the patterns for the cluster. So, if the average mean of the cluster is greater than the decision line of the subset, then instances are labeled as suspicious (1) in that cluster, otherwise, they are labeled normal (0).

Table 6 Characteristics of SB subsets

Subset	1 Day	3 Days	5 Days	7 Days	10 Days
<i>Mean of each pattern per subset</i>					
BT	0.1434	0.1394	0.1419	0.1455	0.1162
BR	0.1287	0.1328	0.1235	0.1273	0.1021
SO	0.0996	0.1047	0.0872	0.1149	0.0620
LB	0.4624	0.4511	0.4676	0.4678	0.4746
EB	0.4314	0.4192	0.4318	0.4348	0.4575
WR	0.3812	0.3718	0.3810	0.3533	0.3496
AB	0.2120	0.1936	0.2403	0.2567	0.2926
ASP	0.5007	0.4301	0.4478	0.4801	0.7123
Avg. means	0.2949	0.2802	0.2901	0.2975	0.3208
<i>STD of each pattern per subset</i>					
BT	0.1973	0.1884	0.1984	0.2019	0.1811
BR	0.1246	0.1330	0.1243	0.1377	0.1165
SO	0.2764	0.2811	0.2583	0.2917	0.2215
LB	0.3773	0.3753	0.3917	0.3783	0.3931
EB	0.3775	0.3742	0.3921	0.3802	0.3968
WR	0.4356	0.4373	0.4402	0.4345	0.4398
AB	0.2323	0.2426	0.2646	0.2658	0.2575
ASP	0.4931	0.4831	0.4863	0.4908	0.4510
Avg. STDs	0.3142	0.3143	0.3194	0.3226	0.3071

Algorithm 1 : Labeling Bidders in a Cluster**Require:** AvgMeans and AvgSTDs of the corresponding SB subset

```

1: Compute MeanCluster
2: if ( $MeanCluster \geq (AvgMeans + \frac{AvgSTDs}{2})$ ) then
3:   for  $x=1$  to NumberBiddersCluster do
4:      $LabelBidder_x = 1$  (Suspicious)
5:   end for
6: else
7:   for  $x=1$  to NumberBiddersCluster do
8:      $LabelBidder_x = 0$  (Normal)
9:   end for
10: end if

```

To validate our approach, we choose randomly 5 auctions among the 7-day duration auctions (in total 309) and select randomly one bidder in each auction (Table 7). As we can observe from this table, the shill bidding instances were successfully labeled by our approach. For example, bidder “g***r” has four fraud patterns with very high values; among them, two have a high weight and one a medium weight. Therefore, the activity of this bidder in auction ID # 2370 is suspicious. On the other

Fig. 3 Decision line of a subset and its labeled clusters

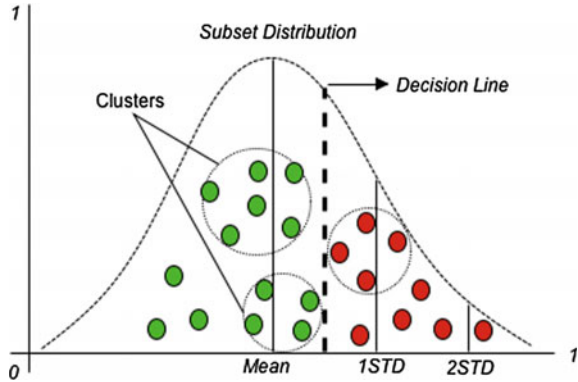


Table 7 SB instances and their labels

AuctionID	1009	900	2370	432	1370
BidderID	z***z	k***a	g***r	0***0	o***-
BT	0.75	0.4705	0.8333	0.5	0.04615
BR	0.3461	0.3076	0.2	0.3333	0.0857
SO	1	0	1	0	0.5
LB	0.5667	0.1909	0.0350	0.2199	0.2966
EB	0.5409	0.1909	0.0239	0.0043	0.2060
WR	0.75	0.4	1	0.5	0
AB	0	0	0.3333	0	0.0526
ASP	0	0	0.9935	0	0
Generated label	1	0	1	0	0

Table 8 Final results of instance labeling

SB subset	1 Day	3 Days	5 Days	7 Days	10 Days	Total
No. of normal instances	1135	1303	975	2098	135	5646
No. of suspicious instances	154	105	85	329	2	675

hand, the bidder “k***a” has all his fraud patterns with very low values; this indicates that this bidder behaved normally in the auction ID # 900. All these results are consistent with the labels produced by our approach.

Table 8 provides all the final results of the labeling task of our SB training dataset. There are 5646 instances categorized as normal and 675 instances as suspicious.

7 Conclusion and Future Work

There are limited classification studies on the SB fraud due to the difficulty of producing training data on one hand and labeling multidimensional instances on the other hand. Our aim in this paper is to effectively label SB instances based on the hierarchical clustering CURE that showed a remarkable capability for partitioning the online behavior of bidders. First, we divide the SB dataset into several subsets according to the different bidding durations of the auctions that we scraped from eBay. Then, we efficiently partition each SB subset into clusters of users with similar bidding behavior. At last, we apply a systematic labeling approach to each cluster to classify bidders into normal or suspicious.

In the following, we highlight two important research directions:

- The generated SB dataset is highly imbalanced, which will negatively impact the performance of classifiers as demonstrated in numerous studies such as [9]. The decision boundary of the fraud classifiers will be biased toward the normal class, which means suspicious bidders will be poorly detected. Handling the class imbalance problem is a continuous area of study [21]. In our research, we will investigate this problem by testing different types of techniques, such as data sampling and cost-sensitive learning, to determine the most suitable technique for our SB dataset.
- Ensemble learning has produced a reliable performance for many practical applications. The goals defined by ensemble learning are lowering the model's error ratio, avoiding the overfitting problem, and reducing the bias and variance errors. The most common ensemble methods are Boosting and Bootstrap Aggregation (Bagging). Thus, we will employ ensemble learning to develop a robust SB detection model, and examine the most fitting ensemble strategy for our SB dataset.

Acknowledgements The first author would like to thank the Saudi Arabian Cultural Bureau in Canada and the Umm Al-Qura University for their financial support.

References

1. Alzahrani, A., Sadaoui, S.: Scraping and preprocessing commercial auction data for fraud classification (2018). [arXiv:1806.00656](https://arxiv.org/abs/1806.00656)
2. Center, I.C.C.: 2015 internet crime report. In: 2015 IC3 Report. IC3 (2016)
3. Chang, J.S., Chang, W.H.: Analysis of fraudulent behavior strategies in online auctions for detecting latent fraudsters. *Electr. Commer. Res. Appl.* **13**(2), 79–97 (2014)
4. Cui, X., Lai, V.S.: Bidding strategies in online single-unit auctions: their impact and satisfaction. *Inf. Manag.* **50**(6), 314–321 (2013)
5. Dong, F., Shatz, S.M., Xu, H.: Combating online in-auction fraud: clues, techniques and challenges. *Comput. Sci. Rev.* **3**(4), 245–258 (2009)
6. Dong, F., Shatz, S.M., Xu, H.: Reasoning under uncertainty for shill detection in online auctions using dempster-shafer theory. *Int. J. Softw. Eng. Knowl. Eng.* **20**(07), 943–973 (2010)
7. Ford, B.J., Xu, H., Valova, I.: Identifying suspicious bidders utilizing hierarchical clustering and decision trees. In: IC-AI, pp. 195–201 (2010)

8. Ford, B.J., Xu, H., Valova, I.: A real-time self-adaptive classifier for identifying suspicious bidders in online auctions. *Comput. J.* **56**(5), 646–663 (2012)
9. Ganguly, S., Sadaoui, S.: Classification of imbalanced auction fraud data. In: *Canadian Conference on Artificial Intelligence*, pp. 84–89. Springer (2017)
10. Ganguly, S., Sadaoui, S.: Online detection of shill bidding fraud based on machine learning techniques. In: *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 303–314. Springer (2018)
11. Guha, S., Rastogi, R., Shim, K.: CURE: an efficient clustering algorithm for large databases. In: *ACM Sigmod Record*, vol. 27, pp. 73–84. ACM (1998)
12. Gupta, P., Mundra, A.: Online in-auction fraud detection using online hybrid model. In: *2015 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 901–907. IEEE (2015)
13. Jank, W., Shmueli, G.: Studying heterogeneity of price evolution in ebay auctions via functional clustering. In: *Business Computing. Handbook of Information Systems Series*, pp. 237–261 (2009)
14. Lei, B., Zhang, H., Chen, H., Liu, L., Wang, D.: A k-means clustering based algorithm for shill bidding recognition in online auction. In: *2012 24th Chinese Control and Decision Conference (CCDC)*, pp. 939–943. IEEE (2012)
15. Ravisankar, P., Ravi, V., Rao, G.R., Bose, I.: Detection of financial statement fraud and feature selection using data mining techniques. *Decis. Support Syst.* **50**(2), 491–500 (2011)
16. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
17. Sadaoui, S., Wang, X.: A dynamic stage-based fraud monitoring framework of multiple live auctions. *Appl. Intell.* **46**(1), 197–213 (2017)
18. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. *Ann. Data Sci.* **2**(2), 165–193 (2015)
19. Yu, H., Liu, Z., Wang, G.: An automatic method to determine the number of clusters using decision-theoretic rough set. *Int. J. Approx. Reason.* **55**(1), 101–115 (2014)
20. Zhang, J., Prater, E.L., Lipkin, I.: Feedback reviews and bidding in online auctions: An integrated hedonic regression and fuzzy logic expert system approach. *Decis. Support Syst.* **55**(4), 894–902 (2013)
21. Zhang, S., Sadaoui, S., Mouhoub, M.: An empirical analysis of imbalanced data classification. *Comput. Inf. Sci.* **8**(1), 151–162 (2015)

Big Data Security Challenges and Preventive Solutions



Nirmal Kumar Gupta and Mukesh Kumar Rohil

Abstract Big data has opened the possibility of making great advancements in many scientific disciplines and has become a very interesting topic in academic world and in industry. It has also given contributions to innovation, improvements in productivity and competitiveness. However, at present, there are various security risks involved in the process of collection, storage and use. The leakage of privacy caused by big data poses serious problems for the users; also the incorrect or false big data may lead to wrong or invalid analysis of results. The presented work analyzes the technical challenges of implementing big data security and privacy protection, and describes some key solutions to address the issues related with big data security and privacy.

Keywords Big data · Big data analysis · Big data security · Privacy protection

1 Introduction

In today’s world, a large number of people share their social information and behavior using the internet and it has led to the explosion of data generated. The constant advance of technologies has allowed an “explosive” growth in the amount of data generated from different sources, for example social networks, mobile devices, sensors, X-ray machines, telescopes, space probes, applications logs, climate predictions, geo-positioning systems and, in general terms, everything that can be classified within the definitions of the internet of things [1]. According to

N. K. Gupta (✉)

Department of Computer Science and Engineering, Jaypee University Anoopshahr, Anoopshahr, India

e-mail: nirmal.gupta@mail.jaypee.ac.in

M. K. Rohil

Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, India

e-mail: rohil@pilani.bits-pilani.ac.in

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_21

285

statistics, an average of 40,000 search queries occur every second; in other words, it can be said that over 3.5 billion searches are processed per day by Google search [2]. At the same time, various monitoring and sensing equipment are also generating data continuously. There is also a large amount of data in various fields such as scientific computing, healthcare, finance and retail. This phenomenon has aroused widespread concern.

The generation of this big data makes data analysis and application more complicated and difficult to manage. According to statistics, the amount of data generated globally over the past 3 years is more than the previous 400 years of data, which include documents, pictures, videos, web pages, e-mail, microblogging, and other types, which include mostly unstructured data as compared to structured data [3]. Gartner had forecasted that 4.9 billion connected objects to be in use by 2015, up 30% from 2014 and will reach 25 billion by 2020 [4]. At present, big data has become another information industry growth point in the field of information technology after cloud computing.

Like other current information systems, big data also involves security risks in the process of storage, processing, transmission and similarly it needs the data and privacy to be protected. The services like storage and management of data are provided in big data and cloud computing by the service providers themselves. But, the problem with big data as compared to cloud computing is that in case of cloud computing the user still has some control over their data to some extent; for example, through the use of cryptographic methods and through the other trusted computational methods. However, in the context of big data the businesses like Facebook not only produce the data but also provide the services like storage and management of data themselves. Therefore, it is extremely difficult to restrict the use of user information by means of technology and to protect the privacy of users' data [5].

At present, many organizations are aware of the big data security issues, and actively take action to focus on big data security issues. This paper focuses on the security challenges brought by the current big data technology and elaborates on the key technologies used for big data security and privacy protection. It should be pointed out that while introducing new security issues and challenges, big data also brings new opportunities in the field of information security. That is, big data-based technologies for information security can be used in turn for big data security and privacy protection.

2 Big Data Research Overview

2.1 Big Data Sources and Characteristics

Big data may have various sources from where it may be generated. Based upon the source of generation of big data, it can be divided into following categories [6]:

1. From the people: all kinds of data generated by people in the process of performing activities over the internet. The generated data can be in the form of text, images, videos or of any other type.
2. From the machines: This includes the data which is generated by different computers and information processing units which may be in the form of files, databases, multimedia, and so on, and also includes automatically generated information such as logs.
3. From the devices: data collected by various types of digital devices, such as the digital signals continuously generated by the camera; the different data related to human beings generated through various medical devices; the large amount of data which is generated by the astronomical telescopes.

2.2 *Big Data Analysis Goal*

At present, big data analysis is applied to various diversified areas such as science, medicine and commerce. Overall, the goals of big data analytics fall into the following categories:

(1) *Gain knowledge through extensive analysis*

People have a long history of data analysis. There may be various reasons. The first and most important reason for analyzing the data is to get knowledge from it. Since there is a large amount of unprocessed real sample information, it can effectively abandon individual differences and help people through the mining, and more accurately grasp the common purpose behind the things. Depending on the knowledge they have discovered, one can predict more accurately the nature or social phenomena that will occur. Typical examples include the ability to retrieve information about the flu through Google's search using data mining [7]; predicting stock quotes based on Twitter information [8]; and so on.

(2) *Grasp individual laws through long-term analysis*

Individual activities have distinct personal characteristics while satisfying certain common characteristics. Through long-term multi-dimensional data accumulation and gaining information through that data, various companies get the insights of users' behavior and this may help them to accurately describe the individual user's profiles. In this way, it helps them to provide more accurate products and services according to users' individual needs. It also helps companies to accurately provide recommendations related to advertisements.

For example, Google analyzes users' habits and hobbies through its big data products, helping advertisers to evaluate the efficiency of their advertising campaigns, and it is estimated that there may be hundreds of billions of dollars in the market in the future [9].

(3) *Control epidemic through analysis*

Many times the timely information obtained through the analysis of big data can provide more valuable information regarding spread of epidemics, than by disease-prevention centers. For example, during the 2009 flu pandemic, big data analysis was performed by Google to get the timely information. Generally, patients do not go to the doctor immediately just after getting infected, but their search and discussion trends can be analyzed to get information about most influenced geographical areas.

2.3 Big Data Technology Framework

Big data processing involves data collection, management, analysis and display. Figure 1 is a schematic view of the relevant technology, including four stages.

1. Data acquisition and preparation

The data sources of big data are diversified, including all kinds of structured, unstructured and semi-structured data such as databases, texts, pictures, videos and web pages. Therefore, the first step in big data processing is to gather data from the data source and pre-process it to provide a consistent, high-quality data set for subsequent processes.

Since there exist various sources of big data, therefore, there may be different models for its description and these may even contradict. Therefore, it becomes important to clean the data during the data integration process so that similar, repetitive or inconsistent data can be removed. In the literature, data cleaning and

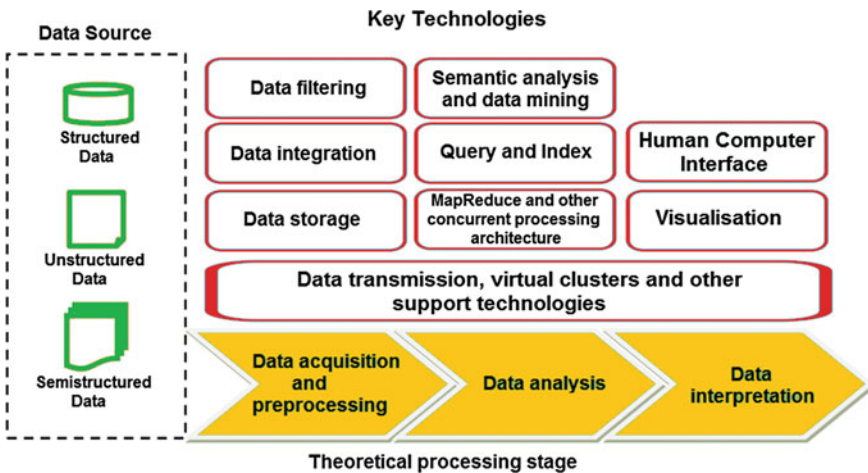


Fig. 1 Big data technology architecture

integration technology is aimed at the characteristics of big data, and proposes the cleaning of unstructured or semi-structured data and the integration of very large-scale data [9].

Data storage and big data applications are closely related. Real-time applications using big data require strong computational ability so that large amount of data could be processed in less amount of time. Therefore, big data processing system for real-time systems should quickly respond to the requests generated. This requires robust computing power for big data. Stream processing mode is more suitable for such applications. Most of the other applications need storage for subsequent deeper data analysis process. This may increase the storage cost. Usually big data uses distributed architecture to increase data throughput and reduce storage cost. Representative studies include file system GFS, HDFS and Haystack; NoSQL database MongoDB, CouchDB, HBase, Redis, Neo4j and so on [10].

2. Data analytics

Traditional data analysis may not work for big data as it was devised for structural data sources, but big data mostly consists of semi-structured or unstructured data. This presents a big challenge to big data analysis, but it is also the core process of big data applications. Based on the different levels, it can be broadly divided into three categories: computing architecture, queries and indexes, and data analysis and processing.

In terms of computing architecture, MapReduce [11] is a widely used big data set computing model and framework. In order to adapt to some analysis requirements that require high task completion time, work in [12] optimizes its performance. In [13], a data flow analysis solution based on MapReduce architecture, MARISSA, was proposed to support real-time analysis task. Dede et al. [14] proposed Mastiff, a big data analysis program based on time. Chandramouli et al. [15] proposed a TiMR framework based on MapReduce to deal with the real-time streaming for applications with high real-time demand such as advertisement push.

In query and indexing, traditional relational database query and indexing techniques are limited due to the large amount of unstructured or semi-structured data contained in big data, and NoSQL database technology received more attention. For example, Chandramouli et al. [16] proposed a hybrid data access architecture, HyDB, as well as a method of concurrent data query and optimization. Wang et al. [17] optimizes the query of key-value type database.

In data analysis and processing, the main technologies involved include semantic analysis and data mining. Owing to the diversification of data in big data environment, it is difficult to unify the terms to mine information when semantic analysis of data is concerned. In [18], for the big data environment, a high-efficiency terminology standardization method for solving the term variation problem is proposed. Keteta et al. [19] studied the heterogeneity of semantic ontology in semantic analysis. Traditional data mining technology is mainly aimed at structured data, so it is very important to study unstructured or semi-structured

data mining technology. Kang et al. [20] proposed a mining technique for image files, and Kang et al. [21] proposed a large-scale TEXT file retrieval and mining technology.

3. Data interpretation

The purpose of data interpretation is to represent data analysis results in a way which can serve the user's purpose. The major technologies which make it possible are visualization and human-computer interaction. There have been some visualization studies for large-scale data [22], which solve the display problem of large-scale data through data projection, dimension degradation or display wall. As human visual sensitivity limits the effectiveness of larger screens, a human-centric human-computer interaction design will also be an important technique to address the display of big data analytics results.

4. Other support technologies (data transmission and virtual cluster)

Although big data applications emphasize data-centric computing and push calculations to data execution, data transmission is still essential throughout the process, such as the transmission of some scientific observations from observation points to data centers. In [23] the authors study efficient transmission architectures and protocols for big data features.

In addition, because virtual clusters have the advantages of low cost, flexible construction and easy management, people can choose more convenient virtual clusters to complete the various processing tasks during big data analysis. Therefore, virtual machine cluster optimization research for big data applications is needed [24].

3 Big Data Security Challenges

Big data provides a great technology which has its significance in various fields and the security requirements in these fields are also changing. During the various activities performed over data during its collection, refinement and mining, many security threats are also associated. During this process the data may be destroyed, leaked, tampered, which can put the user privacy or corporate secrets to be compromised. In general, the security related to big data has the following characteristics and challenges.

3.1 Technical Challenge

The technical challenges of big data applications are mainly in the following three aspects:

First, as we all know, the application of big data is based on the possession of massive data. This involves the challenges of data storage technology and the technical challenges for data processing and analysis, including the data processing capabilities of computer hardware and supercomputer algorithm technology. The report learned from interviews with technical experts indicates that they are not optimistic about the recent overcoming of the technical challenges described above. Second, in the educational application of big data, data collection and problem-solving analysis are the core links, and application developers have to face the challenges of data acquisition technology and problem-solving analysis technology. The third aspect is the data compatibility challenge, the inconsistency of data encoding and format in different data storage systems, resulting in data sharing difficulties between different systems. The main reason for this problem is the lack of unified planning for the construction and purchase of various systems which results in inability to form unified data platform.

3.2 Mobile Data Security Challenge

In the world of today, various new applications related to social media have been emerged; besides this it is not uncommon for e-commerce portals and mobile applications to generate enormous amount of data. Analyzing such large amount of data in addition to data generated through internet of things network is a big challenge for companies. It also put responsibilities over the companies as the demand for data security capabilities on enterprises also increases greatly. In addition, the increase in data secrecy results in more sensitive analysis of data which is being transferred between mobile devices. Personal information may also be compromised because of the malware which may track user location or steals the confidential information. It requires increasing the safety related to user's privacy and data. Since the mobile devices have grown rapidly in recent years, it has created a challenge for big data security that how the samples of malware could be tracked which may enable analyzing these malware samples and finding the relationship between them.

3.3 Easy Target for Attackers

The various resources and personalized services are being provided through the network access in a flexible manner. In such a large networked society, the voluminous data attract the hackers because of the potential value associated with it. Such big data is large and has interrelated information stored, and due to this an attacker finds it easy to get more information by successfully attacking once, which further reduces the attacking cost and increases their profit value. A good example of such situation is when hackers use information in big data and take control of

important systems to take advantage from it for the time being. Hackers can also deploy AI technologies to take full control of the system by slowly penetrating it over time. Such strategy may work, as it would perform minor changes and such changes would appear natural. Thus, its detection could be avoided.

3.4 User Privacy Protection Challenge

Worldwide people use different kinds of web services and while using such services they need to access the network which stores their credentials like name, username, password, contact information, address and so on. It also includes their other personal information, behavior, habit, political or religious inclination and location; this has increased the risk of user privacy and data disclosure. Through the use of various data mining approaches such data can be easily recorded. If these data are found to be useful, then companies in the relevant areas of customer's need and habit can target to the specific users to achieve greater benefits. The traditional approach of privacy protection is based on anonymity of user data which is publicly available. But in reality, the user privacy protection cannot be achieved only through data protection through anonymity. There exist many other requirements and characteristics for user privacy. But the problem is that most of the existing privacy protection models and algorithms are only for traditional relational data, and cannot be directly applied to big data applications.

3.5 Safe Storage of Massive Data Challenge

The quantity of big data has increased tremendously, which includes structured and unstructured data. It has increased to such a level that the previous storage systems are not capable to meet the needs of big data applications. The current disk technology has some limitations and this limit is around 4 terabytes per disk. So for such a massive data of exabyte levels, this limitation creates a problem, because just to store 1 exabyte requires around 25,000 disks. It will really create the hardware issue of how to attach such large number of disks to a single computer, even if the computer is capable of processing that much of data.

To satisfy the demands of data storage of such large amount, various new storage technologies have been developed. Storage technologies such as direct attached storage (DAS), network attached storage (NAS) and SAN (storage area network) are being used to solve the data storage-related issues. Another technology called NoSQL storage technology is also used to capture, manage and process big data. Using NoSQL data storage can be extended and performance also can be improved, but still some issues exist. These issues include access control and

privacy control issues, issues related to technical vulnerabilities, security issues for authorization and verification, data management and confidentiality issues, and so on.

3.6 Challenge of Analysis

With the increase of storage capabilities, prior selection of data became insignificant. This can be seen as a real chance, allowing to keep focus on the potential future uses and which are not always fully defined at the time of their acquisition. In particular, many issues that were considered nonexistent later become accessible before the use of their potential of significant advantage (competitive advantage for example). It should be kept in mind, however, that more data is not always better data. It depends on whether or not they are heterogeneous, and whether they are representative of what is being sought. In addition, as the number of parameters increases, the number of erroneous correlations also increases. The analysis part will have to take into account these essential aspects.

Heterogeneous (structured, unstructured) data or incomplete or uncertain data for which specific treatments are needed will also be stored. Moreover, in this regard specific treatments are already required for the more standard data, even if some old methods remain effective for the volume of the existing data, which may cause theoretical and practical difficulties, unknown in advance. Thus, the simple statistical tests [25] become inoperative for large sample sizes. We can also mention the difficulty of multi-dimensional analysis on large sets of data, which arises during their interpretation. The visualization is considered until now as an extremely powerful tool, but it risks becoming inoperative by simple graphic saturation effect. In addition, real-time analysis of continuous flow of data from different sources also poses specific challenges. All these problems involve the development of new statistics for big data, for example, requiring a review of basic calculations such as statistical tests and correlations [26]. Of course, these technical analysis tools cannot be isolated from the computer tools and techniques dedicated to big data, for example, NoSQL, Hadoop, MapReduce or Spark.

3.7 Big Data Security Trust Challenge

Although big data has provided various opportunities to its users, but still it lacks the complete trust of the people using it. The visibility of social profiles generated by users varies across different types of networks and these are crawled by the search engines and therefore they become visible by the other users whether they have account or not. Therefore, here a trust issue arises from the user that how safe their data privacy is. This requires trust measures to be integrated with big data. These trust measures should not be treated as a static measure. That means, as the

data evolves the trust measures should also be updated accordingly. Yin and Tan [27] in their research have put the fact that semi-supervised learning methods that start with ground truth data are able to provide higher accuracy and trust on the source data. Another fact is that the different people have different personal opinions regarding various factors affecting their life and when there exist differences with statistics it leads to the market doubts about statistics.

4 Key Technologies for Big Data Security and Privacy Protection

At present, it is urgent to carry out research on key technologies of big data security in view of security challenges such as user privacy protection, data content credibility and access control faced by the big data. This section introduces some related key research areas for this.

4.1 Data Anonymity Protection Technology

For structured or unstructured data in big data, the core key technologies and basic means for data protection to achieve its privacy protection are still in the stage of development and improvement. Take the typical k anonymity scheme as an example. The early schemes [28] and their optimization schemes [29] group quasi-identifiers by data processing such as tuple generalization and suppression. The quasi-identifiers in each packet are the same and contain at least k tuples, so each tuple is at least indistinguishable from $k - 1$ other tuples. Since the k -anonymous model is for all attribute collections, it is not defined for a specific attribute, and it is prone to insufficient anonymity of a certain attribute. If the value of a sensitive attribute in an equivalence class is the same, the attacker can effectively determine the value of the attribute. This research is for static, one-time release. In reality, data publishing often faces scenarios in which data is continuously and repeatedly released. It is necessary to prevent an attacker from analyzing the data associations that are published multiple times, and destroying the original anonymity of the data [30].

In big data scenarios, data anonymity protection is more complicated: an attacker can get data from multiple sources, not just the same source. For example, in the Netflix application, people [31] found that an attacker could identify the target's Netflix account by comparing the data to the publicly available imdb. According to this, the user's political inclinations and religious beliefs are obtained (obtained through the user's viewing history and comments and scoring analysis of certain movies). Such issues are subject to further research.

4.2 *Social Network Anonymity Protection Technology*

The data generated by social networks is one of the important sources of big data, and it contains a large amount of user privacy data. Because social networks have the characteristics of graph structure, their anonymous protection technology is very different from structured data.

The social networks also require anonymity protection. Here, some typical requirements are user anonymity and attribute anonymity. While using social networks the user identity and attribute information is also required to be hidden while publishing these. The related data of different users should not disclose their relationship and anonymity between users is required. It is also known as edge anonymity. Hide the relationship between users when publishing. The attacker tries to use the various attributes of the node (degrees, tags, some specific connection information, etc.) to re-identify the identity information.

The current side-anonymity schemes are mostly based on additions and deletions of edges. The method of randomly adding and deleting exchange edges can effectively implement edge anonymity. Among them, Ying et al. [32] keep the eigenvalues of the adjacency matrix and the corresponding second eigenvalues of the Laplacian matrix in the anonymity process. Zhang et al. [33] group according to the degree of the nodes, and select the nodes with the same degree. The problem with this type of method is that the randomly added randomness is too scattered and sparse, and there is a problem of insufficient protection of the anonymous side.

4.3 *Data Watermarking*

Digital watermarking refers to a method in which identification information is embedded in a data carrier in an imperceptible manner without affecting its use, and is more commonly found in multimedia data copyright protection. There are also some watermarking schemes for databases and text files.

The method of adding watermarks in databases and documents is very different from the multimedia carrier, which is determined by the characteristics of data disorder and dynamics. The basic method is that there can be redundant information in the data or it can bear certain precision errors. For example, Agrawal et al. [34] have an error tolerance range based on numerical data in the database, embedding a small amount of watermark information into the least significant bits randomly selected from these data. Sion et al. [35] proposed a scheme based on statistical features of data sets, embedding one-bit watermark information in a set of attribute data to prevent attackers from destroying the watermark. In addition, by embedding database fingerprint information in the watermark [36], the owner of the information and the object being distributed can be identified, which is beneficial for tracking the leak in a distributed environment. Watermarking based on text content [37] depends on modifying the content of the document, such as adding spaces,

modifying punctuation, natural language-based watermarking [38] and so on, through the understanding of semantics to achieve changes, such as word substitution or sentence changes.

4.4 *Data Traceability Technology*

Owing to the diversified sources of data, it is necessary to record the source of the data and its distribution to provide support for later mining and decision-making. Data provenance technique has been extensively studied in the database field long before the big data concept emerged. The basic purpose is to help people determine the source of each data in the data warehouse, for example, which data items in tables are computed, so that it is convenient to check the correctness of the results at a very small cost. The basic method of data tracing is notation, such as marking the data in the data warehouse in [39] to record the query and propagation history of the data in the data warehouse. Data traceability techniques can also be used for traceability and recovery of files. For example, the work in [40] created a prototype system of data origin storage systems by extending the Linux kernel and file system, which can automatically collect origin data.

Further data traceability technologies can play an important role in the field of information security. However, data traceability technology also faces the following challenges in the protection of big data security and privacy:

1. *The balance between data traceability and privacy protection.* Using traceability to provide big data security protection requires first to obtain big data source using analysis of big data. Then the next step becomes to define the security policy and provide the required security mechanism. Often, the source of big data is privacy-sensitive and users are not interested that data to be accessed by the analysts also. Therefore, the problem is how to balance these two requirements simultaneously so that data traceability and privacy protection of the data both can be achieved.
2. *Security protection of data security technology itself.* The data tracing techniques currently employed are unable to handle the security issues correctly. The problem is that how to determine whether the tag associated with the data is itself correct or not. The other problem is that the tag information itself may not be securely bound with the data content and there may be other similar issues. Also in case of big data, since it is implemented on such a large scale, high speed and diverse characteristics such problems become more important.

5 Conclusion

The arrival of the era of big data has opened great opportunities. Big data not only has impact on everyone's social and economic behavior but also has influenced their way of living and thinking. Although big data is an important solution to various problems, it has also brought new security issues into existence. From the perspective of privacy protection, trust and access control of big data, this paper analyzes various security features and problems in the big data environment, namely, mobile data security, attack targets, user privacy protection challenges and security, storage issues, data security evolution, trust security issues and so on, and also discusses the preventive solutions for them. However, generally speaking, the current research on the protection of big data security and privacy is not sufficient. Only through the combination of technical means and relevant policies and regulations combined can better solve the big data security and privacy protection issues.

References

1. Xia, F., Yang, L.T., Wang, L., Vinel, A.: Internet of things. *Int. J. Commun. Syst.* **25**(9), 1101–1102 (2012)
2. Google search statistics. <http://www.internetlivestats.com/google-search-statistics/>
3. Lee, I.: Big data: dimensions, evolution, impacts, and challenges. *Bus. Horiz.* **60**(3), 293–303 (2017)
4. Nguyen, B., Simkin, L.: The Internet of Things (IoT) and marketing: the state of play, future trends and the implications for marketing. *J. Mark. Manage.* **33**(1–2), 1–6 (2017)
5. Boyd, D., Crawford, K.: Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* **15**(5), 662–679 (2012)
6. Guo-Jie, L., Xue-Qi, C.: Research status and scientific thinking of big data. *Bull. Chin. Acad. Sci.* **27**(6), 647–657 (2012)
7. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. *IEEE Trans. Knowl. Data Eng.* **26**(1), 97–107 (2014)
8. Arias, M., Arratia, A., Xuriguera, R.: Forecasting with twitter data. *ACM Trans. Intell. Syst. Technol.* **5**(1), 8 (2013)
9. Arasu, A., Chaudhuri, S., Chen, Z., Ganjam, K.: Experiences with using data cleaning technology for bing services. *IEEE Data Eng. Bull.* **35**(2), 14–23 (2012)
10. Sarma, A.D., Dong, X.L., Halevy, A.: Data integration with dependent sources. In: *Proceedings of the 14th International Conference on Extending Database Technology*, ACM, pp. 401–412 (2011)
11. Elomari, A., Maizate, A., Hassouni, L.: Data storage in big data context: a survey. In: *International Conference on Systems of Collaboration (SysCo)*, pp. 1–4. IEEE (2016)
12. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. *Commun. ACM* **51**(1), 107–113 (2008)
13. Verma, A., Cherkasova, L., Kumar, V.S., Campbell, R.H.: Deadline-based workload management for mapreduce environments: pieces of the performance puzzle. In: *Network Operations and Management Symposium (NOMS)*, pp. 900–905, IEEE (2012)

14. Dede, E., Fadika, Z., Hartog, J., Govindaraju, M., Ramakrishnan, L., Gunter, D., Canon, R.: Marissa: Mapreduce implementation for streaming science applications. In: IEEE 8th International Conference on E-Science (e-Science), 2012, pp. 1–8, IEEE (2012)
15. Guo, S., Xiong, J., Wang, W., Lee, R.: Mastiff: a mapreduce-based system for time-based big data analytics. In: IEEE International Conference on Cluster Computing (CLUSTER), 2012, pp. 72–80, IEEE (2012)
16. Chandramouli, B., Goldstein, J., Duan, S.: Temporal analytics on big data for web advertising. In: IEEE 28th International Conference on Data Engineering, pp. 90–101. IEEE (2012)
17. Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. *Inf. Sci.* **275**, 314–347 (2014)
18. Wang, Y., Lu, W., Wei, B.: Transactional multi-row access guarantee in the key-value store. In: IEEE International Conference on Cluster Computing (CLUSTER), 2012, pp. 572–575. IEEE (2012)
19. Hwang, M., Jeong, D.H., Jung, H., Sung, W.K., Shin, J., Kim, P.: A term normalization method for better performance of terminology construction. In: International Conference on Artificial Intelligence and Soft Computing, pp. 682–690. Springer, Berlin (2012)
20. Ketata, I., Mokadem, R., Morvan, F.: Biomedical resource discovery considering semantic heterogeneity in data grid environments. In *Integrated Computing Technology*, pp. 12–24. Springer, Berlin (2011)
21. Kang, U., Chau, D.H., Faloutsos, C.: Pegasus: mining billion-scale graphs in the cloud. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5341–5344. IEEE (2012)
22. Kola, A., More, H., Soderman, S., Gubanov, M.: Generating Unified Famous Objects (UFOs) from the classified object tables. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 4771–4773. IEEE (2017)
23. Tang, J., Liu, J., Zhang, M., Mei, Q.: Visualizing large-scale and high-dimensional data. In: Proceedings of the 25th International Conference on World Wide Web, pp. 287–297. International World Wide Web Conferences Steering Committee (2016)
24. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of “big data” on cloud computing: review and open research issues. *Inf. Syst.* **47**, 98–115 (2015)
25. Meyer-Schönberger, V., Cukier, K.: *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan, England (2013)
26. Gandomi, A., Haider, M.: Beyond the hype: big data concepts, methods, and analytics. *Int. J. Inf. Manage.* **35**(2), 137–144 (2015)
27. Yin, X., Tan, W.: Semi-supervised truth discovery. In: Proceedings of the 20th International Conference on World Wide Web, pp. 217–226. ACM (2011)
28. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Syst.* **10**(5), 557–570 (2002)
29. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM (2005)
30. Xiao, X., Taom, Y.: M-invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 689–700. ACM (2007)
31. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: IEEE Symposium on Security and Privacy, 2008. SP 2008, pp. 111–125. IEEE (2008)
32. Ying, X., Wu, X.: Randomizing social networks: a spectrum preserving approach. In: Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 739–750. Society for Industrial and Applied Mathematics (2008)
33. Zhang, L., Zhang, W.: Edge anonymity in social network graphs. In: International Conference on Computational Science and Engineering, 2009. CSE'09, vol. 4, pp. 1–8. IEEE (2009)
34. Agrawal, R., Haas, P.J., Kiernan, J.: Watermarking relational data: framework, algorithms and analysis. *VLDB J. Int. J. Very Large Data Bases* **12**(2), 157–169 (2013)

35. Sion, R., Atallah, M., Prabhakar, S.: On watermarking numeric sets. In: International Workshop on Digital Watermarking, pp. 130–146. Springer, Berlin (2002)
36. Guo, F., Wang, J., Li, D.: Fingerprinting relational databases. In: Proceedings of the 2006 ACM Symposium on Applied Computing, pp. 487–492. ACM (2006)
37. Pease, A., Niles, I., Li, J.: The suggested upper merged ontology: a large ontology for the semantic web and its applications. In: Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, vol. 28, pp. 7–10 (2002)
38. Atallah, M.J., Raskin, V., Hempelmann, C.F., Karahan, M., Sion, R., Topkara, U., Triezenberg, K.E.: Natural language watermarking and tamperproofing. In: International Workshop on Information Hiding, pp. 196–212. Springer, Berlin (2002)
39. Cui, Y., Widom, J., Wiener, J.L.: Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.* **25**(2), 179–227 (2000)
40. Muniswamy-Reddy, K.K., Holland, D.A., Braun, U., Seltzer, M.I.: Provenance-aware storage systems. In: USENIX Annual Technical Conference, General Track, pp. 43–56 (2006)

Role and Challenges of Unstructured Big Data in Healthcare



Kiran Adnan, Rehan Akbar, Siak Wang Khor
and Adnan Bin Amanat Ali

Abstract Unprecedented growth in the volume of unstructured healthcare data has immense potential in valuable insight extraction, improved healthcare services, quality patient care, and secure data management. However, technological advancements are required to achieve the potential benefits from unstructured data in healthcare according to the growth rate. The heterogeneity, diversity of sources, quality of data and various representations of unstructured data in healthcare increases the number of challenges as compared to structured data. This systematic review of the literature identifies the challenges and problems of data-driven healthcare due to the unstructured nature of data. The systematic review was carried out using five major scientific databases: ACM, Springer, ScienceDirect, PubMed, and IEEE Xplore. The inclusion of articles in review at the initial stage was based on English language and publication date from 2010 to 2018. A total of 103 articles were selected according to the inclusion criteria. Based on the review, various types of healthcare unstructured data have been discussed from different domains of healthcare. Also, potential challenges associated with unstructured big data have been identified in healthcare for future research directions in the technological advancement of healthcare services and quality patient care.

Keywords Unstructured data · Healthcare · Big data · Systematic literature review · Challenges of healthcare data

K. Adnan (✉) · R. Akbar · S. W. Khor · A. B. A. Ali
Faculty of Information and Communication Technology,
Universiti Tunku Abdul Rahman, 31900 Kampar, Malaysia
e-mail: kiranadnan@iutar.my

R. Akbar
e-mail: rehan@utar.edu.my

S. W. Khor
e-mail: khorsw@utar.edu.my

A. B. A. Ali
e-mail: adnanbinamanat@iutar.my

1 Introduction

Rapid advancement in healthcare systems from disease-centric systems to patient-centric systems involves patient health records, integration of data from various health and social sources, advancement in mhealth, smart health and tele-health devices. Advanced Healthcare systems are required to support clinical decision support system, personalized and precision medicine, diagnostics systems, health and wellness monitoring, patient care and predictive analytics. This data-driven healthcare support leads to orchestrate data preprocessing, analytics, interpretation, management and curation techniques for improved healthcare services. Most of the healthcare data are unstructured such as medical prescriptions (hand-written, free-text), clinical notes, Electronic Health Records (EHR), Electronic Medical Records (EMR), Patient Health Record (PHR), Medical Imaging (Magnetic Resource Imaging, MRI, Photoacoustic imaging, Fluoroscopy, Positron emission tomography), discussion forums, social media, sensor data, medical sounds and video data, genomics, and in many other transactional, biometric and application data. By definitional perspective of unstructured data, it has no schema or predefined model, attitudinal and behavioral, and without specified format. This heterogeneity, variability, and diversity of unstructured data make it difficult to get valuable insights from unstructured data that if analyzed in a controlled manner, can improve healthcare analytics and extraction of more real values. In past years, the unstructured data growth rate is increasing exponentially. 90% of digital universe data are unstructured. 57% of all healthcare data will be useful if it is properly tagged and analyzed whereas only 3.1% of healthcare data are providing the highest value. It is a challenging and daunting task to find a needle in a haystack [1, 2].

The role of multifaceted unstructured big data in healthcare is very challenging and gaining popularity because it contains more rich information that is helpful to provide novel insights as well as improve technological capabilities in healthcare systems. Unstructured data in healthcare have great potential but a lot of challenges remain to be addressed. Several techniques have been proposed in the literature regarding big data management, analysis, and representation of data but very less number of articles dealt with the unstructured data in healthcare. The literature about unstructured big healthcare data emphasizes technological advancement need to be more focused. The main objective of this literature review is to highlight the most affected areas of healthcare big data solutions from unstructured data. This review is an attempt to answer the following question about unstructured data in healthcare, i.e., which domains of healthcare are suffering from the complexity of unstructured data, what challenges are being created by unstructured data, and what is the role of unstructured data in inefficient and poor results in analytics and data-driven decision-making in healthcare. The review discusses the types of unstructured big data in healthcare and the challenges associated with unstructured big healthcare data since 2010. A systematic literature review has been conducted in

this regard to fulfill the objectives of the study. The review of existing literature on the challenges associated with unstructured big healthcare data will help to improve usability and quality of unstructured data in various fields of healthcare. It will highlight the major challenges of unstructured data in healthcare that need advancement in tools and technologies.

The rest of the paper is organized in the following manner: Sect. 2 discusses the methodology used and the results of the systematic literature review. Challenges associated with unstructured big healthcare data in technological advancement have been elaborated in Sect. 3. Next Sect. 4 includes the findings of the review. Finally, Sect. 5 presents the conclusions.

2 Methodology

A systematic literature review has been used to conduct an effective review to highlight the potential challenges in healthcare domain due to unstructured data. The objectives of this research were to discuss the types of unstructured big data in healthcare and to explore the challenges associated with unstructured big healthcare data. This review has comprised of three phases: planning the review, conducting the review and reporting. In the planning phase, the fundamental search has been performed using specific keywords. The second phase involves the searching keywords on specific sources to conduct the review and results are obtained. In the third and last phase, results are presented.

2.1 Searching the Literature

According to the objectives of the study, the best suitable keywords were identified to organize more effective review. “Unstructured data” and “healthcare” were the queries formed from the keywords. This search query was used on some of the popular academic databases: IEEE Xplore, Springer, ScienceDirect, ACM digital library, and PubMed.

2.2 Inclusion Criteria

Articles deal with unstructured data in the healthcare domain, published between 2010 and 2018, using the English language was the inclusion criteria to capture the literature. According to the results of databases and inclusion criteria, a total number of articles (n) selected for review were 1522. Articles were filtered by title reading and the total number of articles reduced to (n) = 566. After that, abstracts

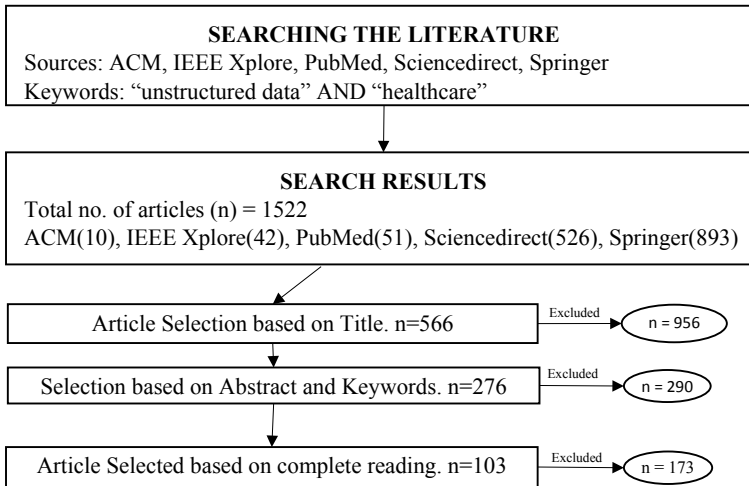


Fig. 1 Literature review process

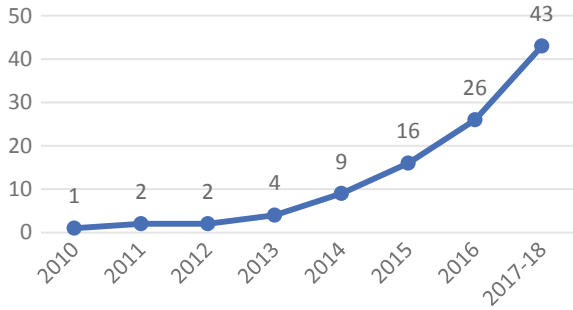
and keywords were carefully studied and n decreased to 276. Based on the complete article study, 103 articles were selected for the study. Type of healthcare technology, type of unstructured data used, the domain of work and identified challenges are the records kept from each selected paper. Following Fig. 1 describes the whole research process in detail.

2.3 Results

This review was conducted to analyze the challenges in technological advancement due to the huge volume of unstructured healthcare data. The following Fig. 2 shows the research trend of selected articles for this review according to the year of publication. The figure shows the number of publications in each year that deals with unstructured data. An exponential rise can be seen in the following diagram that reflects the importance of unstructured data in the healthcare industry.

The results witness several types of healthcare advancement in four major categories: healthcare analytics (51% of selected articles), decision support systems (12%), health and wellness monitoring (18%), precision and personalized medicine (6%). In selected articles, several types of unstructured data have been used in various domains of healthcare such as disease identification from EHR, pharmacovigilance, drug effects identification, phenotyping, genotyping, and medical recommendation systems. Unstructured data are found in EHR, EMR, PHR, medical imaging, discussion forums, social media, sensor data, medical sounds and video data, genomics, and in many other transactional, biometric and application data. A major portion of selected research articles in this literature review was

Fig. 2 Research trend of publications



related to the clinical hand-written notes, free-text in EHR, and text from social media. The main purpose of this literature review was to identify the most affected areas due to unstructured data in healthcare where technological advancement is required in order to improve the patient care. In the following section, challenges identified from the literature have been described in each category along with the requirements.

3 Challenges

This section describes the challenges of unstructured big data in healthcare identified in the literature. Selected articles are divided into five major categories with respect to their relevancy. The major categories of challenges are preprocessing, interoperability, Information Extraction (IE), analytics and result interpretation, and some other issues related to data exfiltration, standardization, and data curation. The following subsections describe the challenges associated with unstructured big healthcare data in detail.

3.1 Data Preprocessing

The variety of data is available in healthcare but these data are not fully usable for analysis and other processing due to noise, errors, missing values, and other quality issues. Preprocessing is a step that can reduce the inefficiency of unstructured data. As, 95% of healthcare data are unstructured data, available in various formats and facing more complexity, quality, usability and dimensionality issues. It arises the importance of preprocessing to improve the efficiency and accuracy of healthcare systems. Clinical decision support systems, health and wellness monitoring, disease identification, precision medicine, and analytics are important areas where unstructured data preprocessing techniques are required. Data preprocessing and

data cleaning are required for improved results in predictive analytics from social media text [3], complaint management in health information systems [4], decision support system from medical text and images [5], disease identification from health and wellness monitoring EHR data [6], in analytics and management of unstructured data [7, 8].

Data aggregation prepares the data in the required format as data is coming from various sources with different formats, frequencies, and representations such as structured EHR, semi-structured logs and unstructured images, audio and video data. The volume of unstructured data is far higher than the structured data. Hence, unstructured data aggregation and acquisition are more challenging. Free-text doctors' notes, handwritten clinical notes, patient care data, transactional, biometric and application data, EMR, and many other types of unstructured data are coming from various sources. Collection of these data found in a variety of formats is more challenging to handle [9]. Transformation engines clean, split, translate, merge, validate and sort unstructured data and make data prepared for analysis. Transformation techniques are required to transform the representation of data from unstructured to structured data in healthcare analytics, drug safety surveillance and clinical decision support systems [10–13]. Healthcare domains like identification of drug safety issues, disease identification systems need an automatic conversion of unstructured documents to structured data to increase the availability and usability of data analytics and databases [14–16]. But context-aware transformation based on user requirement and finding the relationships between the data items is still an open issue in healthcare. Personalized medicine, disease identification, adverse drug events identification, and medical knowledge base are the fields of healthcare where aggregation of structured and unstructured data is found critical. Integration of structured and unstructured data [17], improvement in predictive performance by combining structured and unstructured data [18], combining structured and unstructured data for comprehensive representation in predictive analytics [19], advanced analytical techniques [20], mapping of unstructured data to normalized content [21], complexities arise when outcomes of diverse sources need to be combined [22], and classification of combined data [23] are the challenges found in the literature.

Unstructured data require more advanced techniques than structured data because it has no schema, no predefined model, and no formal representation. Advancement is required to prepare unstructured healthcare data with more efficient data preprocessing methods. New big data technologies are emerging with adapted preprocessing. Apache Spark [24], Hadoop [25], and Flink [26] are performing better in processing big data. Apache Spark is making effort particularly in preprocessing and implemented MLlib to bridge the gap. It is highly desirable to develop new preprocessing techniques for data cleaning, discretization, feature selection, and data profiling using these advanced big data processing tools in healthcare.

3.2 *Interoperability*

Interoperability is another challenge to advance healthcare systems where two or more systems share and use data. Integration of data from different systems without any middleware and exchange of data poses more challenges to healthcare analytics (Table 1).

It requires common interfaces of healthcare systems, standardization of processes and data, and share common characteristics of a variety of data. Lack of uniform coding system across vendors is a challenge to interoperability that needs to be solved in order to manage patient information across hospitals and different applications with considering the ethical and privacy issues of patients' personal data. Following are the challenges identified in this literature review regarding interoperability in healthcare with a variety of healthcare data:

- Integration (Late or Early) of heterogeneous data and common type system for a variety of data types are the common challenges in predictive analytics in precision medicine and clinical diagnostics.
- Conventional databases are not adequate to handle this diversity of data types. Consolidated view of patient-related data using distributed storage is required for advanced information systems.
- Lack of harmonization between structured and unstructured data, evolving interoperability standards and integration of EHRs are creating challenges in analytics, data-driven medicine, precision and personalized medicine, and decision support system.
- All stakeholders need to work together in order to achieve interoperability [46]. Semantic and syntactic interoperability is important and essential for context awareness. Shared use of heterogeneous data, common representation of combined data, and deep understanding of unstructured data is an open issue. However, technical solutions are required for interoperability and integration.

3.3 *Information Extraction*

Disease diagnosis and its complications, communicable and non-communicable diseases' identification, drug surveillance, surgery details, and patient history are some examples of healthcare data where data are stored in the form of natural language text. Information extraction is a technique to extract structured information from these clinical texts with the help of natural language processing and machine learning. These information extraction techniques are helpful in predictive, prescriptive analytics, clinical care, clinical decision support system, disease identification from clinical notes, and precision medicine. Information extraction challenges due to unstructured healthcare data (i.e., free-text and handwritten clinical notes, EHR and EMR) found in the literature are described in Table 2.

Table 1 Related literature regarding interoperability and integration

Healthcare category	Data type used	Domain	Description of challenges
Analytics	Physician's and pharmacy notes, medical images	Systemic Lupus Erythematosus	Integration of various data formats, interoperability, evolving standards are the main challenges for which Common Type System is required for Clinical Natural Language Processing. The accuracy of predictive analytics can be improved using the advancement of these techniques [21, 27–31]
Analytics	EMR Clinical Free Text	Phenotyping	
Predictive Analytics	EHR Text	Clinical Diagnostics	
Analytics	Genomic Data	Evidence-based precision medicine	
Analytics	Omics data	Health data Spectrum	
Distributed Storage	Medical Images and associative text data	Radiology Information System	Conventional databases are not able to handle the complexity of a variety of data in healthcare. Consolidated View of Patient-relevant data from heterogeneous and disperse systems [32, 33]
Distributed Storage	Free-text, images and Proprietary formats	Medical Information Systems	
Data-driven Medicine	Text and medical images	Epilepsy	Lack of harmonization between structured and unstructured data, evolving interoperability standards, and integration of EHRs are the potential challenges and need solutions for efficient and effective curation [24, 25, 34–40]
Healthcare Systems	Genome and Medical images	European Single Market for health	
Data Analysis, Representation	EHR	Health Information Systems	
Precision and Personalized Medicine	Omics data	Molecular characterization of diseases	
Decision Support system	Behavior data and EHR	Medical Recommendations for Diabetes	
Information Management	Clinical Text	Common type of system development	Deep semantic understanding required structured data [41]. Semantic interoperability and data sharing are required to make computers context-aware [42, 43]. Shared use of heterogeneous data sources [44]. Interoperability, standardization, and integration of data sources are open issues [45]
Smart Healthcare	PHR	IohT	
Decision Support system	Structured and unstructured text	Pharmaceutical IE and Integration	
Medical Image Processing	Medical and diagnostic images	Medical diagnosis	
Decision Support system	Text and images of clinical data	Dental Clinics	

Table 2 Related literature regarding information extraction

Healthcare category	Data type used	Domain	Description of challenges
Analytics	Clinical documents	Open sources IE software	Extracting quality insights from healthcare data sources [47]
Analytics	Social Media streams	Drug and its effects detection	IE regarding drug and its effects from social media text streams is challenging due to ambiguous information [48]
Predictive Analytics	Clinical discharge summaries	Prediction of readmission from discharge summaries	Gaining information regarding readmission from unstructured clinical discharge summaries [49]
Analytics	Clinical Text data	Physiotherapy treatment	IE from unstructured EHR clinical notes [50]
Clinical Care	EHR	Phenotyping	Phenotype extraction from unstructured EHRs [51]
Personalized medicine	EHR and biomedical text literature	Genotype and phenotype extraction for cancer	Finding actionable granularity in disease identification and risk score [52]
Adaptive Analytics	Clinical Text trails of EMR and EHR	Cancer Clinical trials	Subject extraction [53]
Analytics	Unstructured medical notes	Disease identification	Semantically informed IE [54]
Sentiment Analysis	Discussion forums	Medical forum Discourse	Semantic extraction [55]
Exploratory Analytics	Physician and Biomedical data	Dashboard for querying heterogeneous data	Standardization, data formats, and their aggregation will improve semantic data processing [56]
Information retrieval and extraction	Clinical text documents	Software implementation in hospital	Finding relevant data [57]
Analytics and Representation	Imaging, Genetic and healthcare data		Multifaceted proprietary, open-source, and community developments [58]
Decision Support system	Medical and clinical references and publication text	Medical Recommendation System for Mental health	Predefined schema, large manually crafted dictionaries and ontologies are required to extract information from medical domain publications [59]
Clinical Decision Support System	EMR, Unstructured Clinical Notes	Automatic diagnosis prediction	Unstructured data combined with structured and semi-structured data will show good results [60]

(continued)

Table 2 (continued)

Healthcare category	Data type used	Domain	Description of challenges
Disease identification	Emergency dept. text notes	Automated identification of Pediatric Appendicitis Score	Unstructured data processing is challenging to extract useful information [61]
Bio surveillance	Clinical Free text	Detailed clinical and epidemiological variable extraction	Temporal IE and its integration for more elaborative and comprehensive representation of information [62]
Precision Medicine	EHR text	Prescreening and automated identification of HFpEF patients	The immense need for automatic data extraction system for the transformation of unstructured data into structured representation [63]
Data Management	Social Network text	Health data mining	Extracting information and associative features from unstructured text (associative analysis) [64]
Database Analytics		Data linkage and administrative databases	Metadata Management and IE [65]

Entities, relations, events, terms, and other important information can be extracted from text data in healthcare. These subtasks of information extraction are valuable to identify important terms and its associative relations in the medical domain. In healthcare analytics, decision support systems, medicine, and data management area, information extraction deal with several challenges due to unstructured big data. This review identifies those challenges to highlight the potential areas that need solutions for improved healthcare. Following is the summary of challenges:

- Extracting quality insights from EHR clinical notes, documents, clinical discharge summaries, and social media streams for healthcare analytics in treatment and medicine.
- Phenotype and genotype information extraction for disease identification from EHR and biomedical text up to actionable granularity and identify risk score.
- Semantic information extraction from various types of analytics to identify the disease from medical data and advanced intelligent IE systems are required to extract semantic information from multifaceted, open, and diverse unstructured health data.
- Automated extraction of various types of information from EHR clinical notes, EMR, medical and clinical publication text for medical recommendation system and automatic disease prediction is an immense need to represent data into the more elaborative structured form.

- Information extraction from healthcare unstructured data and metadata management can facilitate data linkage and databases. Traditional knowledge-based decision support systems deal with the data that have clear structure [66]. Lack of knowledge-based systems for unstructured data requires advance automatic IE system to build knowledge bases.

3.4 Data Analysis and Representation

Healthcare analytics is comprised of automated computer-based processes and workflows to transform raw health data into meaningful information for effective healthcare decision-making and wellness monitoring. Healthcare analytics is not only facilitating patients but also physicians, clinicians, nursing and other administrative staff. This review explores the challenges of unstructured big data to healthcare where analysis of data needs more accuracy because it is directly related to human life. According to the review, healthcare analytics has been identified as the most critical and most prominent research field and most affected by the heterogeneity and diversity of unstructured data. Table 3 briefly describes the related research.

Application of big data analytics techniques on disease diagnosis, pharmacovigilance, intelligent patient care, phenotyping, disease prediction, and patient health and wellness monitoring is helpful for preventive interventions and treatments. Technological development in healthcare is producing huge volume of different types of data. Patient demographic data, laboratory results, patient history, genomic data, EHR, EMR, PHR, and many other types of data are being produced every day. These data can be structured, semi-structured or unstructured. Due to the heterogeneity, different data formats, no predefined schema or model, and diversity of data sources adding more challenges to unstructured big data analytics in healthcare. This section identifies the potential challenges of unstructured data analytics to predictive, prescriptive, descriptive and diagnostic analytics. According to the literature review, the following is a brief description of the area where unstructured data are critical for analytics and data representation. The challenges of each area are also described categorically.

3.4.1 Smart Healthcare

Most of the data generating by sensors in smart healthcare are unstructured. Sensor-based health and wellness monitoring are generating a huge amount of unstructured data which could not be handled and interpreted manually. This huge amount of unstructured data is generating a gap between its potential and usability. The huge amount of streaming data from sensors is not usable due to heterogeneity and dimensionality of unstructured data. The data analytics pipeline for smart

Table 3 Related Literature regarding Healthcare Analytics and Representation of data

Healthcare category	Data type used	Domain	Description of challenges
Smart Healthcare	Sensor data	Patient Monitoring	Data discovery in heaps of unstructured data where streams of data are coming at very high speed, advanced smart health, and management to automate the curation process and improve the infrastructure for context-aware analytics [67–69]
	Heterogenous dataset text, images, sound	Healthcare Analytics pipeline	
		Analysis of healthcare systems	
Data Representation and Predictive Modeling	Clinical text events	Pharmacovigilance	Summarization and representation of data are facing difficulties due to lack of standardization such as abbreviations. Aggregated data from multiple sources contain diversity, sparsity, and multidimensionality due to data formats and representations. These different representations of EHR data making context-aware analytics challenging [70–73]
Data representation	Clinical Text documents	Phenotyping: Cancer Care	
Data representation	Clinical Notes and EMRs	Congestive heart failure	
Decision Support system	EHR text and Genetic data	Phenotyping using HER	
Healthcare System		Israel Health Policy Research	
Predictive Analytics	EHR clinical narratives	Framingham risk factors identification of Coronary artery disease	Due to data quality issues and heterogeneity, unstructured data are considered useless for clean analytics. Missing data and acronyms decreases the efficiency of analytics. Social media data have poor quality and noise, so the analysis of social media data is critical. Hashtag words can be used to improve analytical process but still, it needs health, media, and computer literacy for efficient results [74–79]
Twitter Analytics	Twitter text	Common disease identification	
eHealth	All	Implementation of eHealth platform	
Healthcare big data		Review of healthcare big data	
Twitter Analytics	Twitter hashtag words	Identification of top diseases	
Text Analytics	EHR Text	Russian language medical data of acute coronary syndrome	
Text Analytics	Clinical Notes	Medical Research	
Privacy and Security	EHR text	Conceptual design of healthcare system	
Visual Analytics	Structured and unstructured text	Visual content correlation analysis	
Predictive Analytics	EHR Text	Eczema Disease identification	
Predictive Analytics	EMR-free clinical text	Coronary Artery Disease identification	[80–85]

(continued)

Table 3 (continued)

Healthcare category	Data type used	Domain	Description of challenges
Advanced analytics	EMR	Intelligent patient care	Efficient Unstructured Data Analytical techniques are required for technological development which can be achieved using integrative and combined approaches that can handle a variety of data [86–90]
Content Analysis	Medical images, sound, video, text	Big data Analytics Capabilities	
Text Analytics	Biomedical and clinical datasets	Pharmacovigilance: drug safety surveillance	
Text Analytics	Free clinical text	Patient safety event (PSE) reports	
Unstructured Data Analytics	Medical Imaging data	Deep learning	

healthcare applications deals with data discovery and curation, analysis, and result of interpretation phases similar to the traditional analytical process. But in healthcare, good data discovery, data processing, and data management need sound considerations for further development [67, 68]. Good data discovery means right data at the right time for the right context. In order to improve the context awareness in healthcare applications, there is a need to remove the gap between multiple disciplines such as computer science and medical science [69]. Hence to deal with good data discovery, data curation is more helpful to improve understanding of patient physiological and psychological care.

3.4.2 Data Interpretation

Combining structured and unstructured EHR data can improve the performance of predictive analytics. Clinical events in EHR data can be extracted and semantic space classification of similar words can be performed. This combination of structured and unstructured data by concatenating their representation using semantic space for each category of clinical events is more efficient than shared semantic space. Diverse distributed representations of clinical text from EHR using semantic spaces showed pretty notable results for predictive performance [70]. But semantic classification based on similar words can reduce the efficiency for different datasets because there is no standardization for words, and acronym, abbreviations which add more challenges to the semantic classification of data. Pharmacovigilance, phenotyping, and disease identification from health records are examples where different types of information can be extracted for different purposes from the data. EHR, EMR, PHR, and omics data contain rich information for various medical fields but there is a need to explore the ways to leverage these data

for healthcare advancement. In this regard, clinical data and genomic data have been integrated for deep phenotyping of cancer, and the proposed model was evaluated using interviews with domain experts [71]. Since the research has some limitation, that is, nor the representation standard was defined neither the evaluation of the proposed model was on real-time dataset. A comprehensive knowledge-base and precise data modeling can be helpful to leverage unstructured clinical notes of diverse formats across different facilities [72]. Extracting useful data from various types of health records is not sufficient for efficient unstructured data analytics but the interpretation or representation of data is also important for improved unstructured data analytics [73].

3.4.3 Data Quality

Several quality factors have been identified in the literature to improve and assess the quality of big data such as accuracy, completeness, consistency, timeliness, objectivity, interpretability, and accessibility [91, 92]. Unstructured data are adding more challenges to this task due to heterogeneity, lack of structure, noise, no schema, and no predefined model [74, 75, 77, 78]. Social media analytics are helpful for analyzing the psychological issues of the patient or most common diseases of society. Social media analytics is suffering from quality issues more than any other field because social media posts, reviews or comments cannot be standardized. Abbreviations, acronym, short words, and other language ambiguities generate problems for clean analytics [76]. In this regard, hashtag words can improve the efficiency of analytics but still, it needs health, media, and computer literacy to improve healthcare social media analytics [79]. Lack of structure and data heterogeneity and other quality issues can be reduced by continuous data acquisition and data cleaning for efficient healthcare analytics.

3.4.4 Context-Aware Analytics

Understanding the clinical notes in the right context is one of the challenges for unstructured data analytics. Information in the medical domain is coming from various health examiners, each examiner examines in his own way. Hence, understanding the right context is difficult due to non-standardized approaches for unstructured data while aggregation [81]. The structure is one of the important dimensions for unstructured data in terms of quality and context understanding. But the gap between semantic and syntactic structure leads to inefficient results in unstructured data analytics [80]. In this regard, the time-based topic-oriented visualization technique has been designed to find a correlation between unstructured and structured data fields for context-aware analysis [83]. But the study deals with limited text fields, whereas unstructured data in the medical field come in different formats. Also, the study steps to the initial phase of context understanding

by finding the context of written notes using words correlation. More advanced techniques with improved efficiency are required to extract useful information from unstructured clinical notes and other types of data without sacrificing the privacy of the patient [82]. Feature extraction, feature selection, and dimensionality reduction are important areas where advancement for a variety of medical data can significantly improve information extraction and analytics [85]. Anonymization and de-identification can hide the identity of the patient but still revealing or extracting maximum information from an unstructured text can be harmful to the privacy and security of data. Hence, the advancement of approaches is required to shift from data analytics to context-aware analytics by considering the ethical and privacy issues of patient data.

Efficient unstructured data analytical techniques are immense need for technological development which can be achieved using integrative and combined approaches that can handle a variety of data in healthcare [87, 88]. It means data aggregation and unstructured data analytics are important for improved analytics. Cognitive computing where computers and humans can interact in a natural way to predict, analyze, and diagnose using human cognition. Problem-oriented summary of patient EMR with IBM Watson has been proposed and generated a problem list with the help of NLP and ML. Data summarization and semantic search capability are two important factors in cognitive computing [86]. Patient safety event system has been designed where patient safety reports (containing structured and free-text unstructured data) are summarized and search approach was applied to find the most common events [89]. Unstructured healthcare data do not exist only in text form but also MRI, X-ray, and other types of visual data. Using deep learning, these data can be analyzed by several techniques of machine learning [90] but the selection of classification technique(s) for improved accuracy is critical.

3.5 *Other Issues*

There were some other issues discussed in the literature regarding unstructured data in healthcare.

- The healthcare systems are inadequate to handle data exfiltration due to unstructured data and its constantly evolving nature [93]. Patient data are personal and sensitive data which cannot be shared. But sharing is important to integrate and advanced processing of unstructured data [94]. De-identification and anonymization are used to cure privacy issues, but the deep analysis is required to identify the issues of unstructured big healthcare data and their solutions in technological advancement. Privacy protection of patient data, secure accessibility, extensive query frameworks, and unstructured data management are the challenges that cause inefficiency and inaccuracy of results with a variety of unstructured data in healthcare [95–98]

- Standardization of processes are required to deal with unstructured data quality and care, especially in integration, accessibility, and data extraction [99–103]
- Data curation is more critical for unstructured data than structured data due to several factors such as data quality, usability, heterogeneity, and high dimensionality issues of unstructured data. Automatic data labeling for disease severity in health and wellness monitoring for heterogeneous data [104, 105], semantic data enrichment [106], immense data cleaning of unstructured data for analytics [107], evaluated advanced data curation framework and data fusion techniques for unstructured data [108] are some of the main challenges that need to be considered.

4 Findings

This systematic literature review assessed the potential challenges associated with unstructured data in healthcare. Several literature reviews are available on big data analytics in healthcare, but this literature review is specific to unstructured data in healthcare and explore the challenges being created by unstructured data in healthcare. Best available literature has been identified about the healthcare systems, sources of various types of healthcare unstructured data, domains of healthcare affected by unstructured data, and problems generated by these data. Figure 3 depicts the categorization of unstructured data challenges found in the literature.

The results revealed that unstructured data are critical for healthcare analytics, preprocessing, information extraction, and interoperability, but healthcare analytics is most affected by the unstructured data. These areas need to be addressed to improve healthcare systems. This literature reveals that unstructured data are the most critical challenge for the advancement of healthcare systems because it is affecting almost all areas of healthcare. Following are the main findings of the review:

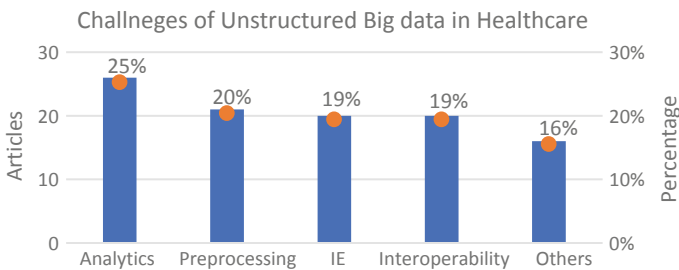


Fig. 3 Challenges of unstructured healthcare data

- Advanced techniques for data preprocessing emerged with big data processing tools are the ultimate requirement for improved healthcare systems. Efficient transformations of unstructured data, semantic data enrichment, profiling, feature selection, cleaning and other activities of preprocessing are to be assessed for multifaceted unstructured data in healthcare. Sequence of preprocessing activities, combining structured and unstructured data, mapping of content, selection of tools are some open issues to develop preprocessing tools in healthcare.
- Semantic and syntactic interoperability is important and essential for context awareness, shared use of heterogeneous data, common representation of combined data, and deep understanding of unstructured data. Standardization of processes and data can reduce the lack of uniform system but privacy and security of patient data should also be considered while achieving interoperability in healthcare systems
- Semantic information extraction in healthcare is a critical challenge to extract disease information from text. It requires a higher accuracy level. Automatic information extraction systems with improved accuracy for unstructured healthcare data with actionable granularity are required to improve the quality of healthcare systems
- Unstructured data analytic techniques are an ultimate requirement of the healthcare industry to improve the quality of care, effective and uninterrupted healthcare services, and diagnostic and decision support tools. Efficient data discovery, context-aware analytics by integrating structured and unstructured data, improving the quality of unstructured data for analysis, and effective interpretation are the ultimate challenges in unstructured data analytics that need to be addressed in future.

Unstructured data exploitation is an important task that can help to reduce the number of technical challenges in healthcare. But the quality, usability, heterogeneity, multidimensionality, scalability, and complexity of unstructured data are making healthcare analytics more challenging. Healthcare systems are unable to achieve higher accuracy and quality in service without considering the importance of unstructured data exploitation.

5 Conclusion

The exponential growth of unstructured data in healthcare forces to adopt more advanced computational systems with more innovative solutions to process the huge volume of unstructured data. Unstructured big healthcare data are rich in content and context as well which will play a pivot role in the development of more advanced and efficient technologies in future healthcare. This review explores the influence of healthcare unstructured data on analytics, clinical decision support systems, health and wellness monitoring, medicine, and smart healthcare.

Various types of unstructured healthcare data such as clinical notes, free-text EHR, handwritten prescriptions, EMR, medical imaging, X-ray, MRI, genomic sequencing, audio, and video medical data are generating challenges in all fields of technology in healthcare. Here, these challenges have been identified and discussed. The results of the literature review have shown that most of the technical challenges of healthcare unstructured data include healthcare analytics, interoperability and integration, preprocessing, and information extraction. Unstructured data have quality and standardization issues which require advanced data cleaning and semantic data enrichment in terms of preprocessing of data. Integration of health data from diverse sources, combining data from various systems to centralize the patient record is future of healthcare which can be achieved using the precise interpretation of data, secure data management, and intelligent healthcare systems. Unstructured data processing and management is an important task in order to improve context-aware analytics and semantic information extraction. The exploitation of unstructured data will maximize the efficiency of healthcare analytics and services.

Acknowledgements This research is funded by Universiti Tunku Abdul Rahman (UTAR) under the UTAR Research Fund (UTARRF): IPSR/RMC/UTARRF/2017-C1/R02.

References

1. Gantz, J., Reinsel, D.: The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. www.emc.com/collateral/analyst-reports/idc-the-digital-universein-2020.pdf (2012)
2. Turner, V., Gantz, J.F., Reinsel, D., Minton, S.: The digital universe of opportunities: rich data and increasing value of the internet of things. IDC White Paper, No. April, pp. 1–5 (2014)
3. Kiran, A., Vasumathi, D.: Predictive methodology for women health analysis through social media. In: Proceedings of the Second International Conference on Computational Intelligence and Informatics, vol. 712, Springer Singapore, pp. 511–520 (2018)
4. Correia, C., Portela, F., Santos, M.F., Silva, Á.: Data science analysis of healthcare complaints. In: Trends and Advances in Information Systems and Technologies, vol. 747, Springer International Publishing, pp. 176–185 (2018)
5. Kim, K.H., et al.: A text-based data mining and toxicity prediction modeling system for a clinical decision support in radiation oncology: a preliminary study. *Korean Phys. Soc. J.* **71** (4), 231–237 (2017)
6. Anzaldi, L.J., Davison, A., Boyd, C.M., Leff, B., Kharrazi, H.: Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC Geriatr.* **17**(1), 1–7 (2017)
7. Saiod, A.K., Van Greunen, D., Veldsman, A.: Electronic health records: benefits and challenges for data quality. In: Handbook of Large-Scale Distributed Computing in Smart Healthcare, Springer, Cham, pp. 123–156 (2017)
8. Gökalp, M.O., Kayabay, K., Akyol, M.A., Koçyiğit, A., Eren, P.E.: Big Data in mHealth. In: Current and emerging mHealth technologies, Springer International Publishing, pp. 241–256 (2018)
9. Austin, C., Kusumoto, F.: The application of Big Data in medicine: current implications and future directions. *Interv. Card. Electrophysiol.* **47**(1), 51–59 (2016)

10. Angelov, P., Sadeghi-Tehran, P.: A nested hierarchy of dynamically evolving clouds for big data structuring and searching. *Procedia Comput. Sci.* **53**(1), 1–8 (2015)
11. Kundeti, S.R., Vijayananda, J.: Clinical named entity recognition: challenges and opportunities. In: *IEEE International Conference on Big Data (Big Data)*, pp. 1937–1945 (2016)
12. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* **2**(1), 3 (2014)
13. Liu, M., Hu, Y., Tang, B.: *Role of Text Mining in Early Identification of Potential Drug Safety Issues*, pp. 227–251. Humana Press, New York, NY (2014)
14. Luo, L., et al.: A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. *BMC Med. Inform. Decis. Mak.* **16**(1), 1–14 (2016)
15. van Ooijen, P.M., Jorritsma, W.: Medical imaging informatics in nuclear medicine. In: *Quality in Nuclear Medicine*. Springer, Cham, pp. 241–267 (2017)
16. Saravana Kumar, N.M., Eswari, T., Sampath, P., Lavanya, S.: Predictive methodology for diabetic data analysis in big data. *Procedia Comput. Sci.* **50**, 203–208 (2015)
17. Marashi, P.S., Hamidi, H.: Business challenges of big data application in health organization. In: *Competitiveness in Emerging Markets*. Springer, Cham, pp. 569–584 (2018)
18. Bandyopadhyay, S., et al.: Modeling heterogeneous clinical sequence data in semantic space for adverse drug event detection. In: *Data Mining and Knowledge Discovery (2015)*, p. 31 (2015)
19. Ling, Z.J., et al.: GEMINI: an integrative healthcare analytics system. *Proc. VLDB Endow.* **7** (13), 1766–1771 (2014)
20. Wang, Y., Kung, L.A., Byrd, T.A.: Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **126**, 3–13 (2018)
21. Schmidt, D., Budde, K., Sonntag, D., Profitlich, H.J., Ihle, M., Staeck, O.: A novel tool for the identification of correlations in medical data by faceted search. *Comput. Biol. Med.* **85**, 98–105 (2017)
22. Ong, K.L., De Silva, D., Boo, Y.L., Lim, E.H., Bodi, F., Alahakoon, D., Leao, S.: Big data applications in engineering and science. In: *Big Data Concepts, Theories, and Applications*. Springer, Cham, pp. 315–351 (2016)
23. Sedghi, E., Weber, J.H., Thomo, A., Bibok, M., Penn, A.M.: A new approach to distinguish migraine from stroke by mining structured and unstructured clinical data-sources. *Netw. Model. Anal. Heal. Bioinf.* **5**(1), 30 (2016)
24. Apache Spark™—Unified Analytics Engine for Big Data (online). <https://spark.apache.org/>. Accessed 09 Oct 2018
25. Apache Hadoop (online). <http://hadoop.apache.org/>. Accessed 09 Oct 2018
26. Apache Flink: Stateful Computations over Data Streams (online). <https://flink.apache.org/>. Accessed 09 Oct 2018
27. Gomathi, S., Narayani, V.: Implementing big data analytics to predict systemic lupus erythematosus. In: *IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication systems (ICIIECS)*, pp. 1–5 (2015)
28. Wu, S.T., et al.: Generality and reuse in a common type system for clinical natural language processing. In: *Proceedings of the First International Workshop on Managing Interoperability and Complexity in Health Systems—MIXHS'11*, p. 27 (2011)
29. Scheurwegs, E., Luyckx, K., Luyten, L., Daelemans, W., Van den Bulcke, T.: Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *J. Am. Med. Informatics Assoc.* **23**(e1), 11–19 (2016)
30. Talukder, A.K.: Big data analytics advances in health intelligence, public health, and evidence-based precision medicine. *Int. Conf. Big Data Anal.* **10721**, 243–253 (2017)
31. Feldman, K., Johnson, R.A., Chawla, N.V.: The state of data in healthcare: path towards standardization. *J. Healthc. Inf. Res.* **2**(3), 248–271 (2018)

32. Yu, W.D., Kollipara, M., Penmetsa, R., Elliadka, S.: A distributed storage solution for cloud based e-Healthcare Information System. In: IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013), pp. 476–480 (2013)
33. Bhaskaran, S., Suryanarayana, G., Basu, A., Joseph, R.: Cloud-enabled search for disparate healthcare data: A case study. In: 2013 IEEE International Conference on Cloud Computing in Emerging Markets, CCEM 2013, pp. 1–8 (2013)
34. Kraus, J.M., et al.: Big data and precision medicine: challenges and strategies with healthcare data. *J. Int. Data Sci. Anal. J.* **6**(3), 1–9 (2018)
35. Genannt Halfmann, S.S., Mählmann, L., Leyens, L., Reumann, M., Brand, A.: Personalized medicine: What’s in it for rare diseases? In: Rare Diseases Epidemiology: Update and Overview, Springer, Cham, pp. 387–404 (2017)
36. Istephan, S., Siadat, M.R.: Unstructured medical image query using big data—an epilepsy case study. *J. Biomed. Inform.* **59**, 218–226 (2016)
37. Auffray, C., et al.: Making sense of big data in health research: towards an EU action plan. *Genome Med.* **8**(1), 1–13 (2016)
38. Cuggia, M., Avillach, P., Daniel, C.: Representation of patient data in health information systems and electronic health records. In: Medical Informatics, e-Health, pp. 65–89 (2014)
39. Cruz-Ramos, N.A., Alor-Hernández, G., Sánchez-Cervantes, J.L., Paredes-Valverde, M.A., del Pilar Salas-Zárate, M.: DiabSoft: a system for diabetes prevention, monitoring, and treatment. In: Exploring Intelligent Decision Support Systems, Springer, Cham, pp. 135–154 (2018)
40. Chen, E.S., Sarkar, I.N.: Mining the electronic health record for disease knowledge. In: Biomedical Literature Mining, pp. 269–286 (2014)
41. Wu, S.T., et al.: A common type system for clinical natural language processing. *J. Biomed. Semant.* **4**(1), 1–12 (2013)
42. da Costa, C.A., Pasluosta, C.F., Eskofier, B., da Silva, D.B., da Rosa Righi, R.: Internet of Health Things: toward intelligent vital signs monitoring in hospital wards. *Artif. Intell. Med.* **89**, 61–69 (2018)
43. Kozák, J., Nečský, M., Dědek, J.: Linked open data for healthcare professionals. In: Proceedings of International Conference on Information Integration and Web-based Applications and Services, p. 400 (2013)
44. Ilyasova, N., Kupriyanov, A., Paringer, R., Kirsh, D.: Particular use of BIG DATA in medical diagnostic tasks. *Pattern Recognit. Image Anal.* **28**(1), 114–121 (2018)
45. Goh, W.P., Tao, X., Zhang, J., Yong, J.: Decision support systems for adoption in dental clinics: a survey. *Knowl. Based Syst.* **104**, 195–206 (2016)
46. Leyens, L., Reumann, M., Malats, N., Brand, A.: Use of big data for drug development and for public and personal health and care. *Genet. Epidemiol.* **41**(1), 51–60 (2017)
47. Malmasi, S., Hosomura, N., Chang, L.-S., Brown, C.J., Skentzos, S., Turchin, A.: Extracting healthcare quality information from unstructured data. In: AMIA... Annual Symposium Proceedings/AMIA Symposium, pp. 1243–1252 (2017)
48. Martínez, P., Martínez, J.L., Segura-Bedmar, I., Moreno-Schneider, J., Luna, A., Revert, R.: Turning user generated health-related content into actionable knowledge through text analytics services. *Comput. Ind.* **78**, 43–56 (2016)
49. Sundararaman, A., Valady Ramanathan, S., Thati, R.: Novel approach to predict hospital readmissions using feature selection from unstructured data with class imbalance. *Big Data Res.* **1**, 1–11 (2018)
50. Delespierre, T., Denormandie, P., Bar-Hen, A., Josseran, L.: Empirical advances with text mining of electronic health records. *BMC Med. Inform. Decis.* **17**(1), 1–15 (2017)
51. Wilcox, A.B.: Leveraging electronic health records for phenotyping. In: Translational Informatics. Springer, London, pp. 61–74 (2015)
52. Simmons, M., Singhal, A., Lu, Z.: Text mining for precision medicine: bringing structure to EHRs and biomedical literature to understand genes and health. In: Translational Biomedical Informatics, vol. 939, pp. 139–166 (2016)

53. Goodman, K., Krueger, J., Crowley, J.: The automatic clinical trial: leveraging the electronic medical record in multisite cancer clinical trials. *Curr. Oncol. Rep.* **14**(6), 502–508 (2012)
54. Kotfila, C., Uzuner, Ö.: A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases. *J. Biomed. Inform.* **58**, S92–S102 (2015)
55. Alnashwan, R., Sorensen, H., O’Riordan, A., Hoare, C.: Accurate classification of socially generated medical discourse. *J. Int. Data Sci. Anal.*, pp. 1–13 (2018)
56. Husain, S.S., Kalinin, A., Truong, A., Dinov, I.D.: SOCR data dashboard: an integrated big data archive mashing medicare, labor, census and econometric information. *J. Big Data* **2**(1), 13 (2015)
57. Jackson, R., et al.: CogStack-experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust Hospital. *BMC Med. Inf. Decis.* **18**(1), 47 (2018)
58. Dinov, I.D.: Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience* **5**(1), 1–15 (2016)
59. Hu, B.V., Terrazas, B.: Building a mental health knowledge model to facilitate decision support. In: *Knowledge Management and Acquisition for Intelligent Systems*, vol. 9806, pp. 198–212. Springer, Cham (2016)
60. Pulmano, C.E., Estuar, M.R.J.E.: Towards developing an intelligent agent to assist in patient diagnosis using neural networks on unstructured patient clinical notes: inaccurate classification of socially generated medical discourse: analysis and models. *Procedia Comput. Sci.* **100**, 263–270 (2016)
61. Norman, B., Davis, T., Quinn, S., Massey, R., Hirsh, D.: Automated identification of pediatric appendicitis score in emergency department notes using natural language processing. In: *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 481–484 (2017)
62. Chapman, W.W., Gundlapalli, A.V., South, B.R., Dowling, J.N.: Natural language processing for biosurveillance. In: *Infectious Disease Informatics and Biosurveillance*, vol. 27, pp. 279–310 (2011)
63. Jonnalagadda, S.R., Adupa, A.K., Garg, R.P., Corona-Cox, J., Shah, S.J.: Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of HFpEF patients for clinical trials. *J. Cardiovasc. Transl. Res.* **10**(3), 313–321 (2017)
64. Kim, J.C., Chung, K.: Associative feature information extraction using text mining from health big data. *Wirel. Pers. Commun.* **105**(2), 691–707 (2018)
65. Clark, A., Ng, J.Q., Morlet, N., Semmens, J.B.: Big data and ophthalmic research. *Surv. Ophthalmol.* **61**(4), 443–465 (2016)
66. Syomov, I.I., Bologva, E.V., Kovalchuk, S.V., Krikunov, A.V., Moiseeva, O.M., Simakova, M.A.: Towards infrastructure for knowledge-based decision support in clinical practice. *Procedia Comput. Sci.* **100**, 907–914 (2016)
67. Sakr, S., Elgammal, A.: Towards a comprehensive data analytics framework for smart healthcare services. *Big Data Res.* **4**, 44–58 (2016)
68. Lee, C., Murata, S., Ishigaki, K., Date, S.: A data analytics pipeline for smart healthcare applications. In: *Sustained Simulation Performance 2017*. Springer International Publishing, pp. 181–192 (2017)
69. Pramanik, M.I., Lau, R.Y.K., Demirkan, H., Azad, M.A.K.: Smart health: big data enabled health paradigm within smart cities. *Expert Syst. Appl.* **87**, 370–383 (2017)
70. Henriksson, A., Zhao, J., Dalianis, H., Boström, H.: Ensembles of randomized trees using diverse distributed representations of clinical events. *BMC Med. Inform. Decis. Mak.* **16** (Suppl 2), 69–79 (2016)
71. Hochheiser, H., Castine, M., Harris, D., Savova, G., Jacobson, R.S.: An information model for computable cancer phenotypes. *BMC Med. Inform. Decis. Mak.* **16**(1), 1–15 (2016)
72. Wang, Y., et al.: NLP based congestive heart failure case finding: a prospective analysis on statewide electronic medical records. *Int. J. Med. Inf.* **84**(12), 1039–1047 (2015)

73. Jackson, K.L., et al.: Performance of an electronic health record-based phenotype algorithm to identify community associated methicillin-resistant *Staphylococcus aureus* cases and controls for genetic association studies. *BMC Infect. Dis.* **16**(1), 1–7 (2016)
74. Lovis, C., Gamzu, R.: Big Data in Israeli healthcare: hopes and challenges report of an international workshop. *Isr. J. Health Policy Res.* **4**(1), 4–9 (2015)
75. Jonnagaddala, J., Liaw, S.T., Ray, P., Kumar, M., Chang, N.W., Dai, H.J.: Coronary artery disease risk assessment from unstructured electronic health records using text mining. *J. Biomed. Inform.* **58**, S203–S210 (2015)
76. Bamwal, A.K., Choudhary, G.K., Swamim, R., Kedia, A., Goswami, S., Das, A.K.: Application of twitter in health care sector for India. 2016 3rd International Conference on Recent Advanced Information Technology, pp. 172–176 (2016)
77. Rinaldi, G.: An introduction to the technological basis of eHealth. In: *eHealth, Care and Quality of Life*. Springer Milan, pp. 31–67 (2014)
78. Persico, V.: Big data for health. In: *Encyclopedia of Big Data Technologies*. Springer International Publishing, pp. 1–10 (2018)
79. Grover, P., Kar, A.K., Davies, G.: ‘Technology enabled Health’—Insights from twitter analytics with a socio-technical perspective. *Int. J. Inf. Manage.* **43**(May), 85–97 (2018)
80. Metsker, O., Bolgova, E., Yakovlev, A., Funkner, A., Kovalchuk, S.: Pattern-based mining in electronic health records for complex clinical process analysis. *Procedia Comput. Sci.* **2017**(119), 197–206 (2017)
81. Khatri, I., Shrivastava, V.K.: A survey of big data in healthcare industry. *Adv. Comput. Commun. Technol.* **452**, 245–257 (2016)
82. Sarkar, B.K.: Big data for secure healthcare system: a conceptual design. *Complex Intell. Syst.* **3**(2), 133–151 (2017)
83. Wei, F., et al.: Visual content correlation analysis. In: *Proceedings of the first International Workshop on Intelligence Visual Interfaces for Text Analysis—IVITA’10*, no. 1, p. 25 (2010)
84. Jayalatchumy, D., Thambidurai, P.: Prediction of diseases using Hadoop in big data—a modified approach. In: *Artificial Intelligence Trends in Intelligent Systems*. Springer, Cham, pp. 229–238 (2017)
85. Buchan, K., Filannino, M., Uzuner, Ö.: Automatic prediction of coronary artery disease from clinical narratives. *J. Biomed. Inform.* **72**, 23–32 (2017)
86. Devarakonda, M.V., Mehta, N.: Cognitive computing for electronic medical records. In: *Healthcare Information Management Systems*, pp. 555–577 (2016)
87. Wang, Y., Kung, L.A., Wang, W.Y.C., Cegielski, C.G.: An integrated big data analytics-enabled transformation model: application to health care. *Inf. Manag.* **55**(1), 64–79 (2018)
88. Maitra, A., Annervaz, K.M., Jain, T.G., Shivaram, M., Sengupta, S.: A novel text analysis platform for pharmacovigilance of clinical drugs. *Procedia Comput. Sci.* **36**, 322–327 (2014)
89. Fong, A., Hettinger, A.Z., Ratwani, R.M.: Exploring methods for identifying related patient safety events using structured and unstructured data. *J. Biomed. Inform.* **58**, 89–95 (2015)
90. Singh, N., Singh, S.: Object classification to analyze medical imaging data using deep learning. In: *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pp. 1–4 (2017)
91. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* **41**(3), 1–52 (2009)
92. Wahyudi, A., Kuk, G., Janssen, M.: A process pattern model for tackling and improving big data quality. *Inf. Syst. Front.* **20**(3), 457–469 (2018)
93. Ullah, F., Edwards, M., Ramdhany, R., Chitchyan, R., Babar, M.A., Rashid, A.: Data exfiltration: a review of external attack vectors and countermeasures. *J. Netw. Comput. Appl.* **101**, 18–54 (2018)
94. Wuyts, K., Verhenneman, G., Scandariato, R., Joosen, W., Dumortier, J.: What electronic health records don’t know just yet. A privacy analysis for patient communities and health records interaction. *Health Technol. (Berl)* **2**(3), 159–183 (2012)

95. Istephan, M.R., Siadat, S.: Extensible query framework for unstructured medical data—a big data approach. In: IEEE International Conference on Data Mining Workshop (ICDMW), pp. 455–462 (2016)
96. Tchagna Kouanou, A., Tchiotso, D., Kengne, R., Zephirin, D.T., Adele Armele, N.M., Tchinda, R.: An optimal big data workflow for biomedical image analysis. *Inf. Med. Unlocked* **11**, 68–74 (2018)
97. Meystre, S.M.: De-identification of unstructured clinical data for patient privacy protection. In: *Medical Data Privacy Handbook*. Springer, Cham, pp. 697–716 (2015)
98. Aqeel-ur-Rehman, Khan, I.U., ur Sadiq ur Rehman, S.: A review on big data security and privacy in healthcare applications. In: *Big Data Management*. Springer International Publishing, Cham, pp. 71–89 (2017)
99. Gaylis, F., Cohen, E., Calabrese, R., Prime, H., Dato, P., Kane, C.J.: Active surveillance of prostate cancer in a community practice: how to measure, manage, and improve? *Urology* **93**, 60–66 (2016)
100. Hardy, L.R., Bourne, P.E.: Data science: transformation of research and scholarship. In: *Big Data-Enabled Nursing*. Springer, Cham, pp. 183–209 (2017)
101. Khennou, F., Khamlichi, Y.I., El Houda Chaoui, N.: Designing a health data management system based hadoop-agent. In: 4th IEEE International Colloquium on Information Science and Technology (CiSt), pp. 71–76 (2016)
102. Vest, J.R., Grannis, S.J., Haut, D.P., Halverson, P.K., Menachemi, N.: Using structured and unstructured data to identify patients’ need for services that address the social determinants of health. *Int. J. Med. Inf.* **107**(August), 101–106 (2017)
103. Hong, N., et al.: Integrating structured and unstructured EHR data using an FHIR-based type system: a case study with medication data. *AMIA Joint Summits on Translational Science Proceedings*, vol. 2017, pp. 74–83 (2018)
104. Rastegar-Mojarad, M., et al.: Using unstructured data to identify readmitted patients. In: IEEE International Conference on Healthcare Informatics (ICHI), pp. 1–4 (2017)
105. Boursalie, O., Samavi, R., Doyle, T.E.: Machine learning and mobile health monitoring platforms: a case study on research and implementation challenges. *J. Healthc. Inf. Res.* **2**(1–2), 179–203 (2018)
106. Zillner, S., Neururer, S.: Technology roadmap development for big data healthcare applications. *KI Künstliche Intelligenz* **29**(2), 131–141 (2015)
107. Giambrone, G.P., Hemmings, H.C., Sturm, M., Fleischut, P.M.: Information technology innovation: the power and perils of big data. *Br. J. Anaesth.* **115**(3), 339–342 (2015)
108. Banos, O., et al.: An innovative platform for person-centric health and wellness support. *Int. Conf. Bioinf. Biomed. Eng.* **9044**, 31–140 (2015)

Zip Zap Data—A Framework for ‘Personal Data Preservation’



K. Arunkumar and A. Devendran

Abstract In the era of Mobile devices, a huge volume of mobile applications produces vast and variety of data in their own formats. There is every chance that most of these data will be lost or gone forever. Let us explain how, technically any digital data becomes irretrievable if there is a change of data format or a change in application interpreting it or a change in platform feature used by the application and your app did not change to accommodate it. In simple terms, if you try to upgrade your mobile—there are a lot of possibilities that data from your previous mobile could be lost forever. Reason for loss could be your newer device comes with some obsolete platform features and hence some of your applications no longer run in the new device as developer decided to discontinue that app. Even more pressing issue is there is no way to transfer application data from old to a new device and this is very common with mobile apps. The root cause for this is the lack of backup and retrieval mechanisms in devices and applications. We studied various approaches and research works related to this problem and proposed ‘Zip Zap Data’, a framework for effective backup and recovery of personal digital data.

Keywords Personal data longevity · Personal data preservation

1 Introduction

We are seeing a huge increase in the smartphone adoption across the globe. This is the main reason we see millions of mobile applications in different mobile platform (Android, iPhone, and Windows). These applications were developed by enormous developers/companies across the globe. Most of the applications produce a lot of

K. Arunkumar
Ppltech, Chennai, India
e-mail: arunkumar@gmail.com

A. Devendran (✉)
Dr. M.G.R Education and Research Institute, Chennai, India
e-mail: devendran.alagarsamy@gmail.com

data related to users, which can be of value to the user or not. Some of them could be really confidential and private. Since there is no standard or protocol in restricting the data representation inside the application, every developer is storing data in their own tool and formats. What it means is even if you get the raw application data file from the device somehow, one cannot make use of it easily. It is only their application code which can access the data and deserialize it properly. The schema of data representation and security of accessing it are the reasons. This part closely ties data with apps on smartphones.

So what is the problem? One fine day when the app developer decides to stop supporting his app in play store, all data of that app is almost gone or at least locked with that current device. There is no way you can get it out meaningfully.

Practical problem: Moving/Upgrading to a new smartphone.

Many of my friends including me have faced the below-stated problem when upgrading to a new model of smartphone. Google, Mac, and device manufacture helps to migrate your contacts, gallery, and platform apps data. But there is no easy way you can get all your apps, especially with data. For example, for the popular messaging app ‘WhatsApp’ you have to configure your backup option and do the backup functionality regularly. Alternatively, there is an option for exporting data to your pc and later you can import back in the new device. If this is a way for migrating data to a famous app with ~ 00 million active users, consider the case of others (especially free ones) out there in the market. Hundreds and thousands of apps out there in play store are left orphans for different reasons.

Security of apps data: Most of the mobile platforms runs every app in a sandbox to ensure the safety of system resources and data. Following diagrams (Fig. 1) shows the sandboxing in iPhone. Every application needs users permission to access additional system resources like using storage, camera, internet(wifi, mobile data), bluetooth, etc. Also, data from platform apps like SMS, Gallery, Calendar, etc.

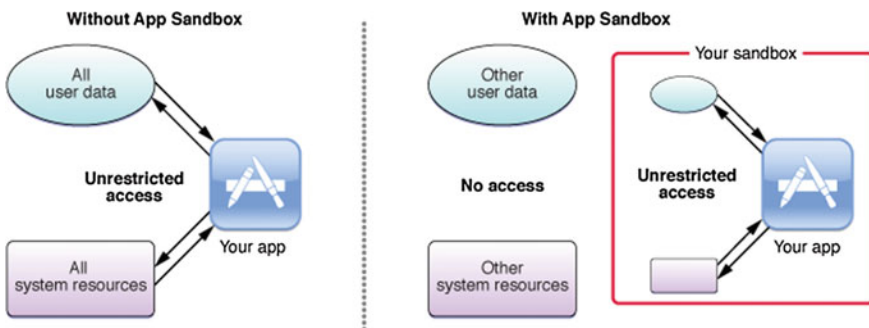


Fig. 1 iPhone App Sandbox model for data security

1.1 *Digital Data Preservation*

Having said the problem in accessing and interpreting mobile data, can solution to the well know problem of digital realm ‘Digital data preservation’ help us here?

What is digital preservation? In the definition medium from American Library Association(ALA) => ‘The goal of digital preservation is the accurate rendering of authenticated content over time.’

In other words, Digital preservation is the mean of ensuring access of digital data across time (10s, 100s or even 1000s of years). Clearly, this impacts the future of ‘history of events’, the future of knowledge, culture, evidence, etc.

Couple of statements to understand the seriousness of ‘Data Preservation’

1. In his interview on BBC @ 13 February 2015—Vint Cerf, one of ‘the fathers of the Internet’, expressed his concern that today’s digital content might be lost forever. His talk mainly points on the fact that—if technology continues to outpace preservation tactics, future citizens could be locked out of accessing today’s digital content enter a ‘digital dark age’.
2. “Digital data lasts forever or 5 years, whichever comes first”. This is from a Scientific American journal released back in 1995 written by Jeff Rothenberg [10].

Discussing in detail about digital longevity is out of scope.

Fundamental problem with digital data: A sequence of bits is meaningless if it cannot be decoded and transformed to some meaningful representation and viewability. Example: consider documents created with ‘WordStar’ (a 1980s software for data handling), are unreadable today as there is no software to interpret this data.

Another important question is what to preserve? Clearly, we need to preserve digital data for our future generation. Should all details available in the digital realm be preserved? If not, what should be preserved? Who will define the archival selection—what and whose story gets told, and whose do not. Again discussion on this direction is out of the scope of our work.

The main concern here is the data is being generated by a wide range of electronic devices and the quantities of data being generated are really huge. Now, digital data comes from personal computers, smartphones, game devices, digital cameras, digital recorders, digital TVs, and so on. The main cause for this exploded data quantity is the increase in dependency of our modern society on digital information and devices. Without proper strategies and technology support for preservation of all these data are under huge risk of being lost.

2 Core Problem

In digital preservation paradigm, one of the most difficult challenges is to maintain the interpretability of data across changing technologies. As the sequence of bits is meaningless if it cannot be decoded and transformed into some meaningful representation.

2.1 *Internet Era Data Problem*

Most of the modern world data be it books, images, videos, music exist as digital data. All these are represented and stored in different formats. Some of these data formats are standard and others are proprietary (purely dependent on software generating them). In the internet-connected world, anyone with proper privileges can access them instantly. The data generated can be static (created and stored once) or dynamic (generated on every time a request for access comes).

The irony is even though most of the content which could be static is often dynamically generated(i.e., through execution of code to create content at runtime).

The challenge is how do we preserve these data over time? It is no longer the three static object or collection of static objects to be preserved. It is the entire stack of technology that is serving content along with its data needs to be frozen and preserved., i.e., We need to freeze and precisely reproduce the execution that dynamically produces the content.

What is the problem there? Preservation and reproduction of software execution are one of the most complex problems. Because this involves perfect alignment of many moving parts to make it work, like OS support for older system libraries, dlls, implementation specific libraries, and hardware and firmware. Preserving this alignment over time is difficult. Many things can change hardware, operating system, linked libraries, configuration and user preferences, geographic location, execution timing, and so on. Even a single change may completely break execution.

2.2 *Mobile Devices Data Problem*

Across the globe, 52.3% of internet traffic comes from mobile devices and it has been increasing year by year. The study and inference on the mobile device adoption and usage show it [12]. And for data generation, even a greater percentage of the share goes to mobile. Mobile data can be generated by native apps or mobile web apps running in a browser. Recently, we see a shift towards PWA (progressive web app) from native apps, coz of its lightness and features or maybe google is

batting for it. Technically speaking, PWA is more of an optimized mobile web application with a shortcut in app screen.

In general, we can group apps under 2 categories, from data generation point. Both brings its own complexity.

- Native mobile apps (running in sandbox)
- PWA or Web apps (running in browser sandbox)

2.3 Personal Data Preservation

Digital data preservation is a long and still open research area. There are various solutions proposed and implemented till date. All the solutions can be broadly classified into 2 categories (migration and emulation), based on the strategy to ensure longevity [11]. We restricted our focus on ‘Personal data preservation’ [14] using the emulation approach and proposed a framework for preserving personal data from modern devices especially smartphones.

2.4 Data Privacy—Huge Concern

Data privacy is a huge concern these days. Every big corporate with personally identifiable information or other sensitive information about users is investing lots in ensuring the data privacy of its data. When you should worry—You are collecting, storing, using above mentioned user data – in digital form or otherwise. Consider your healthcare records, financial transactions, biological traits, such as genetic material, residence and geographic records in wrong hands. The challenge of data privacy is to utilize data while protecting individual’s privacy preferences and their personally identifiable information.

2.5 Secured Documents Problem

The simplest way to achieve secrecy and longevity is to store the secret in a distributed way on many servers and put a constraint that a minimum number of shares be required to reconstruct the secret. From a long term archival point of view, any retrieval of such secured document needs some piece of software along with the data for the construction of secret key from k or more pieces. There is a lot of research work happening in this area.

3 Related Works

3.1 Parity Cloud Service: A Privacy Protected Personal Data Recovery Service

Here the authors mainly talk about the privacy protection of data in the cloud, an important problem in providing personal data recovery as a service. They proposed a novel mechanism ‘data recovery service framework’ built on cloud infrastructure. PCS (Parity Cloud Service) can be easily set up on any cloud provider. The main differentiating factor in the proposed framework is that zero user-specific data to be uploaded/saved to the server for data recovery.

3.2 Algorithm for Backup and Recovery of Data Stored on Cloud Along with Authentication of the User

Authors of this work propose a scheme of storing data on the cloud server along with a copy in the remote server. They propose the process of taking a backup copy in remote server regularly. If the main file is lost, it can be recovered from the remote server. To make backup and recovery more secure, user authentication based on attributes, secret keys are used. Every transaction in the system is authenticated per above scheme for security.

3.3 Securing BIG Storage: Present and Future

In this work, the authors reviewed existing works in cloud computing technologies and security threats of BIG DATA storage in cloud environments. Also, existing solutions that address security threats. They also discussed Internet of Things (IOT) and its influence in BIG DATA and propose solution to various threats identified.

3.4 A Framework for Aggregating Private and Public Web Archives

Here the authors talk about the problem of existing public web archive systems and HTTP memento RFC-7089 [13]. The problem with users personalized web pages (Facebook like), feeds and even instance of a private web application (banking like) in existing archiving systems and ways to mitigate them. They worked on amending existing Memento and Timestamp syntaxes of HTTP archiving to accommodate the private data.

3.5 Syncing Mobile Applications with Cloud Storage Services

The authors analyzed various cloud providers APIs and in the context of mobile application data sync. Their experiments show various options for mobile app developers to choose among the providers to sync their app data across multiple devices.

3.6 Commercial Products

Most of the commercial applications doing entire app backup and restore with data need a rooted device for accessing the data specific to the application.

Following list of applications fall in this category:

Helium, MyBackup Pro, Titanium Backup, Ibackup, etc.

Google Backup & Restore—is a recent feature from google, which will allow the user to store some of users app data like calendar, Browser, Contacts, Gmail, Photos, and Music to his google drive account and retrieve it when needed. However, this is not a complete solution for the backup/restore.

3.7 Limitation of Existing Backup/Restore Apps

1. They need a rooted device (so that root permission can be granted to the backup app)
2. Google vendor lock and not a complete backup and restore for everything
3. Some of them only backup apps (not the data generated by them) and common user data like contacts, photos, videos, downloads, etc.

4 Our Contribution

The fundamental problem with the existing backup and restore system is closely tied with the security model of android applications. All the files created in an app are stored by default in app’s own sandbox, which is not visible to any other app running in the device. Unless you have an app running with root privilege, nothing can see the data created by it.

4.1 Core Idea

The novelty comes in bringing the backup/restore action to the app level and breaking the security barrier of the android system. Also, another important part of the idea is storing data in end user cloud storage and not in any app developer or company cloud storage, which opens doors for privacy vulnerabilities.

To put it simple Zip Zap Data

1. Store user data in the user's cloud space and retrieve it when needed.
2. Store app data along with the app and its metadata like device model, OS, etc.
3. All the data is compressed and encrypted before storing and applied in reverse order @ retrieval time.

4.2 Zip Zap Data Framework

Application specific data:

Most of the useful mobile application has some data starting from user preferences, game scores, to do notes to complex databases. As an app developer, depending on the use case one can store data in a mobile device or some external cloud storage. Some apps were built to operate 100% online and store all its data in the cloud. Even those apps started storing part or some data in a mobile device to provide offline support.

This offline feature in most applications gives seamless behavior to users and crucial for their user engagement in developed/developing countries. The point we try to make here is—every modern app being developed have a lot of application-specific data. It is this data residing in the device needs to be backed up and restored properly to give seamless transition when upgraded.

Synchronize data across multiple device:

These days, it is very normal for a user to have more than one device and very likely that he/she uses the same application in all of them. As an end user—I expect the app should be intelligent to be in the same state across all my devices. This use case brings its own complexity to app developers to write special mechanism to synchronize data across multiple devices of the same user. Even though our work focused on the data backup/retrieval part, a properly implemented system could achieve synchronization across devices with our framework.

4.3 Data in Mobile Application

Almost all the apps even the ones storing data in cloud stores data in device memory (for storing user preferences or to give offline capability). Let us get into

implementation details. As an app developer, storing data in a mobile device can be done in three ways.

- Shared preferences: This is a key, value store provided by android SDK. It is straight forward to use. There are ways in which you can make this data visible outside application sandbox. One has to set ‘Context.MODE_WORLD_READABLE’ or ‘Context.MODE_WORLD_WRITABLE’ on the preference for this to work.
- Local files: Any mobile application can create/update/delete raw files (text or any binary files) with standard file system calls. The files created are completely inside the sandbox of application and no other application can see this. These files cannot survive multiple installations even on the same device.
- Local Databases: From a technical point of view, this is very similar to local file case. All the raw files of DB are stored within the sandbox of app.

5 SDK for Backup & Recovery

As developing backup/retrieve capabilities require a considerable time and additional skills on the part of developers. If an easy to use and off the shelf utility is available for this purpose, developers would be too happy to adapt it. Hence, we developed an open-source SDK for this purpose. Zip Zap Data Framework allows developer separate interfaces to handle these three cases of app data stored in devices mentioned above in Sect. 4.3. This SDK solves the fundamental backup/recovery problem along with other benefits listed below.

Solving—Data preservation problem:

As we are storing ‘app metadata’ along with the data, we are ensuring the retrieval of the data stored any time in future, i.e., the data is readable ‘simply...forever’. The fundamental philosophy is to recreate the original environment [11] for the app and its data @ the time of retrieval. Worst cases might involve making a new application or providing dependencies to a lost/obsolete library component.

Solving—App data sync problem:

Integrate app data sync feature into your android, iOS and web apps with just a few lines of code without any server setup or cost. Enable your users to have a seamless experience across device usage. Allow him to move to new devices without any issue of data loss.

Unlike existing cloud provider solutions like parse or azure or others, which require payment from developers for the sync feature, the proposed solution won’t require any cost from the developer. This is possible as we are going to use end user’s cloud storage (like dropbox, Google Drive, iCloud, SkyDrive, etc.) for storage part and also the data privacy part is taken care.

Avoiding cloud vendor lock in problem:

The key thing in our proposal is to store user's data in user's own cloud space (like dropbox, Google Drive, iCloud, SkyDrive, etc.), which avoids any vendor locking for server-side implementation for giving sync mechanism. As we are storing data in compressed and encrypted form, concerns on the privacy of user data and security are addressed.

5.1 How It Works?

1. The cloud storage of the user is configured in that app once for every device.
2. The SDK part running in every application identifies and send application data be it local files or shared preferences or sqllite DB to the backup and restore app regularly.
3. For optimized data transfer from device to cloud storage, we have version based files maintained across all apps using SDK.
4. The delta of the data to or from the cloud is identified as mentioned above and transfer is done when the user goes online.



5.2 Zip Zap Data Sync Backup and Restore

Zip Zap Data Sync is the open-source app that backups & restores apps and app data. This app needs no rooted devices to backup and restores user app-specific data. It gets the data from the app through SDK and uses it for backup/restore.

Along with the app data, user can get the entire list of installed apps in the device. There is a provision to maintain and retrieve multiple versions of apps something like marking a checkpoint and retrieve from that stable point.

6 Integration and Results

6.1 Integration Steps

SDK is available as an archive file (source code for the same is available in github). Integration can be done in a few steps after configuring the file, shared preference or DB details. It can be done in below four steps.

1. Import the ‘Zip Zap SDK.aar’ file as a new module
2. Add a dependency in the application gradle file ‘compile project(‘:zipzapsdk’)’
3. For restoring data or fetching data from cloud storage use ‘`sync.getZipZapData();`’
4. For data backup or store data to cloud storage use ‘`sync.putZipZapData();`’

6.2 Open-Source Integrations

We have integrated our SDK with some open-source android application and made them available in git after our integration [9]. Now, these apps have Backup/Restore feature along with seamless data sync across multiple devices of the same user.

Application	Local data	Usage
AndroFish	Shared Preference	A simple game, where user can guide fish to eat the other fish and grow bigger
Android-PreferencesManager	Shared Preference	Preferences Manager is an open-source application that allows you to seamlessly edit application’s preferences
Android-pedometer	Local File	App for Android phones that counts your steps
Simple-Draw	Local File	A canvas you can draw on with different colors
LeeCo	Local File	LeeCo is an awesome app for (including unlock) problems, solutions, discuss(from leetcode) and comments
todo.txt-android	Local File	Official Todo.txt Android app for managing your todo.txt file stored in Dropbox

(continued)

(continued)

Application	Local data	Usage
Design game-pegboard	Local File	A simple design game with pins—to demonstrate the usage of ZipZapData SDK
OpenSudoku	Local DB	open-source Sudoku game for Android
CoCoin	Local DB	CoCoin, Multiview Accounting Application
Runnerup	Local DB	An open-source run tracker for Android
Simple-Notes	Local DB	A simple textfield for adding quick notes without ads

7 Conclusion

The main idea of the proposal is moving the responsibility of backup/restore data of an application to application itself rather than a separate stand-alone app on its own. This is working well and addresses the fundamental security model. This works well for real time app sync across the devices. With our integration on 10 + open-source applications, we observed backup/restore functionality can be added with few lines of code. App sync is a bit tricky as put Zip Zap Data () and get Zip Zap Data () needs to be called in proper life cycle of the corresponding activities. But we have seen it working well in ‘OpenSudoku’ game.

The current implementation of android app ‘Zip Zap Data Sync’, giving the feature of synchronize of device files and user’s cloud storage is more like the – ‘One click to Backup all’ feature most user will like to have. Also, the storage choice of users cloud in encrypted form adds great value to the system by addressing the data privacy concerns.

We are planning to integrate more android apps with our SDK both open-source and commercial apps. Our development of SDK for iOS and Windows platform is in progress and will be available in github shortly. Another addition will be the flexibility of user to create multiple checkpoints for user data being preserved similar to memento work, which will let the user replay various instances of data with the application at any point of time.

References

1. https://en.wikipedia.org/wiki/Information_privacy
2. Gupta, V.H., Gopinath, K.: An extended verifiable secret redistribution protocol for archival systems. In: International Conference on Availability, Reliability and Security (ARES) (2006)
3. Song, C., Park, S., Kim, D., Kang, S.: Parity Cloud service: a privacy-protected personal data recovery service. In: IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (2011)

4. Raje, M., Mukhopadhyay, D.: Algorithm for back-up and recovery of data stored on cloud along with authentication of the user. In: International Conference on Information Technology (ICIT) (2015)
5. Bhargavi, I., Padmaja, M., Ammayappan, K.: Securing BIG storage: present and future. In: Online International Conference on Green Engineering and Technologies (IC-GET) (2016)
6. Kelly, M., Nelson, M.L., Weigle, M.C.: A framework for aggregating private and public web archives. In: JCDL ‘18 Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (2018)
7. Pocatilu, P., Boja, C., Ciurea, C.: Syncing Mobile Applications with Cloud Storage Services. *Inform. Econom.* **17**(2/2013) (2013)
8. <http://www.techradar.com/how-to/phone-and-communications/how-to-backup-your-android-device-1298811>
9. ZipZapData SDK and Sample Apps. <https://github.com/zipzapdat/>
10. Rothenberg, J.: Ensuring the longevity of digital documents. *Sci. Am.* **272**(1), 42–47 (1995)
11. Arunkumar, K., Devendran A.: Digital data preservation—a viable solution. In: Data Management, Analytics and Innovation, pp 129–141 (2018)
12. <https://www.statista.com/statistics/241462/global-mobile-phone-website-traffic-share/>
13. <https://tools.ietf.org/html/rfc7089>
14. Poursardar, F., Shipman, F.: What is part of that resource? User expectations for personal archiving. In: JCDL ‘17 Proceedings of the 17th ACM/IEEE on Joint Conference on Digital Libraries (2017)

A Systematic Mapping Study of Cloud Large-Scale Foundation—Big Data, IoT, and Real-Time Analytics



Isaac Odun-Ayo, Rowland Goddy-Worlu, Temidayo Abayomi-Zannu and Emanuel Grant

Abstract Cloud computing is a unique concept which makes analysis and data easy to manipulate using large-scale infrastructure available to Cloud service providers. However, it is sometimes rigorous to determine a topic for research in terms of Cloud. A systematic map allows the categorization of study in a particular field using an exclusive scheme enabling the identification of gaps for further research. In addition, a systematic mapping study can provide insight into the level of the research that is being conducted in any area of interest. The results generated from such a study are presented using a map. The method utilized in this study involved analysis using three categories which are research, topic, and contribution facets. Topics were obtained from the primary studies, while the research type such as evaluation and the contribution type such as tool were utilized in the analysis. The objective of this paper was to achieve a systematic mapping study of the Cloud large-scale foundation. This provided an insight into the frequency of work which has been carried out in this area of study. The results indicated that the highest publications were on IoT as it relates to model with 12.26%; there were more publications on data analytics as it relates to metric with 2.83%, more articles on big data in terms of tool, with 11.32%, method with 9.43% and more research carried out on data management in terms of process with 6.6%. This outcome will be valuable to the Cloud research community, service providers, and users alike.

I. Odun-Ayo (✉) · R. Goddy-Worlu · T. Abayomi-Zannu
Department of Computer and Information Sciences,
Covenant University, Ota, Nigeria
e-mail: isaac.odun-ayo@covenantuniversity.edu.ng

R. Goddy-Worlu
e-mail: rowland.goddy-worlu@stu.cu.edu.ng

T. Abayomi-Zannu
e-mail: temidayoabayomizannu@gmail.com

E. Grant
School of Electrical Engineering and Computer Science,
University of North Dakota, Grand Forks, USA
e-mail: emanuel.grant@engr.und.edu

Keywords Big data · Cloud computing · Internet of Things · Real-time analytics · Systematic mapping

1 Introduction

Cloud is popularly known as a distributed and parallel computing system which comprises a multitude of virtualized and interconnected computers which are provisioned dynamically and shown as at least, one unified computing resource established through negotiations among the service providers and clients which is in the form of a service level agreements (SLAs) [1]. Cloud computing provides storage and compute resources that appear inexhaustible. This is essential to the utilization of Cloud's broad-scale foundation for real-time analytics, IoT (Internet of Things), and big data. Several Cloud service types are known but three main ones stand out, namely Software-as-a-Service (SaaS) which offers customized apps to its users across the Internet and the CSP (Cloud Service Provider) which takes the burden of installation and license from the user. Platform-as-a-Service (PaaS) enables the users to deploy and develop various apps on the CSP's infrastructure, while Infrastructure-as-a-Service (IaaS) allows the Cloud users to have control of some of the Cloud service provider's infrastructure and the consumer benefits from the massive compute and storage capacity available to the CSP. Cloud computing is turning out to be very valuable and the services being offered are continuously being improved on a day-to-day premise due to the strong indispensable applications and infrastructure [2, 3]. Cloud computing also has four deployment models and these are private, public, community, and hybrid Cloud. The private Cloud is hosted on-premises in an organization and could either be run by a third party or the organization itself, but in-house staff is utilized. Public Cloud are hosted by major Cloud service providers that possess the resources to build large data centers across various geographical locations. Community Cloud are usually hosted by institutions and organizations having common interests. Hybrid Cloud are used to take advantage of the different Cloud types. Organizations may want to retain core operations on-premises on a private Cloud at the same time, a lesser amount of essential data and services are migrated to the public Cloud. That being said, because of the multitenancy and virtualization procedure in Cloud, a lot of security issues have begun to surface [4, 5].

Big data deals with large volumes of data, IoT involves data transfers between "things", while real-time analytics relates to the processing of data in real time. The common denominator here is usually the processing of a large amount of data and based on the models and service types, Cloud computing has the infrastructure to support the processing and storage of large volumes of information. The big data paradigm often points to a vast and convoluted datasets which conventional processing does not have the capacity to capture, store, and analyze [6]. The essence is to obtain intelligence from such complex big data, which also requires more computing and storage power which could be administered by the Cloud [6].

The growing interest in IoT and big data requires an understanding of their definitions, challenges, and potentials [7]. Big data analytics is needed in the advancement in IoT services and processes due to the ever-increasing amount of data being generated daily [7]. The pace at which current data is being generated is astonishing and at this rate, it has surpassed the available capability to create the appropriate Cloud computing platform for the analysis of data which is a major challenge for researchers [8]. Again, one may be wondering about the connection between all these and the Cloud. This is simply summarized by Amazon Web Services (AWS). Analyzing a huge set of data requires a powerful compute capacity which can change in size depending on the volume of input data and the type of analyses [9]. Despite the fact that Cloud service providers are striving to offer reliable and efficient services on the Cloud, there is a major problem of trust [10]. The data workloads characteristics are impeccably appropriate to the pay-as-you-go model of the Cloud, whereby applications can easily be scaled down or up depending on the demand [9].

In embarking on a research or writing an article, a technical interest area must be considered by the researcher. This includes understanding the topic by reviewing relevant studies relating to that particular topic. It generally involves reviewing multiple journals, conference proceedings, books, etc. In addition, determining an area of interest may also require a lot of search on digital libraries, attending workshops, seminars, and conferences. Researchers can often stumble onto new and often unanticipated research ideas through long hours spent reviewing other people's research and through the process of conducting research as well. Also, researchers may become interested in particular research in a specific observed phenomenon serving as motivation to carry out a large-scale research in all fields and areas of study. To summarize, the researcher's curiosity about an observed phenomenon can usually provide a satisfactory impetus for selecting a research topic. Clearly, the way toward deciding a research topic is sometimes usually complicated. In the area of Cloud large-scale foundation, a large number of research have been undertaken especially as it relates to IoT, big data, and real-time analytics, hence it has become imperative to sum-up and provide an analysis of some of the work that has been done in this field. A systematic mapping study is the process of categorizing research by utilizing a structure and scheme and reporting the result using a map [11].

In this paper, a systematic mapping study presenting the frequency of studies of Cloud large-scale foundation is presented. Such information allows further insight into research carried out so far and areas requiring further attention. The systematic map was obtained by using three facets because of the areas being investigated and these three facets are as follows: the research facet which focuses on the types of research conducted and result within such areas; the contribution facet that deals with the mode of research, in terms of method and tool; and, the topic facets which extracts core topics in Cloud large-scale foundation. The purpose of this paper is to conduct a systematic mapping study of Cloud large-scale foundation for IoT, big data, and real-time analytics. The remaining paper is organized as follows:

Sect. 2 examines related work. Section 3 discusses the systematic mapping process. Section 4 presents the results and discussion. Section 5 is the conclusion with a suggestion for future work.

2 Related Work

In [12], the planning stage of a systematic mapping study was analyzed and it identified the software patterns as evident all through the requirement engineering phase of multiple projects, seeking for a better understanding of the parts being carried out by these patterns which is usually dependent on the fundamental parameters needed in the development procedure. A protocol was created for the study with the fundamental steps to replicate such a work in the research community in order to ascertain the validity of the research. The digital libraries that were utilized in their work were IEEEExplore, ACM DL, Web of Science, and Scopus. The guidelines laid down in [11] were followed for this study.

In [13], they dwelled on the protocol's depiction for a systematic mapping study as it relates to DSL (domain-specific languages). Their study is diverted toward a better understanding of the domain-specific language research area with an emphasis on future directions and research trend. This study encompasses between July 2013 to October 2014, and it takes advantage of three rules for carrying out a systematic review which are conducting the review, planning, and reporting such.

In [14], the systematic mapping study is dependent on the examination of the utilization of idea maps within Computer Science and their study conveys the result which was centered on the evaluation and collection of previous research on concept maps within Computer Science. For this study, five different electronic databases were utilized. Manual approaches and backward snowballing were utilized in the search procedure. This study also showed a rich investigation and massive interest in concept maps because of teaching and learning support toward that path or area. The search strings of this study were applied on ScienceDirect, Scopus, ACM DL, Compedex, and IEEEExplore.

Within the research work of [15], systematic mapping to inquire how game-related methods have been utilized in software engineering education and how these methods bolster explicit software engineering knowledge domains with the future directions and research gaps recognized. The essential investigations of the work tie down on the assessment of games, their factors, and use on software engineering education. Based on publications from 1974 to 2016, a sum of 156 essential investigations was recognized in this research and the mapping procedure was carried out in tandem with [11].

The research in [16] carried out a power system model mapping centered upon the overview of the power system models and applications being utilized by organizations in Europe, and identification and analysis of both their modeling gaps and features. 228 surveys were sent to power experts for information elicitation,

and 82 questionnaires were completed and the knowledge mapping was done accordingly.

In [17], a domain-specific languages systematic mapping was carried out with an essential enthusiasm for the kind of research, the type of contribution, and the focus area. The work features an inquiry from legitimate sources between 2006 to 2012 with the systematic mapping carried out dependent on highlighting the research questions, data extraction, search, classifying, and conducting the screening. The materials utilized in this research are experience papers, opinion papers, solution proposal, conceptual or philosophical papers, and validation of the research materials.

A systematic mapping was conducted in [18] on legal core ontologies and was based on [19] concepts. The work based its search more on both “legal concepts and theory”. The studies selected were classified based on the contribution as reflected in tool, language, model, and method. Other steps include the identification of focus with a clear recommendation to use the two ontologies, identification of the used legal theories in legal core ontologies building process, and the analysis of every chosen research for cogent deductions about ontological and legal research.

In [11], a systematic mapping study in software engineering was conducted and is a foundation for many systematic mapping studies. It provides guidelines for the conduct of systematic mapping studies. In addition, a comparison of systematic reviews and systematic maps was also carried out. The work revealed that systematic maps and reviews are not the same in terms of goals, breadth, validity measures, and implications, because they employ different analysis methods.

The work in [20] provides an observational research in software Cloud-based testing based on systematic mapping. The procedure involves building a categorization scheme making use of both functional and nonfunctional testing methods, which were examined closely alongside the applications of their methods and peculiarities. They made use of 69 main analyses as discovered in 75 publications and just a small amount of the examination unites exact factual investigation with quantitative outcomes. Many of the analyses made use of a singular experiment for the assessment of their proposed solution and there has not been any research focused on systematic mapping of Cloud large-scale foundation for IoT, real-time analytics, and big data from the literature.

3 The Systematic Mapping Process

A systematic study provides an overview of a field of interest by depicting article frequencies in that area. This research was carried out by utilizing the formal steps on systematic mapping research in [11] and [19]. A systematic mapping research is a replicable process of eliciting and translating the available materials associated with a research objective [21]. There are definite procedures for designing a systematic map [11]. Usually, where the scope of the review is developed, the definition of

research questions is the initial step. An inquiry is normally conducted on each and every paper accessible in the proposed area of study and after the search, such papers are examined to discover those important and relevant to the study. Next is the classification scheme based on the use of keywording that is applied to the abstracts of the searched papers. Finally, there is a process of data extraction that prompts the making of the systematic map and each step is utilized in the making of the systematic map of Cloud large-scale foundations for IoT, big data, and real-time analytics. In the perspective of the paper choice criteria in terms of the requirements and research question, a sum of 112 publications were appropriate for inclusion based on the initial 1,716 papers from 2011 to 2018 and the list of selected main research papers are in the Appendix.

3.1 Definition Research Question

The significance of a systematic map is to offer a better understanding of the frequencies and types of research that has been done in an area under consideration. It might likewise be important to find out where such research articles were distributed. These issues determine the focus of the inquiries that are utilized for the research. In this particular paper, the research questions are

- RQ 1: What and how many articles cover the different areas in Cloud large-scale foundation for real-time analytics, IoT, and big data and how are they addressed?
- RQ 2: What evaluation and novelty do they constitute in particular, and what papers are being published in this field/area?

3.2 Conduct of Research for Primary Studies

The strategy of inquiry of primary studies is mostly accomplished by examining unique digital libraries. This is done through searches carried out on conferences and journal publications. To acquire the necessary papers needed for this systematic mapping paper, a search was performed on some scientific digital libraries that are accessible online and information from printed materials and books were not focused on in the search. The search was conducted on journal and conference

Table 1 Digital libraries used for the systematic mapping study

Digital libraries	Uniform resource locator
ACM	http://dl.acm.org/
IEEE	http://ieeexplore.ieee.org/xplore
Science direct	http://www.seciencedirect.com/
Springer	http://www.springerlink.com/

publications in four major digital libraries due to the high-impact factors of their articles. Table 1 shows the URL and the databases used in this research.

The outcome, population, comparison, and intervention in the focus of study influenced the design of the search string. The keywords utilized in the search string were gotten from every aspect structure of the title for this research on large-scale foundations for real-time analytics, big data, and IoT. The search string being utilized on major digital libraries are

(TITLE(Cloud) AND (TITLE(“Large scale”) OR KEY(“Large scale”)) ((TITLE(“Big Data”) OR KEY(Big Data”) OR (TITLE(IOT) OR TITLE (“Internet of things”) OR KEY(“Internet of things”)) OR TITLE(“Real time analysis”))) AND (ALIMIT-TO(BUBJAREA, “COMP”)).

The searches were carried out by making use of the above-customized search string on document metadata in order to make sure that the important articles were not omitted and appropriate databases significant to Computer Science and the Cloud were considered.

3.3 Screening of the Papers for Inclusion and Exclusion

The essence of a selection process is to discover and include each and every publication important to the review being carried out. It is a normal procedure to use exclusion and inclusion criteria to remove articles which are not important to real-time analytics, IoT, and big data in Cloud computing. The exclusion and inclusion criteria are also utilized to eliminate publications which are not important to the research questions. A few abstracts tend to point out the main focus without any further details, hence such abstracts were not considered in the review process. This study excluded papers on panel discussions, tutorials, summaries, and prefaces because they do not contain abstracts. Articles that discussed the main focus of the paper and some measure of secondary aspects were considered. The list of included papers is in the Appendix. The primary focus of this research is on big data, IoT, and real-time analytics, therefore, the exclusion and inclusion procedure was carried out as shown in Table 2.

Table 2 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
The abstract explicitly mentions real-time analytics, big data, and IoT as it relates to Cloud computing. Furthermore, the abstracts has a secondary focus on large-scale foundation to a certain extent	The paper lies beyond the area of Cloud computing notably as it related to big data, IoT, and real-time analytics

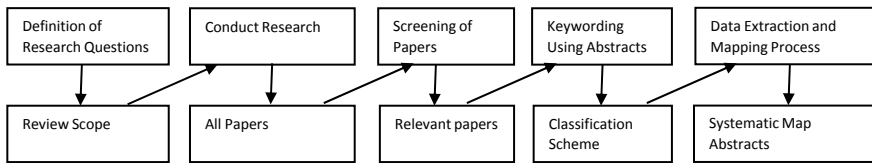


Fig. 1 The systematic mapping process [11]

3.4 *Keywording of Abstracts*

The keywording of abstracts is the process that leads to the design of a classification scheme. The systematic mapping process shown in Fig. 1 involves the entities listed below [11].

- Abstract
- Keywording
 - Article
 - Sort article into scheme
 - Update scheme
- Systematic map

Keywording is obviously required to lower the time needed to create a systematic map for real-time analytics, big data, and IoT in the Cloud. Also, keywording guarantees that the classification pattern covers every important study and the process includes examining the abstracts in order to extract ideas and keywords from different articles on the focus of study. Keywords from different articles are usually combined to provide a better understanding of the contribution and type of research. Subsequently, this was used in order to better comprehend the set of classifications needed for the study, but it was also crucial in some cases to also examine the introduction and conclusion of some publications so as to ensure reliable keywording. A cluster of keywords was used in order to know the types of facets to be used for creating the systematic map for this study and three main facets were eventually utilized for creating the map. The first facet is the topics category that utilizes the keywords that were directly related to the topic of study. Hence, the topic facet used keywords such as real-time, big data, IoT, orthogonal, and data analytics. The second facet focused on the type of contribution available in a research type. It considered the research contribution in terms of metric, model, method, tool, and process as it relates to the field of study under consideration. The third facet sees to the nature of study carried out.

3.5 Research Type Facet with Category and Description

The essence of the third facet is to determine what types of research were conducted and the applicability to the area of study. The style of the categorization of research approaches by [22] was utilized in this facet and has the following categories and description:

- **Evaluation Research:** Techniques adopted here have been implemented and evaluated. The outcome of such implementation also considered advantage and disadvantages.
- **Validation Research:** Processes utilized are novel, though are still to be applied.
- **Solution Proposal:** Research provides a unique answer to an issue and the advantages plus application of such solutions can be verified.
- **Philosophical Papers:** The method introduces a new way of examining a challenge with the use of concepts and frameworks.
- **Opinion Papers:** This kind of study does not adopt any known research methodology; it naturally expresses the researcher's opinion.
- **Experience Papers:** Relate the researcher's unique experience and explain the way things can be accomplished.

These categories were thought to be adequate and suitable for their utilization in this research. The papers included in this research were reviewed on this basis of the categories of research discussed above.

3.6 Data Extraction and Mapping of Study

During the keywording process, pertinent articles were arranged using a classification design. The following step was to extract data from the relevant publications and this procedure determines the nature of the classification scheme. During this process, new categories could be added, some categories could be eliminated, while some categories could be merged. The data extraction process for this study was done using Microsoft Excel tables which contain frequencies of publication based on different research types and contributions. The frequencies of publication are determined by combining either the contribution or topic tables or the research or topic type tables. The analysis focused on presenting the rates of publication as obtained from the Excel tables which help to identify which areas of the study were emphasized more. Furthermore, it facilitates the identification gaps in the publications that could stimulate further research.

By using the results of the Excel tables, bubble plots were created for the purpose of visual presentation of the frequencies of publications and the systematic map was produced using a two x-y scatter plot with bubbles to indicate the number of articles based on the classes under consideration. There were two quadrants in the systematic map that was created in this study as a result of the three facets

employed. Every quadrant presents a visual map which is usually centered at the junctions of either the research types or topics facet with the contribution category. This makes it uncomplicated to visualize multiple facets at the same time. Furthermore, summary statistics were entered into the bubbles to further enhance understanding of the systematic map and provide a quick summary of the study in the area of big data, IoT, and real-time analytics.

4 Results and Discussion

The primary objective of this research on a large-scale foundation for real-time analytics, big data, and IoT is to have a thematic analysis, carry out classification, and possibly identify the areas where the articles were published. From the analysis, gaps were noticeable from the map, thus identifying which topic areas and research types were lacking in papers. On the contrary, the results also point to areas that are sufficiently covered in terms of publications. Within this research, high-level classifications were utilized in order to access the papers mentioned, hence the reliable readings are obtained and used for developing the map.

4.1 *The Contribution Category*

The systematic map created from this study is in Fig. 2. On the x-axis of the left quadrant of Fig. 2 is the result of the contribution facet. The contribution facet indicates the type of inputs in the focus of study. The result showed that publications that discussed model in relation to big data, IoT and real-time analytics was 35.85% out of 106 papers examined in this category. Also, tool had 20.75%, metric had 5.66%, method had 25.47%, and process had 12.26%.

4.2 *The Research Type Category*

On the x-axis of the right quadrant of Fig. 2 is the type of research conducted on the Cloud that is related to big data, IoT and, real-time analytics. The result showed that publications that discussed solution research in relation to big data, IoT, and real-time analytics was 43.75% out of the 112 papers reviewed. Also, evaluation research had 33.93%, validation research had 4.46%, philosophical research had 7.14%, and experience research had 10.71%.

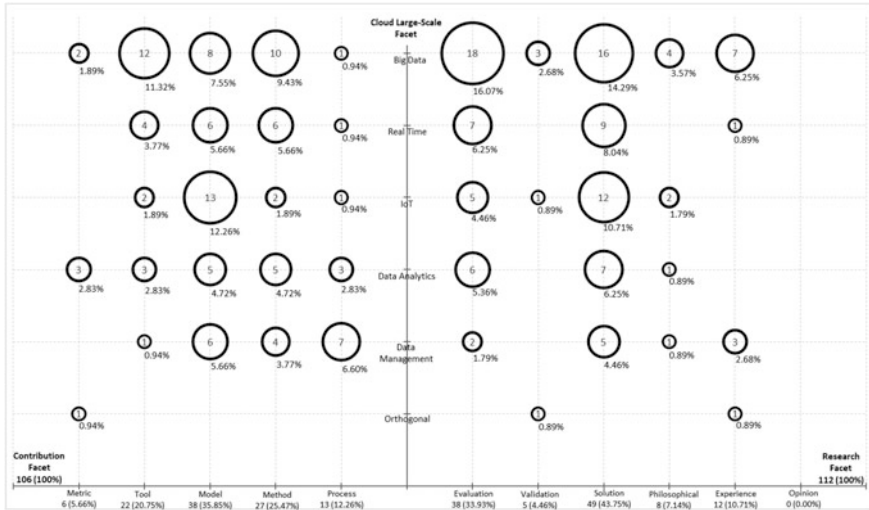


Fig. 2 Systematic map for real-time analytics, big data, and IoT on the Cloud

4.3 Topics and Contribution Facet

The topics used were extracted from the abstracts which related directly to the title of this research work. They are

1. Big data
2. Real-time analytics
3. IoT
4. Data Analytics
5. Data Management
6. Orthogonal

The left quadrant of Fig. 2 displays the relationship between the topics extracted and the contribution facet. Models contributed 38.85% of the papers reviewed, however, a breakdown showed that 7.55% was on big data, 5.66% was on real-time analytics, 12.26% was on IoT, and 4.72% was on data management and no orthogonal contribution. Other aspects of the contribution category as it relates to topics are shown in Fig. 2.

4.4 Topic and Research Facet

The right quadrant of Fig. 2 displays the relationship among the topics and research category and, for example, 33.93% of the papers reviewed in real-time analytics, big data, and IoT in the Cloud were related to evaluation research. The breakdown

indicated 16.07% was on big data, 6.25% was on real-time analytics, 4.4% was on IoT, 5.36% was on data analytics, and 1.79% was on data management. Other aspects of the topics category as it relates to research types are shown in Fig. 2.

4.5 Systematic Map for Big Data, Internet of Things, and Real-Time Analytics

The systematic map for real-time analytics, IoT, and big-data on the Cloud is shown in Fig. 2. The first quadrant is a two x–y scatter plot with bubbles at the intersection of the topic and contribution facet, while the left quadrant is the systematic map depicting a two x–y scatter plot showing the node of the topic and research facet. As discussed earlier, the results made it possible to identify which category of the study had more emphasis in terms of publications. It can be identified from the left quadrant of Fig. 2 that the highest publication was on IoT as it relates to model with 12.26%; there were more publications on data analytics as it relates to metric, (2.83%), more articles on big data in terms of tool, (11.32%), and method (9.43%). Furthermore, there were more research carried out on data management in terms of process with 6.6%. Similarly, on the right quadrant, the highest publication was on big data in terms of evaluation research. More articles also dealt with big data with respect to validation research (2.68%), solution research had 14.29%, philosophical research had 3.57%, and experience research had 6.25%.

On the other hand, to the best of our understanding, there were no any articles on metric in relation to real-time analytics, IoT, and data management on the left quadrant and there were no publications on validation research as it relates to real-time, data analytics and data management on the right quadrant. In addition, there were no publications on experience research in relation to IoT and data analytics. Furthermore, there were no articles on opinion research. Also, the least number of articles were in the study of model and process in the contribution category, while the lowest publications were on validation research, philosophical research, and experience research, in the research category. The visual appeal of the systematic map has been used to summarize the categories of publication making results available to researchers. The essence was to generate interest in the gaps, based on shortages in publications identified and then encourage further research. The value of a systematic map even without a consecutive systematic review cannot be overemphasized. Research gaps could be identified with ease, based on a shortage of publications as shown in frequencies of articles published. This research has created a systematic map that points to the areas lacking in studies in terms of Cloud large-scale foundation. The primary purpose of this study is that analysts at all levels and venture professionals can utilize this as a beginning stage to direct further investigations. This investigation created six classes, namely big data, real-time analytics, IoT, data analytics, data management, and orthogonal in connection to the focus of study. What's more, the six classes of study can be talked

about either regarding the instrument, method, model, process, and metric or in terms of validation, evaluation, philosophical, solution, and opinion research. These fields among others are consequently prescribed for future research. The rundown of included references will likewise help expecting researchers. The significant lessons learnt in this research is that research work is inexhaustible and continuing.

5 Conclusion

Cloud computing is a model that is becoming more relevant by the day. The Cloud enhances the operations of IoT and the processing of big data. Despite the volume of studies, it was identified from the results that so much is happening in terms of big data and IoT compared to real-time analytics. Systematic mapping is a veritable tool in helping to summarize and visually depict the frequencies of publications in an area of research. The mapping process was applied to big data, IoT, and real-time analytics with the attendant results displayed on the systematic map. The results from this research are based on the gaps identified in terms of model, tool, metric, method, and process in relation to real-time analytics, big data, and IoT. This research also identifies some gaps in the area of validation, evaluation, solution, opinion, and philosophical research on real-time analytics, IoT, and big data. To the best of the authors' understanding, there were no articles on metric in relation to real-time analytics, IoT, and data management. In addition, there were no publications on validation research as it relates to data analytics, and data management. Also, there were no publications on experience research in relation to IoT and data analytics. There were no articles on opinion research. Furthermore, the least number of articles were in the study of model and process in the contribution category, while the lowest publications were on validation research, philosophical research, and experience research, in the research category. It is normal that this outcome will fill in as a wide guide into topics that can be looked into on the region of real-time analytics, big data, and IoT. Further research could likewise be completed to approve this investigation or resolve conflicting issues.

Acknowledgements We acknowledge the support and sponsorship provided by Covenant University through the Centre for Research, Innovation and Discovery (CUCRID).

Appendix 1: List of Primary Studies

- [1] Adam, O.Y., Lee, Y.C., Zomaya, A.Y. Constructing Performance-Predictable Clusters with Performance-Varying Resources of Cloud (2016) *IEEE Transactions on Computers*, 65 (9), art. no. 7362012, pp. 2709-2724.

- [2] Agrawal, D., Das, S., El Abbadi, A. Big data and Cloud computing: Current state and future opportunities (2011) ACM International Conference Proceeding Series, pp. 530-533.
- [3] Agrawal, H., Mathialagan, C.S., Goyal, Y., Chavali, N., Banik, P., Mohapatra, A., Osman, A., Batra, D. Cloudcv: Large-scale distributed computer vision as a Cloud service (2015) Mobile Cloud Visual Media Computing: From Interaction to Service, pp. 265-290.
- [4] Al-Ayyoub, M., Jararweh, Y., Tawalbeh, L., Benkhelifa, E., Basalamah, A. Power Optimization of Large Scale Mobile Cloud Computing Systems (2015) Proceedings - 2015 International Conference on Future Internet of Things and Cloud, FiCloud 2015 and 2015 International Conference on Open and Big Data, OBD 2015, art. no. 7300885, pp. 670-674.
- [5] Alfazi, A., Sheng, Q.Z., Zhang, W.E., Yao, L., Noor, T.H. Identification as a service: Large-scale Cloud service discovery over the world wide web (2016) Proceedings - 2016 IEEE International Congress on Big Data, BigData Congress 2016, art. no. 7584980, pp. 485-492.
- [6] Al-Jaroodi, J., Mohamed, N., Jawhar, I., Mahmoud, S. CoTWARE: A Cloud of Things Middleware (2017) Proceedings - IEEE 37th International Conference on Distributed Computing Systems Workshops, ICDCSW 2017, art. no. 7979819, pp. 214-219.
- [7] Al-Quraan, M., Al-Ayyoub, M., Jararweh, Y., Tawalbeh, L., Benkhelifa, E. Power optimization of large scale mobile Cloud system using cooperative Cloudlets (2016) Proceedings - 2016 4th International Conference on Future Internet of Things and Cloud Workshops, W-FiCloud 2016, art. no. 7592697, pp. 34-38.
- [8] Alsmirat, M.A., Jararweh, Y., Obaidat, I., Gupta, B.B. Internet of surveillance: a Cloud supported large-scale wireless surveillance system (2017) Journal of Supercomputing, 73 (3), pp. 973-992.
- [9] Angrisani, L., Ianniello, G., Stellato, A. Cloud based system for measurement data management in large scale electronic production (2014) 2014 Euro Med Telco Conference - From Network Infrastructures to Network Fabric: Revolution at the Edges, EMTC 2014, art. no. 6996651,.
- [10] Apolonia, N., Sedar, R., Freitag, F., Navarro, L. Leveraging low-power devices for Cloud services in community networks (2015) Proceedings - 2015 International Conference on Future Internet of Things and Cloud, FiCloud 2015 and 2015 International Conference on Open and Big Data, OBD 2015, art. no. 7300840, pp. 363-370.
- [11] Auger, A., Exposito, E., Lochin, E. Sensor observation streams within Cloud-based IoT platforms: Challenges and directions (2017) Proceedings of the 2017 20th Conference on Innovations in Cloud, Internet and Networks, ICIN 2017, art. no. 7899407, pp. 177-184.
- [12] Bachiega, J., Reis, M.A.S., De Araujo, A.P.F., Holanda, M. Cost optimization on public Cloud provider for big geospatial data: A case study using open street map (2017) CLOSER 2017 -Proceedings of the 7th

- International Conference on Cloud Computing and Services Science, pp. 54-62.
- [13] Bojan, V.-C., Raducu, I.-G., Pop, F., Mocanu, M., Cristea, V. Cloud-based service for time series analysis and visualisation in Farm Management System (2015) Proceedings - 2015 IEEE 11th International Conference on Intelligent Computer Communication and Processing, ICCP 2015, art. no. 7312697, pp. 425-432.
 - [14] Bosse, S. From the Internet-of-Things to Sensor Cloud - Unified Distributed Computing in Heterogeneous Environments with Smart and Mobile Multi-Agent Systems (2015) Smart Systems Integration 2015-9th International Conference and Exhibition on Integration Issues of Miniaturized Systems: MEMS, NEMS, ICs and Electronic Components, SSI 2015, pp. 297-305.
 - [15] Cao, Y., Sun, D. Migrating large-scale air traffic modeling to the Cloud (2015) Journal of Aerospace Information Systems, 12 (2), pp. 257-266.
 - [16] Chang, B.-J., Lee, Y.-W., Liang, Y.-H. Reward-based Markov chain analysis adaptive global resource management for inter-Cloud computing (2018) Future Generation Computer Systems, 79, pp. 588-603.
 - [17] Chang, V. A cybernetics Social Cloud (2017) Journal of Systems and Software, 124, pp. 195-211.
 - [18] Chen, H., Guo, W. Real-time task scheduling algorithm for Cloud computing based on particle swarm optimization (2015) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9106, pp. 141-152.
 - [19] Chen, T.-Y., Wei, H.-W., Leu, J.-S., Shih, W.-K. EDZL scheduling for large-scale cyber service on real-time Cloud (2011) Proceedings - 2011 IEEE International Conference on Service-Oriented Computing and Applications, SOCA 2011, art. no. 6166234,.
 - [20] Clemente-Castelló, F.J., Nicolae, B., Mayo, R., Fernández, J.C., Rafique, M.M. On exploiting data locality for iterative MapReduce applications in hybrid Cloud (2016) Proceedings - 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, BDCAT 2016, pp. 118-122.
 - [21] Costan, A., Dobre, C. 1st Workshop on Big Data Management in Cloud - BDMC2012 (2013) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7640 LNCS, pp. 1-2.
 - [22] Cuzzocrea, A., Fortino, G., Rana, O. Managing data and processes in Cloud-enabled large-scale sensor networks: State-of-the-art and future research directions (2013) Proceedings - 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2013, art. no. 6546142, pp. 583-588.
 - [23] Da Silva, M.A.A., Sadovykh, A., Bagnato, A., Brosse, E. Taming the complexity of big data multi-Cloud applications with models (2014) CEUR Workshop Proceedings, 1234, pp. 1-12.

- [24] Das, A.K., Koppa, P.K., Goswami, S., Platania, R., Park, S.-J. Large-scale parallel genome assembler over Cloud computing environment (2017) *Journal of Bioinformatics and Computational Biology*, 15 (3), art. no. 1740003.
- [25] Demidov, D. A systematic approach to describing the source code of a Cloud platform with assured security (2017) *Proceedings - 2017 5th International Conference on Future Internet of Things and Cloud Workshops, W-FiCloud 2017*, 2017-January, pp. 31-36.
- [26] Dey, S., Chakraborty, A., Naskar, S., Misra, P. Smart city surveillance: Leveraging benefits of Cloud data stores (2012) *Proceedings - Conference on Local Computer Networks, LCN*, art. no. 6424076, pp. 868-876.
- [27] Dhar, P., Gupta, P. Intelligent parking Cloud services based on IoT using MQTT protocol (2017) *International Conference on Automatic Control and Dynamic Optimization Techniques, ICACDOT 2016*, art. no. 7877546, pp. 30-34.
- [28] Dzik, J., Palladinos, N., Rontogiannis, K., Tsarpalis, E., Vathis, N. MBrace: Cloud computing with monads (2013) *Proceedings of the 7th Workshop on Programming Languages and Operating Systems, PLOS 2013 - In Conjunction with the 24th ACM Symposium on Operating Systems Principles, SOSP 2013*, .
- [29] Feller, E., Ramakrishnan, L., Morin, C. Performance and energy efficiency of big data applications in Cloud environments: A Hadoop case study (2015) *Journal of Parallel and Distributed Computing*, 79-80, pp. 80-89.
- [30] Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M.J., Benítez, J.M., Herrera, F. Big Data with Cloud Computing: An insight on the computing environment, MapReduce, and programming frameworks (2014) *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4 (5), pp. 380-409.
- [31] Fortino, G., Russo, W. Towards a Cloud-assisted and agent-oriented architecture for the Internet of Things (2013) *CEUR Workshop Proceedings*, 1099, pp. 97-103.
- [32] Gebremeskel, G.B., Chai, Y., Yang, Z. The paradigm of big data for augmenting internet of vehicle into the intelligent Cloud computing systems (2014) *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8662 LNCS, pp. 247-261.
- [33] George, J., Chen, C.-A., Stoleru, R., Xie, G.G., Sookoor, T., Bruno, D. Hadoop MapReduce for tactical Cloud (2014) *2014 IEEE 3rd International Conference on Cloud Networking, CloudNet 2014*, art. no. 6969015, pp. 320-326.
- [34] Gomes, H.M., Carvalho, J.M.D., Veloso, L.R., Teixeira, A.G., Jr., Filho, T.B.D.O., Araujo, A.M.C.D., Zhang, T., Trarbach, L., Machado, F. MapReduce vocabulary tree: An approach for large scale image indexing and search in the Cloud (2016) *Proceedings - 2016 IEEE 2nd International*

- Conference on Multimedia Big Data, BigMM 2016, art. no. 7545016, pp. 170-173.
- [35] Hajibaba, M., Gorgin, S. A review on modern distributed computing paradigms: Cloud computing, jungle computing and fog computing (2014) *Journal of Computing and Information Technology*, 22 (2), pp. 69-84.
- [36] Han, H., Lee, Y.C., Choi, S., Yeom, H.Y., Zomaya, A.Y. Cloud-aware processing of MapReduce-based OLAP applications (2013) *Conferences in Research and Practice in Information Technology Series*, 140, pp. 31-38.
- [37] He, S., Cheng, B., Wang, H., Huang, Y., Chen, J. Proactive personalized services through fog-Cloud computing in large-scale IoT-based healthcare application (2017) *China Communications*, 14 (11), art. no. 8233646, pp. 1-16.
- [38] Huang, Q., Li, Z., Xia, J., Jiang, Y., Xu, C., Liu, K., Yu, M., Yang, C. Accelerating geocomputation with Cloud computing (2013) *Modern accelerator technologies for geographic information science*, 9781461487456, pp. 41-51.
- [39] Huang, T.-C., Shieh, C.-K., Huang, S.-W., Chiu, C.-M., Liang, T.-Y. Automatic self-suspended task for a mapreduce system on Cloud computing (2014) *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8643, pp. 257-268.
- [40] Huo, Z., Mukherjee, M., Shu, L., Chen, Y., Zhou, Z. Cloud-based Data-intensive Framework towards fault diagnosis in large-scale petrochemical plants (2016) *2016 International Wireless Communications and Mobile Computing Conference, IWCMC 2016*, art. no. 7577209, pp. 1080-1085.
- [41] Ji, C., Li, Y., Qiu, W., Awada, U., Li, K. Big data processing in Cloud computing environments (2012) *Proceedings of the 2012 International Symposium on Pervasive Systems, Algorithms, and Networks, I-SPAN 2012*, art. no. 6428800, pp. 17-23.
- [42] Jinzhou, Y., Jin, H., Kai, Z., Zhijun, W. Discussion on private Cloud PaaS construction of large scale enterprise (2016) *Proceedings of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2016*, art. no. 7529570, pp. 273-278.
- [43] Kang, S., Veeravalli, B., Aung, K.M.M., Jin, C. An efficient scheme to ensure data availability for a Cloud service provider (2015) *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, art. no. 7004378, pp. 15-20.
- [44] Kanwal, S., Lonie, A., Sinnott, R.O., Anderson, C. Experiences in implementing large-scale biomedical workflows on the Cloud: Challenges in transitioning to the clinical domain (2015) *CEUR Workshop Proceedings*, 1468, .

- [45] Karlsson, J., Torreño, O., Ramet, D., Klambauer, G., Cano, M., Trelles, O. Enabling large-scale bioinformatics data analysis with Cloud computing (2012) Proceedings of the 2012 10th IEEE International Symposium on Parallel and Distributed Processing with Applications, ISPA 2012, art. no. 6280355, pp. 640-645.
- [46] Kaur, B. Optimizing VM provisioning of mapreduce tasks on public Cloud (2016) ACM International Conference Proceeding Series, 12-13-August-2016, art. no. a79, .
- [47] Kaur, R., Chana, I., Bhattacharya, J. Data deduplication techniques for efficient Cloud storage management: a systematic review (2018) Journal of Supercomputing, 74 (5), pp. 2035-2085.
- [48] Kebande, V., Venter, H.S. A functional architecture for Cloud forensic readiness large-scale potential digital evidence analysis (2015) European Conference on Information Warfare and Security, ECCWS, 2015-January, pp. 373-382.
- [49] Kitanouma, T., Nii, E., Adachi, N., Takizawa, Y. SmartFinder: Cloud-based self organizing localization for mobile smart devices in large-scale indoor facility (2017) GIoTTS 2017 - Global Internet of Things Summit, Proceedings, art. no. 8016245.
- [50] Kos, A., Umek, A., Tomaic, S. Comparison of Smartphone Sensors Performance Using Participatory Sensing and Cloud Application (2016) Proceedings - 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things, IIKI 2015, art. no. 7428349, pp. 181-184.
- [51] Król, D., Kitowski, J. Towards adaptable data farming in Cloud (2015) Proceedings - 4th IEEE International Conference on Big Data and Cloud Computing, BDCLOUD 2014 with the 7th IEEE International Conference on Social Computing and Networking, SocialCom 2014 and the 4th International Conference on Sustainable Computing and Communications, SustainCom 2014, art. no. 7034805, pp. 283-284.
- [52] Li, P., Guo, S. Load balancing for privacy-preserving access to big data in Cloud (2014) Proceedings - IEEE INFOCOM, art. no. 6849286, pp. 524-528.
- [53] Li, T., Wang, K., Zhao, D., Qiao, K., Sadooghi, I., Zhou, X., Raicu, I. A flexible QoS fortified distributed key-value storage system for the Cloud (2015) Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015, art. no. 7363794, pp. 515-522.
- [54] Lin, J., Yin, J., Cai, Z., Liu, Q., Li, K., Leung, V.C.M. A secure and practical mechanism for outsourcing ELMs in Cloud computing (2013) IEEE Intelligent Systems, 28 (6), pp. 35-38.
- [55] Liu, L., Liu, L., Fu, X., Huang, Q., Zhang, X., Zhang, Y. A Cloud-based framework for large-scale traditional Chinese medical record retrieval (2018) Journal of Biomedical Informatics, 77, pp. 21-33.

- [56] Liu, X.-F., Zhan, Z.-H., Lin, J.-H., Zhang, J. Parallel differential evolution based on distributed Cloud computing resources for power electronic circuit optimization (2016) GECCO 2016 Companion - Proceedings of the 2016 Genetic and Evolutionary Computation Conference, pp. 117-118.
- [57] Liu, Z., Hu, J., Li, Y., Huang, Y. Toward virtual dataspace for material scientific data Cloud (2016) *Concurrency Computation*, 28 (6), pp. 1737-1750.
- [58] Liu, Z., Hu, C., Li, Y., Hu, J. DSDC: A domain scientific data Cloud based on virtual dataspace (2012) *Proceedings of the 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2012*, art. no. 6270579, pp. 2176-2182.
- [59] Lolos, K., Konstantinou, I., Kantere, V., Koziris, N. Elastic management of Cloud applications using adaptive reinforcement learning (2018) *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018-January*, pp. 203-212.
- [60] Loreti, D., Ciampolini, A. SHYAM: A system for autonomic management of virtual clusters in hybrid Cloud (2016) *Communications in Computer and Information Science*, 567, pp. 363-373.
- [61] Loreti, D., Ciampolini, A. MapReduce over the Hybrid Cloud: A Novel Infrastructure Management Policy (2015) *Proceedings - 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing, UCC 2015*, art. no. 7431408, pp. 174-178.
- [62] Luo, W., Zhang, H. Visual analysis of large-scale LiDAR point Cloud (2015) *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, art. no. 7364044, pp. 2487-2492.
- [63] Luszczek, P., Kurzak, J., Yamazaki, I., Keffer, D., Dongarra, J. Scaling point set registration in 3D across thread counts on multicore and hardware accelerator platforms through autotuning for large scale analysis of scientific point Cloud (2018) *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018-January*, pp. 2893-2902.
- [64] Marwan, M., Kartit, A., Ouahmane, H. Security in Cloud-based medical image processing: Requirements and approaches (2017) *ACM International Conference Proceeding Series, Part F129474*, art. no. 6.
- [65] Mohrehkesh, S., Fedorov, A., Vishwanatha, A.B., Drakopoulos, F., Kikinis, R., Chrisochoides, N. Large Scale Cloud-Based Deformable Registration for Image Guided Therapy (2016) *Proceedings - 2016 IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2016*, art. no. 7545815, pp. 67-72.
- [66] Moses, M.B.S., Ambika, K. A hybrid scheme for anonymous authentication of data storage in Cloud (2016) *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*, art. no. 7380489, pp. 360-364.

- [67] Mouradian, C., Yangui, S., Glitho, R.H. Robots as-a-service in Cloud computing: Search and rescue in large-scale disasters case study (2018) CCNC 2018 - 2018 15th IEEE Annual Consumer Communications and Networking Conference, 2018-January, pp. 1-7.
- [68] Moustafa, N., Creech, G., Sitnikova, E., Keshk, M. Collaborative anomaly detection framework for handling big data of Cloud computing (2017) 2017 Military Communications and Information Systems Conference, MILCIS 2017 - Proceedings, 2017-December, pp. 1-6.
- [69] Mseddi, A., Salahuddin, M.A., Zhani, M.F., Elbiaze, H., Glitho, R.H. On optimizing replica migration in distributed Cloud storage systems (2015) 2015 IEEE 4th International Conference on Cloud Networking, CloudNet 2015, art. no. 7335304, pp. 191-197. Nastic, S., Vogler, M., Inzinger, C., Truong, H.-L., Dustdar, S.
- [70] RtGovOps: A runtime framework for governance in large-scale software-defined IoT Cloud systems (2015) Proceedings - 2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud 2015, art. no. 7130866, pp. 24-33.
- [71] Nunez, D., Agudo, I., Lopez, J. Delegated access for hadoop clusters in the Cloud (2015) Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom, 2015-February (February), art. no. 7037691, pp. 374-379.
- [72] Ochoa, L., González-Rojas, O., Verano, M., Castro, H. Searching for optimal configurations within large-scale models: A Cloud computing domain (2016) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9975 LNCS, pp. 65-75.
- [73] Ogunyadeka, A., Younas, M., Zhu, H., Aldea, A. A Multi-key Transactions Model for NoSQL Cloud Database Systems (2016) Proceedings - 2016 IEEE 2nd International Conference on Big Data Computing Service and Applications, BigDataService 2016, art. no. 7474331, pp. 24-27.
- [74] Ohnaga, H., Aida, K., Abdul-Rahman, O. Performance of hadoop application on hybrid Cloud (2016) Proceedings - 2015 International Conference on Cloud Computing Research and Innovation, ICCCRI 2015, art. no. 7421903, pp. 130-138.
- [75] Pakdel, R., Herbert, J. A Cloud-based data analysis framework for object recognition (2015) CLOSER 2015-5th International Conference on Cloud Computing and Services Science, Proceedings, pp. 79-86.
- [76] Panneerselvam, J., Liu, L., Antonopoulos, N. Characterisation of hidden periodicity in large-scale Cloud datacentre environments (2018) Proceedings - 2017 IEEE International Conference on Internet of Things, IEEE Green Computing and Communications, IEEE Cyber, Physical and Social Computing, IEEE Smart Data, iThings-GreenCom-CPSCoM-SmartData 2017, 2018-January, pp. 496-503.

- [77] Perera, C., Talagala, D.S., Liu, C.H., Estrella, J.C. Energy-efficient location and activity-aware on-demand mobile distributed sensing platform for sensing as a service in iot Cloud (2015) *IEEE Transactions on Computational Social Systems*, 2 (4), art. no. 7397993, pp. 171-181.
- [78] Petrolo, R., Mitton, N., Soldatos, J., Hauswirth, M., Schiele, G. Integrating wireless sensor networks within a city Cloud (2014) *2014 11th Annual IEEE International Conference on Sensing, Communication, and Networking Workshops, SECON Workshops 2014*, art. no. 6979700, pp. 24-27.
- [79] Pham, X.-Q., Man, N.D., Tri, N.D.T., Thai, N.Q., Huh, E.-N. A cost- and performance-effective approach for task scheduling based on collaboration between Cloud and fog computing (2017) *International Journal of Distributed Sensor Networks*, 13 (11).
- [80] Prassanna, J., Ajit Jadhav, P., Neelanarayanan, V. Toward an analysis of load balancing algorithms to enhance efficient management of Cloud data centres (2016) *Smart Innovation, Systems and Technologies*, 49, pp. 143-155.
- [81] Rani, B.K., Babu, A.V. Scheduling of Big Data application workflows in Cloud and inter-Cloud environments (2015) *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, art. no. 7364103, pp. 2862-2864.
- [82] Rastogi, G., Sushil, R. Analytical literature survey on existing load balancing schemes in Cloud computing (2016) *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*, art. no. 7380705, pp. 1506-1510.
- [83] Rea, S., Aslam, M.S., Pesch, D. Serviceware-A service based management approach for WSN Cloud infrastructures (2013) *2013 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2013*, art. no. 6529470, pp. 133-138.
- [84] Requa, M., Vaughan, G., David, J., Cotton, B. Using Cloud bursting to count trees and shrubs in Sub-Saharan Africa (2016) *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, art. no. 7840947, pp. 2960-2963.
- [85] Roopaei, M., Rad, P., Jamshidi, M. Deep learning control for complex and large scale Cloud systems (2017) *Intelligent Automation and Soft Computing*, 23 (3), pp. 389-391.
- [86] Rossigneux, F., Lefèvre, L., Gelas, J.-P., De Assunção, M.D. A generic and extensible framework for monitoring energy consumption of open-stack Cloud (2015) *Proceedings - 4th IEEE International Conference on Big Data and Cloud Computing, BDCloud 2014 with the 7th IEEE International Conference on Social Computing and Networking, SocialCom 2014 and the 4th International Conference on Sustainable Computing and Communications, SustainCom 2014*, art. no. 7034862, pp. 696-702.

- [87] Shao, Y., Luo, Y., Hu, X., Xue, Y., Xiang, Y., Yin, K. FLAX: A flexible architecture for large scale Cloud fabric (2015) Proceedings - 2015 IEEE International Conference on Smart City, SmartCity 2015, Held Jointly with 8th IEEE International Conference on Social Computing and Networking, SocialCom 2015, 5th IEEE International Conference on Sustainable Computing and Communications, SustainCom 2015, 2015 International Conference on Big Data Intelligence and Computing, DataCom 2015, 5th International Symposium on Cloud and Service Computing, SC2 2015, art. no. 7463881, pp. 1151-1154.
- [88] Sheshasaayee, A., Megala, R. A study on resource provisioning approaches in autonomic Cloud computing (2017) Proceedings of the International Conference on IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2017, art. no. 8058325, pp. 141-144.
- [89] Spichkova, M., Thomas, I.E., Schmidt, H.W., Yusuf, I.I., Drumm, D.W., Androulakis, S., Opletal, G., Russo, S.P. Scalable and fault-tolerant Cloud computations: Modelling and implementation (2016) Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS, 2016-January, art. no. 7384320, pp. 396-404.
- [90] Spichkova, M., Schmidt, H.W., Thomas, I.E., Yusuf, I.I., Androulakis, S., Meyer, G.R. Managing usability and reliability aspects in Cloud computing (2016) ENASE 2016 - Proceedings of the 11th International Conference on Evaluation of Novel Software Approaches to Software Engineering, pp. 288-295.
- [91] Sundharakumar, K.B., Dhivya, S., Mohanavalli, S., Vinob Chander, R. Cloud based fuzzy healthcare system (2015) Procedia Computer Science, 50, pp. 143-148.
- [92] Taherkordi, A., Eliassen, F. Poster abstract: Data-centric IoT services provisioning in Fog-Cloud computing systems (2017) Proceedings - 2017 IEEE/ACM 2nd International Conference on Internet-of-Things Design and Implementation, IoTDI 2017 (part of CPS Week), pp. 317-318.
- [93] Taherkordi, A., Eliassen, F. Scalable modeling of Cloud-based IoT services for smart cities (2016) 2016 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2016, art. no. 7457098, .
- [94] Talia, D., Trunfio, P., Marozzo, F. Data Analysis in the Cloud: Models, Techniques and Applications (2015) Data Analysis in the Cloud: Models, Techniques and Applications, pp. 1-138.
- [95] Tang, H., Li, Y., Jia, T., Wu, Z. Evaluating performance of rescheduling strategies in Cloud system (2016) Proceedings - 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 10th IEEE International Conference on Big Data Science and Engineering and 14th IEEE International Symposium on Parallel and Distributed Processing with Applications, IEEE TrustCom/BigDataSE/ISPA 2016, art. no. 7847125, pp. 1559-1566.

- [96] Tian, W., Xue, R., Dong, X., Wang, H. An approach to design and implement RFID middleware system over Cloud computing (2013) *International Journal of Distributed Sensor Networks*, 2013, art. no. 980962, .
- [97] Tunc, C., Hariri, S., Montero, F.D.L.P., Fargo, F., Satam, P., Al-Nashif, Y. Teaching and Training Cybersecurity as a Cloud Service (2015) *Proceedings - 2015 International Conference on Cloud and Autonomic Computing, ICCAC 2015*, art. no. 7312173, pp. 302-308.
- [98] Vaquero, L.M., Celorio, A., Cuadrado, F., Cuevas, R. Deploying large-scale datasets on-demand in the Cloud: Treats and tricks on data distribution (2015) *IEEE Transactions on Cloud Computing*, 3 (2), art. no. 6910293, pp. 132-144.
- [99] Vilaplana, J., Solsona, F., Teixidó, I., Mateo, J., Usié, A., Torres, N., Comas, J., Alves, R. MetReS: A metabolic reconstruction database for Cloud computing (2014) *Proceedings - 2014 International Conference on Intelligent Networking and Collaborative Systems, IEEE INCoS 2014*, art. no. 7057165, pp. 653-658.
- [100] Wang, G.-L., Han, Y.-B., Zhang, Z.-M., Zhu, M.-L. Cloud-based integration and service of streaming data (2017) *Jisuanji Xuebao/Chinese Journal of Computers*, 40 (1), pp. 107-125.
- [101] Wang, T., Yao, S., Xu, Z., Jia, S., Xu, Q. A data placement strategy for big data based on DCC in Cloud computing systems (2015) *Proceedings - 2015 IEEE International Conference on Smart City, SmartCity 2015, Held Jointly with 8th IEEE International Conference on Social Computing and Networking, SocialCom 2015, 5th IEEE International Conference on Sustainable Computing and Communications, SustainCom 2015, 2015 International Conference on Big Data Intelligence and Computing, DataCom 2015, 5th International Symposium on Cloud and Service Computing, SC2 2015*, art. no. 7463793, pp. 623-630.
- [102] Woodworth, J., Salehi, M.A., Raghavan, V. S3C: An architecture for space-efficient semantic search over encrypted data in the Cloud (2016) *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, art. no. 7841040, pp. 3722-3731.
- [103] Xu, Y., Helal, A. Scalable Cloud-Sensor Architecture for the Internet of Things (2016) *IEEE Internet of Things Journal*, 3 (3), art. no. 7155463, pp. 285-298.
- [104] Yang, C., Liu, C., Zhang, X., Nepal, S., Chen, J. A time efficient approach for detecting errors in big sensor data on Cloud (2015) *IEEE Transactions on Parallel and Distributed Systems*, 26 (2), art. no. 6714550, pp. 329-339.
- [105] Yingchi, M., Ziyang, X., Ping, P., Longbao, W. Delay-Aware Associate Tasks Scheduling in the Cloud Computing (2015) *Proceedings - 2015 IEEE 5th International Conference on Big Data and Cloud Computing, BDCloud 2015*, art. no. 7310724, pp. 104-109.

- [106] Yu, T., Wang, X., Jin, J., McIsaac, K. Cloud-orchestrated physical topology discovery of large-scale IoT systems using UAVs (2018) *IEEE Transactions on Industrial Informatics*, 14 (5), pp. 2261-2270.
- [107] Yuan, D., Jin, J., Grundy, J., Yang, Y. A framework for convergence of Cloud services and Internet of things (2015) *Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2015*, art. no. 7230984, pp. 349-354.
- [108] Zhang, J., Li, T., Pan, Y. PLAR: Parallel large-scale attribute reduction on Cloud systems (2014) *Parallel and Distributed Computing, Applications and Technologies, PDCAT Proceedings*, art. no. 6904253, pp. 184-191.
- [109] Zhang, Y.F., Zhang, G., Yang, T., Wang, J.Q., Sun, S.D. Service encapsulation and virtualization access method for Cloud manufacturing machine (2014) *Jisuanji Jicheng Zhizao Xitong/Computer Integrated Manufacturing Systems, CIMS*, 20 (8), pp. 2029-2037.
- [110] Zhou, X., Tang, N., Kuang, Y. A universal framework for flexible Cloud computing (2016) *Proceedings of 2016 IEEE International Conference on Big Data Analysis, ICBDA 2016*, art. no. 7509841, .
- [111] Zhuang, Z., Guo, C. OCPA: An algorithm for fast and effective virtual machine placement and assignment in large scale Cloud environments (2013) *Proceedings - 2013 International Conference on Cloud Computing and Big Data, CLOUDCOM-ASIA 2013*, art. no. 6821001, pp. 254-259.
- [112] Zimmermann, A., Pretz, M., Zimmermann, G., Firesmith, D.G., Petrov, I., El-Sheikh, E. Towards service-oriented enterprise architectures for big data applications in the Cloud (2013) *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC*, art. no. 6690543, pp. 130-135.

References

1. Buyya, R., Broberg, J., Goscinski, A.: *Cloud Computing Principles and Paradigms*, pp. 4–10. Wiley, New York (2011)
2. Odun-Ayo, I., Ananya, M., Agono, F., Goddy-Worlu, R.: Cloud computing architecture: a critical analysis. In: *IEEE Proceedings of the 2018 18th International Conference on Computational Science and Its Applications (ICCSA 2018)*, pp. 1–7 (2018). <https://doi.org/10.1109/iccsa.2018.8439638>
3. Odun-Ayo, I., Odede, B., Ahuja, R.: Cloud applications management-issues and developments. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (LNCS)*, vol. 10963, pp. 683–694. Springer, Berlin, Germany (2018)
4. Odun-Ayo, I., Misra, S., Abayomi-Alli, O., Ajayi, O.: Cloud multi-tenancy: issues and developments. In: *UCC '17 Companion. Companion Proceedings of the 10th International Conference on Utility and Cloud Computing*, pp. 209–214 (2017)
5. Odun-Ayo, I., Misra, S., Omoregbe, N., Onibere, E., Bulama, Y., Damasevičius, R.: Cloud-based security driven human resource management system. *Front. Artif. Intell. Appl.* **295**, 96–106 (2017). <https://doi.org/10.3233/978-1-61499-773-3-96>

6. Stephanopoulos, G., Mavromoustakis, C.X., Mastorakis, G.S., Paragiostakis, S., Pallis, E., Batalia, J.M.: Big Data and cloud computing: a survey of the state of the art and research challenges. In: Mavromoustakis, C., Mastorakis, G., Dobre, C. (eds.) *Advances in Mobile Computing and Big Data in SG Era, Studies in Big Data*, vol. 22. Springer, Berlin (2017)
7. Mohammadi, M., Al-fuqaha, A., Sorour, S., Guizani, M. (2017) Deep learning for IoT big data and streaming analytics: a survey. [arXiv:1712.04301v1](https://arxiv.org/abs/1712.04301v1) [cs.NI]
8. Hashem, A., Yaqoob, I., Anuar, N.M., Mokhtar, S., Gani, A., Khan, S.U.: The rise of big data on cloud computing: review and open research issues. *Inf. Syst.* **47**, 98–115 (2014)
9. Amazon Web Services: Big data analytics option on AWS (2016)
10. Odun-Ayo, I., Omoregbe, N., Odusami, M., Ajayi, O.: Cloud ownership and reliability - issues and developments. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (LNCS)*, vol. 10658, pp. 231–240. Springer, Berlin, Germany
11. Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M.: Systematic Mapping Studies in Software Engineering. In: *EASE'08 Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering, Italy*, pp. 68–77, 26–27 June 2008
12. Barros-Justo, J.L., Cravero-Leal, A.L., Benitti, F.B., Capilla-Sevilla, R.: Systematic mapping protocol: the impact of using software patterns during requirements engineering activities in real-world settings. *Cornell University Library* (2017). [arXiv:1701.05747v1](https://arxiv.org/abs/1701.05747v1) [cs.SE]
13. Kosar, T., Bohra, S., Mernik, M.A.: Protocol of a systematic mapping study for domain-specific languages. *J. Inf. Softw. Technol.* **21**(3), 77–91 (2016)
14. Santos, V., Souza, E.F., Felizardo, K.R., Vijaykumar, N.L.: Analyzing the use of concept maps in computer science: a systematic mapping study. *Inform. Educ.* **16**(2), 257–288 (2017). <https://doi.org/10.15388/infedu.2017.13>
15. Souza, M., Veado, L., Moreira, R.T., Figueiredo, E., Costa, H.: A systematic mapping study on game-related methods for software engineering education. *Inf. Softw. Technol.* **95**, 201–218 (2018)
16. Fernandez-Blanco, C.R., Careri, F., Kavvadias, K., Hidalgo Gonzalez, I., Zucker, A., Peteves, E.: *Systematic Mapping of Power System Models: Expert Survey*, EUR 28875 EN. Publications Office of the European Union, Luxembourg (2017). <https://doi.org/10.2760/422399>. ISBN 978-92-79-76462-2
17. Mernik, M.: Domain-specific languages: a systematic mapping study. In: *International Conference on Current Trends in Theory and Practice of Informatics, Lecture Notes in Computer Science*, vol 10139, pp. 464–472. Springer, Berlin, Germany (2017)
18. Griffo, C., Almeida, J.P.A., Guizzardi, G.: A systematic mapping of the literature on legal core ontologies. In: *Brazilian Conference on Ontologies, ONTOBRAS 15, CEUR Workshop Proceedings*, 1442 (2015)
19. Kitchenham, B., Charters, S.: *Guidelines for performing systematic literature review in software engineering*. Version 2. 2007-01 (2007)
20. Ahmad, A., Brereton, P., Andras, P.: A systematic mapping study of empirical studies on software cloud testing methods. In: *IEEE International Conference on Software Quality, Reliability and Security Companion*, pp. 555–562 (2017)
21. Muhammed, A.C., Muhammed, A.B.: *A Systematic Mapping Study Of Software Architectures For Cloud Based Systems*, IT University Technical Report Series, IT University of Copenhagen (2014)
22. Wieringa, R., Maiden, N.A., Mead, N.R., Rolland, C.: Requirement engineering paper classification and evaluation criteria. A proposal and a discussion. *Requir. Eng.* **11**(1), 102–107 (2006)

Studies on Radar Imageries of Thundercloud by Image Processing Technique



Sonia Bhattacharya and Himadri Bhattacharyya Chakrabarty

Abstract Severe atmospheric event can cause huge damage to civilization. Severe thunderstorm is one of those weather events. Analysis of cloud imageries can be used to forecast severe thunderstorm. Convective clouds are one of the main reasons for the formation of severe thunderstorm. Analysis of such cloud imageries by image processing can be used to predict severe thunderstorm. Analysis of RGB values of pixel of cloud imageries can be used to show the formation of severe thunderstorm. Histogram analysis of such cloud imageries can also be used to predict severe thunderstorm. In this study analysis of RGB values of pixels and histograms of cloud imageries has been used to now cast severe thunderstorm with a lead time of 6 to 8 h. This lead time is necessary to save life and property from huge damages.

Keywords Convective cloud · Histogram · Image processing · Rader imageries · RGB values · Severe thunderstorm

This is here by certified that this paper has not been submitted, accepted or published anywhere and the work done in the manuscript is original.

S. Bhattacharya (✉)

CWTT, Department of Computer Science, Panihati Mahavidyalaya,
Sodepur, Kolkata 700110, India
e-mail: bhattacharyasonia@rocketmail.com

H. B. Chakrabarty

Principal, JRM, University of Calcutta, Kolkata, India
e-mail: himi0201@gmail.com

Head of the Department & PG Coordinator, Department of Computer Science,
Surendranath College (On lien), University of Calcutta, Kolkata, India

UGC sponsored Visiting Professor, Radio Physics and Electronics,
University of Calcutta, Kolkata, India

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_25

1 Introduction

Convective clouds are formed vertically upwards as the result of the instability of atmosphere. In India, the coastal areas (like the west coast, Orissa, Andhra Pradesh, and West Bengal) as well as north east have been affected many times by thunderstorms and heavy rainfall. Natural calamities are unpreventable and incur huge losses both in terms of life and property. Analysis of convective cloud imageries can give correct prediction of such severe weather event. During the last decade, new classes of cloud resolution mesoscale models have been hurriedly developed. They are able to simulate the life cycle of each convective clouds [1]. Some papers [2–4] concerned the propagation and regeneration of cells within a multi cell storm. The three months of March–April–May (MAM) are pre-monsoon season in India. Severe thunderstorms occur mostly in this time period. Four months of June–July–August–September are identified as monsoon time in India. Cyclonic weather features are evolved generally at the post-monsoon season due to formation of convective cloud during October–November months. The observations showed that the mesoscale phenomenon is the beginning of a severe convection before the monsoon and its effect on the climate, [15]. It is often apparent on satellite photos that convective conditions are occurring simultaneously over extensive areas and are the result of the vertical structure of an air mass and the characteristics of the actual dominant large-scale weather systems, [5]. A large number of research works have been done to analyze clouds based on satellite images [5]. Since the number of droplets of cloudy liquid and ice crystals increases in the cloud, reflectivity value increases, this has a direct relation with the rain [6]. Color is an important feature for image representation [5]. Color moments, color histogram and coherent color vector are the important components of the extraction of color characteristics [7]. Any satellite picture of convective cloud fields has a great variety of cloud patterns and cloud sizes, [8]. The size distribution of convective clouds might give hints of the precipitation observed at ground stations [8]. The main aim of this research work is to study color density of the pixel which changes with formation of cloud. Red color in cloud imagery represents water content of the cloud. This study will reveal that color density of pixel increases from image to image as the cloud devolved with respect to time. In this work the imageries of a thunder cloud have been studied stage by stage such as developing stage/Cumulus stage, mature stage, and Dissipating stage.

2 Data

Here in this study ten images of a thundercloud having different stages from time to time are considered for analysis. All these images are taken over Kolkata (22.3°N/88.3°E) on April 10, 2005 from 07:12:04 to 14:12:04. The cloud imageries were obtained from the observations through Doppler Radar of Regional Meteorological Center, India Meteorological Dept., Kolkata.

3 Methodology

Image Processing

Here in this study ten GIF images have been considered for weather forecast. All these imageries are TRUE color image. Here each pixel has a particular color; that color being described by the amount of red, green and blue in it. Since the total number of bits needed for each pixel is 24, these images are also called bit color images. An image of this type can be considered as a stack of three matrices; representing the values of red, green, and blue for each pixel. This means that for each pixel there are three values [14]. Here in this study `impixel (I)` function has been used which returns the value of pixels in image I. The pixel values of two consecutive images have been obtained using `impixel` function which gives two three columnar matrices. A comparison has been done between these two matrices of two consecutive imageries which revealed that there is a noticeable increase between their values. There are a total of nine comparisons between these ten images.

4 Observations on Cloud Imageries

As the first three images showed that formation of convective cloud, in which mainly three types of colors are observed, these are Blue, Red and Yellow. Red color represents the water content of the cloud. In the first image there is no cloud activity is observed, in the second image cloud has been formed over Bankura and Purulia and became denser in the third image moving towards Kolkata (Figs. 1, 2, 3, 4, 5 and 6).

In the fourth, fifth and sixth images it can be observed that density of red color pixels has been increased noticeably and the cloud has been moved more towards Kolkata. It can be observed that from seventh image the density of red colored pixels were decreasing as it moves over Kolkata. Since red colored pixel denotes water content of the cloud, reduction of red color pixel density indicates towards the precipitation (Figs. 7 and 8).

It can be clearly observed that in ninth and tenth image red colored pixel were noticeably reduced leaving yellow colored pixels, this signifies that precipitation has been done (Figs. 9 and 10).

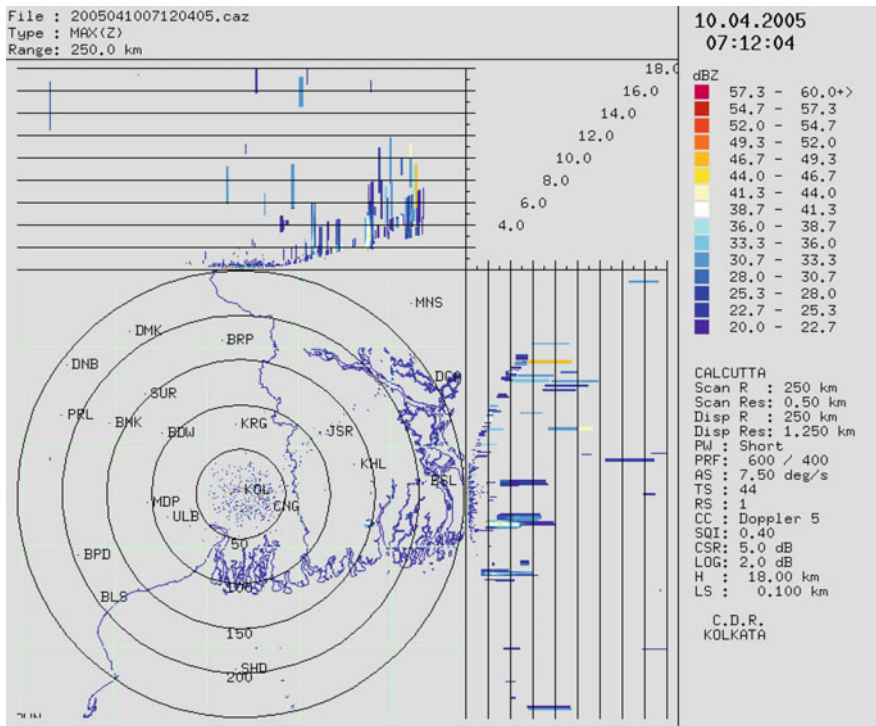


Fig. 1 First image of cloud formation

5 Result

Using impixel function RGB value of each pixel has been obtained and a comparison between the RGB values of the pixels of two consecutive images yield the following result:

Table 1 show that the density of RGB values increases from image to image and has maximized between sixth and seventh images. We have already observed that red color pixel density were most in seventh image. These values are reducing for the next remaining images, though values increase slightly in the last image. Comparing the last two images, it has been found that a few more red colored pixels are present in the tenth image.

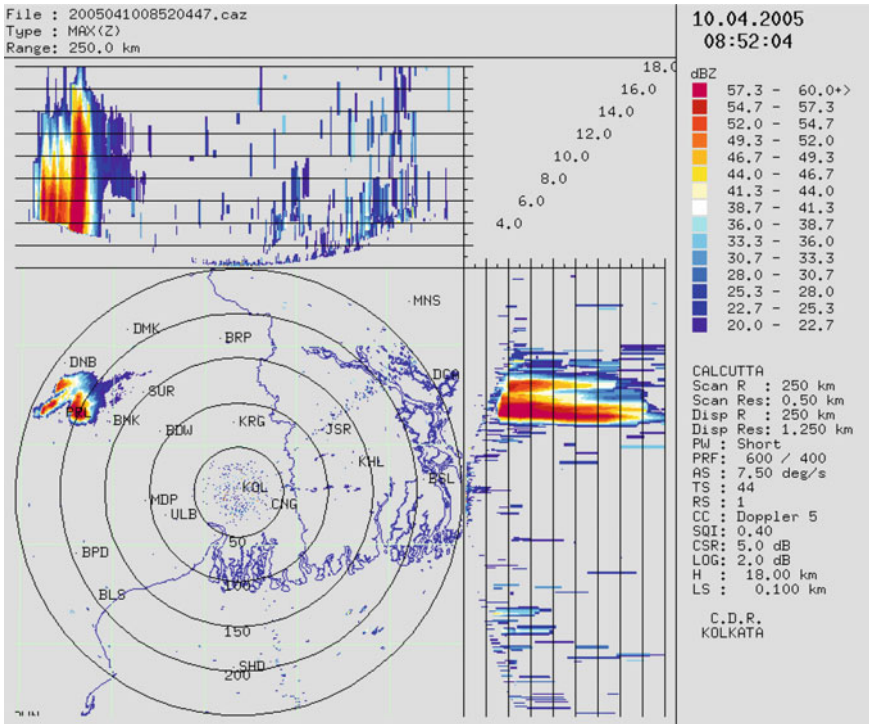


Fig. 2 Second image of cloud formation

6 Analysis of Histograms

It can be further observed that the histograms of these images show that the intensity has increased with formation of the cloud. The $imhist(I)$ function displays plot of the histogram. The histogram shows the distribution of pixel values. The study reveals that frequency of pixels has been increased from one image to next consecutive image. It has been reached most in the seventh image. The seventh image shows maximum presence of red colored pixels indicating towards the intensification of cloud. Precipitation is indicated by increase of yellow colored pixels in image the effect of which can be seen from eighth imagery. The distribution of pixel values has been increased from image to image as cloud has developed with respect to time (Figs. 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20).

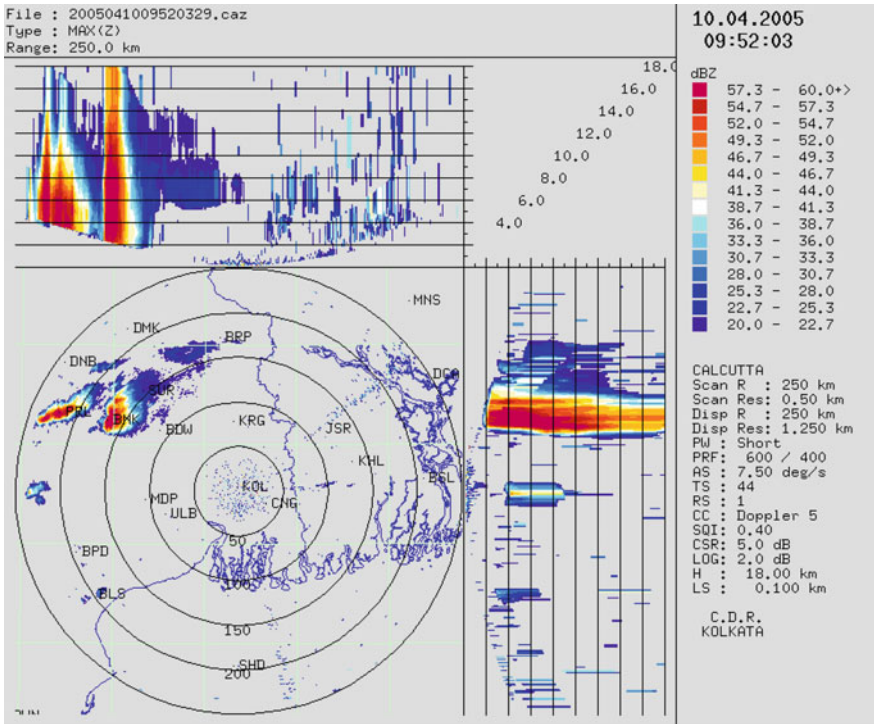


Fig. 3 Third image of cloud formation

7 Discussion

The pixel comparison of clear sky images is used for the determination of colors [9]. It is very important to monitor the analysis of clouds images to forecast when the thunderstorm along with the precipitation occurs. It can also be analyzed how much precipitation can occur by studying cloud images and its duration of rainfall. One can guess the severity of the natural disaster tentatively 6 h before by this type of cloud analysis. People can become alert with a lead time of 6 to 8 h. There are several meteorological satellites that provides large-scale observation of clouds through the day [10, 11, 12]. The patterns of image pixels are used for the classification of images [13].

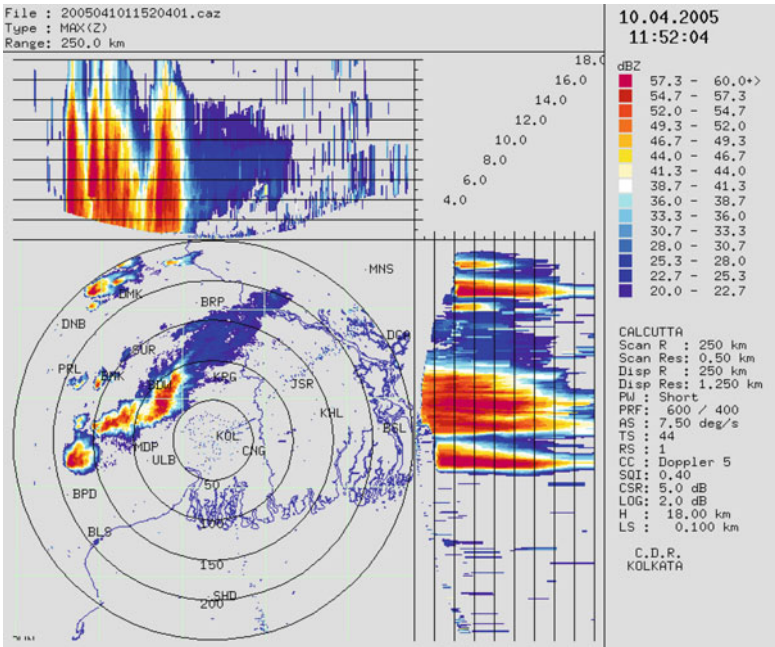


Fig. 4 Fourth image of cloud formation

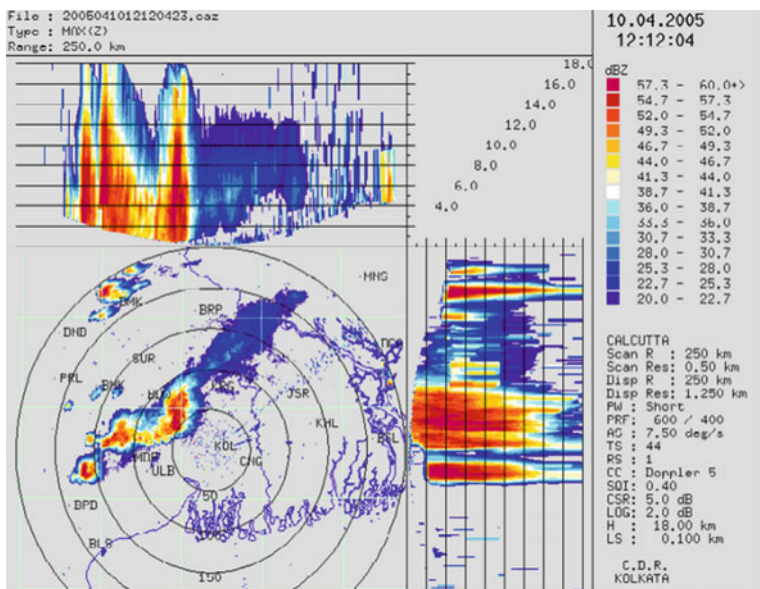


Fig. 5 Fifth image of cloud formation

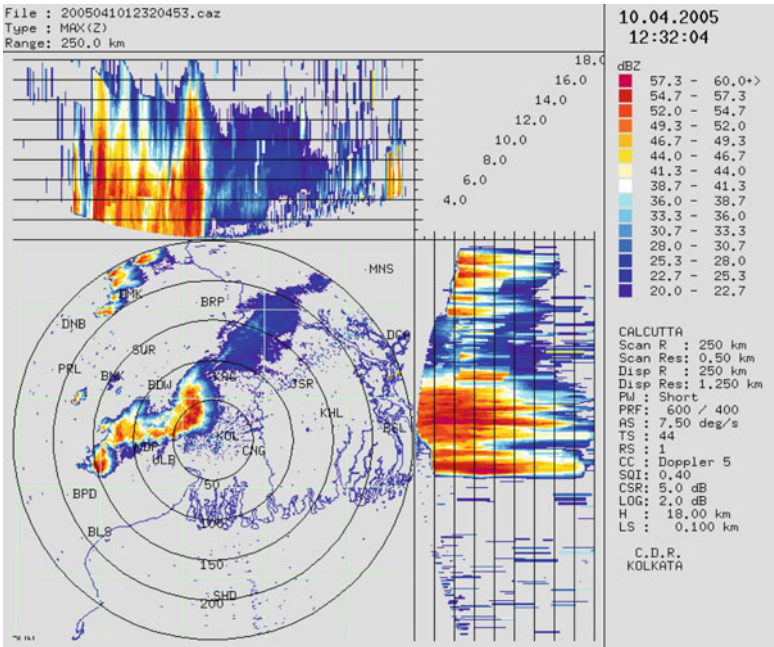


Fig. 6 Sixth image of cloud formation

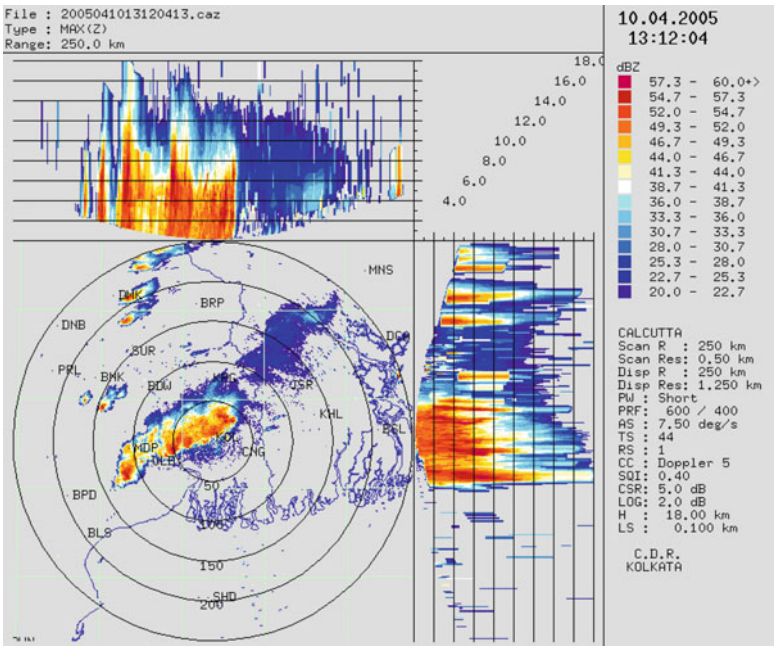


Fig. 7 Seventh image of cloud formation

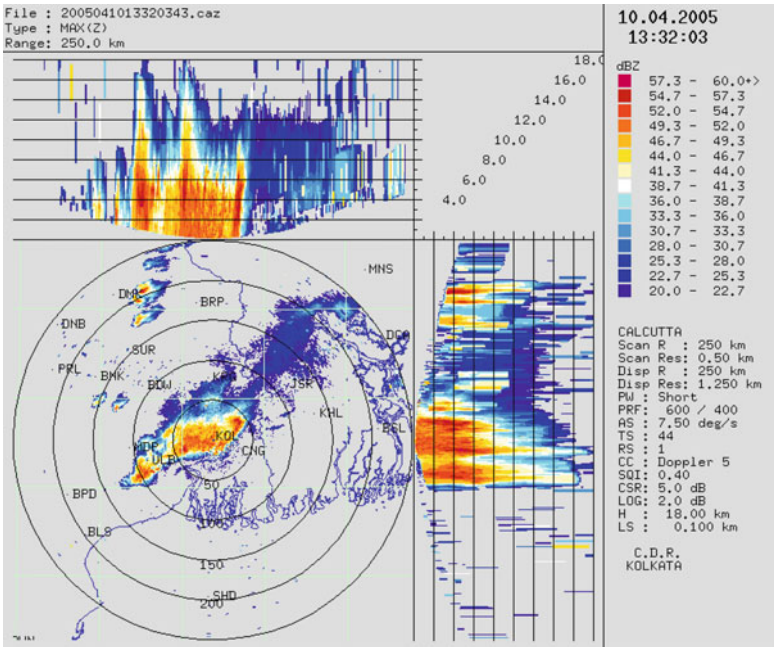


Fig. 8 Eighth image of cloud formation

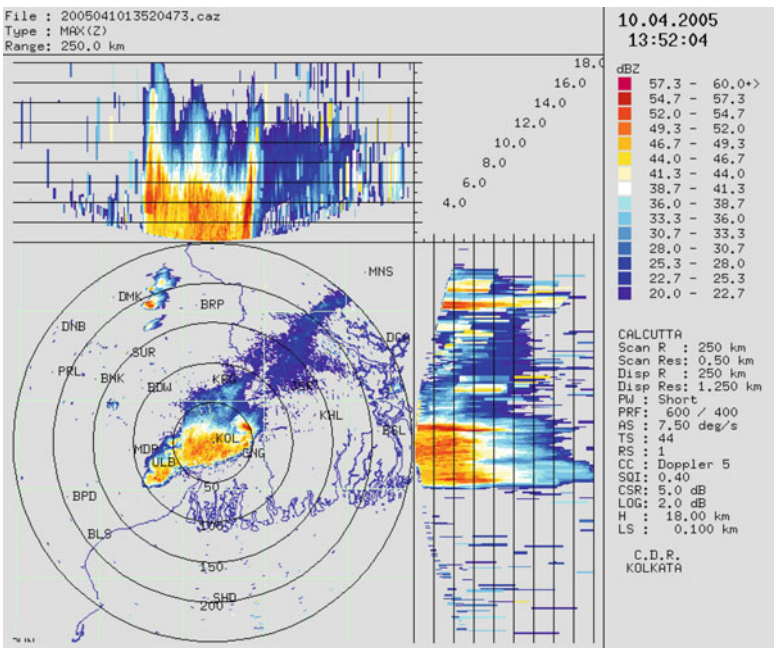


Fig. 9 Ninth image of cloud formation

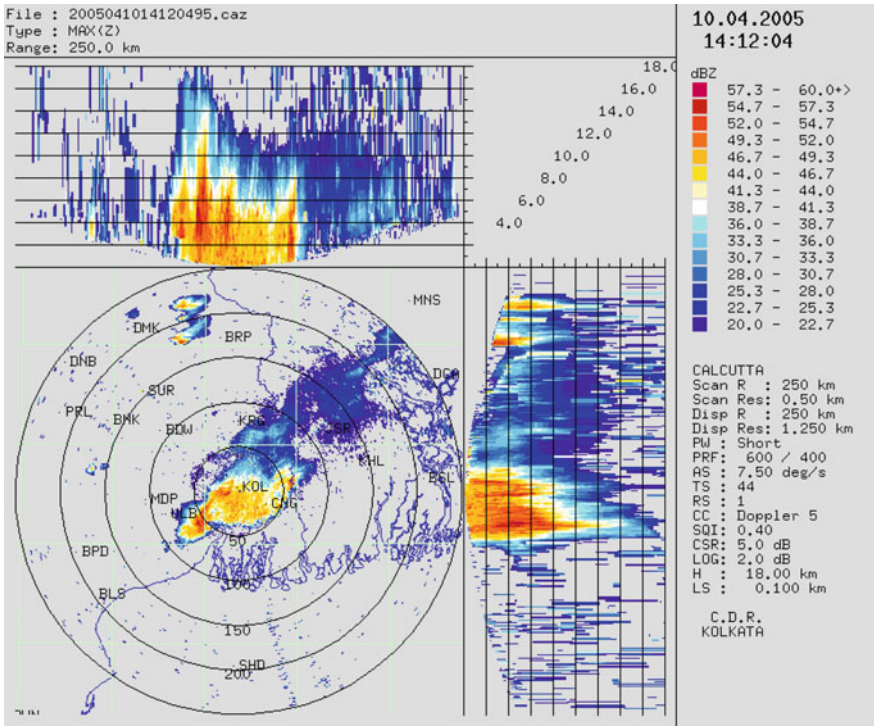


Fig. 10 Tenth image of cloud formation

Table 1 Compared value between two consecutive images

Increase in pixel value	Consecutive images
218	First-second image
240	Second-third image
829	Third-fourth image
1336	Fourth-fifth image
2126	Fifth-sixth image
3472	Sixth-seventh image
2539	Seventh-eighth image
1837	Eighth-ninth image
1845	Ninth-tenth image

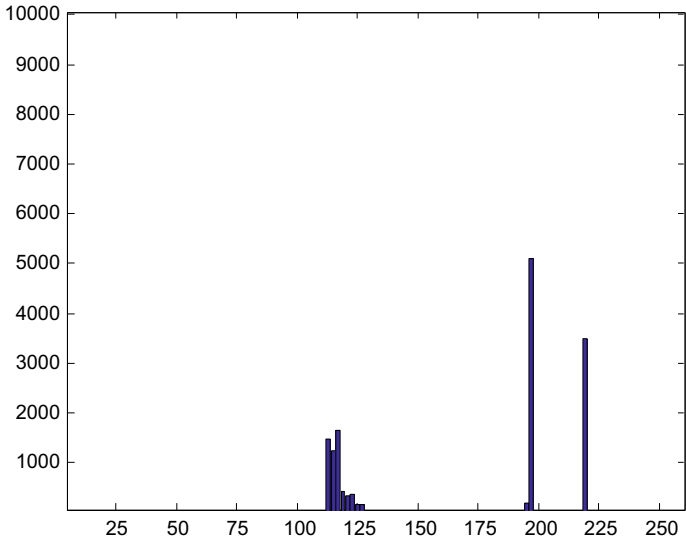


Fig. 11 Histogram of first image

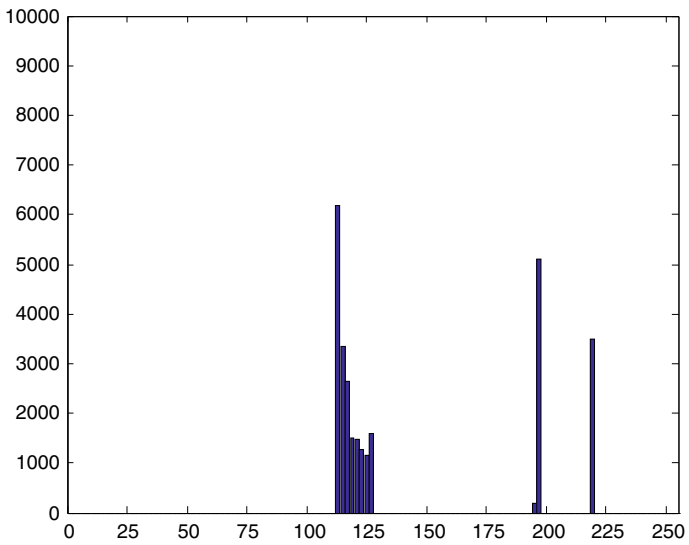


Fig. 12 Histogram of second image

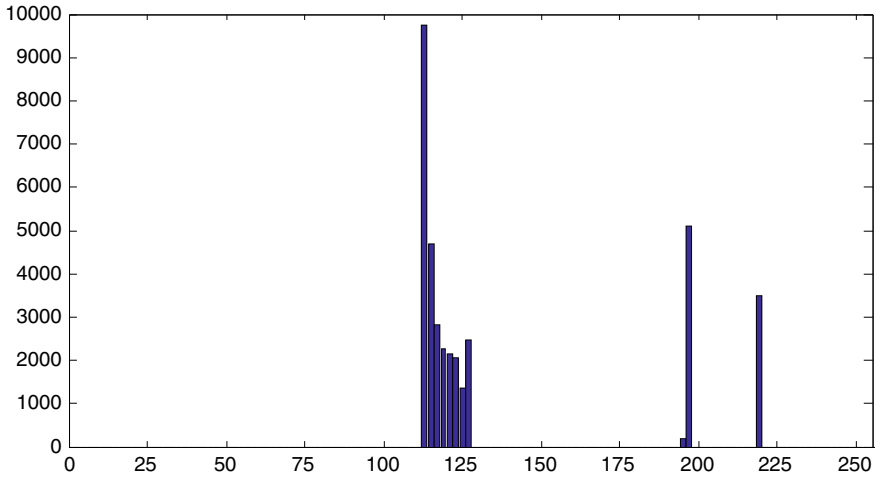


Fig. 13 Histogram of third image

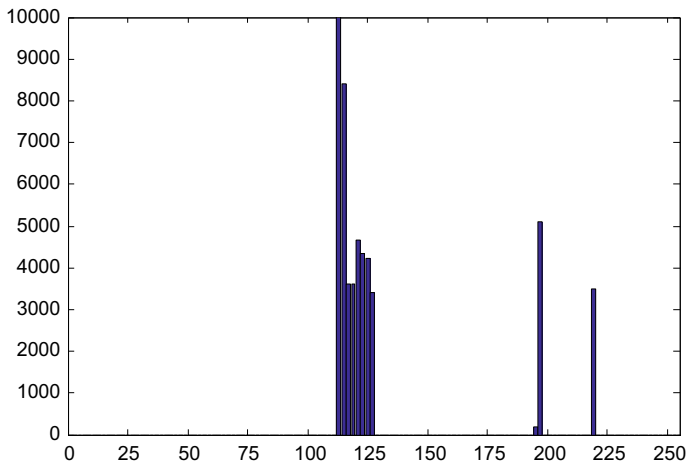


Fig. 14 Histogram of fourth image

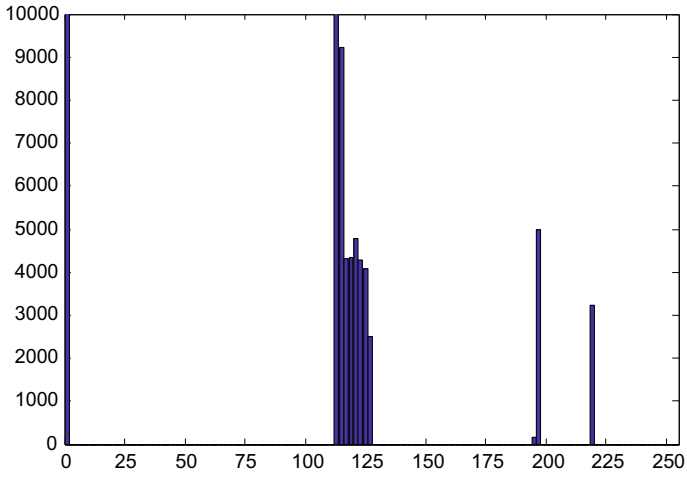


Fig. 15 Histogram of fifth image

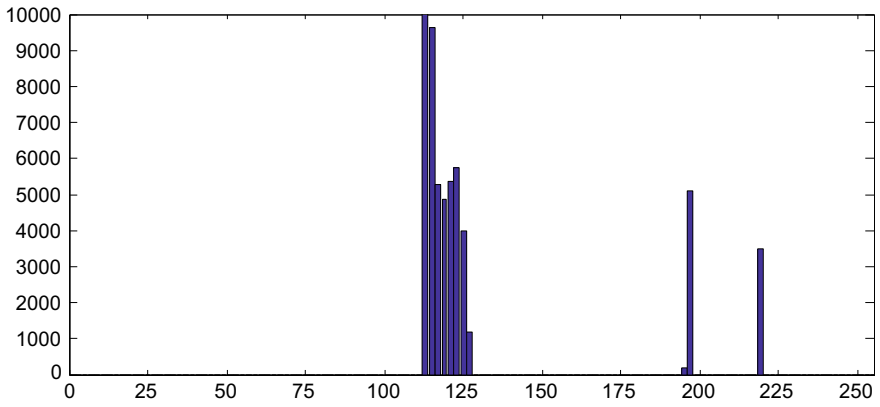


Fig. 16 Histogram of sixth image

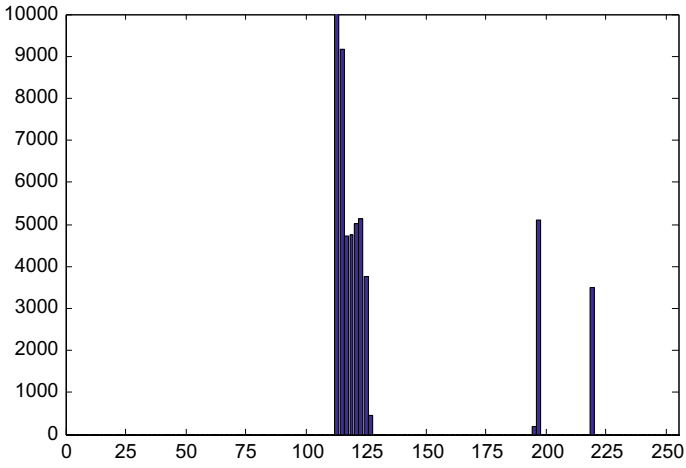


Fig. 17 Histogram of seventh image

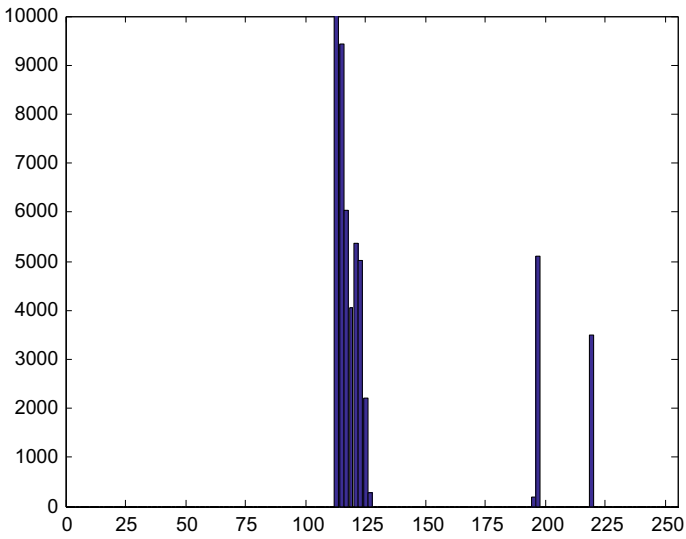


Fig. 18 Histogram of eighth image

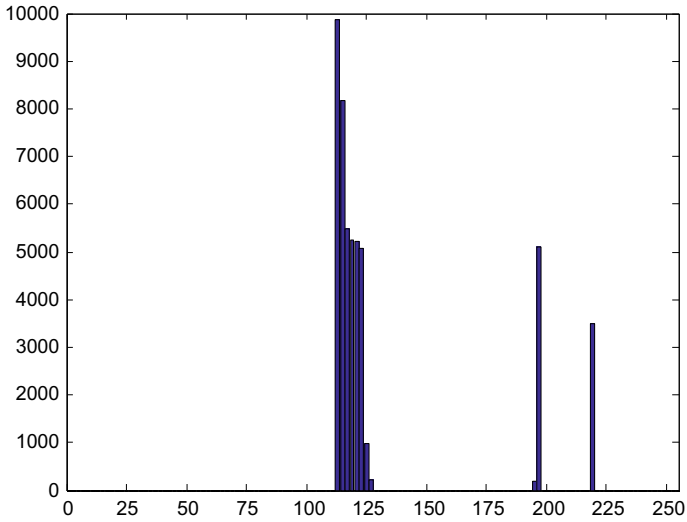


Fig. 19 Histogram of ninth image

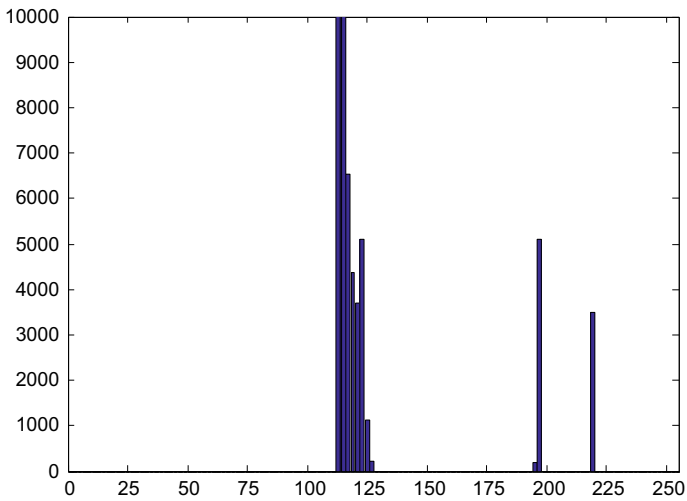


Fig. 20 Histogram of tenth image

References

1. Mladjen, C., Janc, D., Vujovic, D., Vuckovic, V.: The effects of a river valley on an isolated cumulonimbus cloud development. *Atmos. Res.* **66**, 123–139 (2003)
2. Lin, Y.L., Joyce, L.E.: A further study of mechanisms of cell regeneration, development and propagation within a two-dimensional multicell storm. *J. Atmos. Sci.* **58**, 2957–2988 (2001)
3. Lin, Y.-L., Deal, R.L., Kulie, M.S.: Mechanisms of cell regeneration, propagation, and development within two-dimensional multicell storms. *J. Atmos. Sci.* **55**, 1867–1886 (1998)
4. Fovell, R.G., Tan, P.-H.: Why multicell storms oscillate. In: 18th Conference on Severe Local Storms. American Meteorological Society, San Francisco, CA, pp. 186–189. Preprints (1996)
5. Anil Kumar, P., Anuradha, B., Arunachalam, M.S.: Extraction of time series convective cloud profile from doppler weather radar MAX (Z) product using a novel image processing technique. *Int. J. Adv. Eng. Res. Dev.* **4**(7), 2348–4470 (2017)
6. Gil, J.Y., Kimmel, R.: Efficient dilation, erosion, opening, and closing algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(12), 1606–1617 (2002)
7. Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Wang, W., Powers, J.G.: A description of the advanced research WRF, Version 2. NCAR Tech. Note NCAR/TN-4681STR, p. 88 (2008)
8. Gryscha, M., Witha, B., Etling, D.: Scale analysis of convective clouds. *Meteorol. Z.* **17**(6), 785–791 (2008)
9. Souza-Echer, M.P., Pereira, E.B., Bins, L.S., Andrade, M.A.R.: A simple method for the assessment of the cloud cover state in high-latitude regions by a ground-based digital camera. *J. Atmos. Ocean. Technol.* **23**(3), 437–447 (2006)
10. Hutchison, K.D., Hardy, K.R., Gao, B.C.: Improved detection of optically thin cirrus clouds in nighttime multispectral meteorological satellite using total integrated water vapor information. *J. Appl. Meteorol.* **34**, 1161–1168 (1995)
11. Jolivet, D., Feijt, A.J.: Cloud thermodynamic phase and particle size estimation using the 0.67 and 1.6 mm channels from meteorological satellites. *Atmos. Chem. Phys. Discuss.* **3**, 4461–4488 (2003)
12. Glantz, P.: Satellite retrieved cloud optical thickness sensitive to surface wind speed in the subarctic marine boundary layer. *Environ. Res. Lett.* **5**, 034002 (2010). <https://doi.org/10.1088/1748-9326/5/3/034002>
13. Ghonima, M.S., Urquhart, B., Chow, C.W., Shields, J.E., Cazorla, A., Kleiss, J.: A Method for Cloud Detection and Opacity Classification Based on Ground Based Sky Imagery, pp. 4535–4569. Copernicus Publications, Göttingen (2012)
14. McAndrew, A.: An Introduction to Digital Image Processing with Matlab Notes for SCM2511 Image Processing 1. School of Computer Science and Mathematics Victoria University of Technology, Footscray (2004)
15. Himadri Chakrabarty, C.A., Murthy, S.Bhattacharya, Gupta, A.D.: Application of artificial neural network to predict squall-thunderstorms using RAWIND data. *Int. J. Sci. Eng. Res.* **4**(5), 1313–1318 (2013). (ISSN 2229-5518)

Artificial Intelligence and Data Analysis

PURAN: Word Prediction System for Punjabi Language News



Gurjot Singh Mahi and Amandeep Verma

Abstract This paper presents an outline of the PURAN: A state-of-the-art word prediction system for Punjabi language news. Word prediction systems are used to increase the user text composition rate while typing the text. Brief background of the various approaches utilized in the development of word prediction systems, while discussing the various factors affecting the development of such systems is provided. This paper also elaborates the word prediction system architecture in detail. The system performance was tested on Keystroke saving, Hit ratio, Average rank and Average keystrokes benchmark metrics. The paper demonstrates that the PURAN has achieved highest Hit ratio in Regional news genre followed by National news genre by making lowest average keystrokes in the said categories of news. The results show that system has achieved 88.38% Average Hit ratio with 51.42% Average keystroke saving for $N = 10$.

Keywords Word completion · Word prediction · Punjabi text composition

1 Introduction

Word prediction system enables the user to complete the partially entered string known as the prefix, by providing the selection among list of possible words. The word prediction system populates the list of possible words from the provided database for the given prefix. The word prediction systems instinctually help the user with automatic completion of the word. Thus word prediction system in turns helps to increase the text composition rate. Several software applications like PROFET [1], FASTY [2], PAL [24] were proposed to predict the word in English and other European languages. Text composition in Indian languages is relatively

G. S. Mahi (✉) · A. Verma

Department of Computer Science, Punjabi University, Patiala, Punjab, India
e-mail: gurjotmahi28@gmail.com

A. Verma

e-mail: vaman71@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_26

383

more tedious as compared to other languages due to numerous concerns. Large character database, different vowels symbols, and complex language syntax makes text composition difficult in Indian context [4]. The studies like hIndiA [4] and [29] were performed for other Indian languages—Hindi and Bengali.

This paper intends to present a word prediction system—PURAN, designed to predict the words in News items for Punjabi language. PURAN was designed to increase the text composition rate for Punjabi news composition for 5 news genres namely Business, International, National, Regional, and Sports. PURAN system architecture is divided into two phases—Corpus creation and statistical inference phase and word prediction phase. The PURAN system has performed well on all benchmark metrics when tested in simulated environment.

The paper is distributed into eight sections. Section 2 provides the background about the state of art in word prediction systems and various factors affecting its performance. Section 3 presents the proposed system architecture used to design PURAN word predictor. System architecture is elaborated using a working example in Sect. 4. Performance metrics used to test the system architecture are discussed in Sect. 5. Testing dataset and system configuration is given in Sect. 6. Explored results and discussion is mentioned in Sect. 7. Section 8 concludes the study.

2 Background

Many studies [1–5] have been carried out to develop specialized software applications known as word prediction systems to increase the users' text composition rate. The idea of the development of a word prediction system was started way back in 1968 when [9] published the technique to reduce the keystrokes for completing a word [10]. Most of the word prediction systems were developed to help the people with motor and speech disabilities to overcome the communication problems [11]. Garay-Vitoria and Abascal [11] also states that word selection based text-input interfaces helps the people having poor control over the limbs to communicate in regular conversations. The success of word prediction system depends on various factors, i.e., prediction methodology, prediction speed, dictionary structure, user interface used for demonstrating the results, number of suggestions in user interface, etc. Some of the main factors affecting word prediction are discussed as follows.

2.1 Methodology

The success of word prediction system heavily depends on the methodology used for the design of the system. Although, diverse strategies are used for designing prediction systems, but collectively predictors methodology can be divided into two main categories, (1) Statistical Predictor and (2) Syntactic Predictor [4].

Statistical Predictor. The statistical predictors rely on the n-gram language model for language information in statistical word prediction system [4]. Most of the word prediction systems designed were based on frequency calculation, where distinctive word unigram occurrences were taken into account. Garay-Vitoria and Abascal [12] suggested the method of using the frequencies for word prediction, in which a dictionary-based approach is used. Dictionary is composed of words (unique unigrams) with respective frequencies. Each time a word is accepted from the proposed list of words the frequency of the respective word is updated in the frequency. Similar frequency-based approach was used by [13–16] for the development of word prediction system.

Frequency-based completion system concentrates on the unigram language model for prediction of the word. Fazly [17] gave the predictor which was trained on very large English language corpus using unigram and bigram language model. Unigram language model only concentrate on the current word being typed and bigram model approximated the probability of one word in the past for word prediction. Also, trigram (which takes the two previous words into account for prediction) is the baseline language model used for the word prediction systems [4, 18]. The main disadvantage of trigram model is that it requires large training data and demand huge memory space to store the processed corpus [11]. The present study used the frequency distribution method for the development of PURAN word prediction system.

Syntactic Predictor. Rule based syntactic structures of a natural language are followed for the design of syntactic word prediction systems. Frequency remains an important factor in addition to the other approaches during design of syntactic word prediction system, like [2] used the frequency table of words with syntactic knowledge i.e. Part-of-Speech (PoS) for more precision to design the prediction system—FASTY, for people suffering from motor and speech disability, enabling them to type 4 European languages, i.e., German, French, Dutch, and Swedish more precisely. Syntactic predictors like FASTY takes the probability of the grammatically correct Part-of-Speech (PoS) information into account for making any natural language word prediction. Another prediction system named New Profet [1] used lexicons for word completion for individuals suffering from linguistic impairments. [1] developed an efficient lexicon development algorithm, which helped in the creation of word lexicons from Stockholm-Umeå Corpus.¹ Furthermore, this algorithm facilitates the creation of new lexica from untagged or PoS tagged corpus.

In addition to the lexicon- and PoS-based syntactic word predictors, [19] developed a morphologically based word predictor for the Swedish language. An algorithm developed by the [19] allows users to select words by providing the base form of the word in a two-step process. Another syntactic prediction system like [20] and [21] provides a word completion or prediction system for Inflected languages. Syntactic predictors make the prediction using semantic knowledge about the natural language for which the word prediction system is being designed.

¹<https://spraakbanken.gu.se/eng/resources/suc>.

2.2 Dictionary

An arrangement of the words with their frequencies or probability values requires a data structure for enhanced and firm results. Garay-Vitoria and Abascal [11, 22, 23] provide a detailed architecture about importance and usage of dictionaries in the word prediction systems. Dictionary is a general concept that is used to manage the keys and its values in an efficient manner. Keys in a dictionary data structure are implied as words and values are referred as words frequencies or probabilities. Word predictors designed using bigram or trigram language model require large storage capacity, which results in large dictionary size and complexity [22]. A single dictionary structure used for lexicon makes the access easy, but with the increase in number of dictionaries the problem complexity and access time for lexicon also increases [11, 22]. The lexicons are usually stored in the tree data structure, which helps in faster search but also have a complex organization structure.

Garay-Vitoria and Abascal [11] states that adapting dictionaries is an important factor while considering the natural language word prediction system. A single dictionary containing words and their respective frequencies are easy to maintain and update, as these frequencies can be easily updated in accordance to their usage in the real time.

2.3 User Interface

Newell et al. [24] gave detailed arguments about the importance of user interface in the selection of word from the suggestion list. The user interface provides an interactive space for a list of suggestions to appear from which user selects the most appropriate word that matches the user-specified prefix. Cognitive cost and suggestions demonstration are one of the factors related to the user interface. Point-of-gaze² plays an important role in increasing the production rate of a word predictor, as ease to locate the needed word in the list directly affects the hit ratio of the word prediction [24]. The design of user interface unswervingly affects the eye and hand movement during the selection of a word from suggestion list [4, 11, 24, 25]. Minimum eye and hand movement should be supported by the user interface [4]. Newell et al. [24], Swiffin et al. [15] suggests that vertical list of suggestions require less effort to select the word from the suggestion list. Swiffin et al. [15] argues that keeping the user interface immediately below the given prefix help in keeping minimum head and eye movement. Garay-Vitoria and Abascal [22], Koester and Levine [26] suggests that user interface suggestions can be listed in alphabetical order for better results. Garay-vitoria and Abascal [22] also mentions that triangular matrix provides the best average case access time to select the word

²“Eye tracking”, https://en.wikipedia.org/wiki/Eye_tracking.

from the suggestion list, but difficult screen presentation makes them hard to use. Due to the extensive importance of word prediction system, user interface perspective is the uttermost important factor to look into.

2.4 *Number of Suggestions*

Another significant factor affecting the deliverance of words prediction system is the number of suggestions appearing for the user to select from. With the use of more suggestions in the suggestion list, a predictor can avail higher hit ratio, but will also increase the cognitive load on the user [11]. Garay et al. [21] points that suggestion list with ten suggestions provide stability between keystrokes and cognitive cost.

3 System Architecture

To perform the prediction of the word - ω , the system architecture is distributed into two phases, (1) Corpus creation and statistical inference phase and (2) word prediction phase. The phase of corpus creation and statistical calculation is performed only once whereas, phase 2 is initiated whenever the user enters any new prefix or update the prefix with a new set of characters.

3.1 *Corpus Creation and Statistical Inference*

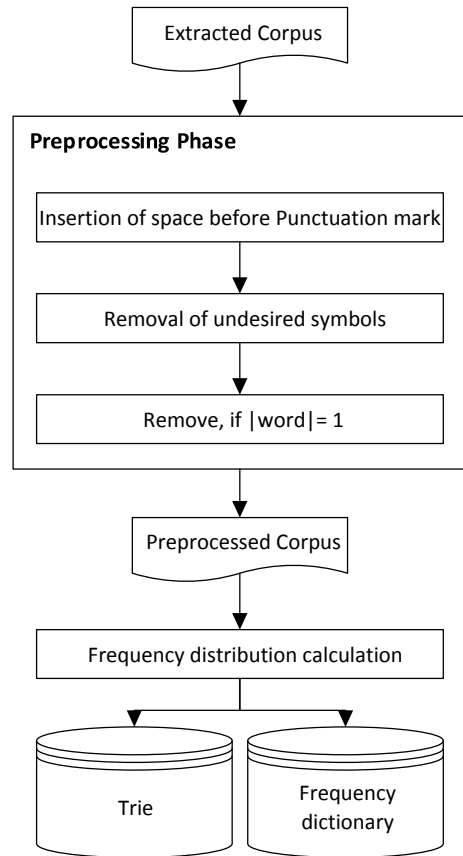
The initial phase of the system architecture is divided into four diverse steps, (1) Extraction of Punjabi news corpus, (2) Preprocessing of the corpus, (3) Calculation of frequency distribution and (4) Formation of words Trie and frequency dictionary. Figure 1 illustrates the Corpus creation and statistical inference phase of system architecture.

Extraction of Punjabi news corpus. Corpora or Corpus³ or Text corpus is referred to as a structured set of text, which is used to formulate some hypothesis about language features or perform statistical analysis on the language data. India is country of 1,324,171,354 people.⁴ Such a large population is usually seen as a huge potential reader base by Indian news industry. In 2013–14, 99660 registered newspapers and periodicals were in circulation in India [6] with the reader base of 450586 [7] in various Indian languages. The findings by [8] suggest that there are more than 300

³“Text corpus”, https://en.wikipedia.org/wiki/Text_corpus.

⁴World Population Prospects: The 2017 Revision”. United Nations Department of Economic and Social Affairs, Population Division.

Fig. 1 Corpus creation and statistical inference



million internet users across India. Advancement of technology and low data rates has widely increased the viewership of e-paper and online newspapers across readers in India. Indian newspapers have also reinvented themselves to stay relevant in changing time and have started publishing news online in local Indian languages. One of the most popular Indian languages is Punjabi⁵ language. Punjabi (ਪੰਜਾਬੀ [pañjābī]) is an Indo-Aryan language with more than 100 million speakers across the world. Punjabi is widely spoken in Punjab state of India and Pakistan, with speakers distributed across the world and is written using two distinct scripts—*Gurmukhi* and *Shahmukhi*. 1212 Newspapers and periodicals are published in Punjabi language (specifically using Gurmukhi script) in India [6] with enormous reader's base.

The Punjabi language news content extracted from Ajit,⁶ a prominent Punjabi news website was used for developing the news corpus. The extracted corpus contains five genres of news, i.e., Business, International, National, Regional, and

⁵“Punjabi Language”, https://en.wikipedia.org/wiki/Punjabi_language.

⁶“Ajit”, <https://www.ajitjalandhar.com/>.

Table 1 Summary of extracted corpus statistics

Statistics	Before preprocessing	After preprocessing
Total words	1408041	1357443
Total characters	8795714	7075784
Average length of words	4.252	4.212
Unique words	65551	48521
Mode of words	4	4

Sports. This extracted corpus is further sent to preprocessing step for additional refinement.

Preprocessing of the corpus. The extracted corpus contains many undesired symbols and words, which does not hold any semantic meaning in the word prediction system. This makes the corpus to be preprocessed before any statistical inference could be obtained from the corpus. The preprocessing phase is divided into three stages, (1) Insertion of Space before a punctuation mark, (2) Removal of undesired symbols, and (3) Removal of words with length equal to one.

Insertion of Space before a punctuation mark. The preprocessing is instigated by inserting a space before Punjabi language punctuation mark—“|” [d̄aṅ̄ d̄a:|.]. In the extracted corpus, many sentences are ended by joining “|” [d̄aṅ̄ d̄a:|.] with no space before the preceding word, which create an ambiguousness in the corpus. This makes many words a diverse entity, even if similar words are present in the corpus. For example, in ਮੇਰੀ ਮਾਂ ਬੋਲੀ ਪੰਜਾਬੀ ਹੈ [me:ri: ma:ṅ bo:li: pañjābī hē:d̄aṅ̄ d̄a:|.], punctuation mark “|” [d̄aṅ̄ d̄a:|.] is joined with the word “ਹੈ” [hē:], which make a word “ਹੈ|” [hē:d̄aṅ̄ d̄a:|.] incorrect in accordance to the Punjabi language semantics, as “ਹੈ” [hē:] is acknowledged as a complete word. Inserting one-character space before punctuation mark makes the words recognizable and accessible for processing in the later stage of statistical inference. For example, in the sentence—ਮੇਰੀ ਮਾਂ ਬੋਲੀ ਪੰਜਾਬੀ ਹੈ | [me:ri: ma:ṅ bo:li: pañjābī hē: d̄aṅ̄ d̄a:|.], in which “ਹੈ” [hē:] is now renewed into an individual word in the corpus and is available for further processing.

Removal of undesired symbols. Another stage in the preprocessing step comprises of the elimination of undesired symbols like @,!,(,), etc., as such symbols do not hold any semantic value during the statistical inference stage. After removal of undesired symbols, the processed corpus is sent to stage 3 for final processing.

Removal of words with length equal to one. All single length words were removed from the list, as such words are merely characters and do not hold any significant value in the corpus.

After completing all the stages of preprocessing on the extracted corpus, the remaining corpus is stated as the preprocessed corpus. Statistical analysis was conducted on preprocessed corpus and summary of the calculated statistics is shown in Table 1.

This preliminary preprocessing step can be stated more formally as,

Let c denotes the extracted corpus, i.e., unprocessed Bag of Words (BoW). The preprocessing phase on c can be stated as

$$\text{Preprocess}(c) = \acute{c} \quad (1)$$

where \acute{c} is the processed corpus with N variables. This processed corpus \acute{c} is forwarded to step 2 to make statistical inference.

Calculating frequency distribution. Subsequently, after the preprocessing step, another significant step in the system architecture is to make the statistical inference. For this Frequency distribution is calculated from the preprocessed corpus (\acute{c}). The frequency distribution is an instantaneous description of the dataset which states the frequency of the discrete variables in the processed corpus. This step can be stated as,

Let n denotes the number of discrete variables x in processed corpus \acute{c} . Then $\text{VAR}_{\acute{c}}$ is denoted as the set of discrete variables as

$$\text{VAR}_{\acute{c}} = \{x_0, x_1, x_2, \dots, x_n\} \quad (2)$$

The frequency of i th discrete variable in the corpus \acute{c} can be given by

$$f_0 + f_1 + f_2 + \dots + f_n = N \quad (3)$$

which furthermore can be given in form of a frequency set $\text{FREQ}_{\acute{c}}$,

$$\text{FREQ}_{\acute{c}} = \{f_0, f_1, f_2, \dots, f_n\} \quad (4)$$

where, f_i is the frequency for i th discrete variable for $i = 0, 1, 2, \dots, n$ in $\text{VAR}_{\acute{c}}$.

Set $\text{VAR}_{\acute{c}}$ and $\text{FREQ}_{\acute{c}}$ are further utilized in step 3 of a system architecture for the generation of word Trie and Frequency dictionary.

Formation of words Trie and frequency dictionary. Set $\text{VAR}_{\acute{c}}$ and $\text{FREQ}_{\acute{c}}$ are utilized in the final step for the formation of word Trie and frequency dictionary. Frequency dictionary named DICT_{PUN} is modeled in form of “<word>”: <frequency>, in which <word> corresponds to the distinct variable or word in the set $\text{VAR}_{\acute{c}}$ and <frequency> corresponds to its relative frequencies which is extracted from set $\text{FREQ}_{\acute{c}}$. DICT_{PUN} is utilized to extract the frequencies of distinct words in $O(1)$ time. Extracted distinct variables in a set $\text{VAR}_{\acute{c}}$ are further saved in the form of word Trie data structure named $(\text{TRIE}_{\text{PUN}})$ for an efficient words organization and extraction. The algorithm given in [27] is applied for the development of words Trie data structure $(\text{TRIE}_{\text{PUN}})$. The TRIE_{PUN} and DICT_{PUN} are further utilized in phase 2 of system architecture for generation of suggestion list and ranking of words.

3.2 Word Prediction

After phase 1 of system architecture, another phase is to predict the user intended word starting with specified prefix, i.e., string of character/s. This phase is termed as Phase 2. This phase is divided into three steps, (1) Initialization of prefix,

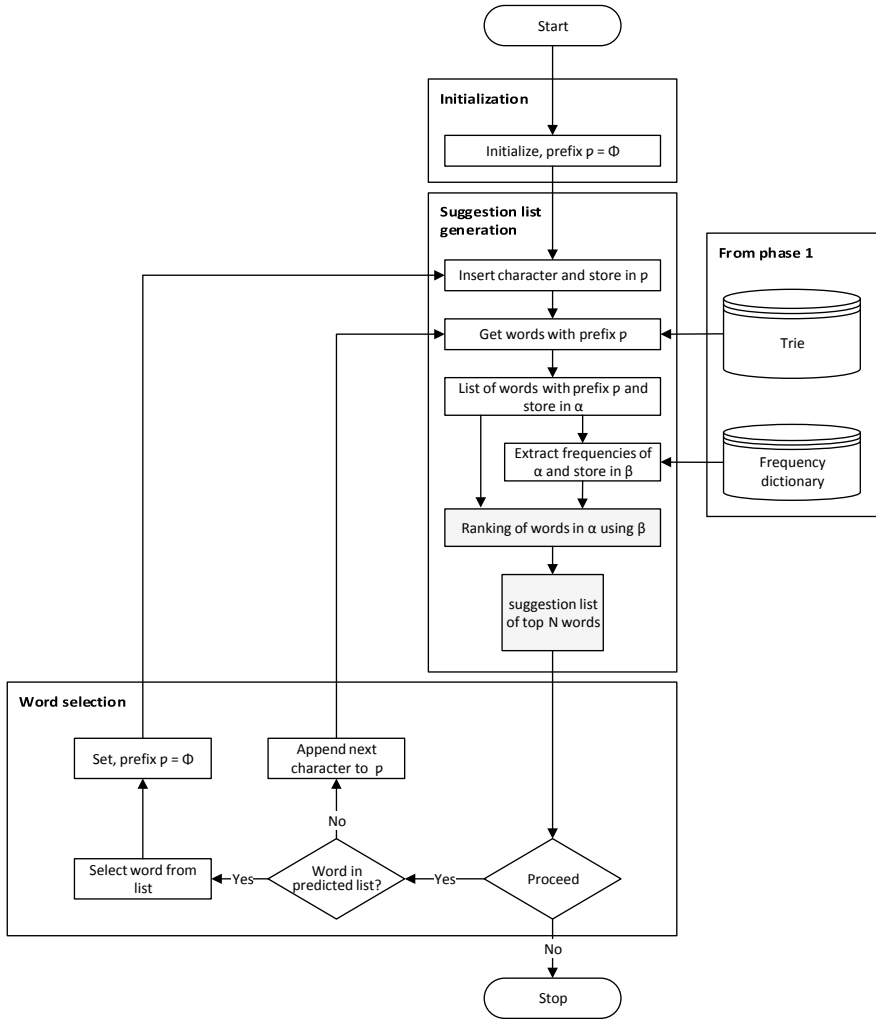


Fig. 2 Word prediction

(2) Suggestion list generation, and (3) Word selection. Figure 2 illustrates the final word prediction phase of system architecture.

Initialization of prefix. The initial phase in this step starts by initializing the prefix with null. If p denotes the user entered prefix, then this step can be stated as,

$$p = \Phi \tag{5}$$

This prefix (p) is further utilized in step 2 of the word prediction phase.

Suggestion list generation. TRIE_{PUN} and DICT_{PUN} generated in step 4 of phase 1 is utilized in this step. Subsequently, user-intended first character is stored in prefix $-p$ and is used to extract all the similar words in list $-\alpha$, from TRIE_{PUN}, which more formally is known as set VAR _{ζ} in Eq. 2. Furthermore, list of words in α are used to extract frequencies in another list $-\beta$, using frequency dictionary DICT_{PUN}, which more formally is stated as set FREQ _{ζ} in Eq. 4. The words in α are ranked against the extracted frequencies in β using the Heap Sort algorithm [27], from higher frequency to lower frequency. Algorithm 1 is used to accomplish this task. This α list of top N suggestions is listed for user selection. This system architecture was tested for 3, 5, 10, 15, 20, and 25 values of N in Sect. 7 of this article.

ALGORITHM 1: List Generation Algorithm

INPUT: Prefix(p) of the word to be predicted and dictionary (punjabi_dict) of Punjabi words as keys corresponding to their frequencies as values

OUTPUT: Returns the list of predicted words(α) to be generated

1. $\alpha = \text{TRIE}(p)$ // List of all the words to be generated with prefix p
 2. $\beta = []$ // Empty list for frequencies
 3. **for** each $\tilde{\omega} \in \alpha$ **do**
 $\beta.append(\text{punjabi_dict}[\tilde{\omega}])$ // extract frequency β for corresponding word $\tilde{\omega}$
 4. HEAPSORT(α, β) // Perform heapsort on α on the basis of β
 5. **Return** α // Return the words in order form of frequencies
-

The problem of suggestion list generalized more formally by [10] as follows, Let $L_{\text{PRED}}(p)$ denotes the set of extracted words from TRIE_{PUN} such that

$$L_{\text{PRED}}(p) \in \text{VAR}_{\zeta} \quad (6)$$

$L_{\text{PRED}}(p)$ is seen as a set of k suggestions from set VAR _{ζ} . The foremost job of said system architecture is to rank the predicted suggestions in set $L_{\text{PRED}}(p)$ to set $R_{\text{PRED}}(p)$ from higher to lower, such that $|R_{\text{PRED}}(p)| = N$, where $N > 0$, represents the top suggestions of set $L_{\text{PRED}}(p)$ to be included in set $R_{\text{PRED}}(p)$, for which cost function COST($R_{\text{PRED}}(p)$) for ranking set $R_{\text{PRED}}(p)$ is formulated as

$$\text{COST}(R_{\text{PRED}}(p)) = \sum_{\omega \in R_{\text{PRED}}(p) \subset L_{\text{PRED}}(p)} \text{cost}(\omega) \quad (7)$$

where $\text{cost}(\omega)$ represent the cost of suggesting ω .

The purpose of step 2 of word prediction phase is to rank the words extracted in set $L_{\text{PRED}}(p)$ with minimum rank in the set $R_{\text{PRED}}(p)$ which is purposely known as the cost of selecting the word $-\omega$.

Word selection. After step 2, the last step is to select the user intended word $-\omega$ from N words in the list α . If the intended word $-\omega$ is found the list, the word is

selected from the list and prefix is again initialized to null, i.e., $p = \Phi$. If user intended word is not found in the list, next character is appended to prefix $-p$.

Step 2 and 3 of phase 2 is iterated till user intended word $-\omega$ is not found for the updated value of p or till system does not exit. The maximum number of iterations made for predicting single word depends upon the length of word needed to be predicted, i.e., if $|\omega| = l$, then the number of iterations needed would be l .

4 Working Example

The intended system functioning is described using Table 2. It is assumed that the proposed system wants to predict the user intended word (ω)“ਸੋਨਾ” [so:na:]. The system initiate by entering initial prefix p_0 to be “ਸ” [s]. The system extracts the words list (α_0) starting with p_0 from TRIE_{PUN} as [‘ਸਿੰਘ’, ‘ਸੀ’, ‘ਸਰਕਾਰ’, ‘ਸੁਰੀ’, ‘ਸਨ’] and list of corresponding list of frequencies β_0 from DICT_{PUN} as [54442, 7132, 5418, 5100, 3982] from higher to lower. For every extracted word x_i in α_0 , the corresponding frequency f_i is extracted in β_0 for every $i = 0, 1, 2, \dots, n$. This initial process is termed as iteration 0.

Table 2 System architecture working example

User intended word (ω)		ਸੋਨਾ
Total characters in ω		4
Iteration 0	Prefix (p_0)	ਸ
	Total characters (c_0)	1
	Extracted words from TRIE _{PUN} (α_0)	[‘ਸਿੰਘ’, ‘ਸੀ’, ‘ਸਰਕਾਰ’, ‘ਸੁਰੀ’, ‘ਸਨ’]
	IPA Format (ipa_0)	[‘sɪn̄ ɡʰ’, ‘si:’, ‘sarka:r’, ‘su:ri’, ‘sn’]
	Extracted frequencies from DICT _{PUN} (β_0)	[54442, 7132, 5418, 5100, 3982]
Iteration 1	Prefix (p_1)	ਸੋ
	Total characters (c_1)	2
	Extracted words from TRIE _{PUN} (α_1)	[‘ਸੋਫੀ’, ‘ਸੋਚ’, ‘ਸੋਠ੍ਹੀ’, ‘ਸੋਧ’, ‘ਸੋਨੀ’]
	IPA format (ipa_1)	[‘so:ɖʰi:’, ‘so:c’, ‘so:nu:~’, ‘so:ɖʰ’, ‘so:ni:’]
	Extracted frequencies from DICT _{PUN} (β_1)	[198, 126, 116, 112, 112]
Iteration 2	Prefix (p_2)	ਸੋਨ
	Total characters (c_2)	3
	Extracted words from TRIE _{PUN} (α_2)	[‘ਸੋਠ੍ਹੀ’, ‘ਸੋਨੀ’, ‘ਸੋਨੀਆ’, ‘ਸੋਨੇ’, ‘ਸੋਨਾ’]
	IPA format (ipa_2)	[‘so:nu:~’, ‘so:ni:’, ‘so:ni:a:’, ‘so:ne:’, ‘so:na:’]
	Extracted frequencies from DICT _{PUN} (β_2)	[116, 112, 76, 48, 44]

If the intended word “सोना” [so:na:] is not found in α_0 , next character is appended to the prefix p_0 and is termed as p_1 . Again system process is repeated for p_1 , the said process is termed to be iteration 1. In iteration 1, α_1 and β_1 is extracted as [‘सोनी’, ‘सोच’, ‘सोनी’, ‘सोय’, ‘सोनी’] and [198, 126, 116, 112, 112]. Again system process is repeated as the intended word is not found in α_1 . Iteration 2 is performed by appending next characters to p_1 and terming it as p_2 . The predicted system is again tested to extract the user intended word using three characters in the prefix p_2 . α_2 and β_2 list is extracted to be, [‘सोनी’, ‘सोनी’, ‘सोनीआ’, ‘सोने’, ‘सोना’] and [116, 112, 76, 48, 44] for iteration 2.

In iteration 2 the user-intended word is found at rank 5 in list α_2 by making three keystrokes. One keystroke is saved using the described system architecture in Sect. 3.

5 Performance Metrics

For the assessment of proposed system architecture, testing metrics proposed by [11] are used. The metrics used for the assessment are discussed as follows:

Hit ratio. It is a metric that is used to describe the reliability of the word prediction system. Higher the hit ratio, higher will be the credibility of the prediction system to predict the correct word. It is defined by formula 8,

$$\text{Hit ratio} = \frac{\text{number of times words is predicted}}{\text{total number of written words}} \quad (8)$$

in which, number of times words are predicted is the calculation of the words prediction before the length actual word.

Keystroke saving. Hit ratio is not only enough to measure the actual saving of keystrokes as the predictor can guess words after entering all the characters. So, to measure the keystroke saving, the following formula 9 is used,

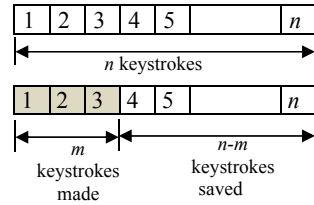
$$\text{Keystroke saving} = 1 - \frac{\text{number of keystrokes made}}{\text{total length of word}} \quad (9)$$

which is defined as the number of keystrokes saved during the course of prediction of the word. It can be defined as follows if the word comprises of n characters, and predictor predicts the word by making m strokes, the number of keystrokes saved can be defined as $1 - (m/n)$. It is depicted using the following Fig. 3.

If total keystrokes m is equal to n , then the resulting response would be 0, as no keystrokes are saved, as the resultant word is completed by making n strokes only.

Average rank. As the vertical list is used to demonstrate the predicted for user selection. The Average rank calculation mechanism is utilized to compute the average rank of the predicted words in the vertical list of suggestions. This helps in

Fig. 3 Keystroke saving mechanism



calculating at which rank the selected word appears in the list. Formula 10 is used for calculation of average rank,

$$\text{Average rank} = \frac{\text{ranks total}}{\text{number of words}} \tag{10}$$

If during the prediction of the word is not found in the list, rank is set to be 0.

Average keystrokes. The average keystroke calculation mechanism is utilized to recognize the average number of keystrokes entered for each word during the prediction testing. Formula 11 is used for the calculation of average keystrokes,

$$\text{Average keystroke} = \frac{\text{total keystrokes made}}{\text{number of words}} \tag{11}$$

6 Testing Dataset and System Configuration

One of the prominent newspaper of Punjabi–Jagbani⁷ was selected to test the performance of system architecture. Jagbani News crawler⁸ was developed to extract the Punjabi news in accordance to the news genres. As the training corpus consists of five major genres of news, i.e. Business, International, National, Regional, and Sports, similar genres of news were extracted from Jagbani website to test the performance of the system. The statistics of the news items in each genre are given in Table 3.

The system architecture was implemented on the Windows 10 Operating system, Intel Core i5-6200U CPU 2.40 GHz with 8 GB RAM. The system is designed using Python programming language. NLTK and URLLIB packages are used for the implementation of said system and design of Web crawlers. The screenshot of the designed system is shown in Fig. 4.

⁷<http://jagbani.punjabkesari.in/> .

⁸“Jagbani News crawler”, https://github.com/GurjotSinghMahi/jagbani_website_crawler.

Table 3 Classification of news in testing corpus

Genre	News items
Business	22
International	65
National	48
Regional	115
Sports	31

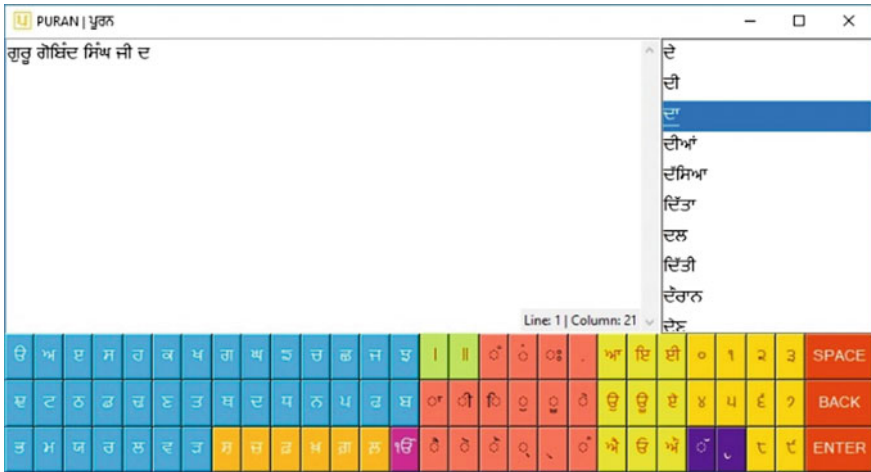


Fig. 4 Punjabi news words prediction system—PURAN

7 Results and Discussion

The performance of PURAN system was rigorously tested in the simulated environment on the basis of metrics mentioned from Eqs. (8) to (11). The extracted news items were passed to the PURAN news words prediction system for evaluation in accordance with the genres. The resultant system is observed by setting the number of words in suggestion list, i.e., N to be 3, 5, 10, 15, 20, and 25. The value of N shows the number of predicted words in accordance to their ranks, to choose from during performance analysis. Table 4 demonstrates the calculated metrics values for the proposed system architecture.

The observed scores are plotted using a line chart. Figure 5 shows PURAN performance for different genres of news in a simulated environment. Figure 5a, depicts the Hit ratio for different Genres of news, Fig. 5b illustrated the Average Rank for different genres of news, Fig. 5c depicts the Keystroke saving for different genres of news and Fig. 5d depicts the Average Keystrokes for different genres of news in a simulated environment.

Table 4 Calculated metrics for proposed system architecture

Metric	Genre	Words in the suggestion list					
		3	5	10	15	20	25
Hit ratio	Business	0.770	0.832	0.881	0.887	0.893	0.897
	International	0.781	0.847	0.878	0.887	0.891	0.895
	National	0.805	0.868	0.901	0.907	0.912	0.916
	Regional	0.829	0.890	0.924	0.932	0.938	0.940
	Sports	0.727	0.792	0.835	0.846	0.854	0.858
Average rank	Business	1.459	2.125	3.475	4.498	5.856	7.097
	International	1.464	2.151	3.315	4.394	5.358	6.349
	National	1.519	2.238	3.456	4.554	5.821	6.869
	Regional	1.544	2.207	3.481	4.649	5.811	6.788
	Sports	1.373	2.027	3.290	4.342	5.480	6.238
Keystroke saving	Business	0.386	0.444	0.503	0.528	0.550	0.565
	International	0.398	0.458	0.510	0.535	0.551	0.563
	National	0.410	0.471	0.524	0.549	0.570	0.583
	Regional	0.429	0.487	0.541	0.569	0.588	0.600
	Sports	0.381	0.440	0.493	0.518	0.538	0.547
Average keystrokes	Business	2.606	2.360	2.112	2.006	1.911	1.848
	International	2.577	2.319	2.099	1.990	1.923	1.870
	National	2.569	2.301	2.072	1.959	1.871	1.814
	Regional	2.423	2.177	1.945	1.829	1.750	1.698
	Sports	2.582	2.336	2.113	2.008	1.928	1.888

The bold text highlight the top values in the particular range

As the above results clearly demonstrate that the PURAN word predictor has performed very well on all the benchmark metrics. From Fig. 5a and d, it can be clearly derived that, PURAN has achieved highest Hit ratio in Regional news genre followed by National news genre by making lowest average keystrokes for predicting the words. These results clearly demonstrate that words can be easily predicted using PURAN words predictor for several mentioned news genres with high hit ratio and low keystrokes in Regional and National news genres. Figure 5c demonstrates that PURAN has given outstanding results for keystroke saving for Regional news genre and trailed by National genre, which is above the benchmark of 49.1%, 47%, and 40.5% keystroke saving given by [1, 16, 28], when N is taken as 10, 15, 20, and 25. Subsequently, it can be clearly stated that with the increase in the number of words in suggestion list the average keystroke saving also increases. Figure 5b states that with the increase in a number of suggestions in a vertical list of predicted words the average rank also increases.

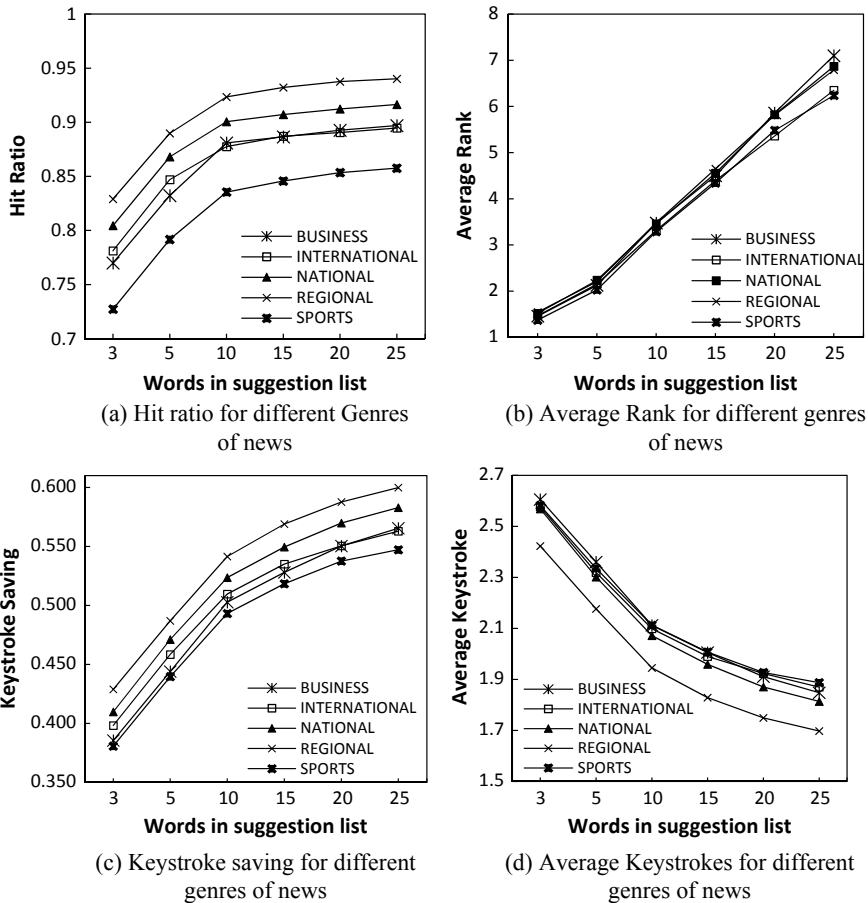


Fig. 5 PURAN performance for different genres of news in a simulated environment

8 Conclusion

Most the word prediction systems earlier designed were for the English and other European languages. This work is an effort to accomplish Punjabi Language with its first prediction system, particularly in News category. This paper presents the reader with the word prediction background with various factors affecting the word predictor accuracy, i.e., predicting methodology, a number of suggestions, user interface, and other features. The overall applied architecture used for suggesting the user the most appropriate word based on the prefix provided by the user was demonstrated. The system performance is tested on the various benchmark metrics like *Keystroke saving*, *Hit ratio*, *Average rank*, and *Average keystrokes* for rigorous review of the proposed system to examine its credibility in Punjabi news category.

PURAN prediction system works well in the categories of Regional and National news genres followed by International, Business, and Sports genres.

Garay et al. [21] states that no more than 10 suggestions must be presented to the user in suggestion list considering more physical movement, i.e., more hand and eye movement during the selection of the desired word. The authors also endorse the idea of [21] and take only 10 words for Final system architecture. The system has achieved 88.38% Average Hit ratio with 51.42% Average keystroke saving for $N = 10$. It is meaningful to mention while concluding that PURAN word predictor can be easily used to write articles for the Punjabi newspapers with an ease and also with high hit ratio and keystroke saving.

Acknowledgements The Authors would like to acknowledge the contribution of Arvinder Singh Kang from UrbanLogiq for providing the Punjabi news corpus for this research. We thank Varun Manchanda, Faculty, RGNIYD Regional Centre, Chandigarh for proofreading the draft.

References

1. Carlberger, A., Carlberger, J., Magnuson, T., Hunnicutt, M.S., Palazuelos-cagigas, S.E., Navarro, S.A.: Profet, a new generation of word prediction: an evaluation study. In: Natural Language Processing for Communication Aids, pp 23–28 (1997)
2. Matiasek, J., Baroni, M., Trost, H.: FASTY—a multi-lingual approach to text prediction. In: International Conference on Computers for Handicapped Persons, pp 243–250. Springer, Berlin (2002)
3. Dunlop, M.D., Crossan, A.: Predictive text entry methods for mobile phones. *Pers. Technol.* **4** (2–3), 134–143 (2000)
4. Sharma, M.K., Samanta, D.: Word prediction system for text entry in Hindi. *ACM Trans. Asian Lang. Inf. Process.* **13**(2), 1–29 (2014)
5. Arnold, K.C., Gajos, K.Z., Kalai, A.T.: On suggesting phrases vs. predicting words for mobile text composition. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology—UIST’16, pp 603–608 (2016)
6. Number of registered newspapers and periodicals by language and periodicity. http://www.mospi.gov.in/sites/default/files/statistical_year_book_india_2015/Table-36.1.xlsx
7. Claimed circulation of registered newspapers. http://www.mospi.gov.in/sites/default/files/statistical_year_book_india_2015/Table-36.4.xlsx
8. 11th annual report 2014–15 (2015). http://www.iamai.in/sites/default/files/annual_report/AnnualReport2014-15.pdf. Accessed: 06 May 2018
9. Longuet-Higgins, H.C., Ortony, A.: The adaptive memorization of sequences. In: Proceedings of the 3rd Annual Machine Intelligence Workshop, pp 311–322 (1968)
10. Cai, F., de Rijke, M.: A survey of query auto completion in information retrieval. *Found. Trends® Inf. Retr.* **10**(4), 273–363 (2016)
11. Garay-Vitoria, N., Abascal, J.: Text prediction systems: a survey. *Univ. Access Inf. Soc.* **4**(3), 188–203 (2006)
12. Garay-Vitoria, N., González-Abascal, J.: Using statistical and syntactic information in word prediction for input speed enhancement. In: Information Systems Design and Hypermedia, pp 223–230 (1994)
13. Heckathorne, C.W., Childress, D.S.: Applying anticipatory text selection in a writing aid for people with severe motor impairment. *IEEE Micro* **3**(3), 17–23 (1983)

14. Hunnicutt, S.: Input and output alternatives in word prediction. *STL-QPSR* **2**(3), 015–029 (1987)
15. Swiffin, A.L., Arnott, J.L., Pickering, J.A., Newell, A.F.: Adaptive and predictive techniques in a communication prosthesis. *Augment. Altern. Commun.* **3**(4), 181–191 (1987)
16. Venkatagiri, H.: Efficiency of lexical prediction as a communication acceleration technique. *Augment. Altern. Commun.* **9**(3), 161–167 (1993)
17. Fazly, A.: The use of syntax in word completion utilities. M.Tech. Thesis, Department of Computer Science, University of Toronto (2002)
18. Jurafsky, D., Martin, J.H.: *Speech and Language Processing*. Prentice Hall, New Jersey (2000)
19. Bertenstam, J., Hunnicutt, S.: Adding morphology to a word predictor. In: *The European Context for Assistive Technology (Proceedings of the 2nd Tide Congress)*, pp 312–315 (1995)
20. Garay-Vitoria, N., Abascal, J.G.: Word prediction for inflected languages application to basque language. In: *Natural Language Processing for Communication Aids*, pp 29–36 (1997)
21. Garay, N., Abascal, J., Gardeazabal, L.: Evaluation of prediction methods applied to an inflected language. In: *International Conference on Text, Speech and Dialogue*, pp 397–403 (2002)
22. Garay-vitoria, N., Abascal, J.: A comparison of prediction techniques to enhance the communication rate. In: *ERCIM Workshop on User Interfaces for All*, pp 400–417 (2004)
23. Garay-Vitoria, N., Abascal, J.: User interface factors related to word-prediction systems. In: *Proceedings of the 7th International Conference on Work with Computing Systems WWCS2004*, pp 77–82 (2004)
24. Newell, A., Arnott, J., Booth, L., Beattie, W., Brophy, B., Ricketts, I.: Effect of the ‘PAL’ word prediction system on the quality and quantity of text generation. *Augment. Altern. Commun.* **8**(4), 304–311 (1992)
25. MacKenzie, I.S., Tanaka-Ishii, K.: Text entry using a small number of buttons. In: MacKenzie, I.S., Tanaka-Ishii, K. (eds.) *Text Entry Systems: Mobility, Accessibility, Universality*, pp. 105–121. Morgan Kaufmann, San Francisco, CA (2007)
26. Koester, H.H., Levine, S.: Model simulations of user performance with word prediction. *Augment. Altern. Commun.* **14**(1), 25–36 (1998)
27. Brass, P.: *Advanced Data Structures*, vol. 1. Cambridge University Press, Cambridge (2008)
28. Lesh, G.W., Moulton, B.J., Higginbotham, D.J.: Techniques for augmenting scanning communication. *AAC Augment. Altern. Commun.* **14**(2), 81–101 (1998)
29. Ghosh, S., Sarcar, S., Samanta, D.: Designing an efficient virtual keyboard for text composition in Bengali. In: *Proceedings of the 3rd International Conference on Human Computer Interaction*, April, pp 84–87 (2011)

Implementation of hDE-HTS Optimized T2FPID Controller in Solar-Thermal System



Binod Shaw, Jyoti Ranjan Nayak and Rajkumar Sahu

Abstract In this paper, a fair approach is interpreted to validate the novelty of Type-2 Fuzzy PID (T2FPID) controller over Type-1 or conventional Fuzzy PID (FPID) and PI controller as secondary frequency controller and Heat Transfer Search (HTS) algorithm is adopted to extract the optimum gains of the controllers. T2FPID controller has a beautiful property to handle large uncertainties of the system with extra degree of freedom. T2FPID controller is implemented in a two area interconnected thermal-PV power system to enhance the system performance. ITAE is adopted as objective function of the system to lessen the undershoot, overshoot, and settling time of frequency and tie-line power deviation. A novel hDE-HTS algorithm is adopted to enhance the system performance by searching the relevant pair of gains of controller. This analysis is executed by implementing a step signal (load disturbance) of magnitude 0.1 in area-2 to study the transiency of the system. The novelty of this work is to implement hDE-HTS optimized T2FPID controller to enhance the Solar-thermal system responses (Frequency and tie-line power deviations).

Keywords Automatic generation control (AGC) · Differential evolution (DE) · Fuzzy PID controller (FPID) · Heat transfer search (HTS) · PI controller · Type-2 fuzzy PID controller (T2FPID)

B. Shaw · J. R. Nayak (✉) · R. Sahu
Department of Electrical Engineering, NIT, Raipur, Chhattisgarh 492010, India
e-mail: bapi.jyoti.2@gmail.com

B. Shaw
e-mail: binodshaw2000@gmail.com

R. Sahu
e-mail: rajkumarsao@gmail.com

1 Introduction

In recent power system, interconnection of areas is an eminent approach to share the load demand along with to maintain stability and reliability of the system. This helps to supply reliable, stable, and economical power to the consumers. In interconnected power system, load deviation in one area causes frequency and power deviations in all interconnected areas. The small deviations are diminished by primary controller of generators. But for large deviations, secondary controller is highly essential to diminish the deviations [1]. Proportional (P), Integral (I), Derivative (D), PID, fuzzy, etc., are used as secondary controller to enhance the capability to handle the load fluctuations in the power system. The elementary purposes of the AGC are to

- i. Settle the frequency deviation as soon as possible (i.e., $\Delta f = 0$).
- ii. Settle the tie-line power deviation to its scheduled value (i.e., $\Delta P_{\text{tie}} = 0$).
- iii. Minimize settling time, undershoot, and overshoot and of the system after any load fluctuation.

Renewable energy is an imperative alternative to fulfill the enormous expanding load demand. Solar system is the most preferable energy system due to its simple implementation. The combination of renewable energy systems enhances the complexity of the power system. In complex power system, intelligent controllers with enormous competence to handle uncertainties should prefer.

AGC is the most important strategy in the power system which concerns the reliable and stable power generation. Various secondary controllers have been implemented as AGC to enhance the system capability to handle the load fluctuation. Ibrahim et al. [2] have portrayed a review of AGC with different schemes to analyze the novelty of controllers. Abraham et al. [3] have implemented superconducting magnetic energy storage (SMES) to portray the effect of storage device on AGC in a two area interconnected hydro-thermal system. Singh and Sen [4] have successfully designed an interconnected power system with thermal units by using PID controller optimized by GA. Dash et al. [5] has introduced cascade controller titled as PI-PD controller optimized by FPA in four area interconnected reheat thermal power plants. Many researchers have enhanced the PID controller by reconstruct the PID controllers entitled as 2DOF PID, 3DOF PID, cascade PD-PID, etc. in [6–8] respectively. Even for better performance the fractional order PID controller is enforced as AGC in [9] and [12]. From last few decades, fuzzy logic controller is generally preferred for various applications. Fuzzy logic controller (FLC) has convinced as a very robust and intelligent controller used as AGC in [13, 14]. The robustness of FPID controller is enhanced with the assemblage of advantages from both PID and FLC controller. FPID controller optimized by different powerful algorithms have depicted in [15–19]. The degree of freedom of FLC is enhanced by providing Foot Of Uncertainty (FOU) entitled as T2FLC [20]. T2FPID controller is adopted to enhance the AGC performance in [21].

In present work, grid connected PV system is enforced in area-1 and reheat thermal plant is enforced in area-2. From last few decades, FPID controller has dominated over conventional controllers. T2FPID controller is an adequate controller over FPID controller to handle nonlinear and uncertain system. HTS algorithm is adopted to extract the controller gains. The superiority of hDE-HTS algorithm is validated over HTS to tune T2FPID controller.

2 System Investigated

In this work, two area interconnected power system is implemented. In area-1 PV system and in area-2 reheat thermal system are resided in the power system network as illustrated in Fig. 1 [10, 11]. The power system parameters are portrayed in Appendix 1. To verify the transiency of the system a 10% load change is enforced in area-2 which propagates an error in each area entitled as Area Control Error (ACE). The basic objective of AGC is to lessen the ACE which concedes deviations in tie-line power and frequency deviation as characterized in Eq. (1).

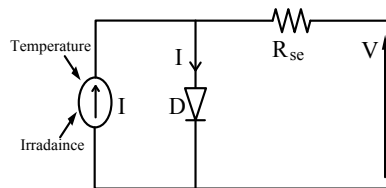
$$ACE = B\Delta f + \Delta P_{tie} \tag{1}$$

where B_1 , and B_2 are the bias factors of frequency. The deviations of frequency with respect to nominal value in area-1 and area-2 are Δf_1 and Δf_2 respectively. The deviation of power in tie-line is ΔP_{tie} and is characterized in Eq. (2).

$$\Delta P_{tie} = \frac{2\pi T_{12}}{s} (\Delta f_1 - \Delta f_2) \tag{2}$$

PI, FPID, and T2FPID controllers are executed in both the areas individually to scrutinize the controller potency to enhance the system performance. Intelligent T2FPID controller is observed as superior controller over other conventional controllers due to its capability to handle nonlinear and uncertainties. The structure of T2FPID controller is as portrayed in Fig. 2. The objective function for this system by considering tie-line power deviation and frequency deviation is characterized in Eq. (3)

Fig. 1 Equivalent circuit solar system [11]



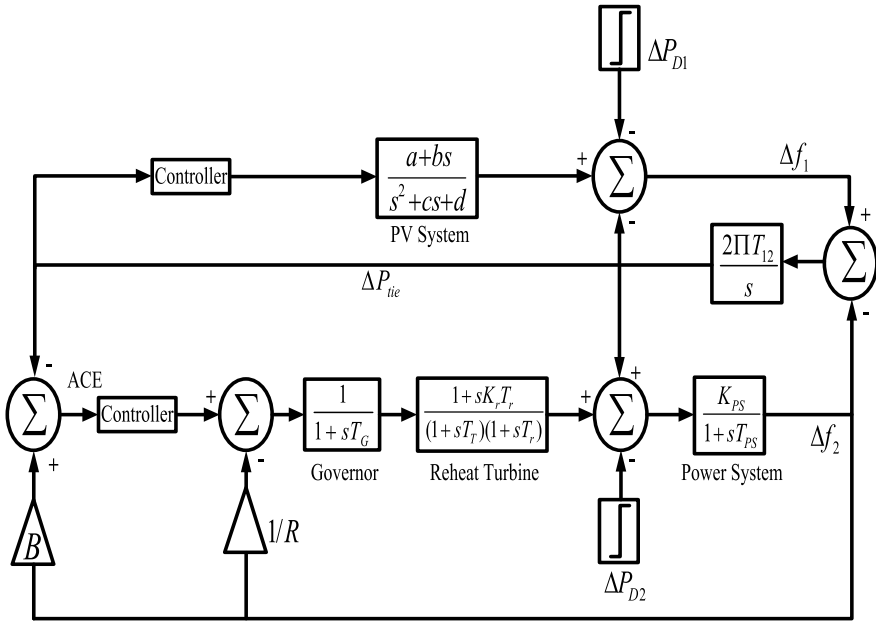


Fig. 2 Transfer function model of solar-reheat thermal power system [10] and [11]

$$J = \text{ITAE} = \int_0^T t(\Delta f_1 + \Delta f_2 + \Delta P_{\text{tie}}) \quad (3)$$

Subject to:

$$-2 \leq K_i \leq 2 \quad i = 1, 2, \dots, n$$

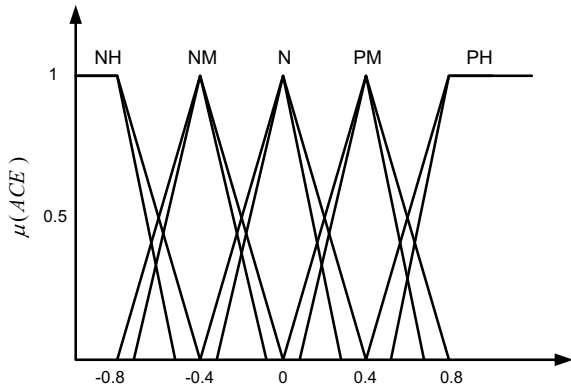
PV System

The PV system is basically depends upon two important factors, i.e., irradiance and temperature. The PV system can be represented with a current source, diode and series resistance as portrayed in Fig. 1.

In present work, PV system is adopted with constant irradiance (1000 w/m²) and temperature (27 °C). The transfer functions by concerning MPPT, inverter, and filter as characterized in Eq. (4).

$$G_{\text{PV}} = \frac{-18s + 900}{s^2 + 100s + 50} \quad (4)$$

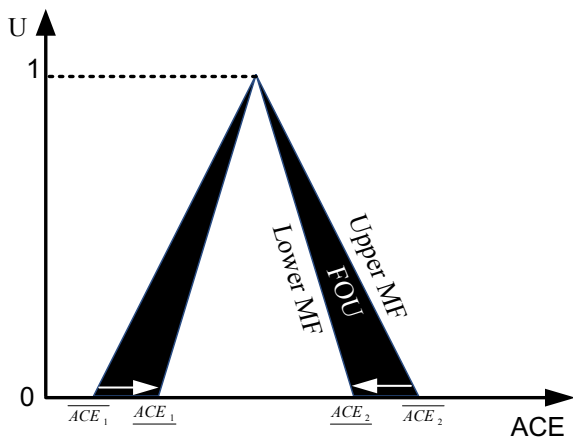
Fig. 3 MFs of Type-2 FLC



3 Type-2 Fuzzy PID Controller

Design of intelligent controllers is an imperative approach to provide reliable and stable power in a highly complex power system. T1FPID controller has substantiated as a better controller over classical controllers. In this work, T2FPID controller is adopted to contribute the supremacy of T2FPID controller over FPID and PID controller. The framework of T2FPID controller is portrayed in Fig. 5. The supremacy of T2FLC is better than T1FLC because of the FOU provided by the controller. FOU is the boundary provided by the two type-1 membership function (MF) entitled as Lower mf (LMF) and Upper mf (UMF) as portrayed in Fig. 4 [20]. The degree of freedom is enhanced by this FOU and also enhances the capability to handle the uncertain information. The MF adopted for T2FPID controller is illustrated in Fig. 3. In this work, five interval type-2 MFs are adopted entitled as negative huge (NH), negative mild (NM), Null (N), positive mild (PM), and positive huge (PH). The rule base is tabulated in Table 1.

Fig. 4 Membership function of interval type-2



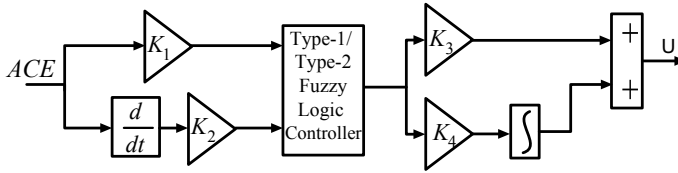


Fig. 5 Type-2 FPID controller structure

Table 1 Rule base

\dot{e} e	NH	NM	N	PM	PH
NH	NH	NH	NM	NM	N
NL	NH	NM	NM	N	PM
Z	NM	NM	N	PM	PM
PL	NM	N	PM	PM	PH
PH	N	PM	PM	PH	PH

4 hDE-HTS Algorithm

The basic purpose of using optimization technique is to design optimum controller which is best-suited for the system. In this work, a novel algorithm entitled as HTS is adopted to contribute highly convenient gain parameters within specified band to minimize “ J ” as illustrated in Eq. (3). Patel et al. [22] have introduced HTS algorithm stimulated from the law of thermodynamics and heat transfer. HTS algorithm is derived from the movement of molecules during three modes of heat transfer, i.e., conduction, convection, and radiation. HTS is a population-based iterative computational technique. The molecules, temperature and energy level are proportionate to the population, design variables and solution respectively. DE algorithm is very influential to enhance the diversity factor of the algorithm with a possibility to trap into local optima [23]. The combination of benefits of both algorithms is proficient enough as compare to individual one. The steps for hDE-HTS algorithm are as:

1. Initialize random population, i.e., $[X]_{NP \times D}$.
2. Evaluate the fitness of the population, i.e., $f(X)$.
3. Initialize the conduction factor (CEF), convection factor (COF) and radiation factor (RF).
4. Donor vector (V_i) is defined as in Eq. (5).

$$V_i = X_{i,r_1} + F(X_{i,r_2} - X_{i,r_3}) \tag{5}$$

where X_{i,r_1} , X_{i,r_2} and X_{i,r_3} are three randomly taken vectors from X_i .

5. Offspring vector (U_i) used for crossover operation is characterized as in Eq. (6).

$$U_i = \begin{cases} V_i & \text{If rand} \leq \text{cr} \\ X_i & \text{Otherwise} \end{cases} \quad (6)$$

6. Selection of target vector is defined as in Eq. (7).

$$X_i = \begin{cases} U_i & \text{If } f(U_i) \leq f(X_i) \\ X_i & \text{Otherwise} \end{cases} \quad (7)$$

7. The solution of DE algorithm is considered as initial vector for HTS algorithm.

8. Initialize a random value R [0 1].

9. Conduction phase (if $R \leq 0.33333$)

Update the particle position as characterized in Eqs. (8) and (9).

$$X_i^{g+1} = \begin{cases} X_k + (-R^2 * X_k), & \text{if } f(X_i^g) > f(X_k^g) \\ X_i + (-R^2 * X_i), & \text{if } f(X_i^g) < f(X_k^g) \end{cases} \quad \text{if } f(X_i^g) \leq f(X_{\text{best}}^g) / \text{CEF} \quad (8)$$

$$X_i^{g+1} = \begin{cases} X_k + (-r * X_k), & \text{if } f(X_i^g) > f(X_k^g) \\ X_i + (-r * X_i), & \text{if } f(X_i^g) < f(X_k^g) \end{cases} \quad \text{if } f(X_i^g) \geq f(X_{\text{best}}^g) / \text{CEF} \quad (9)$$

where r , g and k are the random number, generation and random solution.

10. Convection phase (if $0.3333 \leq R \leq 0.6666$)

Update the particle position as characterized in Eq. (10).

$$X_i^{g+1} = X_i^g + R * (X_{\text{best}} - \text{TCF} * X_i) \quad (10)$$

Where TCF is the temperature change factor.

$$\text{TCF} = \begin{cases} \text{abs}(R - r), & \text{if } f(X_i^g) \leq f(X_{\text{best}}^g) / \text{COF} \\ \text{round}(1 + r), & \text{if } f(X_i^g) \leq f(X_{\text{best}}^g) / \text{COF} \end{cases}$$

11. Radiation phase ($R \geq 0.6666$)

Update the particle position as characterized in Eqs. (11) and (12).

$$X_i^{g+1} = \begin{cases} X_i^g + R(X_k - X_i), & \text{if } f(X_i^g) > f(X_k^g) \\ X_i^g + R(X_i - X_k), & \text{if } f(X_i^g) < f(X_k^g) \end{cases}, \text{if } f(X_i^g) \leq f(X_{\text{best}}^g) / \text{RF} \quad (11)$$

$$X_i^{g+1} = \begin{cases} X_i^g + r(X_k - X_i), & \text{if } f(X_i^g) > f(X_k^g) \\ X_i^g + r(X_i - X_k), & \text{if } f(X_i^g) < f(X_k^g) \end{cases}, \text{if } f(X_i^g) \geq f(X_{\text{best}}^g) / \text{RF} \quad (12)$$

12. The best fitness particles updated for next generation by replacing the worst particles.
13. Repeat steps 4–12 until the termination criteria satisfied.
14. Evaluate the optimum solution.

The parameters of hDE-HTS algorithm are illustrated in Appendix 2.

5 Result and Discussion

HTS algorithm is adopted to design optimum PID, FPID, and T2FPID controller. K_1 , K_2 , K_3 and K_4 are the design variables to be tuned by HTS. A novel hDE-HTS algorithm is opted to tune T2FPID controller to validate the superiority of the hybrid controller. Both HTS and hDE-HTS algorithm are executed individually for different controllers. Each controller is optimized with 50 numbers of population and 100 generations. The basic purpose of execution of HTS algorithm is to lessen ITAE (Integral Time Absolute Error) as cited in Eq. (3). The optimal gain values of all controllers extracted by using both algorithms are tabulated in Table 2.

HTS algorithm optimized PI is validated as better controller than FA optimized PI controller [11]. In the same fashion, supremacy of HTS optimized FPID controller is substantiated by contrasting with ICA-optimized FPID controller [10].

The response parameters by conceding the settling time, undershoot and overshoot of both frequency and tie-line power deviation are tabulated in Table 3. In Table 3, a fair observation is illustrated to demonstrate the supremacy of T2FPID controller and hDE-HTS algorithm. The response parameters and ITAE of the T2FPID controller are minimum.

The deviations in frequency of both areas and tie-line power are portrayed in Figs. 6, 7, and 8 respectively.

Figures 6, 7, 8 and Table 3 precisely represent the superiority of the T2FPID controller. T2FPID controller minimizes the transiency (undershoot, overshoot and settling time) of the system as correlate to PI and FPID controllers.

Table 2 HTS optimized gain parameters

Controllers	Area-1				Area-2			
	K_1	K_2	K_3	K_4	K_1	K_2	K_3	K_4
HTS PI	-1.1861	-0.4544			-1.9474	-1.7060		
HTS FPID	-1.3502	-1.8061	1.7636	1.0369	2.0000	0.2045	-2.0000	-2.0000
HTS T2FPID	-1.0440	-2.0000	2.0000	1.5284	1.3001	1.1986	-2.0000	-1.6456
hDE-HTS T2FPID	-1.2922	-1.7420	1.6962	0.9909	2.0000	0.2031	-2.0011	-2.0000

Table 3 Response parameters such as undershoot (U_{sh}), overshoot (O_{sh}) and settling time (T_s) of the system

Controllers	Undershoot			Overshoot			Settling time			ITAE
	ΔF_1	ΔF_2	ΔP_{tie}	ΔF_1	ΔF_2	ΔP_{tie}	ΔF_1	ΔF_2	ΔP_{tie}	
PI FA [11]	-0.3109	-0.2308	-0.0460	0.1495	0.115	0.0459	25.76	26.91	20.58	5.464
PI HTS	-0.2064	-0.1641	-0.0357	0.0554	0.0859	0.0338	17.19	17.06	10.14	2.312
FPID ICA [10]	-0.1699	-0.1446	-0.0051	0.0207	0.0192	0.0117	16.34	16.35	6.40	1.595
FPID HTS	-0.1255	-0.1192	-0.0004	0	0.0043	0.0029	12.97	11.21	1.86	1.36
T2FPID HTS	-0.0418	-0.0411	-0.0003	0	0.0074	0.0008	12.35	11.16	0.789	1.04
T2FPID hDE-HTS	-0.01387	-0.0142	-0.0001	0	0.0007	0.0006	12.14	10.85	0.76	0.85

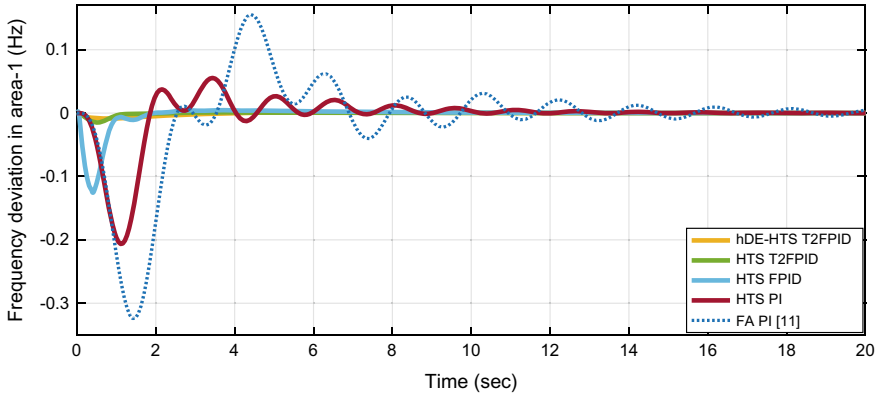


Fig. 6 Frequency deviation in area-1

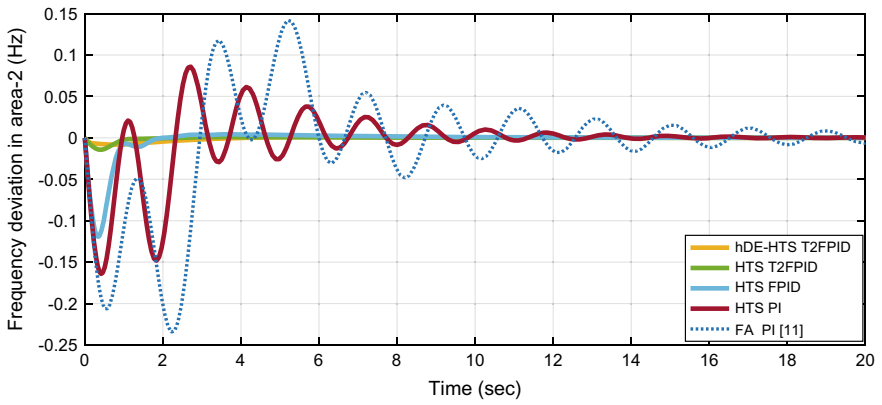


Fig. 7 Frequency deviation in area-2

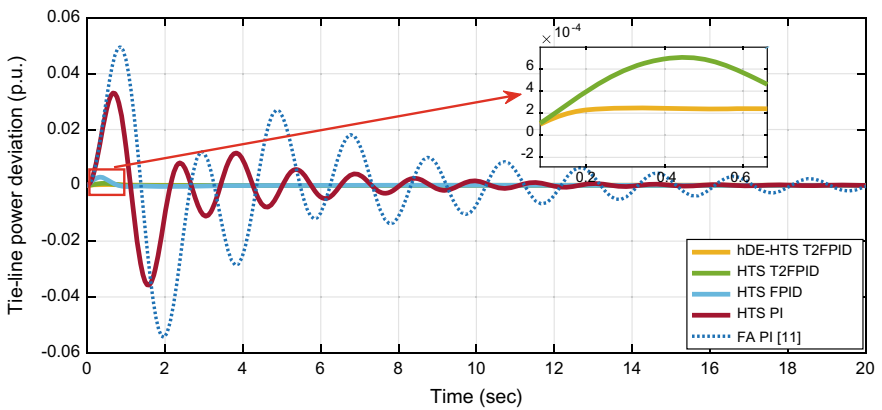


Fig. 8 Tie-line power deviation

6 Conclusion

In this work, the interconnected power system integrated with PV system with constant irradiance and temperature is simulated with various controllers. The purpose to contribute an intelligent control of frequency and power is achieved by employing PI, FPID, and T2FPID controllers. T2FPID controller yields better response over other controllers due to its capability to handle nonlinear and uncertain information. HTS and hDE-HTS algorithms are employed to extract the relevant gain parameters of controller to enhance the potential of the controller. An effort to validate the excellency of T2FPID controller and hDE-HTS algorithm is portrayed clearly through this paper.

Appendix 1: Power System Parameters

$K_{ps} = 120 \text{ Hz/p.u. MW}$, $T_{ps} = 20 \text{ s}$, $B = 0.4249$;
 $R = 2.4 \text{ Hz/p.u. MW}$; $T_G = 0.08 \text{ s}$; $T_T = 0.3 \text{ s}$;
 $K_r = 0.33$; $T_r = 10 \text{ s}$; $T_{12} = 0.0707$;
 $ab/c/d = 900/-18/100/50$;

Appendix 2: Assumptions of Algorithms

COF = 10; CDF = 2; RF = 10;

References

1. Kundur, P.: Power System Stability and Control by Prabha Kundur.pdf, p. 1199 (1993)
2. Ibrahim Kumar, P., Kothari, D.P.: Recent philosophies of automatic generation control strategies in power systems. IEEE Trans. Power Syst. **20**(1), 346–357 (2005)
3. Abraham, R.J., Das, D., Patra, A.: Automatic generation control of an interconnected hydrothermal power system considering superconducting magnetic energy storage. Int. J. Electr. Power Energy Syst. **29**(8), 571–579 (2007)
4. Singh, R., Sen, I.: Tuning of PID controller based AGC system using genetic algorithms, vol. 3, pp 531–534 (2004)
5. Dash, P., Saikia, L.C., Sinha, N.: Flower pollination algorithm optimized PI-PD cascade controller in automatic generation control of a multi-area power system. Int. J. Electr. Power Energy Syst. **82**, 19–28 (2016)
6. Sahu, R.K., Panda, S., Rout, U.K.: DE optimized parallel 2-DOF PID controller for load frequency control of power system with governor dead-band nonlinearity. Int. J. Electr. Power Energy Syst. **49**(1), 19–33 (2013)

7. Sinha, N., Saikia, L.C., Rahman, A.: Maiden application of hybrid pattern search-biogeography based optimisation technique in automatic generation control of a multi-area system incorporating interline power flow controller. *IET Gener. Transm. Distrib.* **10**(7), 1654–1662 (2016)
8. Dash, P., Saikia, L.C., Sinha, N.: Automatic generation control of multi area thermal system using Bat algorithm optimized PD-PID cascade controller. *Int. J. Electr. Power Energy Syst.* **68**, 364–372 (2015)
9. Alomoush, M.I.: Load frequency control and automatic generation control using fractional-order controllers. *Electr. Eng.* **91**(6), 357–368 (2010)
10. Arya, Y.: Automatic generation control of two-area electrical power systems via optimal fuzzy classical controller. *J. Frankl. Inst.* **355**(5), 2662–2688 (2018)
11. Abd-Elazim, S.M., Ali, E.S.: Firefly algorithm-based load frequency controller design of a two area system composing of PV grid and thermal generator. *Electr. Eng.* **100**(2), 1253–1262 (2018)
12. Nayak, J.R., Shaw, B.: Application of group hunting search optimized cascade PD-fractional order PID controller in interconnected thermal power system. *Trends Renew. Energy* **4**(3), 22–33 (2018)
13. Indulkar, C.S., Raj, B.: Application of fuzzy controller to automatic generation control. *Electr. Mach. Power Syst.* **23**(2), 209–220 (1995)
14. Çam, E., Kocaarslan, I.: Load frequency control in two area power systems using fuzzy logic controller. *Energy Convers. Manag.* **46**(2), 233–243 (2005)
15. Sahu, B.K., Pati, T.K., Nayak, J.R., Panda, S., Kar, S.K.: A novel hybrid LUS-TLBO optimized fuzzy-PID controller for load frequency control of multi-source power system. *Int. J. Electr. Power Energy Syst.* **74**, 58–69 (2016)
16. Nayak, J.R., Pati, T.K., Sahu, B.K., Kar, S.K.: Fuzzy-PID controller optimized TLBO algorithm on automatic generation control of a two-area interconnected power system. In: *IEEE International Conference on Circuits, Power and Computing Technologies ICCPCT 2015*, pp 4–7 (2015)
17. Pati, T.K., Nayak, J.R., Sahu, B.K.: Application of TLBO algorithm to study the performance of automatic generation control of a two-area multi-units interconnected power system. In: *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems SPICES 2015* (2015)
18. Nayak, J.R., Shaw, B., Das, S., Sahu, B.K.: Design of MI fuzzy PID controller optimized by Modified Group Hunting Search algorithm for interconnected power system. *Microsyst. Technol.* **24**, 3615–3621 (2018)
19. Nayak, J.R., Shaw, B., Shahu, B.K.: Load frequency control of hydro-thermal power system using fuzzy PID controller optimized by hybrid DECRPSO algorithm. *Int. J. Pure Appl. Math.* **114**(9), 147–155 (2017)
20. Biglarbegian, M., Melek, W.W., Mendel, J.M.: Design of novel interval type-2 fuzzy controllers for modular and reconfigurable robots: theory and experiments. *IEEE Trans. Ind. Electron.* **58**(4), 1371–1384 (2011)
21. Nayak, J.R., Shaw, B., Sahu, B.K.: Application of adaptive-SOS (ASOS) algorithm based interval type-2 fuzzy-PID controller with derivative filter for automatic generation control of an interconnected power system. *Eng. Sci. Technol. Int. J.* **21**(3), 465–485 (2018)
22. Patel, K., Savsani, V.J.: Heat transfer search (HTS): a novel optimization algorithm. *Inf. Sci. (NY)* **324**, 217–246 (2015)
23. Storn, R., Price, K.: Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **11**(4), 341–359 (1997)

Design of Sigma-Delta Converter Using 65 nm CMOS Technology for Nerves Organization in Brain Machine Interface



Anil Kumar Sahu, G. R. Sinha and Sapna Soni

Abstract In this paper, an overview of the present related works in the field of neuroscience is determined. The parts of the neural interface using sigma-delta converter are examined the overall ADC is driven with low voltage to improve the control utilization in the nerves organization. In this work, a basic parts of nervous system is demonstrated which is conquer to solve the problem by delivering power and transmitting data in a minimized manner. For these the above signal is first transmitted to the brain using the help of electrode into the central nervous system of the brain where the signal is diagnosed using brain computer interface by analyzing the data from analog-digital converter. Sigma-delta converter is used for visualizing low frequency signal. Major advantage of this converter is that firstly, clocking circuit need not be design and secondly, it provides good accuracy. Actually, there is no role of digital to analog converter (D/A) connected wirelessly in proposed design. A new topology based on A/D converter, which plays a wide role to minimum supply voltage, and an inactive integrator to decrease control utilization is exhibited, for empowering an in-channel advanced converter plot in a large-scale neural recording implant.

Keywords Large-scale neural recording implants • Low power sigma-delta A/D • Nerves organization • Comparator circuit

A. K. Sahu (✉)
SSGI (FET), SSTC Bhilai, Bhilai, India
e-mail: anilsahu82@gmail.com

G. R. Sinha
IIIT Bangalore, Bangalore, India
e-mail: drgsinha@ieee.org

S. Soni
SSTC Bhilai, Bhilai, India
e-mail: mesapna93@gmail.com

1 Introduction

In this work, a basic parts of nervous system is demonstrated which is conquer to solve the problem by delivering power and transmitting data in a minimized manner. For these the above signal is first transmitted to the brain using the help of electrode into the central nervous system of the brain where the signal is diagnosed using brain computer interface by analyzing the data from analog-digital converter. Sigma-delta converter is used in this paper for visualizing low frequency signal; a comparator based sigma-delta converter has designed in order to change over a few input nerves path into an advanced representation with low obstruction. Sigma-delta A/D converter has been successfully designed so that a couple of info nerves way into an upgraded model with low barrier. In this type of converters clocking circuit is avoided that is why accuracy is better to maintained from low frequency signal into high resolution digital value. Right when the need is to restrict control utilization, an in-channel execution one A/D per channel can provoke huge change. The input signal is determined by utilizing an analog to digital converter by considers as an outsourcing the benefit and control of converter part on the outer collector side when joined with radio transmitter, which have a successful part for nerves organization implementation.

A new topology based on simple to advance converter that play a wide role to minimum supply voltage and an inactive integrator to decrease control utilization is exhibited, for empowering an in-channel advanced converter plot in a large-scale neural recording implant.

2 Literature Review

This section identifies and formalizes the problem faced while implementing the references. Noteworthy contributions in the proposed area have been summarized in previous section. However, some significant contributions are reported here as noteworthy contribution.

Sigma-delta converter has been used in channel medium for digitalizing small amplification of brain signal for large-scale implantation of neural implants because dynamic range of sigma-delta modulator is still less in [1]. Dynamic range will be improved depends on design parameter of operational amplifier comparator circuit and switch capacitor used in the design specification. 3-D neural chronicle micro-system has been created in [2] to drives the signify number of neural channels that can be recorded over a trans-conductions for remote connection. Some places front-end processor for multichannel neuronal narrative is depicted in [3]. It gets 12 differential-input channels of embedded account cathodes.

Low power consistent time sigma-delta modulators circuit has been used to configuration subtle elements in [4]. Lithographically portrayed microelectrode clusters currently allow high-thickness recording for recreation in the cerebrum and

are empowering new bits of information into the affiliation limit of the focal sensory system in [5]. In [6] exhibited a coordinated circuit for remote neural incitement, alongside bench top and in vivo trial comes about. The chip can drive 100 individual incitement terminals with consistent current heartbeats of shifting adequacy, span, interphase deferral, and redundancy rate. In [7] presented a multichip structure accumulated with a therapeutic survey treated steel microelectrode group expected for neural records from different channels. Transmission limit diminishing is capable through movement potential disclosure and finish catch of simulation by strategies for on chip data buffering in [8].

The SAR ADC has been utilized to controlled 32 bit channel remote incorporated in system on-a-chip (SoC). The estimated control utilization of each chronicle channel; consume less power in [9]. Ultra-low-power 32-channel neural-recording coordinated circuit chip comprises of eight neural account modules in [10] where every module contains neural recorder unit, multiplexer(MUX) unit, an A/D converter, and a serial programming interface unit. Introduced a third request incremental sigma-delta converter (ADC) for time-multiplexed signals. Incremental sigma-delta modulation has been used to convey medium to high assurance necessities of multichannel applications, while a third demand steady time execution be inspected as a possibility for low-control plans in [11]. Inside neural interface IC for neural activity dynamic figure of A/D is key factor for electronic control modules and power organization circuits in [12]. The sigma-delta is used to control the voltage controlled oscillator circuit used in neural prosthetic [13]. Different sigma-delta modulator has been proposed to diminish control utilization and size in implantable bio-interfacing frameworks in [14].

Bidirectional neural interfacing applications with filter information rates: higher rates are required for recording of uplink signals compare to downlink signals in [15]. In [16] Exhibited a mind development in vivo requires gathering bioelectrical signals from a couple of microelectrodes in the meantime in demand to get neuron associations. Brain machine interface setup for neural implants is shown in Fig. 1.

The fundamental problem are examined and reported in following:

- The basic neural-recording implants using sigma-delta converter is designed for storage purpose, in this design the area of the chip is less but the power dissipation and delay is quite high.
- In order to reduce the area, power and delay of the chip, the basic chip is designed using sigma-delta converter with different technologies and with using switch capacitor.
- The four bit sigma-delta converter has been designed and their performance is compared, in result we observed that only power is reduced, there is not much affect in area.
- We also noticed that in designing of many CMOS circuits the use of switch capacitor is very important, which plays a major impact on designing the chip by achieving a great resolution.

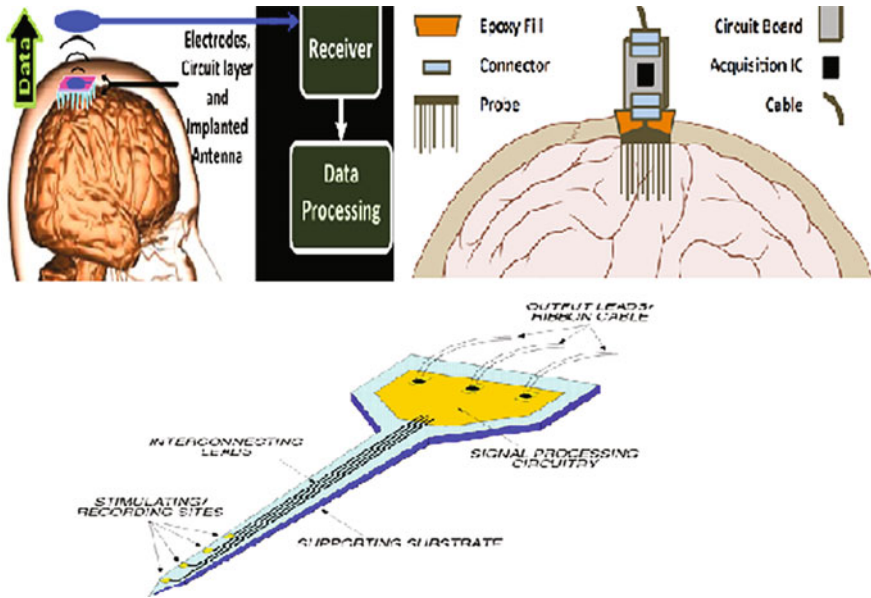


Fig. 1 Brain machine interface setup for neural implants

3 Methodology

In this part, we will examine about first neuron cell, neural action, and neural chronicle interface, second stage portray the execution of four info comparator circuit for neural account and the proposed circuit usage of the sigma-delta converter. In this postulation, proposed approach is outlined and recreated by utilizing Tanner EDA at CMOS BSIM4 innovation. To start with area clarifies the neural chronicle inserts, next segment clarifies the sigma-delta converter and last segment clarifies the proposed technique utilizing BSIM4 innovation.

The objectives of this proposed work are

- (1) To design and characterize sigma-delta modulator based on comparator circuit using Quad input.
- (2) To improved performance of sigma-delta modulator data converter.

Figure 2 shows the design flow chart of sigma-delta converter.

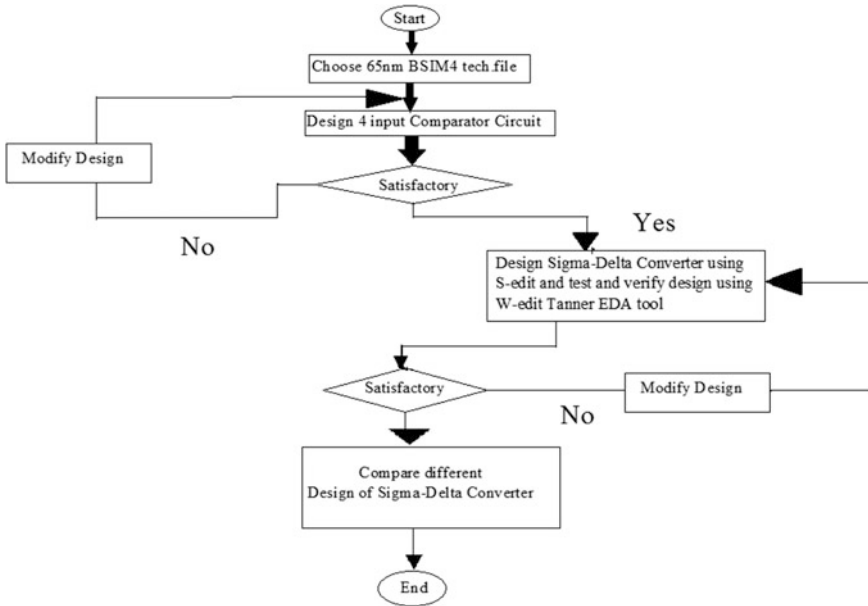


Fig. 2 Flow chart of sigma-delta converter

3.1 Design of Comparator

From Fig. 3 it can be seen that Comparator is design using 12 transistor such as M1, M2, M3, M4, M5, M6, M7, M8, M9, M10, M11, M12 and Two Inverter connected as output. It is simply differential amplifier with two input fw- and fw+. Depend on clack signal $\phi 1$ Comparator will compare the input and produce output.

From Fig. 4, it can seen that comparator circuit using four input is implemented in tanner using 65 nm technology. Figure 5 shown top view of sigma-delta converter using four-input comparator circuit used in Nerves Organization system shown in Fig. 6. It can be seen that the nerves signal read by the terminal hold little a motivation in solitude. Warm noise from the cathode and the tissue interface and natural commotion from the neighboring neurons are accessible. These signs ought to be adjusted and arranged before the information they contain can be considered. This is where the neural interface accepts a critical part (Figs. 7 and 8).

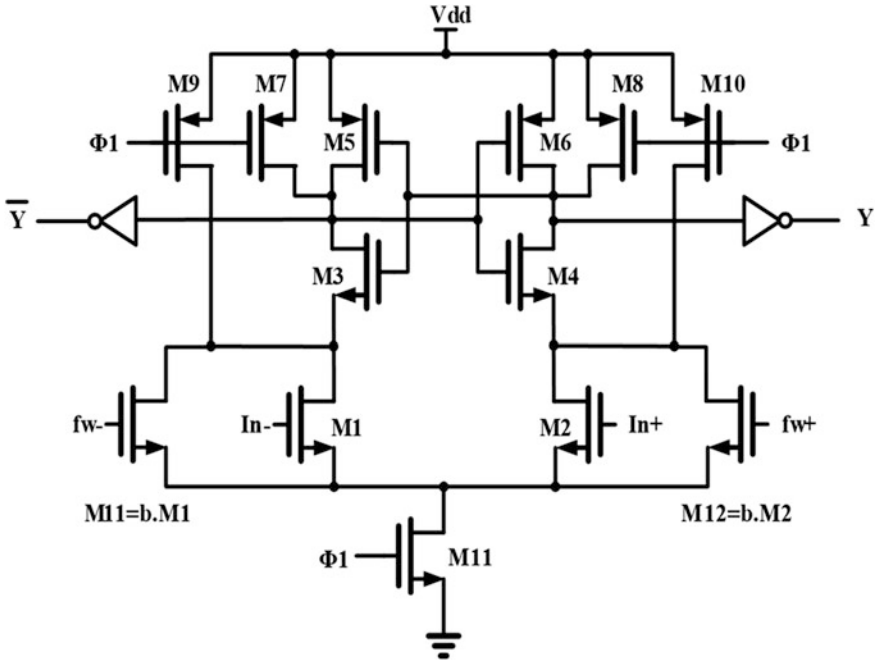


Fig. 3 Schematic of comparator circuit using four input used in sigma-delta ADC

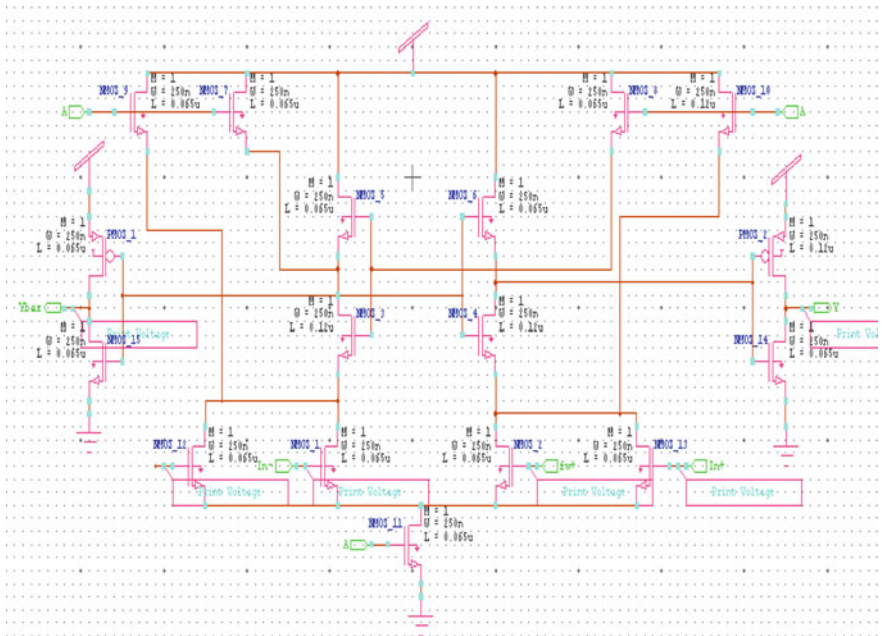


Fig. 4 Implementation of comparator circuit using quad input sigma-delta A/D

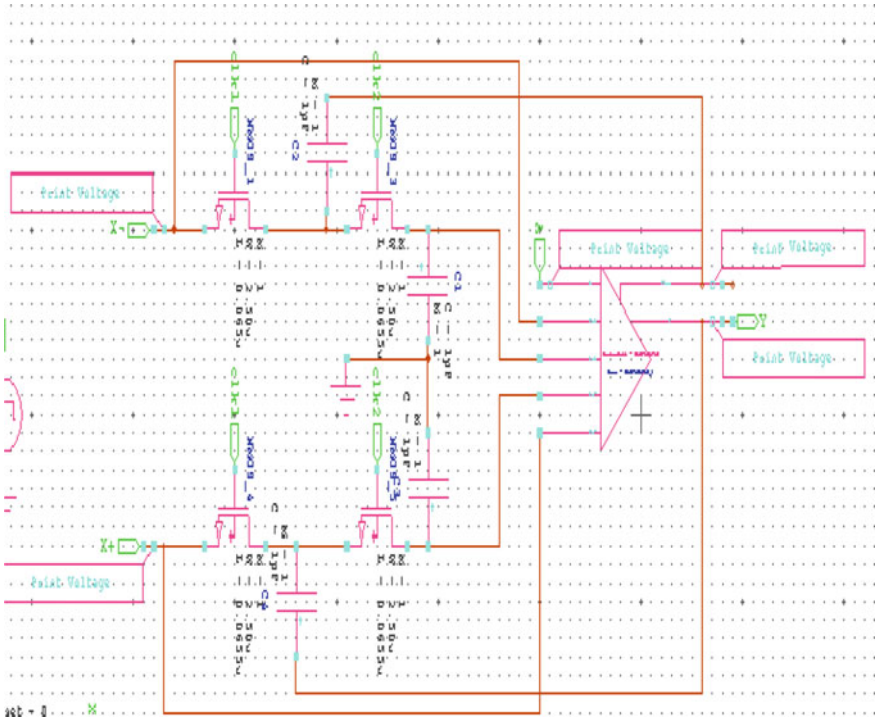


Fig. 5 Top view of sigma-delta converter using quad input comparator circuit

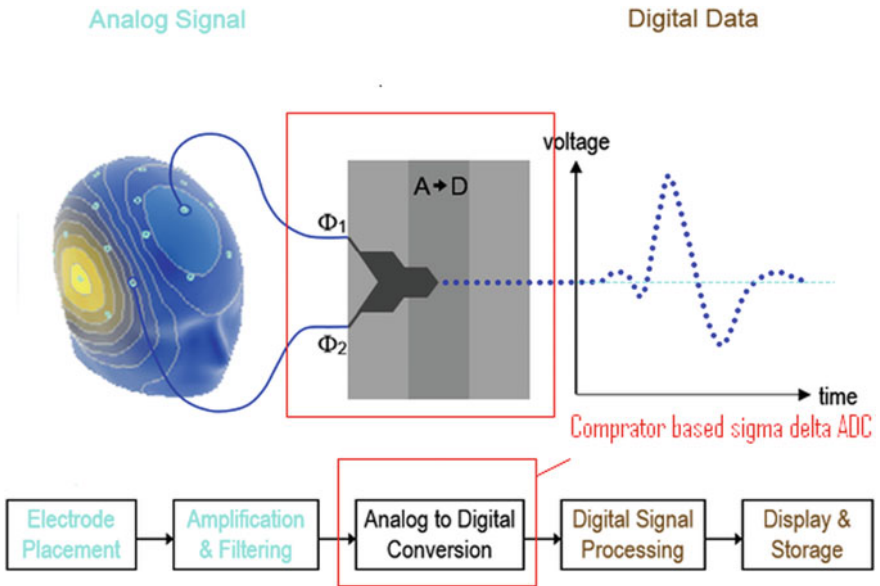


Fig. 6 Schematic view of nerves organization implementation

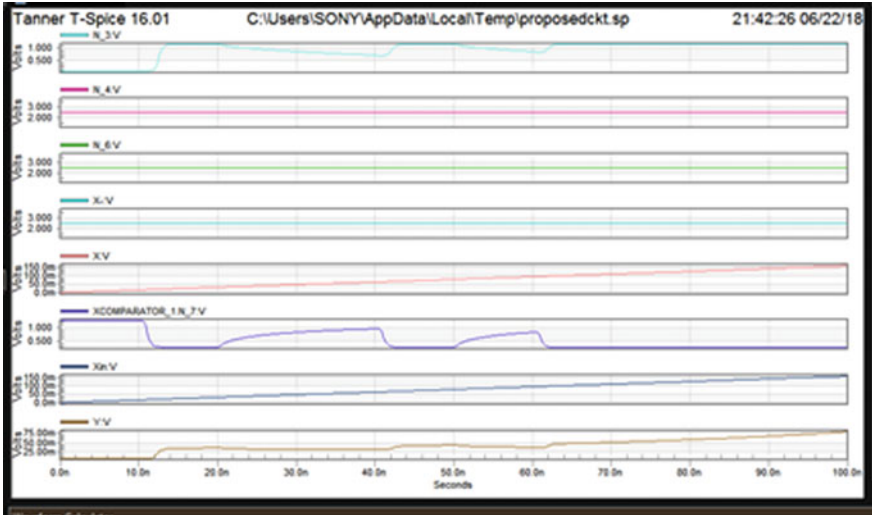


Fig. 7 Simulation result of comparator circuit using four input

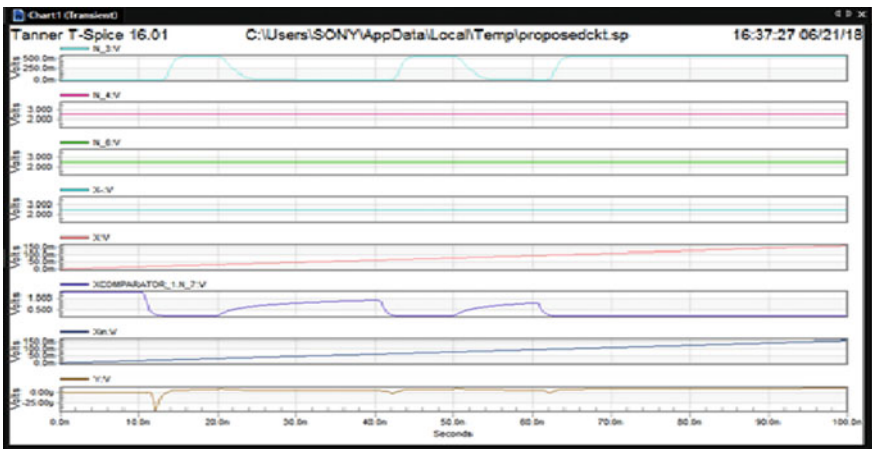


Fig. 8 Simulation results of sigma-delta converter

Table 1 Performance comparisons

Performance parameter	Ref. [3]	Ref. [1]	This work
Technology (nm)	180	180	65
Topology	DT- $\Sigma\Delta$	DT- $\Sigma\Delta$	DT- $\Sigma\Delta$
BW (KHz)	5	10	10
SNDR (dB)	105	55.2	50.2
ENOB (bits)	17.15	8.88	7.58
OSR	–	50	–
Power (μ W)	280	0.11	0.4
FOM	190	11	45

4 Results and Discussion

The proposed circuit has been made in a TSMC 65-nm CMOS circuit. In this paper the input in the above circuit is feeded into the central nervous system where the flexible power coil is implanted using implantable flexible board in which the array of electrode is fully distributed inside the cerebellum by the help of four-input comparator circuit and the output driven signal of the above comparator circuit is used to drive the sigma-delta converter which holds a great impact of driving low frequency signal with low power consumption is determined in this paper, The power utilizes in the above circuit is above $0.4 \mu\text{W}$ that include only two non-covering clock generator. In this paper, the circuit utilizes real time user interface, which holds a great impact of determining damage nerves inside the brain to cure various chronic disease, brain tumor, various trauma inside the cerebellum (Table 1).

5 Conclusion

In this paper, we presented outline of a novel sigma-delta converter that can be used for high-thickness neural record and biotelemetry applications. A FOM has been relate and figured for a couple of determined diagrams, and the implemented circuit poses great impact with relative data converters. The over all circuit inside a multichannel recording neural testing activity is implemented successfully.

Analyzing of data on recorded. To fulfill the predetermined unique fluctuate, in any case utilization of low consumption of control power. Here the style investigations, circuit execution and stream of digital signal determination whole power utilization of the modulator is $0.4 \mu\text{W}$ that compares to a FOM of 45 fJ/change step. The focused circuits details make this style a good possibility for building high accuracy neurosensory. Highlights drawback is to control the input relate to sigma-delta converter using comparator circuit as shown above. The proposed arrangement includes speed frequency range of 1 kHz of 10 kHz, and gives 7.58

bits value of precise division by using another feed forward topology, while including minimum number of additional portions and switches. A four-channel straightforward front-end has been executed in this proposed structure and made in a CMOS 65-nm innovation.

The major limitation of our proposed methodology is to operate switch capacitor and overall circuit design is quite complex and increases the chip size. Here the overall power delay of the entire circuit may be vary when operated with different technology. In future, this sigma-delta A/D design approach using comparator circuit can be designed for n number of bits, which results in low power and lower delay. It is beneficial for larger circuits, in order to reduce the delay.

References

1. Razaei M., Maghsoudlo E., Sawan M., Gosselin B.: A 110-nW in channel sigma-delta converter for large-scale neural recording implants. in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), April, pp. 5741–5744 (2016)
2. Roy, H.O.I.I., Kensall, D.W.: A three-dimensional neural recording microsystem with implantable data compression circuitry. *IEEE J. Solid-State Circuits* **40**(12), 2796–2804 (2005)
3. Perelman, Y., Ginosar, R.: An integrated system for multichannel neuronal recording with spike/LFP separation, integrated A/D conversion and threshold detection. *IEEE Trans. Biomed. Eng.* **54**(1), 130–137 (2007)
4. Pavan, S.: Excess loop delay compensation in continuous-time delta-sigma modulators. *IEEE Trans. Circuits Syst. II Express Briefs* **55**(11), 1119–1123 (2008)
5. Wise, K.D., Sodagar, A.M., Yao, Y., Gulari, M.N., Perlin, G.E., Najafi, K.: Microelectrodes, microelectronics, and implantable neural microsystems. *IEEE J. Mag.* **96**(7), 1184–1202 (2008)
6. Thurgood, B.K., Warren, D.J., Ledbetter, N.M., Clark, G.A., Harrison, R.R.: A wireless integrated circuit for 100-channel charge-balanced neural stimulation. *IEEE Trans. Biomed. Circuits Syst.* **3**(6), 405–414 (2009)
7. Gosselin, B., Ayoub, A.E., Roy, J.F., Sawan, M., Lepore, F., Chaudhuri, A., Guitton, D.: A mixed-signal multichip neural recording interface with bandwidth reduction. *IEEE Trans. Biomed. Circuits Syst.* **3**(3), 129–141 (2009)
8. Lee, S.B., Lee, H.M., Kiani, M., Jow, U.M., Ghovanloo, M.: An inductively powered scalable 32-channel wireless neural recording system-on-a-chip for neuroscience applications. *IEEE Trans. Biomed. Circuits Syst.* **4**(6), 360–371 (2010)
9. Shahrokhi, F., Abdelhalim, K., Serletis, D., Carlen, P.L., Genov, R.: The 128-channel fully differential digital integrated neural recording and stimulation interface. *IEEE Trans. Biomed. Circuits Syst.* **4**(3), 149–161 (2010)
10. Wattanapanitch, W., Sarpeshkar, R.: A low-power 32-channel digitally programmable neural recording integrated circuit. *IEEE Trans. Biomed. Circuits Syst.* **5**(6), 592–602 (2011)
11. Garcia, J., Rodriguez, S., Rusu, A.: A low-power CT incremental 3rd order Σ - Δ ADC for biosensor applications. *IEEE Trans. Circuits Syst.* **60**(1), 25–36 (2012)
12. Yin, M., Borton, D.A., Aceros, J., Patterson, W.R., Nurmikko, A.V.: A 100-channel hermetically sealed implantable device for chronic wireless neurosensing applications. *IEEE Trans. Biomed. Circuits Syst.* **7**(2), 115–128 (2013)

13. Song, H., Chen, C., Lin, M.W., Li, K., Christen, J.B.: A neural rehabilitation chip with neural recording, peak detection, spike rate counter, and biphasic neural stimulator, pp. 415–419 (2014)
14. Rezaei, M., Maghsoudloo, E., Sawan, M., Gosselin, B.: A novel multichannel analog-to-time converter based on a multiplexed sigma delta converter. In: 2015 IEEE 13th International New Circuits and Systems Conference (NEWCAS), Grenoble, pp. 1–4 (2015)
15. Mirbozorgi, S.A., Bahrami, H., Sawan, M., Rusch, L.A., Gosselin, B.: A single-chip full-duplex high speed transceiver for multi-site stimulating and recording neural implants. *IEEE Trans. Biomed. Circuits Syst.* **10**(3), 643–653 (2016)
16. Rezaei, M., Maghsoudloo, E., Bories, C., Koninck, Y., Gosselin, B.: A low-power current-reuse analog front-end for high-density neural recording implants. *IEEE Trans. Biomed. Circuits Syst.* **12**, 1–10 (2018)

Performance Comparison of Machine Learning Techniques for Epilepsy Classification and Detection in EEG Signal



Rekh Ram Janghel, Archana Verma and Yogesh Kumar Rathore

Abstract Epilepsy is a neurological affliction that in impact around 1% of humankind. Around 10% of the United States populace involvement with minimum a solitary convulsion in their activity. Epilepsy has recognized respectively tendency of the cerebrum outcomes unforeseen blasts of weird electrical action which disturbs the typical working of the mind. Since spasms by and large happen once in a while and are unforeseeable, seizure identification frameworks are proposed for seizure discovery amid long haul electroencephalography (EEG). In this exploration, we utilize DWT for highlight extraction and do correlation for all kind of Machine learning order like SVM, Nearest Neighbor Classifiers, Logistic relapse, Ensemble classifiers and so on. In this examination classification accuracy of Fine Gaussian SVM recorded as 100% and it has better as compare to other existing machine learning approaches.

Keywords Decision tree · Ensemble classifier · Electroencephalogram (EEG) · Epileptic seizure · k -Nearest neighbor · Support vector machine

1 Introduction

Epilepsy problem is a bunch of typical neurological conditions of the mind distinguished by recurrent unprovoked convulsions which are consequence of sudden bursts of abnormal electrical discharges in the brain [1–3]. According to a report by WHO [4] around 50 million people are in effected by epilepsy worldwide [5–8].

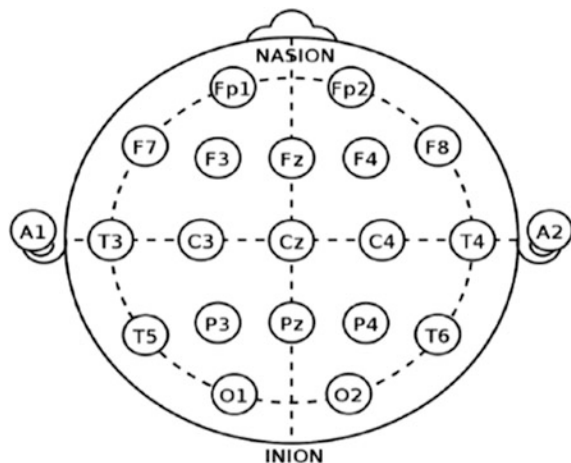
R. R. Janghel (✉) · A. Verma · Y. K. Rathore
Department of Information Technology, National Institute of Technology, Raipur, India
e-mail: rjanghel.it@nitrr.ac.in

A. Verma
e-mail: archana1694@gmail.com

Y. K. Rathore
e-mail: yogeshrathore23@gmail.com

Nearly one in every 100 persons effected by a convulsion in their lifespan [8], generally 2.4 million new instances of epilepsy accounted for consistently internationally [6, 9]. Until now, the manifestation of an epileptic is eccentric and its course of process has been less understood [10], Epileptic patients experience varying symptoms depending on the section and the expanse of the brain affected, Epileptic seizures can induce dull physical, social consequences and psychological, which include loss of consciousness, injury and abrupt death. Epilepsy is of two samples which are depending upon the degree of involvement of the brain tissue, which is, generalized seizures and, partial involves closely to the complete brain, partial seizures create in a specific area of the mind and stay constrained to that segment [8]. Electroencephalogram (EEG) distinguished by the electrophysiological technique to appreciate the problematic activity of the human brain [11]. EEG exactly gauges and registers the electrical action of the mind. Natural EEG signals have classified into many rhythms based on their frequencies, which are a band (3–4 Hz); band (4–8 Hz); band (8–13 Hz); band (13–30 Hz) [12, 13]. An EEG is particularly helpful on occasion when the cerebrum is in danger by giving a delicate sign of cerebral working. Such interim is as a rule of long time ranges, thus a broadened EEG recording required early investigations have demonstrated proof of this unusual movement to be an advantageous guide in the identification of epilepsy and cerebral tumors. These days EEG signals are utilized to get data significant to the determination, visualization, and cure of these anomalous case. EEG can be register using electrodes arranged on the head and have tiny amplitudes in the range of 20 V [14, 15, 16]. The anodes are put according to the 10–20 international system which is shown in the Fig. 1.

Fig. 1 Standardized electrode placement scheme



2 Proposed Work

SVM algorithm is used to a classification of healthy and seizures in 17 full term dataset of newborns baby with seizure by the author [17]. Using PCA with help of decomposing WPD getting an accuracy of 99% with a less computational cost for extracting features on Bonn University data set [18]. Using Novel wavelet-base and SVM for repeated seizure detection and getting sensitivity is 94.46% and specificity is 95.26% [19]. Using the Expectation–Maximization algorithm used for learning the Mixture of experts network structure getting an accuracy of 93.17% for classification seizure and non-seizure [20]. Fuzzy c-mean algorithm and MLP sub-network are used for ECG beat classification in this MIT/BIH dataset are used and enhance their performance [21, 22].

2.1 Dataset

The data elucidated by Andrzejak et al. [8] was employed for the current research. The total data set is comprised of five subsets, (denoted as Z, O, N, F, S) every each subset having 100 single-channel EEG fragments each being on 23:6 s time span sampled at 173:6 Hz. The fragments were chosen and hand-picked from regular continue multi-channel EEG recordings after ocular examination for artifacts for e.g., eye movements or muscle action etc. shown in Fig. 2.

2.2 Discrete Wavelet Transforms

DWT is best for DE noising and it think about signs and pictures as it helps indicates numerous regular happening signs and pictures with the assistance of fewer coefficients this empower a sparser introduction base scale in dwt is organized by you can get diverse scales by raising.

$$2^j, \text{ where } (j = 1, 2, 3, 4 \dots)$$

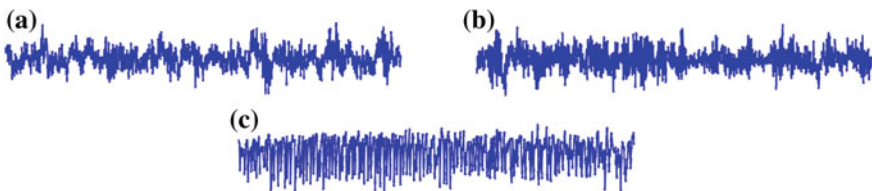


Fig. 2 Raw EEG data belonging to each of subsets a Z, b O and c S

Base scale to integer value shows in below

$$2^j m, \quad \text{where } (m = 1, 2, 3, 4, \dots)$$

The interpretation comes at whole numbers numerous speaks to in this condition this procedure is frequently alluded to as power scaling and moving this sort of inspecting wipes out repetition in coefficients. The output of the change yields the same number of coefficients from the length of the discrete wavelet transform process is proportional to contrasting a signal with discrete multi-rate filter banks conceptually.

The discrete wavelet change is a flexible flag handling device that searches many building and logical functions. One zone in which the DWT has been especially fruitful is the epileptic seizure identification since it catches transient highlights and restricts them in both time and recurrence content precisely. Different data sets are shown in Fig. 3.

After applying Discrete Wavelet transform we apply multiple methods have been put forward for epileptic seizure detection and compare their accuracy with help of confusion matrix. Dataset divided in the 10-k fold for training and testing. Multiple classifications we use for comparing data accuracy of various algorithms.

2.3 *Different Machine Learning Techniques*

A Detailed flowchart of the proposed approach is shown in Fig. 4.

2.3.1 **Support Vector Machine**

SVM is a supervised learning Model. It associated learning algorithms analyzing data used for classification and also regression analysis [23]. In this paper we use 2-class. Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, and Coarse Gaussian SVM is prediction speed is fast in binary input and easy interpretability. It uses medium memory usage shown in Fig. 4.

2.3.2 **KNN**

The k -closest neighbor's calculating (k -NN) is a non-parametric approach utilized for characterization and relapse [24]. In the two cases, the data comprises of the k nearest preparing cases in the element space shown in Fig. 5. In this paper, we use Fine KNN, Cubic KNN, Coarse KNN, Weighted KNN, Medium KNN, Cosine KNN and calculate accuracy and compare each other (Fig. 6).

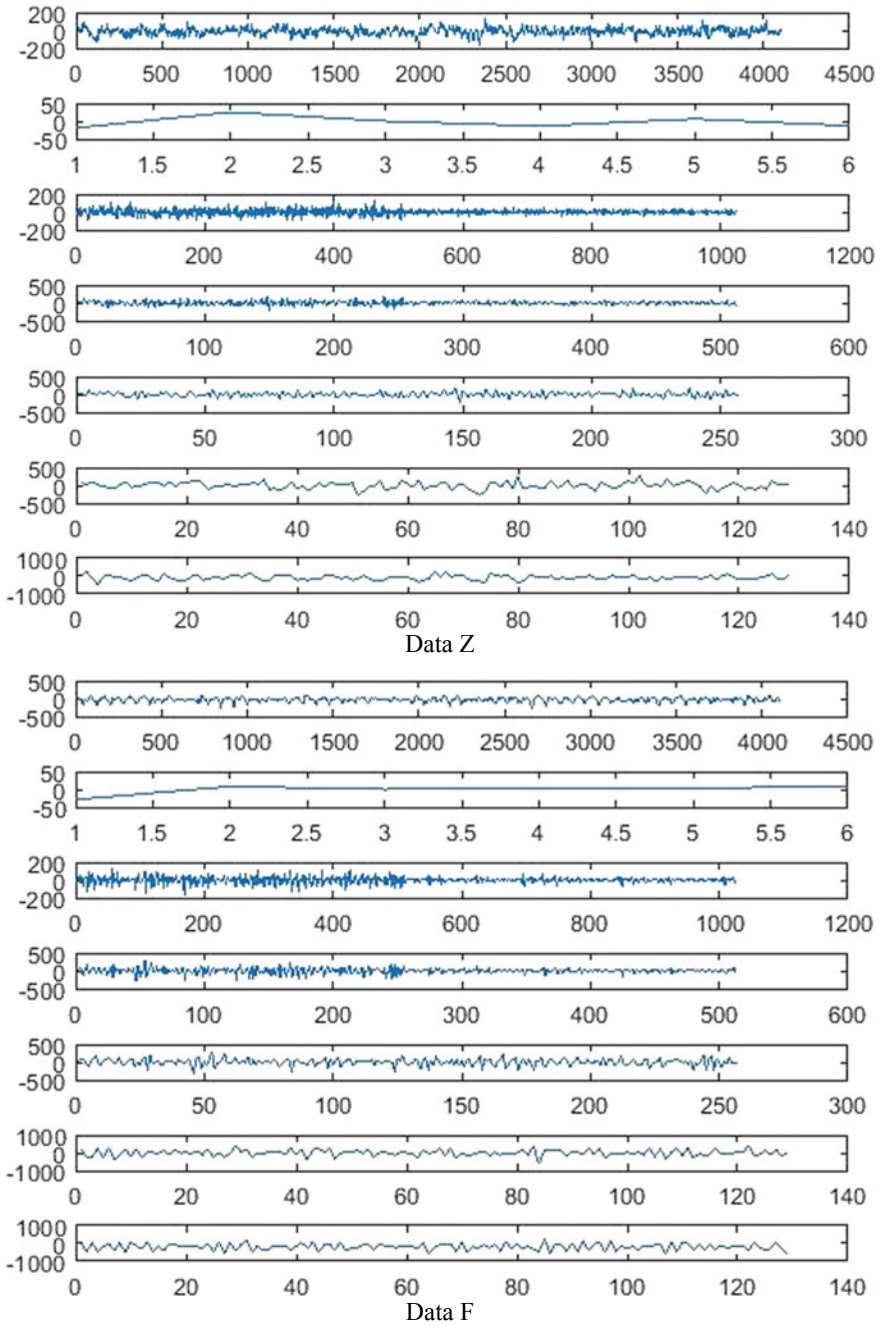


Fig. 3 Decompose dataset by using DWT

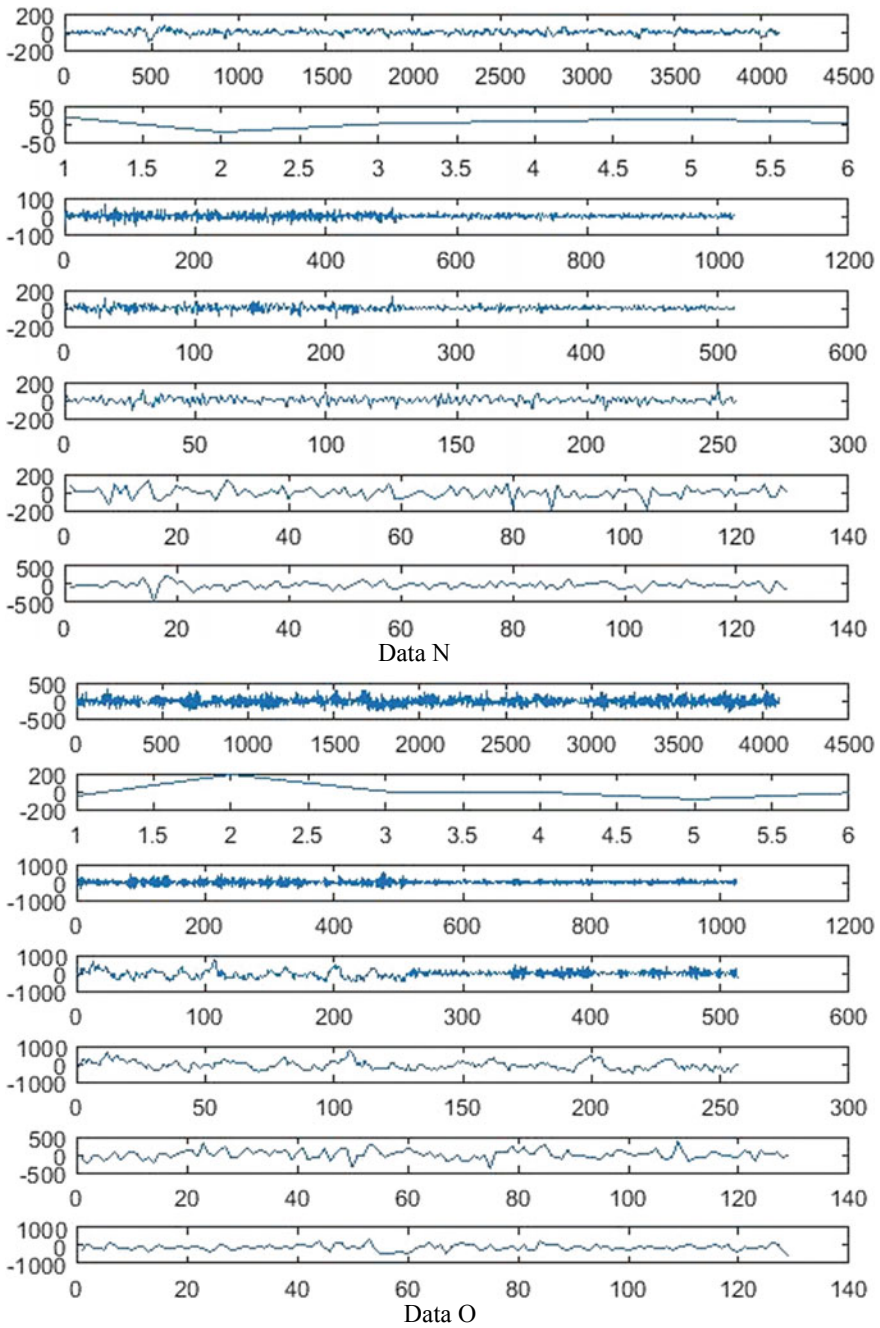


Fig. 3 (continued)

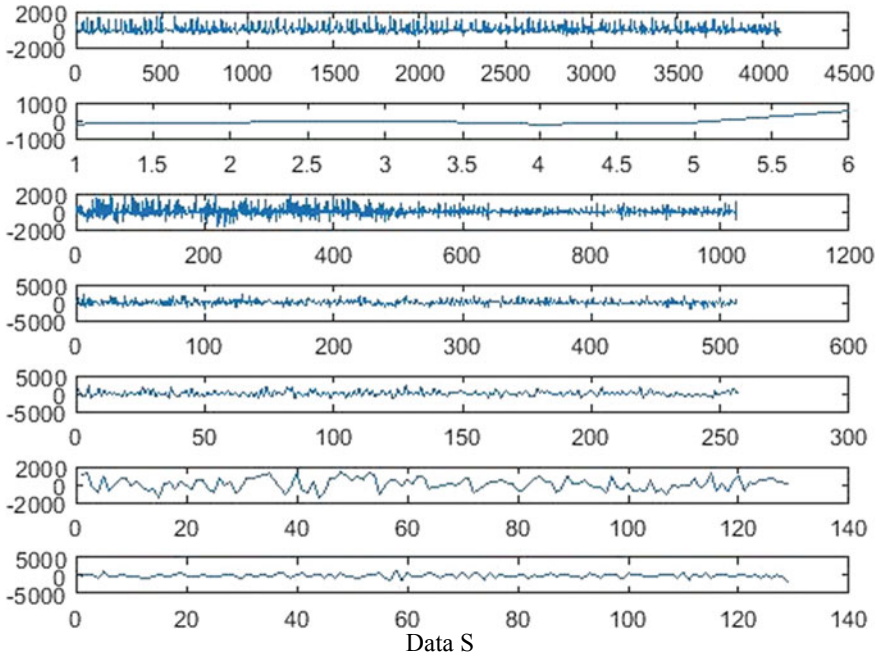


Fig. 3 (continued)

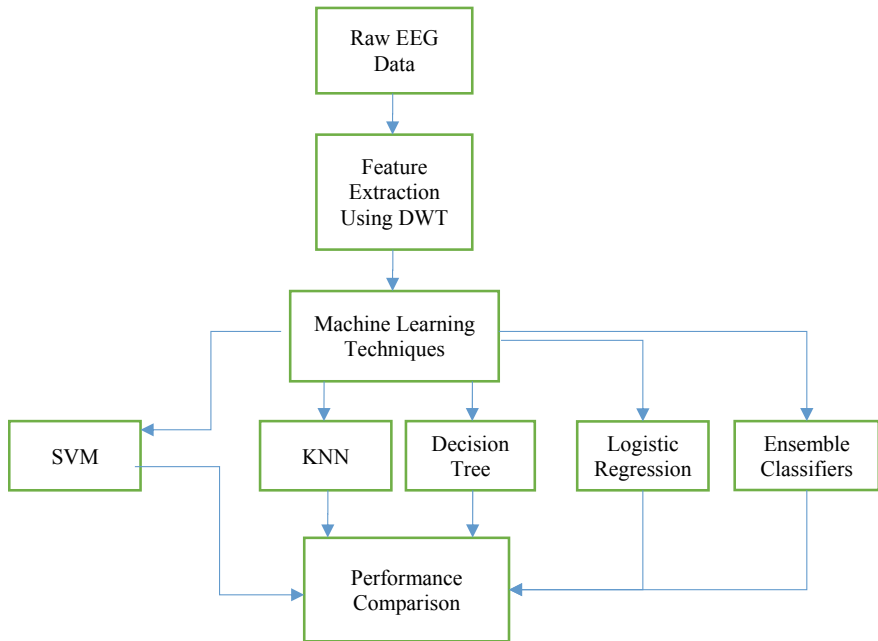


Fig. 4 Flowchart of Proposed Approach

Fig. 5 SVM Classifier [6]

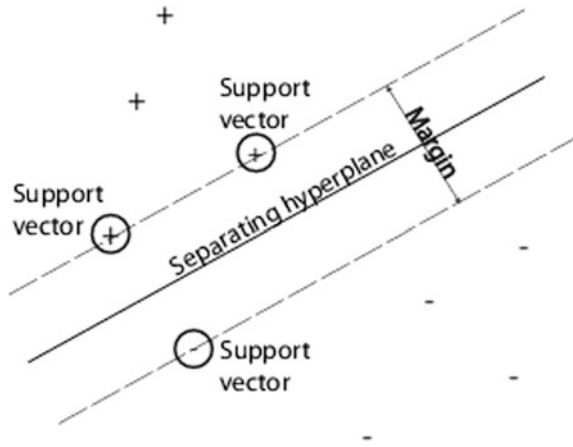
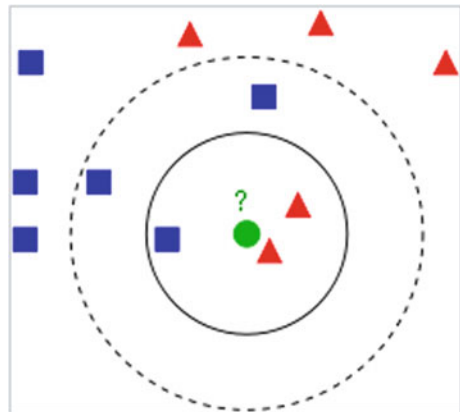


Fig. 6 KNN Classifier [10]



2.3.3 Logistic Regression

Logistic regression is a generally utilized factual displaying procedure in which the probability, $P1$, of dichotomist result occasion is identified with an arrangement of illustrative factors in the shape

$$\begin{aligned} \text{logit}(P1) &= \ln\left(\frac{p1}{1-p1}\right) \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n \\ &= \beta_0 + \sum_{i=0}^n \beta_i X_i \end{aligned} \tag{1}$$

In Eq. (1), β_0 is the intercept and $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients combine with the analytical variable X_1, X_2, \dots, X_n . These info factors are the norm of the wavelet coefficients (D3—D5 and A5) of four-channel EEG signals [25]. A dichotomous variable is confined to two qualities, for example, yes/no, on/off, survive/bite the dust or 1/0, for the most part speaking to the event or non-event of some occasion (for instance, epileptic seizure/not). The illustrative (free) factors might be consistent, dichotomous, and discrete or blend. The utilization of common straight relapse (OLR) in light of slightest squares strategy with a dichotomous result.

2.3.4 Decision Tree

Decision tree learning is a standout amongst the usually broadly utilized and practical strategies for inductive inference. It is a technique for estimation of discrete-valued function that is hearty to uproarious information and fit for learning disjunctive expression [26].

Decision tree study is a route for looking like discrete-valued function, in which the prepared function is spoken to by a decision tree. Decision trees can likewise be spoken to an asset or it at that point guidelines to enhance human intelligibility. Decision tree study is an analytic, one-advance look forward (slope climbing), non-backtracking seek through the space of all conceivable Decision trees.

2.3.5 Ensemble Classifier

It is used for multiple learning algorithms to enhance better predictive performance. It can do increments with the number of training or splitting model flexibility. It's prediction better and memory usage is average. Here, we work on Boosted Tree, Bagged Tree, Subspace Discriminant, Subspace KNN, and RUSBoost tree ensemble classifiers.

3 Experimental Results and Analysis

From the Table 1, we can see that italic value shows the highest performance of a classifier whereas bold value shows the best classification result for a particular feature. On below graphs (Figs. 7, 8, 9, 10) and tables (Table 1 and 2) comparison of classifiers on the basis of accuracy is shown.

Table 1 Results of the implementation

Feature/method	Z versus F					Z versus O					Z versus N					Z versus S				
	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
Cross tree	56.7	62.9	50.6	53.9	77.5	56.2	67.4	68.0	71.9	93.3	65.7	66.3	58.4	53.4	61.2	87.6	82.9	80.9	84.8	88.2
Fine tree	59.0	67.4	63.5	60.7	83.1	57.9	68.5	68.0	70.2	94.4	69.7	66.3	61.8	52.8	65.2	89.9	85.4	87.1	87.1	91.6
Medium tree	59.6	67.4	56.7	60.7	83.1	56.2	68.5	68.0	70.2	94.4	69.1	66.3	61.8	52.8	65.2	89.9	85.4	87.1	87.1	91.6
Linear SVM	54.5	69.7	71.9	70.2	69.1	51.1	57.3	78.7	56.7	69.1	55.1	68.0	78.1	77.5	78.7	60.7	57.3	55.6	58.4	62.4
Quadratic SVM	74.2	68.5	74.7	77.0	85.4	64.0	61.8	79.8	61.8	84.4	75.8	62.4	77.5	82.6	83.7	87.1	62.9	62.9	63.5	75.8
Cubic SVM	71.3	65.2	78.1	74.7	83.1	68.0	62.9	78.7	64.0	82.6	69.7	57.3	78.7	82.0	86.5	89.9	63.5	64.6	66.9	79.2
Fine Gaussian SVM	74.7	70.2	69.7	68.5	91.6	65.2	64.6	65.7	62.4	<i>100</i>	69.1	65.7	64.0	65.2	64.0	95.5	99.4	99.4	99.4	<i>100</i>
Medium Gaussian SVM	68.5	74.2	78.7	81.5	89.9	68.0	87.1	83.1	84.3	93.3	74.2	77.5	83.1	80.9	89.3	86.5	93.8	89.9	90.4	93.8
Coarse Gaussian SVM	55.6	55.1	56.2	56.7	68.5	57.3	57.3	66.7	56.7	68.5	56.7	59.6	62.9	61.8	66.9	56.7	63.5	64.6	68.5	70.2
Logistic regression	50.0	59.6	60.7	55.6	75.3	47.2	45.5	56.7	49.4	66.3	56.2	59.0	57.3	59.0	60.7	66.9	53.4	56.7	53.4	66.9
Fine KNN	73.0	56.2	61.2	68.0	83.1	59.6	62.9	59.6	68.0	80.3	66.3	55.1	57.3	62.9	82.0	65.2	71.3	74.7	75.8	82.0
Medium KNN	69.7	50.0	50.0	60.7	53.4	58.4	50.0	50.0	50.0	53.9	71.3	50.0	50.0	50.6	66.9	80.3	50.6	51.1	53.4	52.8
Coarse KNN	50.0	50.0	50.6	50.0	50.0	50.6	50.0	51.1	50.0	50.0	50.0	50.0	48.9	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Cosine KNN	59.0	80.9	78.7	70.8	78.7	46.6	70.2	84.8	61.8	79.2	65.2	76.4	86.0	79.8	75.3	71.9	69.2	67.4	69.1	62.9
Cubic KNN	69.1	49.4	48.3	56.2	53.4	59.6	50.0	50.0	50.0	53.4	71.3	50.0	50.0	51.1	65.7	79.2	52.8	51.1	52.8	50.6
Weighted KNN	71.9	50.0	51.1	62.9	57.9	61.8	50.0	50.0	50.0	59.6	71.9	50.0	50.0	51.7	78.1	83.1	55.6	56.2	56.7	55.6
Boosted trees	63.5	49.4	49.4	49.4	49.4	66.3	49.4	49.4	49.4	49.4	69.1	49.4	49.4	49.4	49.4	49.4	49.4	49.4	49.4	49.4
Bagged trees	72.5	79.2	77.5	75.8	96.1	63.5	83.1	83.7	79.8	98.3	71.9	77.0	80.3	78.7	81.5	92.7	96.6	97.8	97.8	98.9
Subspace discriminant	50.0	58.4	67.4	70.8	72.5	45.5	57.9	70.2	51.4	71.9	56.7	60.1	69.7	78.1	78.7	64.6	69.1	70.2	70.8	75.8
Subspace KNN	69.1	56.2	63.5	70.2	85.4	65.7	61.8	60.7	67.4	86.0	71.3	55.1	58.4	62.9	84.4	92.7	70.2	86.5	75.8	86.5
RUSBoosted tree	59.6	52.8	49.4	49.4	49.4	60.1	49.4	49.4	53.9	49.4	69.1	49.4	52.2	51.7	49.4	49.4	52.2	74.7	52.8	49.4

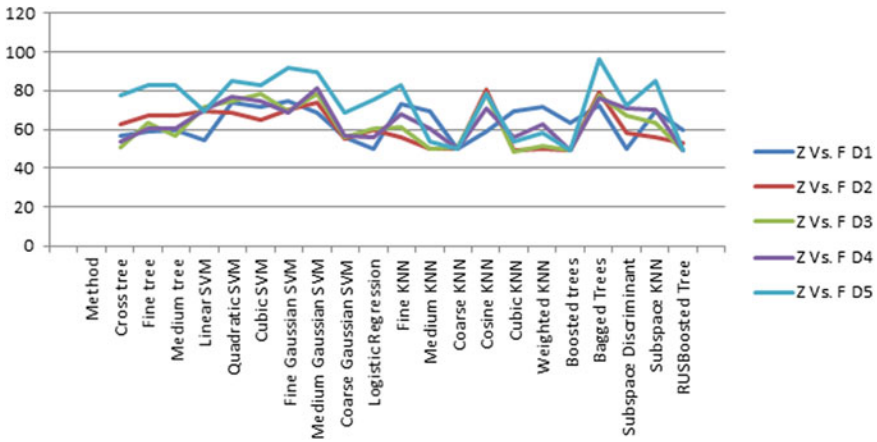


Fig. 7 The performance of different classifiers on different DWT features (Z vs. F)

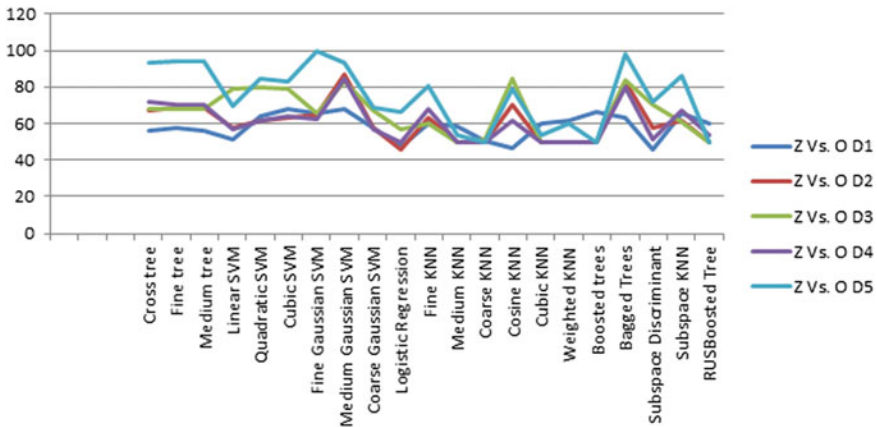


Fig. 8 The performance of different classifiers on different DWT features (Z vs. O)

4 Conclusion and Future Work

In this work Bonn University dataset which consist of five sub samples is individually subjected to all machine learning techniques i.e. SVM, KNN, Decision Tree, Logistic Regression and Ensemble classifiers. For feature extraction to increase performance we have used DWT. Our experimental result shows highest accuracy on using Fine Gaussian SVM machine learning technique on this data set.

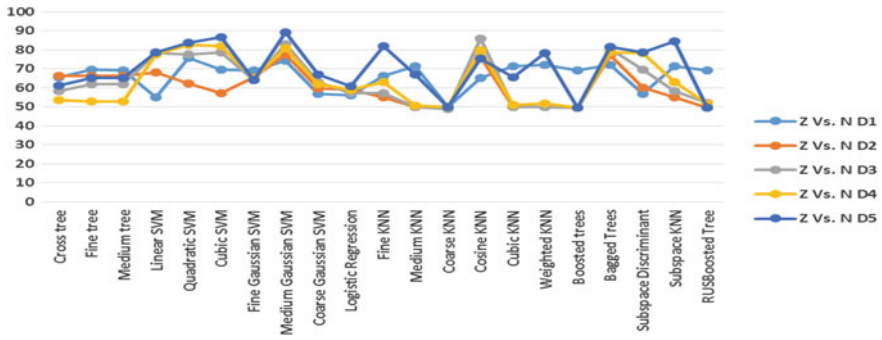


Fig. 9 The performance of different classifiers on different DWT features (Z vs. N)

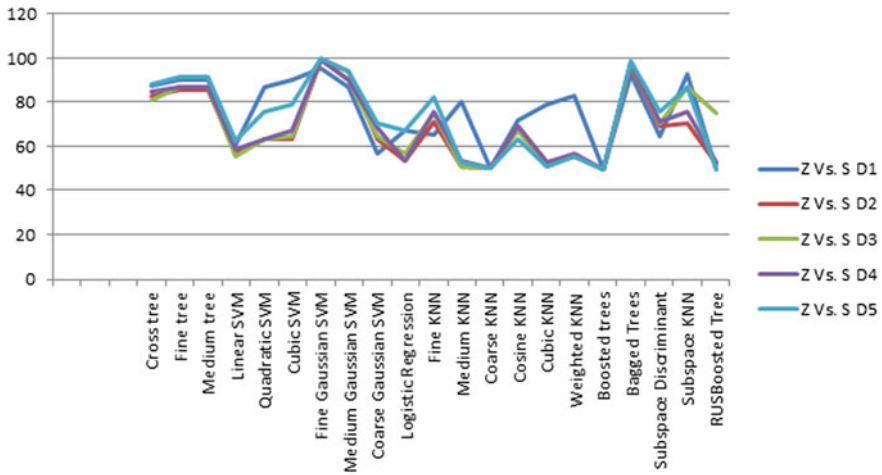


Fig. 10 The performance of different classifiers on different DWT features (Z vs. S)

Table 2 Compare research for detection of normal and epileptic classes

Author	Classifier	Accuracy (%)
Inan and Elif [27]	SVM	99.28
Varun and Ram [28]	Bandwidth features (BAM and BF M) and the LS-SVM	99.50
Abdulhamit [29]	Discrete wavelet transform (DWT), mixture of expert model	95.00
Kamal and Salih [30]	FFT-decision tree classifier	98.72
Yuan and Qi [31]	Nonlinear features + ELM	96.5
Our work	Fine Gaussian Support Vector Machine (SVM)	100.00

Here, we achieved highest accuracy with 100% for seizure classification and detection. Since the proposed framework was designed for limited number of EEG signals and machine learning modalities, it can also be applied and executed for a large data set and different machine learning or deep learning architectures.

References

1. Guo, P., Wang, J., Gao, X.Z., Tanskanen, J.M.: Epileptic EEG signal classification with marching pursuit based on harmony search method. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 283–288 (2012)
2. Guo, L., et al.: Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks. *J. Neurosci Methods* **191**(1), 101–109 (2010)
3. Selvan, S., Srinivasan, R.: Removal of ocular artifacts from EEG using an efficient neural network based adaptive filtering technique. *IEEE Signal Process. Lett.* **6**(12), 330–332 (1999)
4. WHO Report. http://www.who.int/mental_health/neurology/epilepsy/en/. Accessed Jan 2018
5. Talathi, S.S.: Deep Recurrent Neural Networks for Seizure Detection and Early Seizure Detection Systems. arXiv preprint [arXiv:1706.03283](https://arxiv.org/abs/1706.03283) (2017)
6. Acharya, U.R., et al.: Automated EEG analysis of epilepsy: a review. *Knowl. Based Syst.* **45**, 147–165 (2013)
7. Salanova, V., et al.: Long-term efficacy and safety of thalamic stimulation for drug-resistant partial epilepsy. *Neurology* **84**(10), 1017–1025 (2015)
8. Cook, M.J., et al.: Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *Lancet Neurol* **12**(6), 563–571 (2013)
9. Iasemidis, L.D., et al.: Adaptive epileptic seizure prediction system. *IEEE Trans. Biomed. Eng.* **50**(5), 616–627 (2003)
10. Orosco, L., Agustina, G.C., Eric, L.: A survey of performance and techniques for automatic epilepsy detection. *J. Med. Biol. Eng.* **33**(6), 526–537 (2013)
11. Guo, L., et al.: Automatic feature extraction using genetic programming: an application to epileptic EEG classification. *Expert Syst. Appl.* **38**(8), 10425–10436 (2011)
12. Selvan, S., Srinivasan, R.: Removal of ocular artifacts from EEG using an efficient neural network based adaptive filtering technique. *IEEE Signal Process. Lett.* **6**(12), 330–332 (1999)
13. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv preprint [arXiv:1409-1259](https://arxiv.org/abs/1409.1259)
14. Zhou, W., Gotman, J.: Removal of EMG and ECG artifacts from EEG based on wavelet transform and ICA. In: 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEMBS'04, vol. 1, pp. 392–395 (2004)
15. Parvez, M.Z., Paul, M.: Epileptic seizure prediction by exploiting spatiotemporal relationship of EEG signals using phase correlation. *IEEE Trans. Neural Syst. Rehabil. Eng.* **24**(1), 158–168 (2016)
16. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv preprint [arXiv:1409-1259](https://arxiv.org/abs/1409.1259) (2014)
17. Ammar, S., Senouci, M.: Seizure detection with single-channel EEG using Extreme Learning Machine. In: 2016 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), pp. 776–779 (2016)
18. Temko, A., et al.: EEG-based neonatal seizure detection with support vector machines. *Clin. Neurophysiol.* **122**(3), 464–473 (2011)
19. Acharya, U.R., et al.: Use of principal component analysis for automatic classification of epileptic EEG activities in wavelet framework. *Expert Syst. Appl.* **39**(10), 9072–9078 (2012)

20. Liu, Y., et al.: Automatic seizure detection using wavelet transform and SVM in long-term intracranial EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **20**(6), 749–755 (2012)
21. Übeyli, ElifDerya: Wavelet/mixture of experts network structure for EEG signals classification. *Expert Syst. Appl.* **34**(3), 1954–1962 (2008)
22. Engin, Mehmet: ECG beat classification using neuro-fuzzy network. *Pattern Recognit. Lett.* **25**(15), 1715–1722 (2004)
23. Garrett, D., et al.: Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **11**(2), 141–144 (2003)
24. Yazdani, A., Ebrahimi, T., Hoffmann U.: Classification of EEG signals using Dempster Shafer theory and a k-nearest neighbor classifier. In: 4th International IEEE/EMBS Conference on Neural Engineering, NER'09, IEEE, (2009)
25. Subasi, Abdulhamit, Ercelebi, Ergun: Classification of EEG signals using neural network and logistic regression. *Comput. Methods Programs Biomed.* **78**(2), 87–99 (2005)
26. Polat, Kemal, Güneş, Salih: Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Appl. Math. Comput.* **187**(2), 1017–1026 (2007)
27. Panda, R., et al.: Classification of EEG signal using wavelet transform and support vector machine for epileptic seizure diction. In: 2010 International Conference on Systems in Medicine and Biology (ICSMB), IEEE (2010)
28. Bajaj, V., Pachori, R.B.: Classification of seizure and nonseizure EEG signals using empirical mode decomposition. *IEEE Trans. Inf. Technol. Biomed.* **16**(6), 1135–1142 (2012)
29. Subasi, Abdulhamit: EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Syst. Appl.* **32**(4), 1084–1093 (2007)
30. Polat, Kemal, Güneş, Salih: Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Appl. Math. Comput.* **187**(2), 1017–1026 (2007)
31. Yuan, Q., et al.: Epileptic EEG classification based on extreme learning machine and nonlinear features. *Epilepsy Res.* **96**(1–2), 29–38 (2011)

Novel Approach for Plant Disease Detection Based on Textural Feature Analysis



Varinderjit Kaur and Ashish Oberoi

Abstract The image processing is the technique which can propose the information stored in the form of pixels. The plant disease detection is the technique which can detect the disease from the leaf. The plant disease detection algorithms have various steps like preprocessing, feature extraction, segmentation, and classification. The KNN classifier technique is applied which can classify input data into certain classes. The performance of KNN classifier is compared with the existing techniques and it is analyzed that KNN classifier has high accuracy, less fault detection as compared to other techniques. This paper presents methods that use digital image processing techniques to detect, quantify, and classify plant diseases from digital images in the visible spectrum. In plant leaf classification leaf is classified based on its different morphological features. Some of the classification techniques used are neural network, genetic algorithm, support vector machine, and principal component analysis. In this paper results are compared between KNN classifier and SVM classifier.

Keywords GLCM · KNN · K-means · WDDIP-KNN · Plant disease detection

1 Introduction

Agriculture helps to create wealth, improvise farmer's livelihood and decrease the rate of hunger and poverty especially in the remote areas. Agriculture plays an important role to improve the food and life security. But nowadays farming is suffering from various diseases, parasites, shortage of irrigation facilities and lack of best-suited fertilizers. Diseases are not the only reason for the downfall of farm marketing but unhygienic and improper sanitation at farms are also the major issues

V. Kaur · A. Oberoi (✉)

Department of CSE, RIMT University, Mandi Gobindgarh, India
e-mail: a_oberoi01@yahoo.co.in

V. Kaur

e-mail: Vari006rupi@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_30

for the diseases to human. Various methods are used for the detection of plant diseases in which a vast image processing techniques are used. The plant diseases can be analyzed by the texture, color, and structure of the leaf [1]. The information regarding the diseases and species of any plant or crop is given by the leaves, by studying the structure of a leaf researcher can generalize about the diseases of any crop. Since there are huge impacts of plant disease detection on the quality and quantity of the agricultural products, there is a need to study this field and present improvements such that the issues that are being faced can be eliminated. Agriculture is the field on which around 70% of the population of India relies. Appropriate fruit and vegetable crops to be grown can be chosen by the farmers from diverse choices available. However, the technicality level with which these crops are cultivated such that optimum yield and quality product can be produced, is very high. Technical support is highly useful in fulfilling these requirements. There is a need to monitor the fruit crops very closely so that they can be managed efficiently such that no diseases affect their productivity. Depending upon the high resolution multispectral and stereo images, the leaf diseases are to be detected and classified automatically and to do so several researchers have proposed various techniques. Providing a direct economical optimization of the agricultural products is not the only requirement here [2]. There is a need to ensure that the living beings or the environment is not harmed due to them. The crop production particularly needs to reduce the level at which the water, soil, and food resources are being contaminated due to the pesticides. Further, the proposed methods also need to have higher speed and accuracy to provide efficient results as per the technology advancements. The plant disease detection process includes some basic steps which are

- a. *Input image* Capturing a sample image using a digital camera is the initial step of this process. Further, the features are extracted from the image to perform operations. The important features are stored initially within the database and the further processing will be performed in the next steps.
- b. *Image database* In the next step, an image database is generated that includes all the images which will further be used to train and test the extracted data. The application is clearly the most important factor on which the image database is constructed. The efficiency of the classifier is responsible to decide the robustness of an algorithm and it completely relies on the image database.
- c. *Image preprocessing* The operations that are performed on the images at the lowest level of abstraction are collectively known as image preprocessing. The image data is improved and the distortions that are unnecessarily present are suppressed within this process [3]. The processing and task analysis are performed with the help of image features which can be improved through this process. The information content on the image is not increased due to this process. Sufficient amount of redundancy is provided in the images with the help of this method. The brightness value, of neighboring pixels that belong to

the same real object, is also same. In the form of an average value of the neighboring pixels, it is possible to restore a pixel that has been distorted within an image. The image that has been stored within the image database is captured with the help of image preprocessing techniques as well.

- d. *Feature extraction* The features that help in determining the meaning of a given sample are used for the identification and extraction of features from it. Color, shape and texture features are mainly included as image features in case of image processing.
- e. *Recognition and classification* Training and classification are the two broader phases in which the complete recognition process is divided. The numerical characteristics of several image features are analyzed using image classification. Further, the data is categorized in an organized way through this process [4]. Training and testing are the two phases that are used by classification algorithms. The isolation of properties of basic image features is done within the initial training process. Further, a training class is introduced with the help of these isolated features, which has a unique description for each classification category. The image features are classified within the further testing phase with the help of these feature-space divisions created here.

A nonlinear classifier that is used to solve several pattern recognition issues is known as Support Vector Machine (SVM). For achieving better performance of classification, mapping of nonlinear input data is done to the linearly separated data within certain high dimensional space in SVM. The marginal distance present among various classes is increased through SVM [5]. Various kernels are used to divide the classes. Only two classes are used by SVM and to partition them, a hyper plane is defined by it. To do so, the margin that exists between the hyper plane and the two classes is maximized. Support vectors are known as the samples that are nearest to the margin and were chosen such that the hyper place could be determined.

2 Literature Review

Mattihalli et al. [9] proposed a novel approach for detecting the diseases from leaves. Here, the leaf images are taken such that few important features can be extracted which can further be used in this proposed work. A device named as Beagle bone black is used in this work which also consists of a web camera that helps in identifying diseases from the leaf samples. The device includes a database in which there are few pre-stored leaves images. This paper compares the pre-stored images with the images captures from the Beagle bone black device. GSM is used to transmit the information related to detect the diseases and operate the valves. As per the simulation results, it is seen that the proposed approach provides very low cost, user friendly and highly efficient results.

Tichkule and Gawali [10] reviewed various techniques through which plants diseases can be detected. Image processing technique is used for the detection and identification of diseases caused by bacteria, fungi, virus, and excessive use of insecticides. Therefore, this method is used to classify the diseases caused in agricultural field. The authors accurately detected and classified diseases on various plants using all above techniques. K-Means Clustering method is used to detect infected plants and Neural Networks provides the accuracy in detection and classification of diseases. So, these methods have ability to be used in Agrobot system.

Tlhobogang and Wannous [11] proposed the study that investigated the problems related to unavailability, irrelevant, and less accurate farming information. The main objective was to deal with infected plants and to diagnose them. Image analysis, convolutional neural networks and the knowledge of machine learning offer a stable and movable solution. A Science Research Methodology was used in framing the prototype. Proper and systematic studies on farmers' agricultural techniques and effect of recently on-going projects are used to understand the benefits of mobile based agricultural services. The contribution of ICT in agricultural fields, collecting, storing, and disseminating the development are required by the farmers to increase their market value.

Gandhi et al. [12] presents an image-related classification which identifies the plant diseases. As, there are many datasets used in other countries but none of them are connected to India. So, there is a need to develop a local dataset for the benefits of Indian farmers. Generative Adversarial Networks (GANs) are used to limit the number of local available images. A Convolutional Neural Networks (CNN) is employed on the smart mobile applications. This application can be easily installed in any smart phone and the farmer needs to move through the field and capture the images. This could capture the several images which can be further send to the server through which the model runs and gives the classification.

Khan et al. [13] presented multilevel segmentation method in which initial segmentation is performed using expectation maximization algorithm. In this method, there is a very few chances of information loss. Finally, a region is taken out from the quantized image by empowering the binary partitioned tree which is further used in principle Eigen vector. Post processing methods are used to eradicate the useless fragments and to resolve the image labeling problem. When the salient region of the image belongs to single class this technique is mostly used. The new cascaded design in primary color segmentation with the confirmation of infected regions extractions gives the experimental analysis.

3 Research Methodology

This research work is based on the detection of plant disease. To detect plant disease with the proposed technique following steps are followed

1. *Preprocessing* The image is taken as input on which disease need to detect. The image is converted to the grayscale for the feature extraction.
2. *Feature extraction* In the second phase the algorithm of GLCM is applied which can extract features of the input image and store in the database. A tabular description which shows the number of times various combinations of pixel brightness values occur within an image is known as Gray-Level Co-occurrence Matrix. The locations of pixel that has alike gray level values are presented by GLCM [7]. The pairs of gray level values are provided as input from GLCM calculation units. When the original and predictive images are compared, the deviation present in them is presented here. The relationship among two neighboring pixels is considered within GLCM. The initial pixel is called the reference pixel whereas the neighbor pixel is known to be the second one. A two dimensional array in which a set of possible image values are presented in rows and columns is known as a co-occurrence matrix.
3. *Segmentation* The k-mean segmentation algorithm is applied which can segment input image into certain parts. K-means is the data clustering algorithm in which the numbers of clusters within the data are pre-specified is known as k-means algorithm. A trial-and-error method in which appropriate numbers of clusters for particular data set are identified has made it difficult to define correct clustering method [8]. Through the selection of K value, the performance of a clustering algorithm is affected. Thus, there is a need to adopt a set of values instead of using a single predefined K. For reflecting the special properties of data sets, the consideration of large number of values is important. Further, in comparison to the number of objects present in the data sets, there need to be less number of selected values.
4. *Classification* The classification detects the name of the disease with which plant is affected. The system is training with the number of images and test set is given as input after extracting image features. The Proposed Model (WDDIP-KNN) classifier is applied to detect the disease name and flow chart is blow. A classifier, that performs classification through the recognition of neighbors that are closest to the query examples, is known as k-Nearest Neighbor. The class of query is determined with the help of these identified neighbors. In this algorithm, on the basis of the least distance present between the given point and other points is calculated which is then used to perform classification that helps in determining the class within which a particular point belongs [6]. There is no training process involved within this classifier. Since, the robustness to noisy data for this classifier is zero it is not applied within very large number of training examples. The calculation of Euclidean distance between the test and training samples is done which is further used within plant leaf classification. Thus, similar measures are identified in this manner which further helps in identifying the class for test samples as well (Fig. 1).

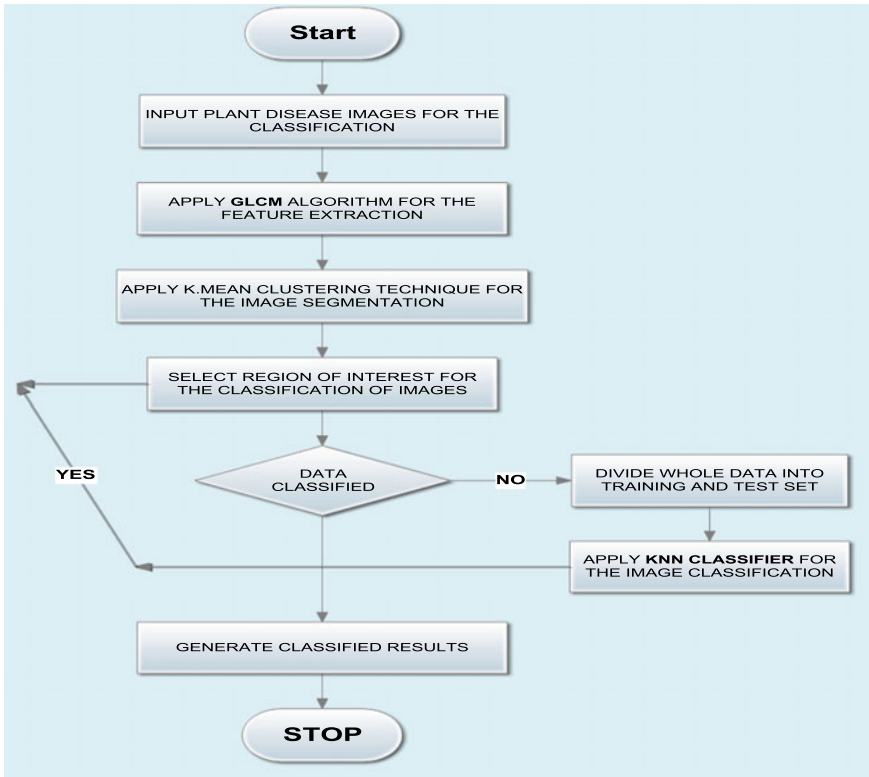


Fig. 1 Proposed model (WDDIP-KNN) control flow

4 Results and Discussion

The technique is proposed for the plant disease detection based on feature detection and classification. The GLCM algorithm is applied for the feature extraction. The k-mean clustering is applied for the image segmentation and KNN algorithm is applied for the disease classification. The Data set of about 25 images is taken as input to prepare the training set. The results are analyzed in terms of certain parameters which are described below.

As shown in Fig. 2, the fault of the proposed and existing algorithm is compared for the performance analysis. It is analyzed that proposed KNN technique has less faults as compared to existing technique SVM (Table 1).

$$\text{Fault Detection Rate} = 100 - \text{accuracy}$$

Fig. 2 Fault detection comparison

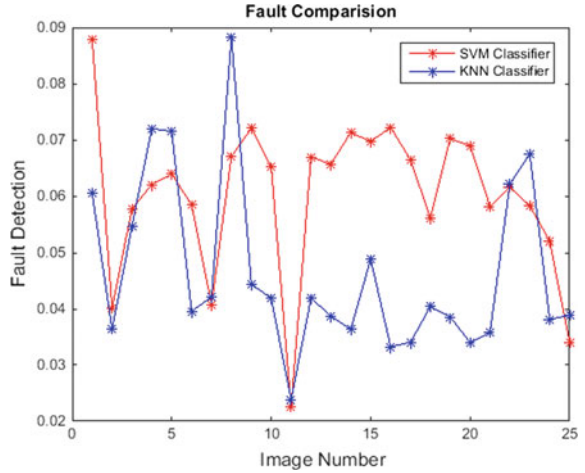
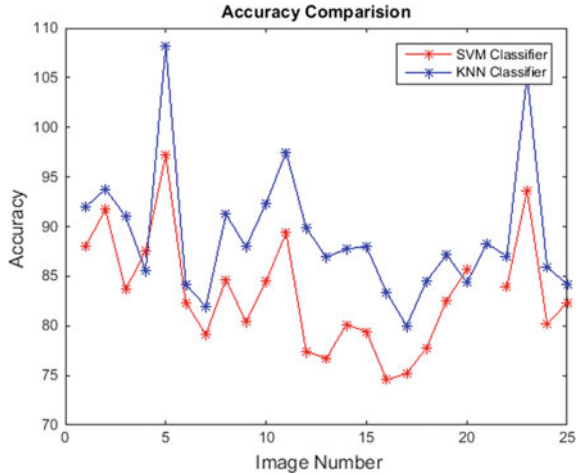


Table 1 Fault detection rate comparison

Image no	Existing technique (SVM classifier)	Proposed technique (WDDIP-KNN)
1	0.087951	0.060472
2	0.040103	0.036336
3	0.057707	0.054626
4	0.062027	0.07203
5	0.063903	0.071472
6	0.058424	0.039559
7	0.040812	0.042147
8	0.067034	0.088288
9	0.072044	0.044439
10	0.065199	0.041873
11	0.022471	0.023817
12	0.066916	0.041897
13	0.065681	0.038756
14	0.071313	0.036336
15	0.069679	0.048892
16	0.072199	0.033249
17	0.066446	0.034057
18	0.056053	0.040433
19	0.070209	0.038563
20	0.068922	0.033993
21	0.058104	0.035843
22	0.061631	0.062198
23	0.058188	0.067419
24	0.051977	0.038162
25	0.033993	0.038839

Fig. 3 Accuracy comparison



As shown in Fig. 3, the accuracy of the proposed KNN technique is compared with the existing SVM technique. It is analyzed that accuracy of the proposed algorithm is high as compared to existing algorithm.

As shown in Fig. 4, the accuracy of the SVM technique is 80.02954 and fault rate is 0.84811 (Fig. 5).

As shown in Fig. 4, the accuracy of the KNN Classifier is 92.25974 and fault rate is 0.94918 (Table 2).

$$\text{Accuracy} = \frac{\text{Total no of points classified}}{\text{Total no of points}} * 100$$

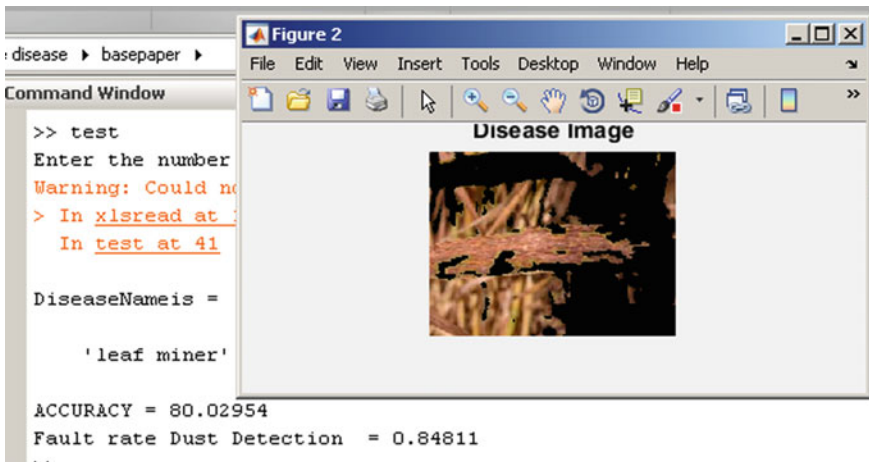


Fig. 4 Disease Name, accuracy and fault rate for one image by SVM classifier

```
ans =
Affected Area is: 15.0101%
Warning: Could not start E
> In xlsread at 187
   In test at 67
Warning: KNNCLASSIFY will
using ClassificationKNN.pr
> In knnclassify at 80
   In test at 69

CLASS =
'Yellow mosaic virus'

ACCURACY = 92.25974
Fault rate Dust Detection = 0.94918
```

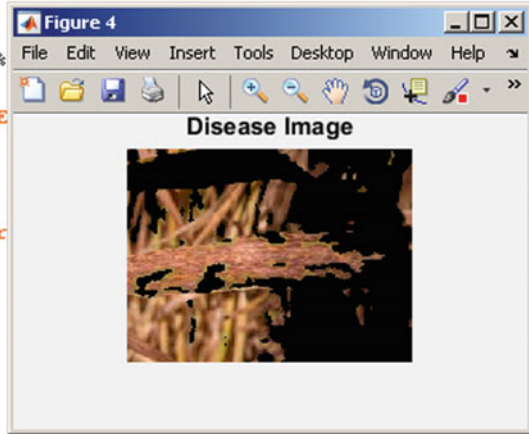


Fig. 5 Disease name, accuracy and fault rate for one image by KNN classifier

Table 2 Accuracy comparison

Image no	Existing technique (SVM classifier)	Proposed technique (WDDIP-KNN)
1	91.97587	88.0339
2	93.75906	91.69698
3	91.02909	83.65844
4	85.55519	87.54323
5	108.1948	97.18758
6	84.1623	82.33363
7	81.95833	79.13637
8	91.23152	84.67711
9	87.96284	80.39575
10	92.37004	84.52256
11	97.47806	89.3719
12	89.84999	77.41904
13	86.92375	76.71102
14	87.75028	80.12171
15	87.96963	79.43393
16	83.36096	74.55252
17	80.01366	75.19481
18	84.559	77.79646
19	87.15097	82.59238
20	84.35514	85.65128
21	88.27116	85.46842

(continued)

Table 2 (continued)

Image no	Existing technique (SVM classifier)	Proposed technique (WDDIP-KNN)
22	87.02477	83.97361
23	105.2089	93.60827
24	93.37007	95.35002
25	90.85422	92.46523

5 Conclusion

In this work, it is concluded that plant disease detection is the technique (WDDIP-KNN) which is applied to detect disease from the plants. The technique of plant disease detection has various steps like feature extraction, segmentation, and classification. In this paper, novel technique(WDDIP-KNN) is proposed based on the GLCM algorithm for the feature extraction, K-mean segmentation is applied for the image segmentation and KNN classification technique is applied for the disease classification. The performance of proposed algorithm is compared with the existing algorithm in terms of accuracy and fault rate. It is analyzed that KNN classifier has high accuracy and less fault rate as compared to existing technique.

References

1. Camargo, A., Smith, J.S.: An image-processing based algorithm to automatically identify plant disease visual symptoms. *Bio Syst. Eng.* **102**, 9–21 (2008)
2. Camargo, A., Smith, J.S.: Image processing for pattern classification for the identification of dis-ease causing agents in plants. *Comput. Electron. Agric.* **66**, 121–125 (2009)
3. Guru, D.S., Mallikarjuna, P.B., Manjunath, S.: Segmentation and classification of tobacco seedling diseases. In: *Proceedings of the Fourth Annual ACM Bangalore Conference* (2011)
4. Zhao, Y.X., Wang, K.R., Bai, Z.Y., Li, S.K., Xie, R.Z., Gao, S.J.: Research of maize leaf disease identifying models based image recognition. In: *Crop Modeling and Decision Support*, pp. 317–324. Tsinghua University Press, Beijing (2009)
5. Fury, T.S., Cristianini, N., Duffy, N.: Support vector machine (SVM) classification and validation of cancer tissue samples using microarray expression data. *Proc. BioInform.* **16** (10), 906–914 (2000)
6. Al-Hiaryy, H., Bani Yas Ahmad, S., Reyalat, M., Ahmed Braik, M., AL Rahamnehiah, Z.: Fast and accurate detection and classification of plant diseases. *Int. J. Comput. Appl.* **17**(1), 31–38 (2011)
7. Mohanaiah, P., Sathyanarayana, P., GuruKumar, L.: Image texture feature extraction using GLCM approach. *Int. J. Sci. Res. Publ.* **3**(5), 1 (2013)
8. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 881–892 (2002)
9. Mattihalli, C., Gedefaye, E., Endalamaw, F., Necho, A.: Real time automation of agriculture land, by automatically detecting plant leaf diseases and auto medicine. In: *32nd International Conference on Advanced Information Networking and Applications Workshops* (2018)

10. Tichkule, S.K., Gawali, D.H.: Plant diseases detection using image processing techniques. In: Online International Conference on Green Engineering and Technologies (IC-GET) (2016)
11. Thobogang, B., Wannous, M.: Design of plant disease detection system: a transfer learning approach work in progress. IEEE (2018)
12. Gandhi, R., Nimbalkar, S., Yelamanchili, N., Ponkshe, S.: Plant disease detection using CNNs and GANs as an augmentative approach. IEEE (2018)
13. Khan, Z.U., Akra, T., Naqvi, S.R., Haider, S.A., Kamran, M., Muhammad, N.: Automatic detection of plant diseases; utilizing an unsupervised cascaded design. IEEE (2018)

Novel Approach for Brain Tumor Detection Based on Naïve Bayes Classification



Gurkarandesh Kaur and Ashish Oberoi

Abstract The brain tumor detection is the approach which can detect the tumor portion from the MRI image. To detect tumor from the image various techniques has been proposed in the previous times. The technique which is adapted in research work is based upon morphological scanning, clustering, and Naïve Bayes classification. The morphological scanning will scan the input image and clustering will cluster similar and dissimilar patches from image then Naïve Bayes classifier spot the tumor portion from magnetic resonance imaging. The advance algorithm is implemented in MATLAB and results are analyzed in terms of PSNR, MSE accuracy, and fault detection and also calculate overlapping area with dice coef. The proposed method has been tested on data set with more than 25 slide scanned images. This proposed method achieved accuracy with 86% best cell detection.

Keywords Brain tumor · Clustering · MRI · Morphological scanning · Naïve Bayes NBC-BTD model

1 Introduction

Several lives have been affected because of a common brain disease known as brain tumor. The patients suffering from this disease have not survived in most of the cases. For fighting this disease, several techniques have been proposed such that the knowledge related to medicine can be expanded and one can understand calculations in a better manner such that the tumor can be detected. Due to the high complexity of brain images and the fact that only expert physicians can analyze the tumors, brain tumor detection is a challenging task within medical image processing [1]. To detect brain tumor from various images, the two most common tests that are

G. Kaur · A. Oberoi (✉)

Department of CSE, R.I.M.T University, Mandi Gobindgarh, India
e-mail: a_oberoi01@yahoo.co.in

G. Kaur

e-mail: deesh.mangat@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_31

451

applied are Magnetic Resonance Image (MRI) and Computer Tomography (CT) scan of brain. Further, to perform various treatments, the location of tumor is also identified through this technique. To heal brain tumor, several treatment techniques are proposed today such as radiation therapy, chemotherapy as well as surgery. On the basis of size, type of tumor as well as its grade, the treatment type is chosen. To check whether other parts are being affected by this tumor or not, it is also important to perform certain analysis. When the appropriate treatment method has to be chosen by the doctor, there are certain factors that are to be considered. The possible side effects of a treatment, consideration of complete health and checking whether the central nervous system is affected due to the tumor or not, are few of these factors. Radio imaging is the most commonly applied technique within MRI due to its dynamicity and flexibility [2]. Various pulse sequences and modification in imaging parameters that are based on Longitudinal Relaxation Time (LRT/T1) and Transverse Relaxation Time (TRT/T2) are used to perform acquisition of variable image contrast. Particular tissue properties are provided in relevance to signal intensities provided on the weighted images T1 and T2. On the basis of pulse sequence parameters, the contrast on MR images is provided. For knowing the details of structures of various organs of the body such as liver, chest, and brain, MRI imaging of the body is done. The treatment can be monitored in the patient efficiently with the help of this approach. There are certain steps performed in order to identify the tumor in the patient's body [3]. Preprocessing, segmentation, feature extraction as well as classification are the commonly applied steps. The MRI samples are gathered at the initial stage.

- a. Preprocessing and Enhancement: The chances that a suspicious region can be detected can be improved through this initial step being performed in image processing. From the image, the noise is eliminated and finer details are extracted. The accuracy of an image is minimized when noise is present within the MRI image. The noise is removed by applying different filters on the image. The filters are also applied to sharpen the image. Since the detection of boundary of tumor can be done more effectively and easily, it is important to sharpen the image with the help of various low pass filters once the noise has been completely eliminated from the images.
- b. Segmentation methods: The procedure where the image is broken down to smaller parts and segments is known as image segmentation. The analysis can be performed in easy manner through this step. Several image segmentation methods have been developed over the time. The approach in which the object boundaries are assumed to be defined by the detected edges and which further helps in recognizing these objects is known as edge-based segmentation approach [4]. There is a need to achieve very distinct and closed boundaries to perform direct segmentation which can be done through this approach. False edge detection can occur many times and the partial edges can be joined within an object boundary through edge linking process. The approach in which the bordering pixels present in one area assume to have similar values is known as region based approach. Instead of identifying the edges, the identification of

object region is more important in this case. The pixels are compared with the neighboring pixels. The pixels is said to belong to the cluster in the form of one or more of its neighbors in case when the congruence criteria is satisfied.

- c. Feature Extraction: To detect brain tumor from images, the extraction of exact tumor image is very important since the structure of brain is very complex [5]. In order to extract certain features, it is important to consider few parameters. The tumor can easily be classified with the help of results achieved from feature extraction process.

2 Literature Review

Kaur et al. [11] analyze technique of Magnetic Resonance Image (MRI) for brain tumor detection. It shows difficult structure of brain cells with thin network. It also considers solid growth. If we want to study the growth we need to study the fragmentation process, which is a huge disadvantage. This disadvantage can be solved by clustering technique. For this extraction of segmented brain tumor from its area a sobel edge detection is used. In clustering technique, the no of clusters is counted by computing them on the peak of histogram. The size and location can be analyzed by the segmented part of the binary image. The final fragmented part is then use to analyze size and perimeter of the tumor. It concludes that. The brain tumor can be detected using MRI and clustering techniques. So, it is used on the nature of image and the number of peaks, the clusters can be computed.

Hazra et al. [12] reviewed detection and localization of tumor region present in the brain by using patient's MRI. It contain three levels namely, preprocessing, segmentation, and edge detection. Preprocessing converts the original image into grayscale image and eradicates noise if any which further followed by the edge detection using Sober and Canny algorithms with technique of image enhancement. The segmentation is applied to display the tumor affected region. Lastly, the clustering algorithm is used for the image clusters. It results identification of the brain tumor is done efficiently using MRI and K-means algorithms. In order to detect the tumor more accurately the algorithms can be improvised.

Chauhan et al. [13] proposed preprocessed median filtering MRI brain images. In order to separate the area from image- and color-based segmentation and edge detection is done. Histogram of oriented gradients and gray level co-occurrence matrix is used to represent the images. The respected extracted features are stored in the transactional database to classify the tumor into normal benign. The classified accuracy is being calculated 86/6%. This summarized that the proposed system help to know about the type of brain tumor and its further treatment. This system has been successfully tested on the large-sized brain scanned images of brain tumor.

Reema Mathew et al. analysed that the Magnetic Resonance Image (MRI) is effective technique of the brain tumor detection and classification. This classification is done in various steps like preprocessing, filtration of sound, feature

extraction, and segmentation. These methods preprocessed the MRI brain image using anisotropic diffusion filters. The discrete wavelet transforms are extracted in the feature extracted step. These features are further given as the input to the segmentation step. A support Vector Machine was used segmentation and tumor classification. Hence, it concludes that the accuracy of proposed system is 86%. The validation of this method with the recent results can be used in the future proposals.

3 Research Methodology

This research work is relies on brain tumor detection.

The technique to detect tumor are based on following steps:

Step 1: Morphological Operations:

The process through which the structure or shape of an object can be deformed or reconstructed is known as morphology. For the representation of shape of an object, the operations that are applied on binary images are known as morphological operations. While performing pre or post processing, these operations are applied such that the shape of objects or areas can be known in more appropriate way. Following are few of the most commonly used morphological operations:

- a. Erosion: The operation with the help of which the boundaries of areas of front-end pixels are eroded from the binary images is known as erosion. In terms of size and holes present within it, the regions of foreground pixels are shrunk. There are two inputs given here [6]. The image is eroded within the initial input and the structuring element is given as the second input. The structuring element place upward of given image such that the erosion of binary image can be calculated. Thus, the origin of structuring element and input pixel coincide with each other.
- b. Dilation: The approach through which the holes are filled by adding the pixels to the boundaries of objects present within the image is known as dilation. Two pieces of data are taken as input in this operator. Image is dilated in the initial one and elements are structured in the second one. On the input image, the structuring element is placed for every background pixel such that the given image pixel position and structuring element coincide [7]. Increase the area of foreground pixels is the basic effect of dilation on the binary image. There is a complete closing up of the operation however, in this operation which is its only demerit. There are several classifiers used in the process of detection brain tumor from images. A data structure in the form of a tree is created within a decision tree classifier. On the basis of one particular feature, each interior node that includes decision criteria is based. The entropy reduction that presents the purity of samples is used to calculate the features that are in relevance to classification [8]. The classifier through which two classes are separated using a hyper plane is known as Support Vector Machine (SVM). From the empirical data, an optical

function can be calculated in case when the classes are separated by hyper plane. A basic feed forward based artificial neural network classifier was introduced known as multi-layer perceptron classifier. For performing simple functions, a single hidden layer is used here at first. Further, to improve the classification performance, two hidden layers were included here. For every data set, different hidden units were selected. Across a number of trails, the numbers of hidden neurons were identified. Back propagation algorithm was used to train the neural network.

- c. Clustering: In image processing is basically defined as the technique in which groups of identical image primitive are identified. Clustering is a method in which objects are unified into groups based on their characteristics. A cluster is basically an assembly of objects which are similar between them and are not similar to the objects fitting to additional clusters. C-mean clustering is mainly assigning points to cluster or class. In this clusters are mainly identified by similarity measure, in terms of distance, connectivity and intensity. Moreover, in this technique, each data point belongs to more than one cluster.

Step 2: Naïve Bayes Classifier: Naive Bayes algorithm is effective method of text classification. It works on large training sample set and gives an accurate result. It is a probabilistic classifier based on Bayes theorem with independent assumption which assumes the presence of particular features of a class is unrelated to presence of other features.

Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes’ theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be “independent feature model”. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. The morphological scanning technique will scan the image and technique of Naïve Bayes is applied which mark the tumor in the image. The classifier that includes all independent attributes when the value of class variable is given is known as Naïve Bayes classifier. Conditional independence is another name for this classifier and it is known to be the easiest form of Bayesian network [9]. Here, the Bayes’ theorem is applied along with the naive assumption that shows the independence among every pair of features within the set of supervised learning algorithms. Following relationship is stated by the Bayes’ theorem:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)} \tag{1}$$

Here, y is a class of variable and from x_1 to x_n a dependent feature vector is included.

PSNR and MSE: The PSNR signal is used to measure the quality of loss and lossless compression (e.g., for image compression). The peek signal used original data. The noise is the error introduced by compression. When comparing

compression codecs, PSNR is an approximation to human perception of reconstruction quality. Although a higher PSNR generally indicates that the reconstruction is of higher quality. When PSNR signal is maximum, the MSE signal is minimum (Fig. 1).

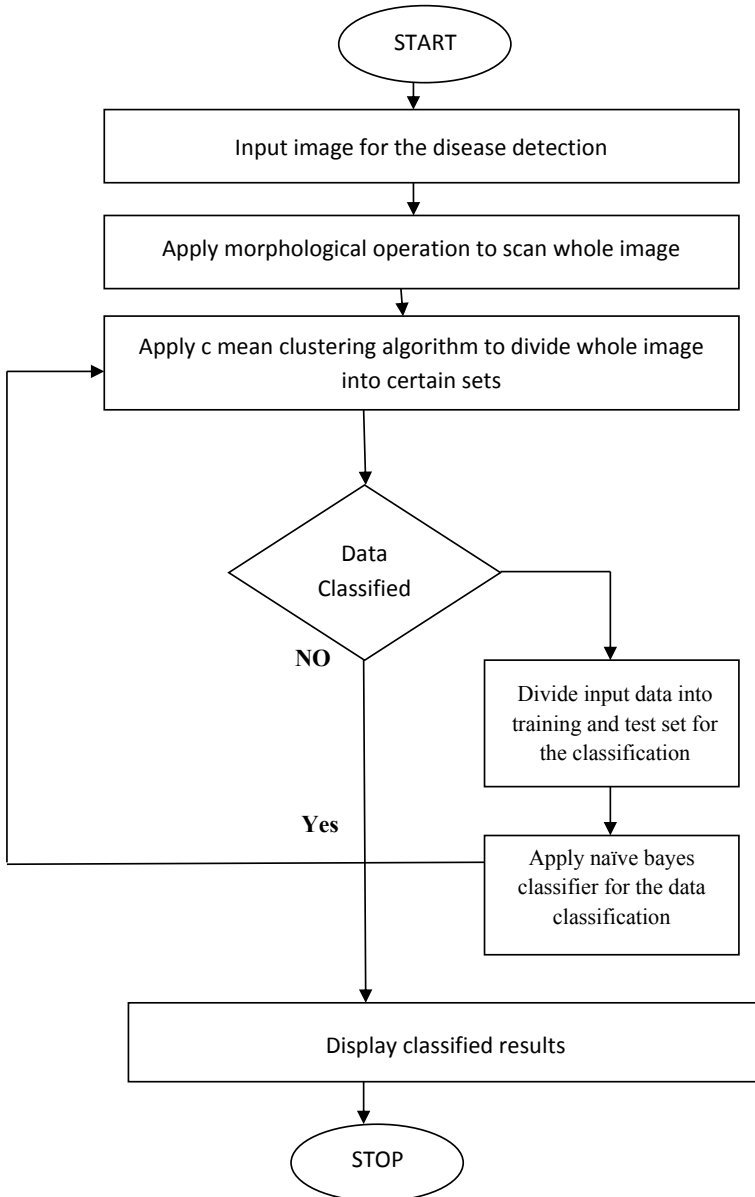


Fig. 1 Proposed flowchart NBC-BTD (brain tumor detection)

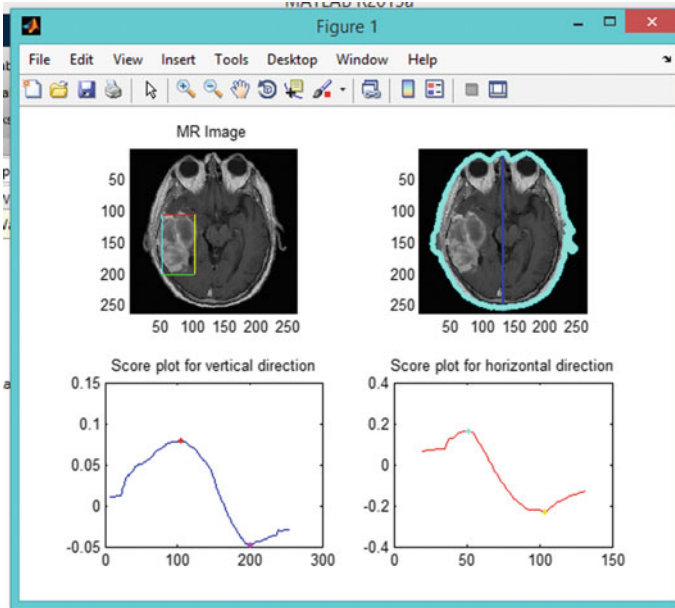


Fig. 2 Naïve Bayes classifier

Result and Discussion: This research work was based on the brain tumor detection; the 20 images taken from data respiratory set. To detect tumor from the MRI images technique of classification was applied. Input the test data for tumor detection, applied in the morphological operations; divided input class into training set; then classified the data. The technique of Naïve Bayes classifier NBC-BTD model marks the tumor portion in the image with horizontal and vertical plot with segmented tumor area. The tumor region grew which segmented tumor would portion from non-tumor region. That gave false positive and negative rate.

As shown in Fig. 2, the technique of Naïve Bayes classifier was applied which marked the tumor portion on the image. The vertical and horizontal position was also calculated from the input MRI Image (Fig. 3; Table 1).

The PSNR value of the Advance and previous research algorithm has compared for the performance analysis. It has analyzed that PSNR value of advance algorithm was high as that to previous research algorithm (Fig. 4; Table 2).

The MSE value of advance and previous research algorithm is compared for the performance analysis. It is analyzed that MSE value of advance algorithm is less as compared to previous research algorithm (Fig. 5; Table 3).

The accuracy value of the advance and previous research algorithm was compared for the performance analysis. It was analyzed that proposed algorithm has high accuracy as compared to previous research algorithm (Fig. 6; Table 4).

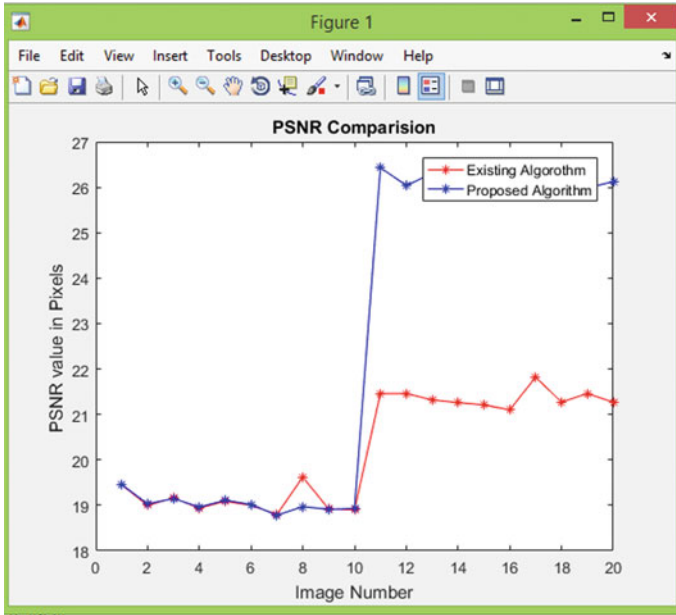


Fig. 3 PSNR comparison

Table 1 PSNR comparison

Image no.	Existing technique (S.V.M classifier)	Proposed technique NBC-BTD
1.	19.45	19.46
2.	19	19.03
3.	19.16	19.15
4.	18.94	18.96
5.	19.09	19.11
6.	19	19.02
7.	18.8	18.78
8.	19.62	18.97
9.	18.92	18.91
10.	18.9	18.93
11.	21.46	26.43
12.	21.46	26.04
13.	21.32	26.3
14.	21.26	26
15.	21.21	26.31
16.	21.1	26.14
17.	21.82	25.83
18.	21.27	26.05
19.	21.46	25.99
20.	21.27	26.12

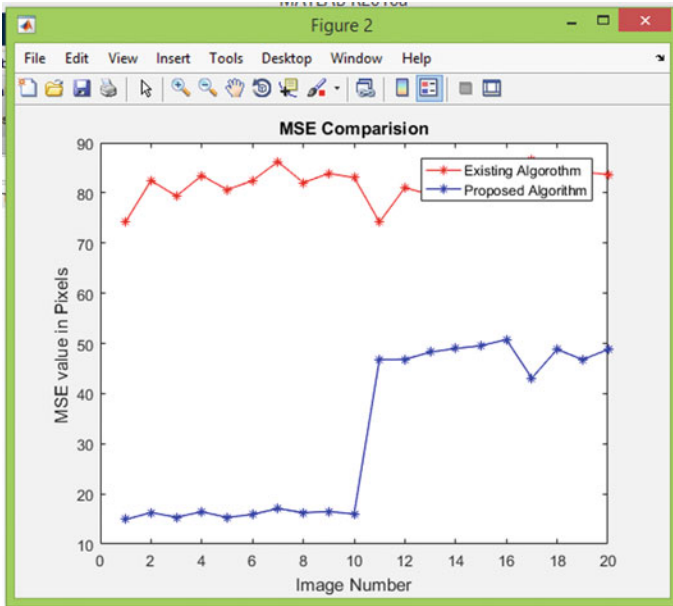


Fig. 4 MSE comparison

Table 2 MSE comparison

Image no.	Existing technique (S.V.M classifier)	Proposed technique NBC-BTD
1.	74.25	14.9
2.	82.45	16.29
3.	79.37	15.36
4.	83.46	16.45
5.	80.63	15.31
6.	82.44	15.92
7.	86.23	17.09
8.	82.03	16.25
9.	83.89	16.48
10.	83.05	16.01
11.	74.18	46.8
12.	81.03	46.81
13.	79.66	48.27
14.	83.19	48.99
15.	80.37	49.55
16.	81.98	50.76
17.	86.69	43.08
18.	83.05	48.83
19.	84.14	46.75
20.	83.68	48.81

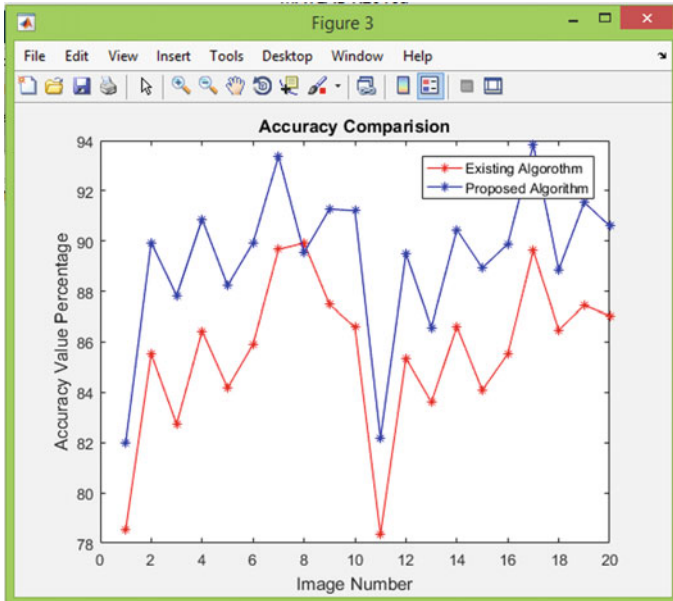


Fig. 5 Accuracy comparison

Table 3 Accuracy comparison

Image no.	Existing technique (S.V.M classifier)	Proposed technique NBC-BTD
1.	78.53	82
2.	85.53	89.94
3.	82.74	87.83
4.	86.44	90.87
5.	84.18	88.23
6.	85.91	89.93
7.	89.67	93.37
8.	89.92	89.54
9.	87.48	91.27
10.	86.59	91.21
11.	78.34	82.18
12.	85.36	89.5
13.	83.61	86.58
14.	86.61	90.45
15.	84.09	88.93
16.	85.53	89.89
17.	89.62	93.83
18.	86.48	88.86
19.	87.45	91.54
20.	87.04	90.6

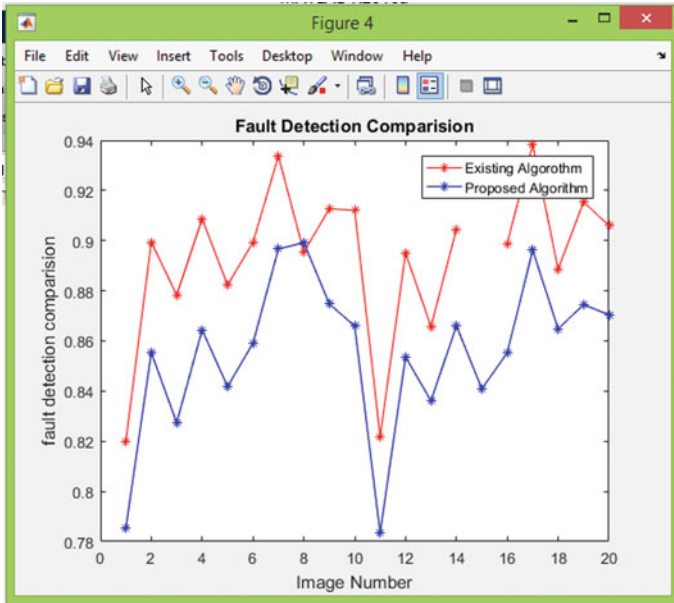


Fig. 6 Fault detection comparison

Table 4 Fault detection

Image no.	Existing technique (S.V.M classifier)	Proposed technique NBC-BTD
1.	0.7853	0.82
2.	0.8553	0.8994
3.	0.8274	0.8783
4.	0.8644	0.9087
5.	0.8418	0.8823
6.	0.8591	0.8993
7.	0.8967	0.9337
8.	0.8992	0.8954
9.	0.8748	0.9127
10.	0.8659	0.9121
11.	0.7834	0.8218
12.	0.8536	0.895
13.	0.8361	0.8658
14.	0.866	0.9045
15.	0.8409	0.8893
16.	0.8553	0.8989
17.	0.8962	0.9383
18.	0.8648	0.8886
19.	0.8745	0.9154
20.	0.8704	0.906

The fault detection rate value of the present and previous research algorithm was compared for the performance analysis. It was analyzed that present algorithm has high fault detection as compared to previous research algorithm.

4 Conclusion

In this work, it is concluded that image processing is the technique which can process information stored in the form of pixels. The brain tumor detection is the technology which can detect tumor portion from the MRI image of brain. In this research work, novel technique is proposed which is based on the morphological operation and Naïve Bayes classifier and clustering techniques. The performance of present algorithm is collate with past and it is analyzed that present algorithm performs well in terms of PSNR, MSE, and accuracy and fault detection with 86% ratio.

References

1. Dhanwani, D.C., Bartere, M.M.: Survey on various techniques of brain tumour detection from MRI images. *IJCER* **4**(1), 24–26 (2014)
2. Joshi, M.A., Shah, D.H.: Survey of brain tumor detection techniques through MRI images. *AJRFANS* **9** (2015)
3. Poonam, J.P.: Review of image processing techniques for automatic detection of tumor in human brain. *IJCSMC* **2**(11), 117–122 (2013)
4. Kowear, M.K., Yadev, S.: Brain tumor detection and segmentation using histogram thresholding. *Int. J. Eng. Adv. Technol.* **1**, 16–20 (2012)
5. Patil, R.C., Bhalchandra, A.S.: Brain tumor extraction from MRI images Using MAT Lab. *IJECSCSE* **2**(1), 1 (2012)
6. Parmeshwarappa, V., Nandish, S.: A segmented morphological approach to detect tumor in brain images. *IJARCSSE* **4**(1), 408–412 (2014)
7. Karuna, M., Joshi, A.: Automatic detection and severity analysis of brain tumors using GUI in matlab. *IJRET Int. J. Res. Eng. Technol.* **2**(10), 586–594 (2013)
8. Kamil, M.Y.: Brain tumor area calculation in CT-scan image using morphological operations. *IOSR J. Comput. Eng. (IOSR-JCE)* **17**(2), 125–128 (2015)
9. Wan, E.W.: Neural network classification: a Bayesian interpretation. *IEEE Trans. Neural Netw.* **1**(4), 303–305 (1990)
10. Chew, K.M., Yong, C.Y., Sudirman, R., Wei, S.T.C.: Bio-signal processing and 2D representation for brain tumor detection using microwave signal analysis. In: *IEEE* (2018)
11. Kaur, N., Sharma, M.: Brain tumor detection using self-adaptive K-means clustering. In: *IEEE* (2018)
12. Hazra, A., Dey, A., Gupta, S.K., Ansari, M.A.: Brain tumor detection based on segmentation using MATLAB. In: *IEEE* (2017)
13. Chauhan, S., More, A., Uikey, R., Malviya, P., Moghe, A.: Brain tumor detection and classification in MRI images using image and data mining. In: *IEEE* (2017)

Automatic Classification of Carnatic Music Instruments Using MFCC and LPC



Surendra Shetty and Sarika Hegde

Abstract With a large collection of digital music in recent days, the challenge is to organize and access the music efficiently. Research in the field of Music Information Retrieval (MIR) focuses on these challenges. In this paper, we develop a system which automatically identifies the instrument in a given Carnatic music on ten different types of instruments. We extract the well-known features namely, MFCC and LPC, and analyze the capability of these features in distinguishing different instruments. Then, we apply, the classification techniques like, Artificial Neural Network, Support Vector Machine, and Bayesian classifiers on those features. We compare the performances of those algorithms along with different features for Carnatic music instruments identification.

Keywords Audio data mining · Carnatic music · Instrument classification · MIR

1 Introduction

Music Information Retrieval (MIR) research explores the different applications involving the retrieval of contents or information from a given music which can be useful for categorizing and indexing music. Due to rapid increase in digital media collections, organization and accessing of the music from large collection has become one of the challenges. Usually, metadata is used to label and segment the songs and is used to search and access a particular music. But this process needs lot of manual intervention for preparing metadata and labeling process which is not so

S. Shetty (✉)

Department of Computer Applications, NMAM Institute of Technology Nitte, Udupi District, Karnataka 574110, India

e-mail: hsshetty@nitte.edu.in

S. Hegde

Department of CSE, NMAM Institute of Technology Nitte, Udupi District, Karnataka 574110, India

e-mail: sarika.hegde@yahoo.in

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_32

463

efficient. Hence, this process can be automated using Machine Learning techniques including both labeling and searching. The digital music may be labeled based on instrument type, singer, genre of music, etc.

Lot of research has been done in this area on western music for content-based information retrieval. However, research on Indian music started very recently with mainly Raga identification as focused application. Instruments like Harmonium, Tabla, etc., are integral part of Indian music. The instrumental music is very melodic and pleasant, played in important Indian occasions like, marriage, festivals, funeral. The Indian instruments can be divided into three types, rhythmic instruments, wind instruments and string instruments. In our paper, we have selected the wind and string types of instruments and applied machine learning technique for automatic identification of instrument. The basic steps include preprocessing and feature extraction from audio, visualizing the features and applying classification algorithms. The set of standard audio features namely, Mel-frequency cepstral coefficient (MFCC) and Linear Predictive Coefficient (LPC) are used. We compare the capability of these features in representing an instrument. The research work in this paper explores the following objectives,

1. Extraction of MFCC and LPC features
2. Analysis of suitability of features through visualization
3. Classification of audio features using Artificial Neural Network (ANN), Bayesian classifiers, and Support Vector Machine (SVM)
4. Comparison of classifiers performance based on classification accuracy.

The content discussed in this paper is organized in five sections. We briefly discuss the literature works on MIR in second section. The methodology and the techniques used in our work are described in third section. The method of experiments and the corresponding results is covered in fourth section which is followed by conclusions in fifth section.

2 Background

Martin and Kim [13] have applied statistical pattern recognition technique for instrument identification. They have demonstrated the usefulness of organizing the musical instruments hierarchically for identification and illustrated the usefulness of timbre feature for this application. An automated system for singer identification from western music is proposed by Kim and Brian [12]. The music is first analyzed to extract the vocal segments and then the singer is identified using GMM and SVM algorithms. Similar work is proposed in Zhang [25] using a statistical model with 80% accuracy. MFCC feature with 32 coefficients is used to characterize the voice of singer and further classified using Neural network in Mesaros and Astola [15]. SVM algorithm along with LPC feature is used by Chetry and Sandier [4] for classifying six instruments successfully. Benetos et al. [3] have performed

instrument identification experiments on large dataset with 20 instruments using variety of classifiers and have reported accuracy ranging between 88.7 and 95.3%. Usefulness of MFCC coefficients for identification of instruments in monophonic signals is investigated in Sturn et al. [21]. A new nonlinear method involving the fractal theory is proposed by Zlatintsi and Maragos [26] for instrument identification. Polyphonic signals are considered in Giannoulis and Klapuri [7] using local features and missing features for instrument identification. Gaussian mixture model based semi supervised learning algorithm is applied on instrument identification by Diment et al. [5].

A continuous polyphonic music is analyzed to perform two tasks namely, automatic music transcription and instrument assignment by Giannoulis et al. [8]. Factorial Gaussian Mixture-HMM (F-GM-HMM) algorithm is used by Hg and Sreenivas [11] for solving the problem instrument detection in a given polyphonic music. Uhlich et al. [24] have used deep neural network for extracting instrument from music. Isolated Instruments identification from noisy clips has been proposed by Mukherjee et al. [16] using the statistical feature LPCC-S which is computed well-known method, Linear Predictive Cepstral Coefficient. Banerjee et al. [2] have proposed a novel approach for string instrument identification using frequency and wavelet domain analysis and the classification algorithms namely, k-NN and random forest method. A very detailed review on Music Information retrieval, its applications and the related work is described in Murthy and Koolagudi [17].

3 Methodology and Techniques

The instrument classification problem can be solved using Machine Learning algorithms. The input is an instrumental audio and the output is the label corresponding to the identified instrument. Before giving the input to Machine Learning algorithm, the input needs to be processed using Digital signal processing technique for extracting the features. In this section, we discuss briefly about the considered dataset, the feature extraction techniques and classification algorithms used in our experiment.

3.1 Dataset

There is no standard benchmark dataset for Indian instrumental music recordings. We collected few CD's consisting of music. From these songs, we extracted the first portion with duration in the range of 2 s–3 min based on the availability. The collection of recordings prepared in this manner consists of instruments namely, *Flute, Harmonium, Mandolin, Nagaswara, Santoor, Saxophone, Sitar, Shehnai, Veena and Violin*. We collected up to 15 audio clips of each instrument category with total number of audios as 150.

3.2 Audio Data Preprocessing and Feature Extraction

To extract useful information from audio which is a non-stationary data, we apply “framing”. It is a mechanism of dividing the audio into smaller chunks of duration of 10–30 ms along with overlapping and windowing. Overlapping is done to ensure continuity in the data and windowing is done for smoothening the data at the edges. Each frame is then processed to extract the audio features namely, Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Coefficients (LPC). We get a features vector of size “ d ” after applying feature extraction technique. The feature for each frame may be denoted as $FV = \{fv_1, fv_2, \dots, fv_d\}$. The set of features extracted from all the recording is further used as input for Machine Learning algorithms.

a. Mel Frequency Cepstral Coefficients (MFCC)

It is a well-known feature used in the areas of speech recognition, MIR applications [9, 20]. This feature simulates the behavior of human ear where the time domain signal is first converted into frequency domain signal using discrete Fourier algorithm. The computed power spectrum from DFT coefficients are filtered using a set of triangular filters with mel-scale spacing. In the last step, inverse Fourier transform is applied by computing DCT from logarithm of Mel-spectrum given as

$$C_i = \sum_{k=1}^K (\log S_k) \cdot \cos\left(\frac{i\pi}{K} \left(k - \frac{1}{2}\right)\right) \quad i = 1, 2, \dots, K \quad (1)$$

Here, the value C_i is the i th MFCC coefficient computed, S_k is the output from k th filterbank channel and K is the number of filterbanks used in the set of triangular filters [19].

Using this technique, 12 MFCC coefficients and its first/second derivative can be computed. The number of coefficients to be used is not fixed. Depending on the requirements of application, number of MFCC can be changed. Most of the time 12 MFCC’s are used. Less number of MFCC makes the model simple. Aucouturier and Pachet [1] in their work have used *eight* MFCC coefficients. In our work, we have experimented with less than eight. The MFCC feature is computed using Auditory Toolbox [20]. Out of the 12 coefficients computed, we consider first *six* coefficients.

b. Linear Predictive Coefficients (LPC)

The time domain signal of a given audio can be analyzed using Linear predictive analysis tool to extract features from it. This tool is found to be very useful in signal processing applications. In this technique, each sample is defined with the weighted sum of previous “ d ” samples. The weights assigned to each previous sample are termed as LPC coefficient which is unknown term. The predicted value for a sample is given as [18]

$$\widehat{s(k)} = \sum_{i=1}^d a_i s(k-i) \quad (2)$$

Here, the $\widehat{s(k)}$ is a predicted value for k th sample, $s(k-i)$ is the sample value at index $(k-i)$. The coefficient a_i is an unknown term. We get such equations with linear combinations for all the samples in the frame where a_i 's are unknown and to be computed. The constraint for computing a_i 's is that the difference between the actual sample $s(k)$ and the predicted sample value $\widehat{s(k)}$ termed as Mean Squared Error (MSE) must be as minimum as possible. This is a least square problem which can be solved using normal equations.

Using this mechanism, we get “d” LPC coefficients which are equal to 13 by default. To keep the number of MFCC's and LPC's same, we consider only first *six* LPC coefficients. For a given frame f_i , the set of six LPC coefficients is represented as, LP = (lp₁, lp₂, lp₃, lp₄, lp₅, lp₆). This LP is used as the feature vector (FV) for discriminating different classes of audio signal in further experiments.

3.3 Machine Learning Algorithms

Classification is a task under Machine Learning algorithm with an objective to assign a given feature vector to one of the known class labels [6]. In our system, the given input is a feature extracted from audio and the label assigned is one to ten corresponding to the ten instruments. The dataset is first divided into training and testing set. The training set is used to construct ML model and the testing set is used to evaluate the performance of ML model. We use three algorithms, namely, Artificial Neural Network, Support Vector Machine, and Bayesian classifier which is discussed in his section.

a. Artificial Neural Network (ANN) Classifier

Artificial Neural Network is one of the popular Machine Learning algorithms mimics the behavior of human brain. It is a network of computing elements known as Neurons. The working of single neuron has two components, summation and activation function given as,

$$y = \frac{1}{1 + e^{-(\sum_{j=1}^p w_j x_j)}} \quad (3)$$

where x is an of “ p ” dimension and w is the weight corresponding to each input. The neurons can be arranged in layers. The three types of layers in ANN are, input layer, hidden layer, and output layer [6]. The number of hidden layers may be zero or more. The number of neurons in input layer is same as the number of feature (p) and in output layer is same number of class labels (10). The number of hidden

layers and number of neurons in hidden layer need to be decided through trial and error. In our work, we have one hidden layer with same “ p ” neurons. The following algorithm shows the usage of ANN in our work.

ALGORITHM 3.1: Training and Testing steps using ANN

```
// The algorithm takes dataset of feature vectors as input and generates the accuracy of classification using ANN
algorithm
//Input: D is a dataset (features vector along with target label)
//Output: m (Confusion matrix)
dataset=D;
[training_set testing_set]=hold_out(dataset);
//Training the ANN
x=Store all the set of feature vectors of training set
y=Sequence of class labels in training set
c=no_of_class
//Each row in T represents the binary pattern of the corresponding value of y
For each row in y
    Initialize the T(i) to 'c' number of zeros
    Set the value of T(y(i))=1
End
network=trainANN(x,T);
//Testing the ANN
For each row from testing set
    FV=feature vector
    output_array=testANN(network, FV); // list of values from Output neuron
    pos=index of maximum value in output_array
    predicted_target_label=pos;
    m(actual_target_label, predicted_target_label)= m(actual_target_label, predicted_target_label)+1
end
```

b. *Support Vector Machine (SVM)*

This is a supervised machine learning algorithm which constructs a hyperplane that can separate the feature vector of one class with feature vectors of all other classes. Out of the different possibilities of hyperplane, the one with widest margin is selected as classifier called Maximal Margin Classifier [23]. Let us consider a set of training observations, given as (\mathbf{x}_i, y_i) ($i = 1, 2, \dots, k$) where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ corresponds to the feature vector (MFCC or LPC) for the i th input observation and $y_i \in \{-1, 1\}$ denote its class label. SVM is a two-class classifier which constructs a hyperplane separating the observations from two class. In our paper, *ten* hyperplanes will be constructed. For each instrument, a constructed hyperplane will separate the observations of that instrument from all other instruments [10]. The decision boundary (hyperplane) can be defined as follows,

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

The parameters \mathbf{w} and b need to be estimated in training phase for a given dataset. Since the hyperplane need to maximal, the constraint and objective function to be minimized for computing the parameters is given as,

$$\begin{aligned}
 w \cdot x + b &\geq 1 && \text{if } y_i = 1 \\
 w \cdot x + b &\leq -1 && \text{if } y_i = -1
 \end{aligned}
 \tag{4}$$

$$f(w) = \frac{\|w\|^2}{2}$$

The following algorithm illustrates the steps followed for using SVM algorithm in our work.

ALGORITHM 3.2: Pseudo code for SVM algorithm
 // The algorithm takes dataset of feature vectors as input and computes the accuracy of classification using SVM algorithm
 //Input: D is dataset (feature vector along with target label)
 //Output: m (Confusion matrix)
 dataset=D;
 [training_set testing_set]=hold_out(dataset);
 //Training the SVM
 x=Store all the set of feature vectors of training set
 y=Sequence of class labels in training set
 c=no_of_class
 Apply Gaussian kernel function to x
 [xsupport, w, b]=trainSVM(x,y);
 //Testing the SVM
 For each row from testing set
 FV=feature vector
 predicted_target_label=testSVM(FV, xsupport,w,b);
 m(actual_target_label, predicted_target_label)= m(actual_target_label, predicted_target_label)+1
 end

c. Bayesian Classifier

Bayesian classifier works based on conditional probability. Posterior probability using Bayes theorem is computed for an feature vector x given as,

$$p(c_j|x) = \frac{p(c_j)p(x|c_j)}{p(x)} \tag{5}$$

where

$$p(x) = \sum_{j=1}^c p(w_j)p(x|w_j) \tag{6}$$

The posterior probability $p(c_j|x)$ indicates the probability that feature vector x belongs to j th class where $j = 1, 2, \dots, 10$ in our work. To compute the class label for a given x , all the 10 posterior probabilities are evaluated and the class corresponding to the highest probability is chosen [22]. The prior probability for class j , i.e., $p(c_j)$, represents the initial degree of belief that a given set of feature vector is a

data from j th class. The class-conditional probability $p(x|c_j)$ is also called as class-conditional probability representing the probability distribution of the features for each class [14].

4 Experimental Results and Discussion

Each audio from the collected dataset is retrieved and processed to extract MFCC and LPC features. The feature vector corresponding to each frame is considered as an observation (an input) for the next stage. Total number of observations for each class is equivalent to number of audios of that class multiplied by number of frames in the audio. We get up to 600–1000 observations for each class. We have considered four different frame size, 512, 1024, 2048, and 4096 for the experiment. For the final classification, we selected the best frame size among this.

In this section, we discuss three types of experiments and the corresponding results. The first one is visualization of features, second is to select the best frame size, and the third one is the classification of instruments. For visualization on 3-D plot, we select the first three coefficients ($fv_{i1}, fv_{i2}, fv_{i3}$) of MFCC and LPC. Randomly selected 25 observations are plotted on 3-D graph with different color/symbols for different instruments. Figures 1 and 2 show the plot for MFCC and LPC features respectively. It can be clearly observed that the instruments of same type are grouped together in both the plots. We can also observe that similar types of instruments like *Saxophone*(inverted triangle) and *Nagaswara*(circle) groups are nearer to each other.

This result of this experiments just gives a brief idea about the potentiality of a feature to distinguish the various instruments and the similarity among some instruments. As per the results, both MFCC and LPC features capable to that. Such experiments are useful in Machine learning applications as a preprocessing step to

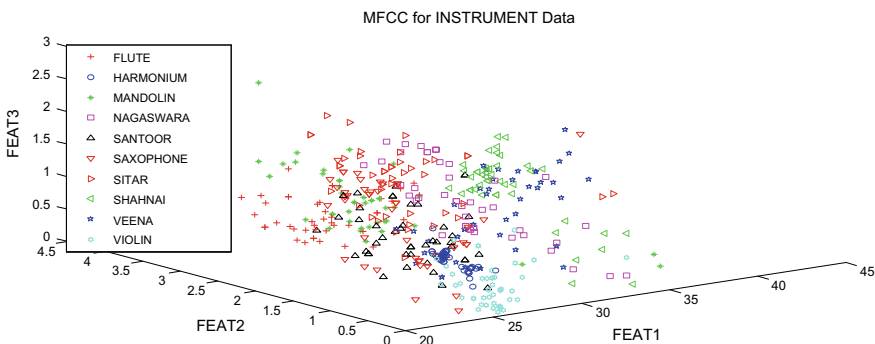


Fig. 1 Visualization of MFCC feature for Instrument data

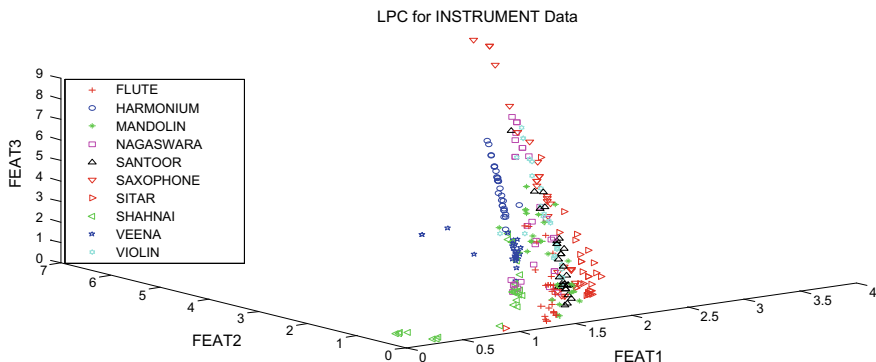


Fig. 2 Visualization of LPC feature for Instrument data

understand the features at preliminary level. The discrimination of instruments through MFCC is slightly clearer than LPC features.

The next experiment done is to select the best frame size suitable for instruments classification. For this experiment, we randomly retrieve 50 observations from each class and apply all the classification techniques using twofold cross-validation evaluation method. Various frame sizes considered are 512, 1024, 2048 and 4096 sample per frame. The features extracted with various frame size are stored as separate datasets. The frame size which gives the highest classification accuracy is selected and considered for further evaluation of classification experiments. Table 1 shows the results of varying frame size across two features and the different classification algorithms for the instrument dataset.

We can observe from the table that the performance of classification is better for the dataset constructed with the frame size of 2048 sample observations. For the further analysis, we consider the dataset which is constructed with the frame size 2048.

The last experiment is classification of ten instruments using, ANN, SVM, and Bayesian classifier. For this, we retrieve 500 observations from each class of instruments, totally constituting 5000 observations from ten classes. The dataset is divided into two sets, set1 and set2 in the ratio of 50:50 randomly. In the first iteration, set1 is used as training set and set2 is used as the testing set and the

Table 1 Comparison of class accuracy for different frame sizes for Instrument data

Classifier		512	1024	2048	4096
ANN	MFCC	56%	31%	72%	55%
	LPC	25%	61%	61%	66%
SVM	MFCC	88%	82%	86%	85%
	LPC	75%	77%	76%	70%

Table 2 Comparison of classification accuracy for Instrument data

Classifier/feature	ANN (%)	SVM (%)	Bayesian (%)
1. MFCC	83.07	69.13	67.12
2. LPC	81.81	67.91	51.25

accuracy of classification is calculated for the test set. Same thing is repeated in the second iteration where set2 is used as the training set and set1 is used as the testing set. Finally, the average accuracy of classification of the two iterations is calculated. Classifiers are evaluated by repeating each of the experiments for *fifteen* times and the average accuracy of classification is taken as the final accuracy of classification. Table 2, shows the accuracy of classification for the dataset of 5000 observations with MFCC and LPC features and the ANN, SVM and Bayesian classifier algorithms. We can observe from the results that MFCC feature is better representative for instrument data and the ANN algorithm classification is better compared to other algorithms. But LPC feature is also useful in representing the instrument data.

The highest accuracy rate for 500 observations is 83.07% with MFCC feature and ANN classifier and Fig. 3 shows the graphical view of the result. The *x* axis gives the name of the instrument (actual target class) and the *y* axis gives the number of observations considered. In visualization, we claimed that MFCC discriminates instrument data better than LPC. This is found to be same in classification also. Even though there is not much difference in the accuracy, MFCC has shown slightly better performance than LPC in all the cases.

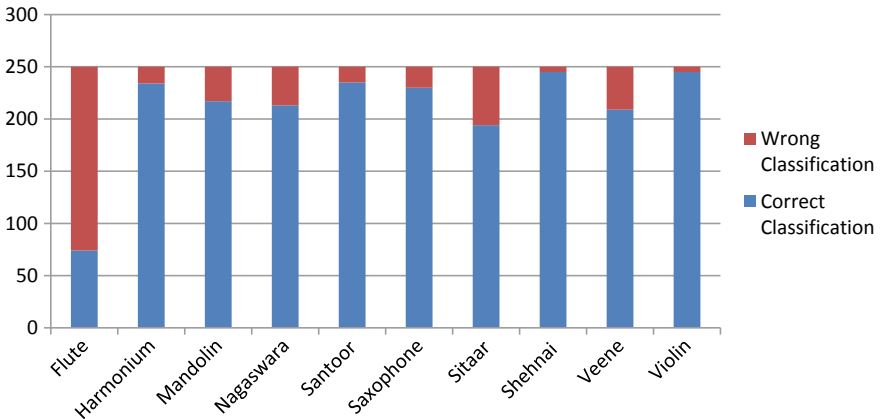


Fig. 3 Graphical representation of confusion matrix for Instrument data classification with ANN and MFCC feature

5 Conclusions

Even though, lot of research is carried for western music, the research in Indian music has lot of scope due to versatile nature of Indian music. Instruments are integral part of Indian music. The automatic identification of instruments from Carnatic music can ease the process of indexing and retrieval. We have developed a system for instrument identification using MCC and LPC features. The visualization helped in understanding the capability of MFCC and LPC features in distinguishing the various instruments. The training dataset and test dataset are selected randomly in each of the iteration and it is found that MFCC feature is shown to be more discriminative than other features. The highest rate of classification accuracy for 10 classes of instrument data is 83% with ANN classifier. More variety of instruments may be considered in future to develop an instrument identification system using deep neural network.

References

1. Aucouturier, J.J., Pachet, F.: Improving timbre similarity: how high's the sky? *J. Negat. Results Speech Audio Sci.* **1**(1), 1–13 (2004)
2. Banerjee, A., Ghosh, A., Palit, S., Ballester, M.A.F.: A novel approach to string instrument recognition. In: *International Conference on Image and Signal Processing*, pp. 165–175. Springer, Cham (2018)
3. Benetos, E., Kotti, M., Kotropoulos, C.: Large scale musical instrument identification. In: *Proceedings SMC'07, 4th Sound and Music Computing Conference, Greece (2007)*
4. Chetry, N., Sandier, M.: Linear predictive model for musical instrument detection. In: *IEEE Conference on Acoustics, Speech and Signal Processing*, vol. 5 (2006)
5. Diment, A., Heittola, T., Virtanen, T.: Semi-supervised learning for musical instrument recognition. In: *2013 Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE (2013)
6. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*. Wiley Publications, Hoboken (2006)
7. Giannoulis, D., Klapuri, A.: Musical instrument recognition in polyphonic audio using missing feature approach. *IEEE Trans. Audio Speech Lang. Process.* **21**(9), 1805–1817 (2013)
8. Giannoulis, D., Benetos, E., Klapuri, A., Plumbley, M.D.: Improving instrument recognition in polyphonic music through system integration. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5222–5226. IEEE (2014)
9. Gold, B., Morgan, N.: *Speech and Audio Signal Processing—Processing and Perception of Speech and Music*. Wiley India Pvt. Ltd., ISBN: 81–265-0822-1 (2006)
10. He, X., Zhou, X.: Audio classification by hybrid support vector machine/hidden Markov model. *UK World J. Model. Simul.* **1**(1), 56–59 (2005)
11. Hg, R., Sreenivas, T.V.: Multi-instrument detection in polyphonic music using Gaussian Mixture based factorial HMM. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 191–195. IEEE (2015)
12. Kim, Y.K., Brian, Y.: Singer identification in popular music recordings using voice coding features. In: *Proceedings of ISMIR (2002)*

13. Martin, K.D., Kim, Y.E.: 2pMU9: musical instrument identification: a pattern-recognition approach. In: 136th Meeting of the ASA (1998)
14. Martinez, W.L., Martinez, A.R.: Computational Statistics Handbook with Matlab. Chapman & Hall/CRC Publications, ISBN: 1-58488-566-1 (2008)
15. Mesaros, A., Astola, J.: The mel-frequency cepstral coefficients in the context of singer identification. In: Proceedings of the International Conference on Music Information Retrieval (2005)
16. Mukherjee, H., Obaidullah, S.M., Phadikar, S., Roy, K.: MISNA—a musical instrument segregation system from noisy audio with LPCC-S features and extreme learning. *Multimed. Tools Appl.* **77**, 27997–28022 (2018)
17. Murthy, Y.V., Koolagudi, S.G.: Content-based music information retrieval (CB-MIR) and its applications toward the music industry: a review. *ACM Comput. Surv. (CSUR)* **51**(3), 45 (2018)
18. Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., Sorsa, T.: Computational Auditory Scene Recognition. In: IEEE International Conference on Acoustics Speech and Signal Processing, vol. 2 (2002)
19. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice Hall, ISBN:10:013051572 (1993)
20. Slaney, M.: Auditory toolbox: a MATLAB Toolbox for auditory modeling work. Technical Report 1998-010, Interval Research Corporation, Palo Alto, CA, USA, 1998, Version 2 (1998)
21. Sturm, B.L., Morvidone, M., Daudet, L.: Musical instrument identification using multiscale mel-frequency cepstral coefficients. In: 18th European Signal Processing Conference (EUSIPCO-2010) (2010)
22. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison Wesley, ISBN: 978-81-317-1472-0 (2006)
23. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, London (2009)
24. Uhlich, S., Giron, F., Mitsufuji, Y.: Deep neural network based instrument extraction from music. In: ICASSP, pp. 2135–2139 (2015)
25. Zhang, T.: System and method for automatic singer identification. In: Proceedings of the IEEE Conference on Multimedia and Expo, vol. 1, pp. 33–36 (2003)
26. Zlatintsi, A., Maragos, P.: Multiscale fractal analysis of musical instrument signals with application to recognition. *IEEE Trans. Audio Speech Lang. Process.* **21**(4), 737–748 (2013)

Semiautomated Ontology Learning to Provide Domain-Specific Knowledge Search in Marathi Language



Neelam Chandolikor, Pushkar Joglekar, Shivjeet Bhosale,
Dipali Peddawad, Rajesh Jalnekar and Swati Shilaskar

Abstract In this research work, our goal is to build a self-sustainable, reproducible, and extensive domain-specific ontology for the purposes of creating a knowledge search engine. We have used online data as the primary information store using which we construct ontology by identifying concepts (nodes) and relationships between concepts. The project encompasses preestablished ideas gathered from successful NLP trials and presents a new variation to the task of ontology creation. The system, for which the ontology is being created, is a knowledge search engine in Marathi. This aims at building semiautomated ontology whose target demographic is primary school children and the selected domain is science domain. This project proposes a method to build semiautomated ontology. We use a combination of natural language processing method and machine learning method to automate the ontology learning task. Automatically learned ontology is further modified by language and domain experts to enrich the contents of ontology. Unlike, standard search engines, our knowledge search engine attempts to provide learned resources directly to the user rather than website links. This approach enables the user to directly get information without having to spend time on browsing indexed links.

N. Chandolikor (✉)

Department of IT and MCA, Vishwakarma Institute of Technology, Pune, India
e-mail: neelam.chandolikor@vit.edu

P. Joglekar · S. Bhosale · D. Peddawad

Department of Computer Science, Vishwakarma Institute of Technology, Pune, India
e-mail: pushkar.joglekar@vit.edu

S. Bhosale

e-mail: shivjeet.bhosale14@vit.edu

D. Peddawad

e-mail: dipali.peddawad16@vit.edu

R. Jalnekar · S. Shilaskar

Department of E&TC, Vishwakarma Institute of Technology, Pune, India
e-mail: rajesh.jalnekar@vit.edu

S. Shilaskar

e-mail: swati.shilaskar@vit.edu

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_33

Keywords Knowledge search engine · Natural language processing (NLP) · Ontology · Resource description framework (RDF) · Semantic search

1 Introduction

Knowledge sources are rarely available in Marathi language on World Wide Web. This paper proposes to provide domain-specific search facility in Marathi. It also aims at converting the extracted data into knowledge and provides output in more semantically coherent form. The proposed knowledge search in Marathi would be based on semiautomated ontology-based framework.

1.1 *Ontology*

Ontology has the ability to capture the structure of a domain. Ontology has a knowledge base which contains structures that are defined by the allowed constructs of ontology [1, 2]. Ontologies are building blocks of semantic web-based systems [3]. Creating ontologies is an intricate process [4, 5]. Ontology defines a common vocabulary for researchers who need to share information in a specific domain. Ontology of any domain is a specification of a conceptualization of that specific domain. Ontologies are concerned with the meaning of terms. Ontology typically consists of concepts and relationship between those concepts. It includes definitions of basic concepts in the specific domain and relations among them. Ontologies are used to represent domain knowledge in a digital form. Ontologies can be automated or semiautomated. Semiautomated ontology is created automatically and edited manually.

1.2 *Knowledge Search Engine*

Knowledge search engine [6–8] provides semantically correlated information to the input keyword. Knowledge search engine is somewhat similar to semantic search engine. A semantic search engine interprets the meaning of a given keyword and provides information either in the form of relevant links or in the form of summarized information. Ontologies allow smart linking of knowledge resources. This gives a computer the potential to use intelligent agents to access knowledge and deliver good quality information to the user. Agents also have the ability to use the ontology semantics [9, 10] to communicate with other agents and thus making computer communications smarter.

1.3 Resource Description Framework (RDF) and Web Ontologies

Ontologies are stored in RDF format [11]. RDF is the widely accepted standard framework for representing information about a certain system. RFD is built around the concept of uniform resource identifier (URI) which is used to describe website metadata. A typical RDF graph consists of RDF triples (subject, predicate, and object) which is displayed as a graph. Web ontology complements the RDF and RDF Schema by giving it an ability to express reasoning and convey logical structure.

2 Related Work

In 2007, author Horacio Saggion et al. proposed the system on ontology-based information extraction for business intelligence in which they used information from business graphics and tabular data and they applied OCR analysis of images and then output of the OCR analysis was corrected by the exploitation of collateral information found around the graphics [12]. In 1992, author Marti A. Hearst described a method for the automatic acquisition of the hyponymy lexical relation from large text corpora first they identified a set of lexico-syntactic patterns and then described a method for discovering these patterns [13]. In 2006, author Patrick Pantel et al. proposed a weakly-supervised, general-purpose algorithm, called Espresso, for harvesting binary semantic relations from raw text. In which they have extracted several standard and specific semantic relations like is a, part of, succession, reaction, and production [14].

In 2004, author Patrick Pantel et al. used top-down approach for labeling semantic classes. They used co-occurrence statistics of semantic classes discovered by algorithms like CBC to label concepts. In this paper, they proposed an algorithm for automatically inducing names for semantic classes and for finding instance/concept (is-a) relationships [15]. In 2013, author Tomas Mikolov proposed two novel model architectures for computing continuous vector representations of words from very large datasets. The author worked on distributed representations of words learned by neural networks in which they proposed two models Continuous Bag-of-Words Model and Continuous Skip-gram Model [16]. Natural language processing methods [17–20] of pos tagging and chunking is widely used by many researchers for identification of concepts for domain-specific ontology. Machine learning techniques like classification [21–24] is used widely for relationship classification.

3 Proposed Methodology

In the proposed methodology, we have focused on creating ontology for a science domain which includes science subject terms and concepts understandable to primary school-going children of Marathi medium. However, proposed methodology can be applied to any domain. In the proposed methodology we at first define ontology structure followed by this ontology learning method is proposed and finally we provide knowledge search through knowledge search engine.

3.1 *Defining Structure of Ontology*

For building ontology and its nodes, we have initially defined ontology structure as per the requirement of project. Ontology ideally has terms and relationships. In this research work for every term, one ontology node is created. Structure of ontology node which is used is as follows:

1. Keyword in English.
2. Definition in English.
3. Automated Translated Marathi keyword.
4. Marathi definition by domain and language expert.
5. Image link.

Once this ontology node structure is defined, all the terms (concepts) are stored using the same node structure.

3.2 *Methodology to Create Semiautomated Ontology*

We have presented a combined approach to natural language processing and machine learning classification for ontology learning. Ontology learning has six individual modules, namely, seed word resolution, data scraping, pos tagging and chunking, predicate classification for identification of relationship, ontology writer and graph intelligence, and text simplification. Figure 1 shows the block diagram of ontology creation. The detailed explanation of the ontology creation are as follows.

3.2.1 **Seed Word Resolution**

The semantic search engine takes an input seed word for learning ontology. The search engine operates on one seed word at a time so as to limit the divergence of

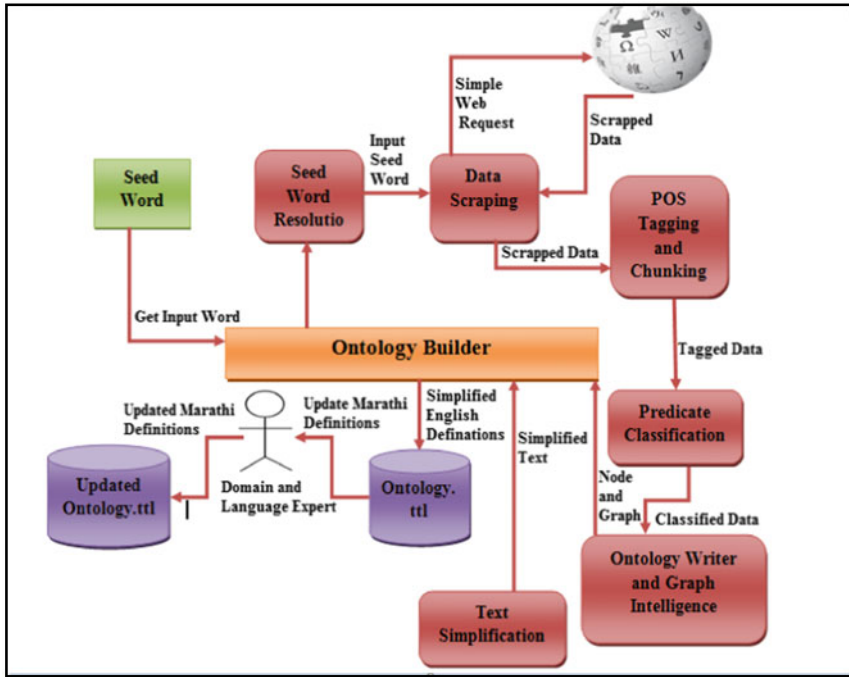


Fig. 1 Block diagram of ontology builder

the data that is gathered. Traditional encyclopedias such as Wikipedia or Britannica are replete with example of entries with topics that can be interpreted in a much broader sense. Our algorithm, in order to avoid this problem, gets additional five words that are semantically similar to the seed word. This is done using Word2vec. Word2Vec is a vector embedding space that encodes semantically similar words in an N-dimensional vector space such that their cosine similarity is minimum. We used Google’s dataset which is trained on roughly 100 billion words with a vocabulary of 3 million words. The top five words similar to the seed word make sure that the algorithm will always have relevant information when displaying to the user. However, to be able to know the meaning of millions of words there is no practical classifier that will give us good description of any word. A dictionary can be used instead but it does not capture the linking between two or more terms. To be able to do this, our algorithm treats Word2vec model as the classifier for the word. This is the prime reason behind the extraction of top five similar words. They act as an index as to what words will point wherein the ontology.

3.2.2 Web Data Scraping

Web Data Scraping is a computer science application, where a programmer automates the process of retrieval of information from websites by the use of intelligent scripts. Every website typically has an HTML format which tells the browser how the raw data must look to the user. Now, say a user wants to analyze product prices on a web store, manual extraction is impractical. In such cases, a simple script can be written to talk to the website's server and retrieve the required data directly into user-defined data structures. Our search engine relies heavily on web scraping to eliminate the need for storing data offline. The top five similar words to the seed word will be the basis of getting raw text data. The algorithm will first scrape data for each word from <https://simple.wikipedia.org/>, which is a simplified version of Wikipedia. At this stage, the algorithm has all the relevant data for each wiki pages for each of the five words and the seed word. The scraper will extract definitions, images, (links) and other paragraphs from each web page. The scraped data objects will be provided to a module that performs a series of operations to further refine the data.

3.2.3 Part of Speech (POS) Tagging

Every word used in a sentence is assigned a specific part of speech. There are several parts of speech and the most important and commonly recognized ones are nouns, verbs, and adjectives. We have to dive a little deeper into the grammar of English language to understand the importance of this module.

A sentence in English falls into three categories, namely, simple, compound, and complex. A machine, however, does not understand these categories. To allow the computer to understand this, we exploit the patterns that the POS tags that appear in all three kinds of sentences. We use NLTK's inbuilt POS tagger for this purpose.

NLTK's POS tagger works on a saved per-trained POS tagging machine learning models. The NLTK POS tagger recognizes the various types of POS tags.

3.2.4 Chunker

The next phase that comes is chunking. Chunking is the process of converting a POS tagged sentence into a list of tree structures that recognizes a certain user-defined grammar. The root of all trees are defined by the letter 'S' and in the absence of grammar, all the words are leaves of this tree. The chunking process, allows us to recognize the subject, predicate, and the object of the sentence. NTLK's Chunker is a regular expression-based function that parses the sentences into the tree from the user-defined grammar. Our algorithm attempts to recognize

the subject, predicate, and object of every sentence. To do this, we defined our own grammar rules to recognize the subject, predicate, and object in a sentence. The grammar used is as follows.

“”

```

NP : {<PRP|PRP$>?<DT>?<JJ>*<NN.*>+}
      {<JJ>?}
      {<PRP|PRP$>}
REL : {<V.*><TO>?}
      {<DT>?<JJ>*<NN.*>+}
      }<VBZ|VBP>{
REL_BOOL : {<RB|RBR|RBS>?<VBZ|VBP*>?<RB.>?<TO>?<IN>?}
      Commas : {<,><NP>}
      NPlist : {<NP><Commas>*<CC>?<NP>}
SVO: {<NP|NPlist><MD>?<REL_BOOL>+<REL>*<IN>*<CD>*<NP|NPlist>}
Prep : {<IN>+<N.|VBG|PRP.>}
Simple: {<SVO><Prep>*}
tag : {<CC><Simple>}
Compound: {<Simple><tag>+}

```

“”

In the above-defined grammar, “NP” is the grammar used to chunk noun phrases, “REL” and “REL_BOOL” are used to chunk verb phrases from sentences. When the chunker spots the grammar defined by “SVO” tag, it will indicate that the algorithm has now safely found an SPO triple. This triple will be further analyzed to generate the ontology.

3.2.5 Concept Identification

After POS tagging and chunking phase, subjects and objects are identified as concept (term) of ontology. After identification of terms, for every term following details are stored in the node of ontology: keyword in English, definition in English, automated translated Marathi keyword, and Marathi definition by domain and language expert and Image link for term.

3.2.6 Predicate Classification

As humans, understanding the meaning of any sentence is very easy. When trying to replicate the same task for computers, the biggest challenge is where to start in order to solve this problem. Ideally one would think that there is no problem in storing the relationship directly as it would make any difference when it is presented to the user. However, the opposite is true for machines. Machines have not been taught to recognize the relationship between two or more entities. When dealing with computers, ultimately, it comes down to etching strict borders to these abstract tasks that humans perform. Ideally, a human will be able to understand if two entities are related, but to know how they are related is quite a huge and complicated task. For instance, a human can easily declare that the word “dog” and the word “golden retriever” are related. As humans, we can relate “dog” and “golden retriever” as class and instance of the class, respectively. However, it is important to note that we can do this without the presence of a predicate. Humans can also express more than one relation between the two entities. This is unique only to humans.

In order to replicate the same result in machine, we first need to define how the relationships are to be specified. This is where the ontology comes into the view. As we discussed in the initial section of this paper, ontology is a specification for conceptualization. The exact syntax and rules that are specified in RDF will form the basis for recording the knowledge in the ontology. For our purposes, we have tested four classifiers, namely, Decision Tree, SVM, Naïve Bayesian, and Maximum Entropy.

In this paper, we propose the use of maximum entropy techniques for ontology building for knowledge search. Experimental results show that max entropy gives more accurate result. Maximum entropy is a method used for a natural language processing task like text segmentation. The first step for constructing model using max entropy is to collect a large number of training data which consists of samples represented in the following format: (x_i, y_i) where the x_i includes the contextual information of the document (the sparse array) and y_i its class. The second step is to summarize the training sample in terms of its empirical probability distribution:

$$\tilde{p}(x, y) = \frac{1}{N} \times T$$

where T is the number of times that (x, y) occurs in the sample and N is the size of the training dataset.

3.2.7 Graph Intelligence

Simply storing the data in the ontology is not very useful if the user has to link the data using their own intelligence. Our graph intelligence module generates a graph that records all the knowledge in terms of subject, predicate, and object represented

as the nodes. The graph intelligence automatically integrates definitions and images for the nodes from the ontology.

After the machine has recognized the word, the next step of the algorithm is to describe the word. To describe a word, we as humans have a brain wiring that allows us to record the features of the word. For example, we can say that to describe a dog, we humans will recall the image of the dog and then describe the features of the dog that we imagine. A particular node in the ontology is the feature of that word that the algorithm has learned and stored by analyzing all the raw text data. Over time all these nodes automatically form connections and which the algorithm can describe. The algorithm allows the user to traverse along the edges of the ontology nodes thus simulating a primitive thought process. This graph is well connected to the other related entities but with a few erroneous relationships. The graph currently traverses two levels deep. The graph can be visualized as follows (Fig. 2).

3.2.8 Text Simplification

Text simplification [25] is a technique in natural language processing which is used to modify text in such a way, that the grammar and structure of the given text is

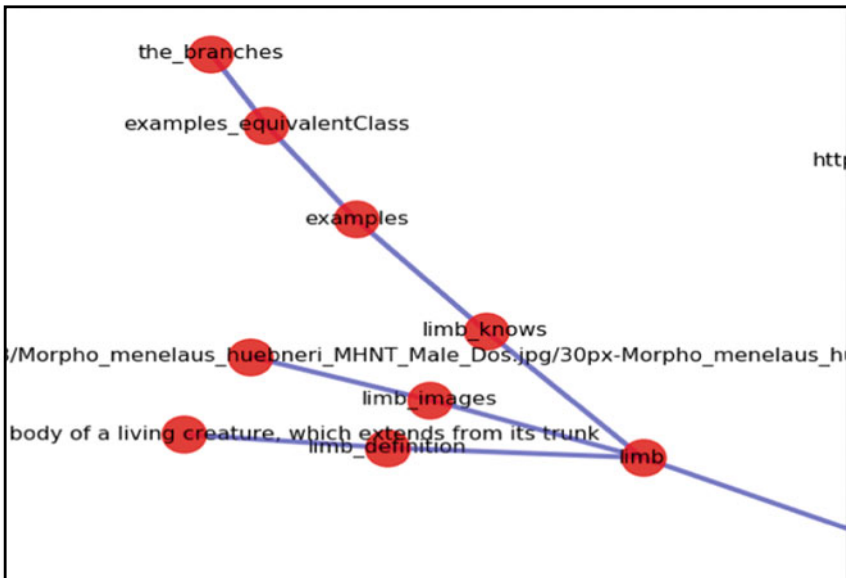


Fig. 2 A part of the subgraph, partly describing the node “limb”

greatly simplified, while the meaning and information remain the same. The target demographic of the knowledge search engine is school-going children so; it is necessary to create definitions which will be understandable by children. In-text simplification we converted complex words to easily understandable words and complex sentences to simple sentences. In this work English definition which is retrieved from World Wide Web and stored in ontology node are further simplified, so that it becomes understandable to target demographics. Further, Marathi definitions are added by domain and Marathi language expert, to complete ontology node as per defined structure.

3.3 Knowledge Search Engine

In the second part, we created a web application using Django framework. This application will help student to search the science terms. This application will give direct results to the students instead of giving a link as a result. So, it will help student to get the result without wasting time on indexed link. Figure 3 shows a block diagram for ontology search.

We will enter a science keyword in Marathi as input. Then that Marathi word is converted to English word then semantic search engine will search that keyword in updated ontology file and will return output definitions and images.

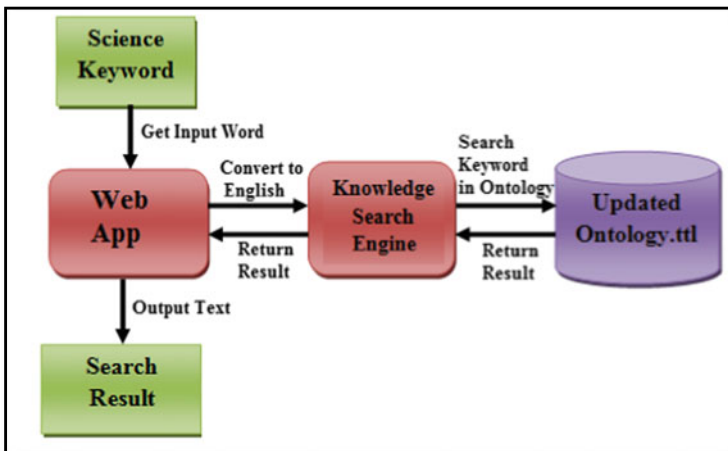


Fig. 3 Block diagram of ontology search

4 Experimental Result

During the course of the project, it has been found out that to generate good quality ontology; the most significant task identifies relevant triples. A relevant triple is a triple in which the subject, predicate, and object talk about the seed word. Current chunking methods perform fairly well for simple and complex sentences.

4.1 Comparative Analysis of Classifiers

The initial approach was to use the Naive Bayes classifier to classify the predicate. In later comparative analysis of the dataset, the following were the accuracies for each classifier (Table 1).

4.2 Classification of Predicate

The predicate classification task requires a huge amount of data along with good labels. The predicate classification works fairly well on maximum entropy classifier with an average accuracy of 76%. The feature of the predicate can be a dictionary of POS tags as the keys and the corresponding word in the predicate as the value of the dictionary.

Maximum entropy classifier gives the highest accuracy. Therefore, this method is used for building ontology nodes and relationships. Relationships are mainly classified into subclass, equivalent, object property, and irrelevant. The relationships that are classified as irrelevant are of varying nature. They do not directly fall into subclass, equivalent, or object property relations but they usually are not completely useless. They are stored as a friendly relation.

Table 1 Comparative analysis of classifiers on our dataset

Classifier	Accuracy (avg of 20 shuffled iterations) (%)
Naive Bayes classifier	72.42
Maximum entropy classifier	76.39
Support vector machine (SVM)	75.15
Decision tree	73.18

4.3 Created Ontology

The proposed method provides a good mechanism to generate domain-specific ontology. Given a seed word, relevant terms are added into ontology. Created ontology has terms and relationships. For every term one ontology node is created. Following examples illustrates ontology node and ontology structure.

(a) Example of ontology nodes

1. Keyword in English:- “Sunlight”.
2. Definition in English (simplified definition):- “It is the light of the sun. Sunlight, also called sunshine. What we experience as sunlight is actually solar radiation. It is the radiation and heat from the Sun in the form of waves. The atmosphere affects the amount of solar radiation received. When solar radiation travels through the atmosphere, some of it is soaked up by the atmosphere (16%)”.
3. Marathi definition (provided by language and domain expert):- “सूर्यप्रकाश म्हणजे सूर्याचा प्रकाश. आपण जो सूर्यप्रकाश अनुभवतो, ती खरे म्हणजे सौर प्रारणे आहेत. सूर्यापासून नघालेली सौर प्रारणे आणि उष्णता तरंगांच्या रूपात असतात. सौर प्रारणांमुळे वातावरणावर परिणाम होतो. सौर प्रारणे जेव्हा वातावरणातून जातात तेव्हा त्यातील काही प्रारणे (सुमारे 16%) वातावरणात शोषली जातात”.
3. Image link: https://upload.wikimedia.org/wikipedia/commons/thumb/7/71/Antelopocanyonjh1.jpg/220px-Antelope_canyon_jh1.jpg.

5 Conclusion

In the context of the current project, we are planning to use the ontology for knowledge search engine where users are school-going children. The semantic search engine is mainly focused on science keywords which will provide a Marathi definition of a keyword as output. Creating ontologies is not an easy task and there is no unique correct ontology for any subject. So, in this project, we used a combined approach to natural language processing and machine learning method for creating ontology. The predicate classification task requires a huge amount of data along with good labels. The predicate classification works fairly well on maximum entropy classifier with an average accuracy of 76%. The maximum entropy classifier is the most suitable for our ontology creation. This paper proposes the building of ontology which provides knowledge search in Marathi, which takes a keyword as an input and uses ontology to provide semantically coherent information and present summarized information in the structured form.

Acknowledgements This work was supported by the Maharashtra Government Project funded by the Rajiv Gandhi Science & Technology Commission Mumbai. We also thank to Marathi Vidnyan Parishad, Pune Vibhag for their immense help in writing Marathi definitions.

References

1. Maedche, A.: *Ontology Learning for the Semantic Web*. Kluwer Academic Publishers, Amsterdam (2002)
2. Maynard, D., Funk, A., Peters, W.: Using Lexico-syntactic ontology design patterns for ontology creation and population. In: *Proceedings of the 2009 International Conference on Ontology Patterns*, vol. 516 (2009)
3. McDonald, R.: *Extracting relations from unstructured text*. UPenn CIS Technical Report (2004)
4. Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y.: Named entity recognition from diverse text types. In: *Recent Advances in Natural Language Processing Conference* (2001)
5. Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F.: Automatic ontology learning: supporting a per concept evaluation by domain experts
6. Maynard, D., Funk, A., Peters, W.: SPRAT: a tool for automatic semantic pattern-based ontology population
7. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, Jr., E.R., Mitchell, T.M.: *Toward an architecture for never-ending language learning*. In: *Association for the Advancement of Artificial Intelligence* (2010)
8. Davulcu, H., Vadrevu, S., Nagarajan, S.: *OntoMiner: bootstrapping and populating ontologies from domain specific web sites*
9. El-gaya, M.M., Mekky, N., Atwan, A.: Efficient proposed framework for semantic search engine using new semantic ranking algorithm. *Int. J. Adv. Comput. Sci. Appl.* **6**(8), P136–P143 (2015)
10. Cai, B., Li, Y.: Design and development of semantic-based search engine model. In: *7th International Conference on Intelligent Computation Technology and Automation*, pp. 145–148 (2014)
11. Dilek, S., Karacan, H., Jahangiri, N., Afzali, S.: *Ontology Creation for an Educational Center* 978-1-4673-1740-5/12/©2012 IEEE
12. Saggion, H., Funk, A., Maynard, D., Bontcheva, K.: *Ontology-based information extraction for business applications*. In: *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*, Busan, Korea (2007)
13. Hearst, M.A.: *Automatic acquisition of hyponyms from large text corpora*. In: *Conference on Computational Linguistics (COLING'92)*, Nantes, France. Association for Computational Linguistics (1992)
14. Pantel, P., Pennacchioni, M.: Espresso: leveraging generic patterns for automatically harvesting semantic relations. In: *Proceedings of Conference on Computational Linguistics/ Association for Computational Linguistics (COLING/ACL-06)*, Sydney, Australia, pp. 113–120 (2006)
15. Pantel, P., Ravichandran, D.: *Automatically labeling semantic classes*. In: *Proceedings of HLT/NAACL-04*, Boston, MA, pp. 321–328 (2004)
16. Mikolov, T., Yih, W.T., Zweig, G.: *Linguistic regularities in continuous space word representations*. In: *NAACL HLT (2013)*
17. Bach, N., Badaskar, S.: *A review of relation extraction. Literature review for language and statistics II* (2007)
18. Saranya, K., Jayanthi, S.: *Onto-based sentiment classification using machine learning techniques*. In: *International Conference on Innovations in information Embedded and Communication Systems* (2017)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: *Efficient estimation of word representations in vector space*. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
20. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: *Natural language processing (almost) from scratch*. *J. Mach. Learn. Res.* **12**, 24932537 (2011)

21. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. *Mach. Learn.* **29**(2–3), 131–163 (1997)
22. Liu, J., Qin, L., Wang, H.: An ontology mapping method based on support vector machine
23. Johnson, I., Ab, J., Charnomordic, B., Destercke, S., Thomopoulos, R.: Making ontology-based knowledge and decision trees interact: an approach to enrich knowledge and increase expert confidence in data-driven models
24. Chieu, H.L., Ng, H.T.: Named entity recognition: a maximum entropy approach using global information. In: *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1 (2002)
25. Siddharthan, A.: An architecture for a text simplification system. In: *Proceedings of the Language Engineering Conference (LEC'02)* 0-7695-1885-0/02© 2002 IEEE

Identifying Influential Users on Social Network: An Insight



Ragini Krishna and C. M. Prashanth

Abstract The advancement in the speed of the internet connection on handheld devices has led to an increase in the usage of social media. This drew the attention of advertisers to use social media as a platform to promote their products thus leading to an increase in the sales of their product, increasing the brand awareness. To increase the rate of information dissemination within a short period of time, influential users on social media were targeted, who would act as the word-of-mouth advertisers of the product. However, there are various parameters on which the influence of a user has to be determined. The parameters can be (1) the connectivity of the user in the network (2) knowledge/interest of the user on a particular topic/product/content (3) activity of the user on the social media. This survey focuses on the various methods and models for identifying influential nodes and also the effect of compliance, where a user falsely agrees to the content of another influential user by retweeting, just to gain status or reputation and thus increasing his influential score. Thus, the list of influential nodes of a social network can be faked upon, due to this issue.

1 Introduction

It is an era of automation and people's digital presence is increasing and finding the influence of the user based on their digital presence is the talk of the time. The digital presence of the user on various social media like facebook, twitter, linkedIn, google+, etc., are used to find the influence of the user on the digital social networks.

R. Krishna (✉) · C. M. Prashanth
Department of Computer Science and Engineering, Acharya Institute of Technology,
Bangalore 107, India
e-mail: raginis.17.pfcs@acharya.ac.in

C. M. Prashanth
e-mail: prashanthcm@acharya.ac.in

With the vast usage of the social network [1] to connect to friends and pass information across the network of friends, the social networks are being targeted for the use of business. Thus, social media is used for promoting a product or for passing on information or news. To speed up the process of the spread of information, people target the influential users.

It is already proven and now it is a fact that social media marketing affects the brand awareness and the purchase intentions [2] but the author “Romy Sassine” [1], also discusses that the effect of influencers for a brand marketing is decreasing due to the overuse of the social media.

The reasons for such a trend can be the overuse or wrong placement of the hashtags in a particular message or it may be due to the competition among the influencers to get more brands contract instead getting connected to the audience.

According to [3], the advertizing done using the word-of-mouth strategy increased the sales of the product by twice and the customers gained from this method had a higher retention rate of 37%. The author also states that just having a large number of followers does not indicate an influence, moreover to determine the influence of a person, his credibility towards the topic and strength of the relationship with followers along with the total numbers of followers should be considered.

To determine the influential users in the social network is one of the most trending research areas. There have been many research work carried to find the influential users in a social network. In this paper, Sect. 2 introduces the methods used to find the centrality of the nodes in a network. In Sect. 3, the various methods used to find the influence on social media is discussed. Section 4 lists out the various shortcomings of the methods and in Sect. 5, we discuss the prospect into the future of identifying the influence on social media.

2 Background

A large number of methods have been proposed to find the influential users on the social network so far. Initially, researchers proposed methods which were based on the centrality [4, 5] of the nodes in the network, such as degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, these methods mostly assumed that the network was static, as they never considered the interactions among the nodes of the network. Thus, to consider the dynamicity of the network various researchers started considering the interactions among nodes in the network [6, 7], such as α -centrality and β -centrality. Thus, link strength was determined based on the level of interactions among the nodes in the network to consider the dynamicity of the network. The link strength indicates the ratio of interaction with the concerned node as compared to the interaction with other nodes, which gives the weightage to the concerned link. The link strength between nodes u and v is the ratio of interaction between u and v to the sum of total interactions of u and total interactions of v .

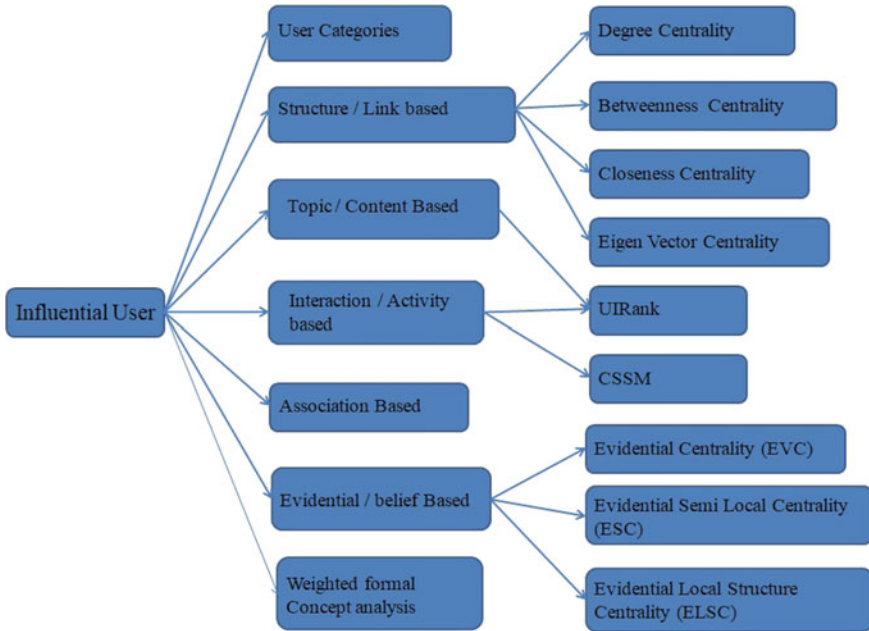


Fig. 1 Categorization of the various methods to find influential users in social networks

The influence of a node in the network can be determined based mainly on its link strength, its interest in the topic of discussion or its activity in the network. Figure 1 shows the categorization of the various methods which determine influential users on the social network.

3 Methods for Finding Influence on Social Network

In this section, we will discuss methods (Fig. 1) to find the influential users based on various parameters like the level of interaction among the nodes, interest of the user in the topics, etc.

3.1 Structure or Link Based Influence

These methods used the connectivity of the nodes in the network to find the influential nodes. However, with the increasing usage of the social media, the connectivity among the nodes kept changing at a frequent interval, thus making them more dynamic, which was not addressed in the link-based methods of finding influence [4, 5, 7].

3.2 *User Categories Based Method to Optimize the Popularity of Twitter Messages*

There are number of users on twitter who generate a million of content everyday however many of the contents go unnoticed or unread which can reduce the popularity of the applications among the users. This may happen because different categories of people will be interested in different topics/content. To detect the effect of the content in the tweet, on its popularity, the tweets were categorized on topics like brands, celebrities, community, entertainment, media, sport, places, and society [8]. The activities of the users were monitored for all these categories. The data was collected and five features-(1) number of mentions, (2) number of hash-tags, (3) number of URLs, (4) the number of pictures, and (5) the sentiment of the tweet were collected, where the first four were directly derived from the twitter API but the fifth feature, the sentiment of the tweet, was extracted from the tweet using labMT word list. The sentimental analysis was done on the tweets and then segregated into neutral and positive or negative, 0 being neutral and 1 being extreme either positive or negative.

1. Analyzing the Influence of the content on the popularity of the user's tweets:

The popularity of the user's tweets was determined based on the features of the content and how the popularity differs for a tweet of different user categories. Each feature of the content was ranked based on its importance for the popularity of each and every user categories.

To minimize the sum of squared error in the prediction for finding the most popular content, LASSO (least absolute shrinkage and selection operation) [9] was used.

From the results in [8], it was observed that the sentiment feature plays a very insignificant role towards the popularity of the tweets for every user categories, whereas when considering retweets and favorites, the feature-mentions were ranked high for favorite but had lower ranking for retweets.

When considering the user categories for the most number of retweets, the feature -picture was ranked highest for communities and brands, whereas the feature-URLs was ranked highest for media and places.
2. Predicting the popularity of the tweets

This work predicts the popularity of a tweet given its features. A prediction model was built using a regression model consisting of 50% of the tweets for the training and 25% to validate and another 25% for evaluating the model.

There were three ways in which the prediction was done, Generic model, the category-specific model, and the smoothed model. In the generic model, the training was done on all the data, whereas in the category-specific model, the training was done based on the segregation of the categories of the tweets.

The Generic model gave high prediction errors for different categories of users. In the category-specific model, the category to which, the tweet in the test set,

belongs was found out and then it was used in the prediction model. According to the arrived results [8], the category-specific model showed improvement in the prediction of the number of retweets, however, it did not outperform the generic model in many other cases, the reason for this may be because the test set considered was very small.

Thus, to get the best from both the models, the best elements of both the models were combined to get the smoothed model. This was achieved by considering the weighted average of both the generic model and the category-specific model. The smoothed model proved to be giving better results as compared to generic or category-specific model.

3.3 Interaction or Activity Based

1. Time slice-based interaction:

To calculate the level of interaction of nodes in the network, methods like indegree and outdegree centrality were used. High indegree of a node represented an inactive user as he/she consumes the information but does not help in the information dissipation.

The steps involved to calculate the influence of a user based on its interaction with its neighbors are [10]:

- Indegree and outdegree centrality
- Link strength
- Clustering value based on indegree and outdegree
- Calculating the influence

Let's take an example to illustrate it:

Step 1: The indegree for the nodes of the network from Fig. 2 at t_1 are:

$$A = 0, B = 2, C = 1, D = 2 \text{ and } E = 2$$

Thus, the outdegree for the nodes at time t_1 are:

$$A = 1, B = 1, C = 2, D = 1 \text{ and } E = 1$$

Thus, if there are users with high indegree and zero or low outdegree, it represents a very low activity.

Step 2: To find the influence of the nodes among friends of a node, the strength of the link that the nodes share across the network is calculated. If the link between nodes is found to be strong, then they can be influenced by their friends with a very little effort from the influencer. Thus, the link strength between the two friends 'a' and 'b' can be calculated by:

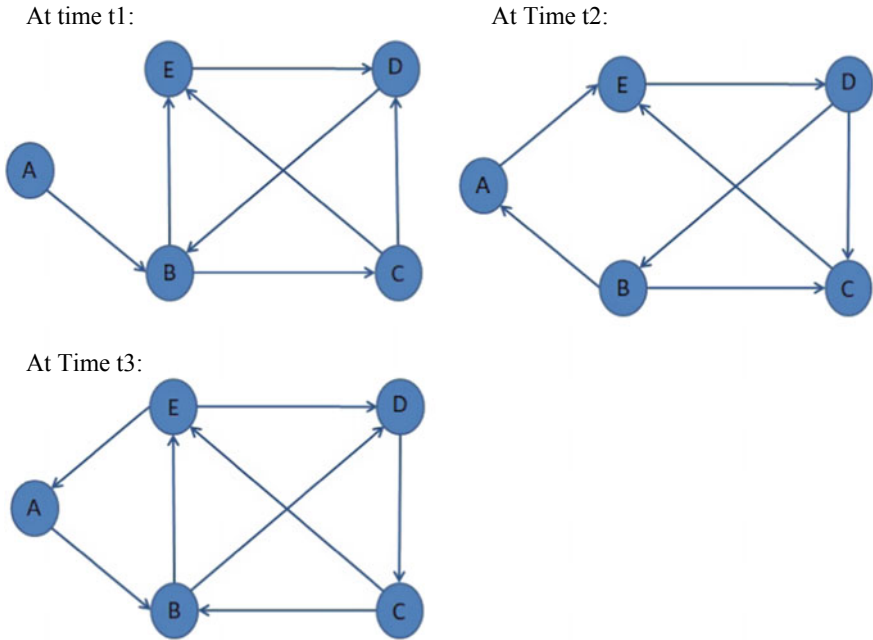


Fig. 2 Representation of nodes interaction in a network at the time slots t1, t2, and t3

$$L(b, a) = \frac{|\bar{e}(b, a)| + |\bar{e}(a, b)|}{d_o(b) + d_o(a)} \tag{1}$$

In the example that has been taken, there are three time periods that are considered as the time of interactions between the various nodes in the network.

Thus, the indegree and outdegree for the nodes across all the three-time slots are:

Indegree centrality:

$$d_i(A) = 2 \quad d_i(B) = 4 \quad d_i(C) = 4$$

$$d_i(D) = 5 \quad d_i(E) = 6$$

Outdegree centrality:

$$d_o(A) = 3 \quad d_o(B) = 4 \quad d_o(C) = 5$$

$$d_o(D) = 4 \quad d_o(E) = 3$$

Thus the link strength is calculated and the values are as follows:

Link strength between node A and B:

$$L(B,A) = 0.428$$

Link strength between Node A and E:

$$L(E,A) = 0.333$$

In a similar way, the link strength between various nodes is calculated. Thus it is inferred that the link strength between the node A and B and the nodes D and E, with the strength value of 0.428, seems to be stronger as compared to the other nodes.

Step 3: clustering value based on indegree and outdegree:

The connectivity of the node within a group of nodes is represented by the cluster value. The higher the cluster value of the node, the more connected it is, to a number of clusters and thus has more reachability to pass on the information to more number of people in the network.

The high indegree clustering value indicates that a number of messages are being transmitted from the neighboring nodes to the node in question, which implicitly means that the concerned node is involved in more number of discussions and thus there will be a high chance of him/her adopting other’s behavior. The indegree clustering value is given by:

$$C_i(a) = \frac{\sum_{b \in S} d_i(b)}{d_i(a) * (d_i(a) - 1)} \tag{2}$$

From Eq. (7), the indegree clustering value for node A is 5, node B is 1.75, node C is 1.25, node D is 0.7, and node E is 1.25. It is evident that node A has high indegree clustering value and thus is connected to more number of clusters.

The number of messages transmitted from the concerned node to its neighboring nodes is indicated by outdegree clustering value. Thus a node with high outdegree clustering value has a high chance of being influential. The outdegree clustering value for a node is given by:

$$C_o(a) = \frac{\sum_{b \in S} d_o(b)}{d_o(a) * (d_o(a) - 1)} \tag{3}$$

After calculating the outdegree clustering values of the nodes in the network, A has a value of 1.167, B has 1.25, C has 0.55, D has 1, and E has 1.167. This clearly shows that the nodes A and E can be influencers in the network. Any node in the network with a high outdegree but a zero indegree value can be a spammer.

Step 4: Thus to find the influencers in a network, the values of the indegree centrality, outdegree centrality and indegree – outdegree clustering values will be used. Thus, from the above discussions, it is quite clear that the user who has high outdegree centrality with high outdegree clustering value will be influential. But the weightage of these parameters depends on the case. For example, maybe the user does not have high outdegree cluster value but he generates a lot of content which make him quite active and thus the interaction of the user with it neighbor will go high, which may make him influential. Thus, to normalize these parameters, two normalizing factors, α_1 and α_2 are introduced. The influencer of the network can be calculated by:

$$Infr(a) = [\tanh(d_0(a))] * (\alpha_1 d_0(a) + \alpha_2 C_0(a)) * \sum_{t \in S} L(b, a) \tag{4}$$

After calculating the influencer values for the nodes, it is found that the node B is most influential with an influencer value of 3.690 after assuming the value for both α_1 and α_2 as 0.5.

The users with high indegree centrality and high indegree cluster value are the users who can be influenced by its neighbors. Thus, they should not be mistaken for inactive users. To remove the inactive users in finding the influencer or the influenced user, the hyperbolic tangent is used to translate the outdegree value of the concerned node 'a' to either zero or one.

The influenced value of a node is:

$$Inf d(a) = [\tanh(d_0(a))] * (\alpha_3 d_i(a) + \alpha_4 C_i(a)) * \sum_{t \in S} L(b, a) \tag{5}$$

Furthermore, the influenced value for each node in the above example is calculated and thus node B has the highest value with 4.087 with the assumption of α_3 and α_4 as 0.5. Thus, from the example network, it can be inferred that the node B creates influence on its neighboring nodes and it is influenced too by its neighbors.

The authors [10] have worked towards finding the influence of users using interaction among them in the network, which clearly avoids the problem of spammers or inactive users in the network.

2. Discovering influence using community scale-sensitive maxdegree (CSSM)

The community in a network is detected based on the interaction among the nodes in the network which represent similarity among the nodes present in the same community. The communities can be identified using various clustering algorithms.

Influence of a node depends on the parameters like the centrality based on the degree of the node, sum of neighbor's degree, and the attributes of the node [11].

The outdegree centrality (ODC) of the node is given by:

$$ODC(x) = \frac{o(x)}{m - 1} \tag{6}$$

where $o(x)$ is the number of neighbors adjacent to x and m is the total number of nodes in the network.

The sum of neighbor's degree (SND) is given by:

$$SND(x) = \sum_{i=1}^{o(x)} o(y_i) \tag{7}$$

where y_i is the i th neighbor node of x . The value represented by SND denotes the number of followers of the node x at level 2.

The results obtained from ODC and SND are summed together for each community and they are finally given as the most influential nodes in the network.

The above method considers only the interaction among the nodes in the network, however, the interaction can be using a negative tweet, which will give an inaccurate result in identifying influence. This issue is rectified by considering the content of the interaction.

3.4 Content Based User Influence Rank (UIRank)

The influence of user on a microblogging site is calculated by a method called UIRank where the node's information disseminating ability (connectivity) and the contribution of the user's tweet (content) is considered [12].

The content generated by the user, that is, the tweet, retweet, and the comment is considered and the influence of the tweet is calculated by considering the retweet and comments on that tweet and is given by:

$$sp(u) = \sum_{t \in \text{Tweets}(u)} Rr(t) + Cr(t) \quad (8)$$

where $sp(u)$ is the probability that the user's u tweet propagates from user u to the neighbors of the fans of user u , $Rr(t)$ is the retweet to the read ratio of the tweet t and $Cr(t)$ is the comment to the read ratio of the tweet t .

The user's influence in the network is found out by finding the connectivity of the user in the follower relationship network. The connectivity of the user will decide the rate of diffusion of the user's tweet in his network and then his follower's network and so on until it reaches all the nodes in the connection. The rate at which the user's tweet is spread depends on the connectivity of the user in his follower relationship and then his fans influence their respective neighbor's network.

The centrality of the node gives the importance of the node in the relationship network. The centrality of the node can be measured by degree centrality, betweenness or closeness centrality.

$sa(u)$ is defined to measure the network influence of a node in the network and it is represented by:

$$sa(u) = C_d(u) + C_b(u) + C_c(u) \quad (9)$$

where $C_d(u)$, $C_b(u)$, and $C_c(u)$ are the degree centrality, betweenness centrality, and the closeness centrality of the user respectively.

The nodes disseminating ability is given by:

$$\pi_{uv} = \begin{cases} \frac{sp(u) + sa(u)}{u_{out}} & \text{if } u \text{ follows } v \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Thus, all the nodes disseminating ability is calculated and put in the form of a transition matrix. Using a random walk method and the page rank algorithm, the UIRank is defined as:

$$IR(u) = \alpha \sum_{v \in \text{followers}(u)} (IR(v) * \pi_{uv}) + (1 - \alpha) \quad (11)$$

where α is a delay factor.

The performance of UIRank is compared to the other four influence finding methods like retweet Rank, outdegree rank, tunk Tank, and fans Rank. The results derived show that UIRank performs best on precision, recall, and F measure, as compared to the other four.

Further, the work can be improved by considering the topic of the tweet posted by the user to find the influence of the user on that topic and also finding the trust factor of the tweet posted by the user.

3.5 *Based on Evidence Theory or Belief Based*

Evidence theory or Dempster–Shafer theory (DST) works on the principle of reasoning with uncertainty. The theory combines, evidence from many sources and then arrive at a belief.

The evidential centrality (EVC) uses the nodes degree and the weight strength of each node in the network to find the influential nodes in the network [13]. However, EVC assumed a uniform distribution regarding the connectivity of the nodes, however, in the complex networks, the interconnectivity of the nodes in the network is unpredictable, and thus it may not be suitable for the complex weighted network. Thus, to overcome the assumptions of the EVC method, semi local structure centrality (ELSC) along with modified EVC [14] was used to find out the influential nodes in weighted networks.

The most common relationships in the twitter social network are follow, mention, and retweet. The belief function is used [15, 16] to overcome the data imperfection in the twitter dataset due to the limitation on the data extraction from twitter. It considers the strength of the link between the direct neighbors, information exchange, and propagation between the users and the assumption that if the user is connected to an influential node, becomes influential. Thus, the influence calculation has been taken up in two sections, the first section where the influence of the node on the directly connected nodes are calculated and the section calculates the influence of the user's neighbors in their respective network. The results of this

method show that it fetches better seed sets of influential users as compared to the previous methods.

3.6 Formal Weighted Concept Analysis

Formal concept analysis (FCA) [17] uses binary objects and attributes relationships to build the knowledge hierarchy which very well represents the relationship between the objects and attributes. A social network can be represented in terms of the formal context by representing the connectivity of the network where nodes indicate the objects and the attributes whereas the edges indicate the binary relationship between them.

To calculate the important nodes using FCA includes three steps. First, the concepts are computed from the adjacency matrix of the formal concepts. Then, the weight of each node is calculated based on the concepts. Thus, to rank the nodes in the network based on its importance, the weighted formal concept Analysis (WFCA) [18] is used, where the hierarchical tree is generated from the cluster of nodes and then the concepts of the network are computed.

The adjacency matrix in Fig. 3b of the graph from Fig. 3a is considered as the formal concept where a row represents an object and a column represents an attribute. Thus, the weight of each node is calculated and then ranked according to this weight, given by

$$w_i = \sum_{k=1}^n \frac{O_{ik}}{A_{ik}} \tag{12}$$

where O_{ik} is the object number of the concept C_{ik} and A_{ik} is the attribute number of the concept C_{ik}

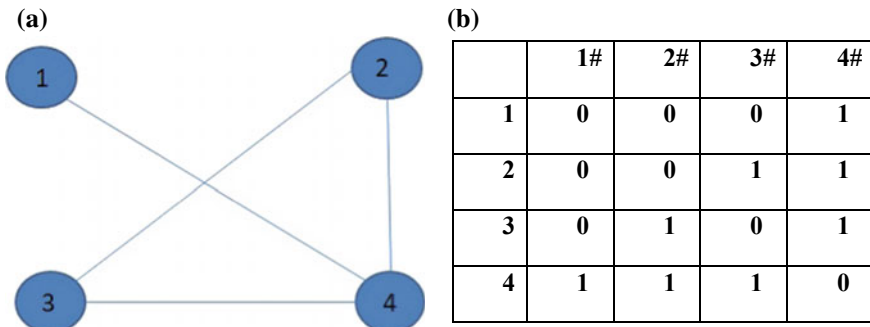


Fig. 3 An example network **a** Graph G for the network. **b** Formal context of the network G

Once the weights of the nodes are calculated, they are ranked depending on their weights.

The adjacency matrix for the network in Fig. 3a is constructed and shown in Fig. 3b. The entries in the adjacency matrix are either 1 or 0 representing the presence of an edge between the nodes or absence of it, respectively.

Figure 4 represents the Hasse Diagram which contains all the concepts of the formal context. The weights, W_i , of all the nodes are calculated using the equation 17 to find the rank of the nodes. From the example network of Fig. 3a, the weight of the node 1 is calculated by using the value of its attributes and objects as follows:

$(\{4\}, \{1\#}), (\{4\}, \{1\#, 2\#}), (\{4\}, \{1\#, 3\#}), (\{4\}, \{1\#, 2\#, 3\#, \})$. Thus, according to the equation 17, the weight of the node 1, W_1 , is: $\frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{3} = 2.33$. Similarly, the weight of node 4 is: $\frac{3}{1} + \frac{1}{2} + \frac{1}{2} = 4$.

After calculating the weights of each node in the network, they are ranked. So from the above example, using WFCA, node 4 is considered to be most influential among other nodes in the network.

4 Outcome of the Review

- Since the social network’s size is increasing exponentially, thus to find an influential node, a sample of dataset representing the social network is

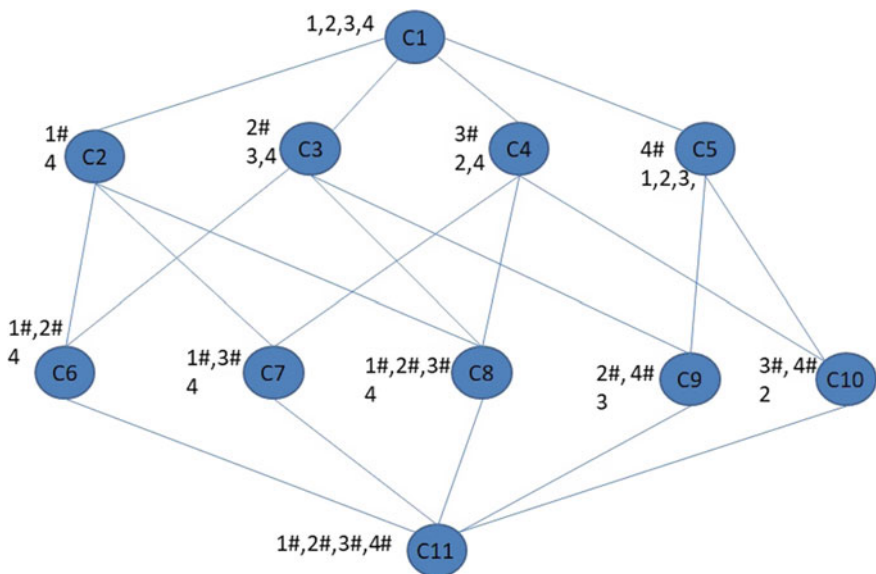


Fig. 4 Concept lattice (Hasse diagram) for the graph in Fig. 3a

considered. Thus, the accuracy may suffer as the result depends on the sample set that is chosen.

- The user's interest can be determined based on the activity of the user, like tweet or retweet. However, if a particular topic is not of a user's interest, he/she may not react to it at all or may give a negative comment. Thus for any business model trying to use influence on social media to promote their business, can't blindly use the user (influential) without knowing their sentiment regarding the product. For example, a user is a well-known wrestler who has won a gold medal at Olympics (thus influential), if an artificially flavored juice product is given to him to promote among his followers, he may not react to it at all.
- Influence of web users is determined by (1) connectivity of the nodes [4–6, 19, 20], (2) through user categories [8], (3) interaction-based [10, 21, 13], and (4) based on the content of the tweet as well as the connectivity [14]. It is observed that the two methods, the content of the tweet and the connectivity is considered, yields a better accuracy, thus there is a need to find the influence based on a method which considered all the above parameters and thus accuracy of the result will be improved.

5 Conclusion

The determination of influential users in the network can be done in many ways; however, the result of finding the influence by combing two or more methods is better as compared to determining by just one method as shown in [12, 14, 15]. Thus, to obtain an accurate and improved result, it is advisable to combine two or more appropriate methods. The study also reveals that one can create a fake influence by retweeting/commenting/following the already established, highly influential user on social media, without being actually influential. The author proposes to explore this issue in future work. A method to detect such fake influencers may be devised so that such users are excluded from the list of the influential user.

References

1. Sassine, R.: What is the impact of social media influencers? Retrieved on February 28th 2018, from <http://digital-me-up.com/2017/04/06/impact-social-media-influencers/>
2. Duiverman, C.: Does influencer marketing have an impact on actual sales? Retrieved on February 28th 2018, from <https://thecircle.com/blog/2017/9/5/influencer-marketing-impact-sales>
3. Wong, K.: The explosive growth of influencer marketing and what it means for you. Retrieved on February 28th 2018, from <https://www.forbes.com/sites/kylewong/2014/09/10/the-explosive-growth-of-influencer-marketing-and-what-it-means-for-you/#1227987752ac>
4. Bonacich, P.B.: Power and centrality: a family of measures. *Am. J. Soc.* **92**, 1170–1182 (1987)

5. Bonacich, P.: Eigenvector-like measures of centrality for assymmetric relations. *Soc. Netw.* **23** (3), 191–201 (2001)
6. Freeman, L.C.: Centrality in social networks conceptual clarification. *Soc. Netw.* **1**(3), 215–239 (1978–1979)
7. Lerman, K., Ghosh, R., Kang, J.H.: Centrality Metric for Dynamic Networks. University of Washington, Washington, DC (2010)
8. Lemahieu, R., et al.: Optimizing the popularity of Twitter messages through user categories. In: IEEE International Conference on Data Mining Workshop (ICDMW). IEEE (2015)
9. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodological)* **58**, 267–288 (1996)
10. Rad, A.A., Benyoucef, M.: Towards detecting influential users in social networks. In: MCETECH (2011)
11. Hao, F., et al.: Discovering influential users in micro-blog marketing with influence maximization mechanism. In: 2012 IEEE on Global Communications Conference (GLOBECOM). IEEE (2012)
12. Jianqiang, Z., Xiaolin, G., Feng, T.: A new method of identifying influential users in the micro-blog networks. *IEEE Access* **5**, 3008–3015 (2017)
13. Wei, D., et al.: Identifying influential nodes in weighted networks based on evidence theory. *Phys. A Stat. Mech. Appl.* **392**(10), 2564–2575 (2013)
14. Gao, C., et al.: A modified evidential methodology of identifying influential nodes in weighted networks. *Phys. A Stat. Mech. Appl.* **392**(21), 5490–5500 (2013)
15. Ren, J., et al.: Identifying influential nodes in weighted network based on evidence theory and local structure. *Int. J. Innov. Comput. Inf. Control* **11**(5), 1765–1777 (2015)
16. Jendoubi, S., et al.: Two evidential data based models for influence maximization in Twitter. *Knowl. Based Syst.* **121**, 58–70 (2017)
17. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: *Ordered Sets*. Springer, Dordrecht, pp. 445–470 (1982)
18. Sun, Z., et al.: Identifying influential nodes in complex networks based on weighted formal concept analysis. *IEEE Access* **5**, 3777–3789 (2017)
19. Chen, D., et al.: Identifying influential nodes in complex networks. *Phys. Stat. Mech. Appl.* **391**(4), 1777–1787 (2012)
20. Erlandsson, F., et al.: Finding influential users in social media using association rule learning. *Entropy* **18**(5), 164 (2016)
21. Li, Y.M., Shiu, Y.L.: A diffusion mechanism for social advertising over microblogs. *Decis. Support Syst.* **54**(1), 9–22 (2012)

Factex: A Practical Approach to Crime Detection



Rachna Jain, Anand Nayyar and Shivam Bachhety

Abstract The crime on roads is a major problem faced today by all the modern cities. Road Transport is the most common escape route for many criminals. Thefts and many other crimes remain unregistered and unsolved due to lack of evidence. Effective tracking of vehicles and criminals is still a big problem and involves plenty of resources. To evade such a condition, we have proposed a machine learning-based practical crime detection system using the text and face recognition techniques. Such systems will be proved useful in parking lots, toll stations, airports, border crossings, etc. In the proposed system, the text recognition involves extracting the characters present in the Indian number plates and the predicted output will be compared with the registered vehicle database. Simultaneously, Face recognition feature constitutes identifying criminal faces based on certain face regions and then mapping the respected coordinates with the criminal database. The proposed system presented in this research paper targets to deliver improvised outcomes considering the time constraints and accuracy with more than 85% successful recognitions in normal working conditions with the goal to accomplish the successful detection of crime using machine learning algorithms such as KNN, SVM, and face detection classifiers to present a practical real time detection.

Keywords Crime detection · Dlib · Haar cascades · KNN · License plate · Text recognition · Face recognition · OpenCV · SVM

R. Jain · S. Bachhety
Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering,
New Delhi, India
e-mail: rachna.jain@bharatividyaapeeth.edu

S. Bachhety
e-mail: shivambachhety97@gmail.com

A. Nayyar (✉)
Graduate School, Duy Tan University, Da Nang, Vietnam
e-mail: anandnayyar@duytan.edu.vn

1 Introduction

The recent works in computer vision and machine learning, text recognition, and face recognition can be visualized as effective measures against rising criminal activities. Using video surveillance, license plate recognition has been an effective approach utilized in the past for traffic management and crime detection. With the integration of text recognition from number plate, an added feature of face recognition can prove to be a boon to the overall system. These techniques involve certain considerations of subjects, posture, emotions, and light for its smooth implementation. Text recognition [1] involves the basic steps of extraction, recognition, and identification of characters, whereas Face recognition [2] involves detecting faces, matching the input face images based on pre-stored face data points and its coordinates and then identifying it. The implemented method with a limited dataset can be seen as a milestone for better surveillance if extended to actual case databases [3]. In the proposed system, KNN is used as text recognition technique and HAAR Faces Classifier and SVM technique for the performing face recognition.

In India, the problems related to traffic are increasing day by day. A typical number plate in India has black foreground color and white background color for private cars. The number on the license plate consists of the two-digit letter showing “state code” followed by two numeral digits, followed by a single letter and then the last four digits. The system consists of the identification of number plate present in the car and then segmentation of all the characters on the number plate. The identification task is interesting because of the nature of the light. Noise in the image background and other environmental factors can result in low accuracy of output. In the proposed system, the edge detection method and mid-filtering are used for noise removal. The next aspect of face recognition consists of mining some significant features of the image and then transforming them for classification. It mainly involves geometric computation of facial features as regions. The face recognition process consists of making a database of faces with numerous facial images for each person. Next, is to detect faces present in the database and then train the face classifier. The last step is testing the classifier to recognize faces it was trained for.

This paper highlights a novel approach to recognize the license plate and criminal face in real time and then matches the output in the vehicle registration database and criminal record face database respectively. The output result produces an email alert to the respective concerned authorities with the license plate number and recognized face details from the database.

1.1 Organization of Paper

Section 2 elaborates the related works highlighting the contributions by several other authors in the corresponding domain. Section 3 highlights the proposed

system, methodology as well as the algorithm used for the live implementation of the proposed system. Section 4 gives a detailed overview of the results produced by the system. Section 5 concludes the paper with future scope.

2 Related Works

Various techniques are proposed and implemented in the past few years regarding license plate detection using visual image processing. Vuong et al. [4] focus on license plate detection based on features like color feature, edge detection, equilibrium, and other operators. Gou et al. [5], proposed license plate detection system in which the structure is regarded as a crucial factor to verify characters from number plate and extraction is done via projected segmentation. For most of the algorithms, Sobel as edge detection operator is frequently used and marked as the best method to transmit from grayscale image to gradient image. Ahmad Radmanesh [6] focused chiefly on the vertical edge detection via the Sobel detection, in comparison to the horizontal edge density of background objects. The template matching and neural networks proved to be an effective technique for character segmentation and recognition as used by Zhu Wei-gang et al. [7] but template matching feature has many limitations when compared to neural networks. After segmentation, the neural network method is expected to be the template match one.

The first approach established for proficiently demonstrating faces was using PCA. The aim of this method was to present a human face as a coordinate system. The vectors setup on the coordinate system were known as eigenfaces vector. Later, Turk and Pentland used this to develop an eigenface-based algorithm for face recognition [2]. SVMs as linear classifiers were first used by Osuna et al. [8] to feature the border between the resulting hyperplane and the training set examples. The ultimate goal of optimal hyperplane was to minimize the error in the classification of the unseen test patterns. Schneiderman and Kanade [9] labeled an object recognition algorithm and projected the results via Bayesian classifier. The result lies in computing the probability of a face to be present in the image based on the count of the frequency of occurrence of a series of patterns over the training images. Saraswathi et al. [10] proposed face recognition authentication process via feature extraction technique, i.e., Linear Discriminant Analysis (LDA). It generates facial regions and then classifies using Euclidean Distance. The experimental results were as high as about 93.7% using the LDA feature set. Extending the PCA to GPU power, Bhumika Agrawal et al. [11] with the help of NVIDIA CUDA performed computations for faster processing of face recognition on Principal Component Analysis algorithm. In the later phase, the work is compared to varied CPU implementations. Jian Yang et al. [12] used the one-dimensional, pixel-based error model based on regression analysis to characterize two-dimensional error image matrix, namely nuclear norm-based matrix regression (NMR) and the alternating direction method of multipliers (ADMM) to estimate the regression coefficients.

3 Proposed System

The present system is mostly based on license plate number recognition using video surveillance, but it has several limitations. The systems based on OCR are highly time-consuming lacks any facial identification for crime detection. The system proposed in this research paper is highlighted in Fig. 1 for observing and supervising real time traffic in the parking lots of private and public organizations, airports, toll booths, and border crossings. Apart from identifying any stolen vehicles, it is highly proficient to track escaping criminals. With minimal processor units and cameras, the system can be installed with much ease. A 24/7 all-time working solution can eliminate human dependency and prevent any crime in the case of negligence also. It can also help keep evidence of crime by capturing images and be used in future investigations. Other benefits of the proposed system include fake number plate detection, tracking criminal activities and better surveillance.

The proposed system has certain limitations in terms of light illumination, masked face identity or makeup, predicting accuracy, facial emotions, and expressions.

3.1 Methodology: Algorithms and Implementation

3.1.1 Text Recognition

K-Nearest Neighbors (KNN) Algorithm

KNN is used for classification of characters present on the number plate. In classification, the class of an object is calculated using the class of its neighbors as

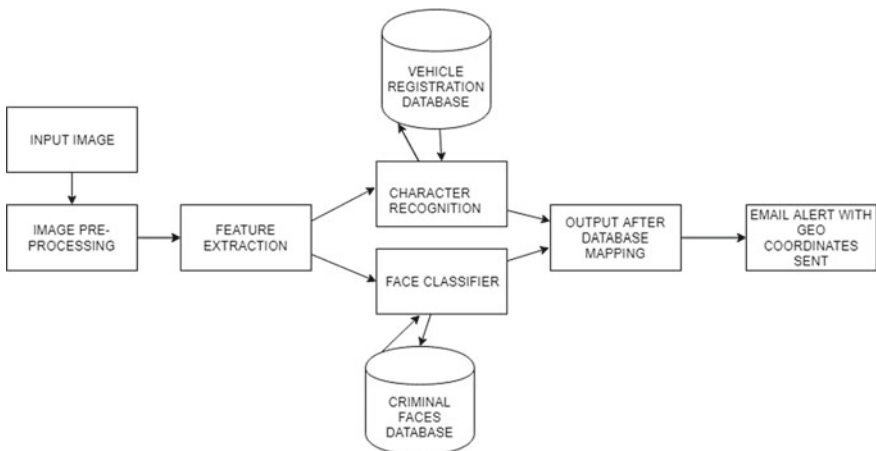


Fig. 1 Proposed method working flow diagram

shown in Fig. 3. It involves two main steps: 1. Character Segmentation [13] and 2. Character Recognition [14]. Characters including alphabets and digits are trained and then recognized by extracting features from the image. The characters are segmented first and then recognized by the algorithm as shown in Fig. 2 (Table 1).

Number Plate Recognition: Steps

Text extracted from the number plate comprises of various steps:

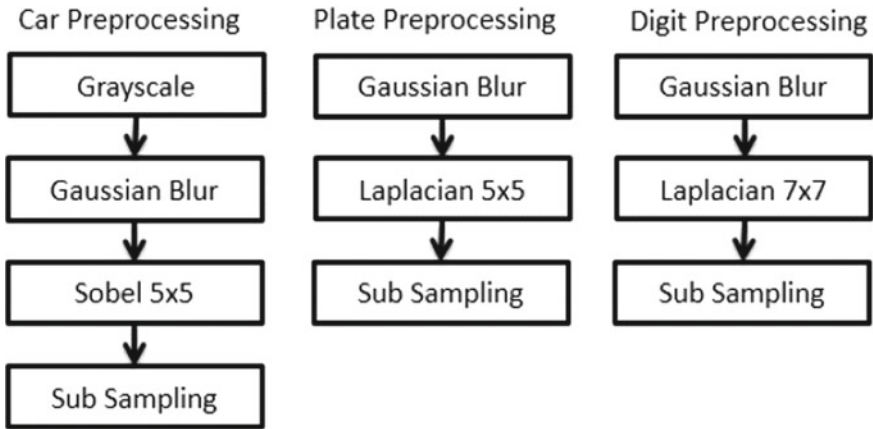


Fig. 2 KNN algorithm flowchart of steps in text recognition

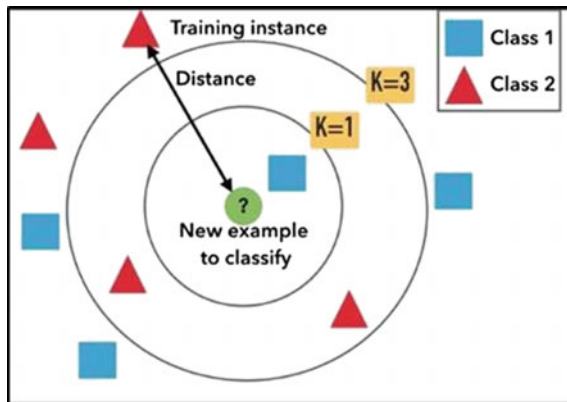


Fig. 3 Green circle is the sample for which classification has to be performed. It has to be assigned either of the two classes, i.e., either class 1 of blue squares or class 2 of red triangles. With $k = 1$, class 1 is assigned to it as its 1 nearest neighbor is a blue square. Now when $k = 3$, class 2 is assigned to the sample because there are 2 triangles in comparison to only 1 square inside the inner circle. Source https://cdn-images-1.medium.com/max/800/0*Sk18h9op6uK9EpT8

Table 1 k-NN Algorithm comparison

Pros	Cons
Output is not affected by those sets that are outside the selected boundary	Storing all the training data makes this algorithm computationally expensive
Assumptions are not made by KNN	More memory is required
One of the simplest algorithms	If the value of k is high, then output may be slow
Accuracy is high but not as high as some of the other better supervised learning algorithms	Output is affected by unrelated features and noise

- *Step 1* Capturing the image from a camera.
- *Step 2* Converting an original image to a grayscale [15] image.
- *Step 3* Converting the grayscale image to binary image and applying morphological transforms using OpenCV library for the ease of detection of a license plate from the image.
- *Step 4* Applying Gaussian blur filter [16] to smoothen the image.
- *Step 5* Listing all contours in the image to remove the boundary of the image that has the same color of intensity.
- *Step 6* Applying KNN algorithm to list all matching characters [17] in the image.
- *Step 7* Listing all possible plates that can be found in the image.
- *Step 8* Selecting the highest probability [18] license plate that is matched from the standard license plate measurement and extracting it.
- *Step 9* Reapplying/Repeating steps 2 to 8 to the number plate after extracting it so as to recognize the characters in the number plate. Recognition [19, 20] is done with the help of the KNN algorithm with the training data set. Recognized characters with KNN results in the output [6].

Face Recognition

For face recognition [21], dlib and OpenCV are used to spot all facial landmarks. Relevant regions of the face are represented by facial landmarks [22]. These regions are:

- Eyes
- Nose
- Jaw
- Mouth
- Eyebrows.

Detection of Facial landmarks is a subset problem of the shape prediction problem. In shape prediction problem, the aim is to identify and localize important points of interest.

In the problem of face recognition, facial landmarks are localized by identifying key points in the face. With problems such as face alignment, head pose estimation, face swapping, blink detection, etc., facial landmarks have been effectively used.

For detecting facial landmarks, the following two steps are used:

- *Step 1* In the image face is identified and localized.
- *Step 2* Key facial structures are detected in the face.

Step 1 Different ways of performing Face detection (Step 1):

- OpenCV's built-in Haar cascades.
- Pre-trained HOG + Linear SVM object detector.
- Deep learning-based algorithms designed for face localization.

Irrespective of the algorithm used, our main aim in this step is to find a closed bounding box around the detected face. This is achieved by finding the set of (x, y) coordinates that bound the face.

Step 2 No matter which face landmark detector is used the basic thing they all do is localizing key facial regions. These regions include:

- Right eyebrow
- Left eyebrow
- Right eye
- Left eye
- Mouth
- Nose
- Jaw
- In this method, images in which facial landmarks are manually labeled are used as training images. These specify $(x-y)$ coordinates of regions surrounding each facial structure.
- Without feature extraction [23], from pixel intensities, the facial landmark positions are projected and a group of regression trees is trained.
- All this results in creating a detector that can detect facial landmarks in real time with good efficiency.

dlib's facial landmark detector.

In the proposed system, the dlib's facial landmark detector is used. It estimates the location of 68 points that locates various facial structures in the human face.

Figure 4 demonstrates the indexes of the 68 coordinates of the human face.

3.1.2 Face Recognition: HAAR Faces Classifier and SVM

Considering face recognition, the complete process can be divided into three main steps:

1. The first step involves finding a decent database of faces which contain several images of each person.



Fig. 4 Picturing the 68 facial landmark coordinates. Source <https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs169436>

2. The second step is to find faces in the database images and make use of them to get face recognizer trained.
3. The last step is to check face recognizer whether it identifies faces it was trained for or not.

Database

The database is created using webcam capturing images of an individual (Fig. 5).

Face Detection Classifiers

A classifier [24] is a computer program that finds out whether an image is a positive image (face image) or negative image (non- face image). We are using OpenCV [25] for this purpose.

Every file starts with the name of the classifier, it belongs to. For example, a Haar cascade classifier starts off as *haarcascade_frontalface_alt.xml*.

Haar Classifier

The Haar Classifier [4, 5] is an algorithm made by Paul Viola and Michael Jones which uses a machine learning-based approach.

It gets started by digging out Haar features [5] from each image as shown in Fig. 6.

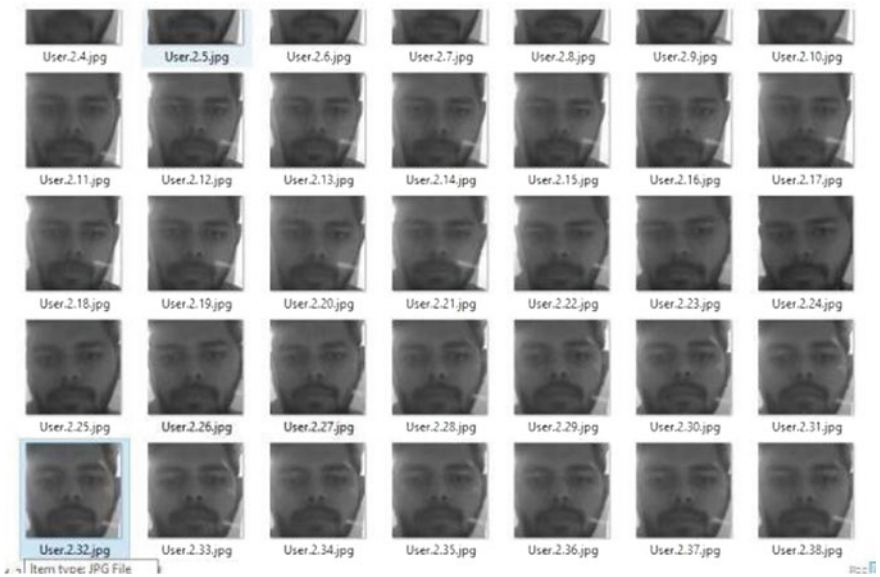


Fig. 5 Image database consisting of 50 images of individual people. Each image has an altered facial expression [JPG]

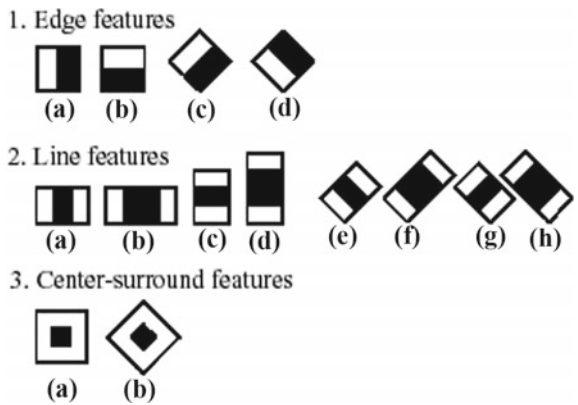


Fig. 6 A couple Haar-like edge, line and center-surround features. *Source* https://docs.opencv.org/2.4/_images/haarfeatures.png

A particular feature is determined by placing each window on the picture. This feature is denoted by a single value which is calculated by subtracting the sum of pixels below the white part of the window from the sum of pixels below the black part of the window [26, 27].

Now, plenty of features are determined by placing all possible sizes of each window on all possible locations of each image. In the end, generally the fact which is considered by the algorithm is non-face region [28, 29] covers most of the area in an image:

- Step 1: Adjusting Tolerance/Sensitivity [30].
- It is done with the tolerance parameter [31]. The default tolerance value is 0.6 and lower the numbers, stricter become face comparisons.
- Step 2: All the faces in an image are automatically found.
- Step 3: The facial features of a person in an image automatically located.
- Step 4: Faces in images are recognized and identified who they are.

4 Result Analysis

The following results were produced using Python programming language. Every detailed step from grayscale conversion to possible vector contours for number plate recognition is shown in the Fig. 9 for the number plate highlighted Fig. 7. Figure 8 shows the predicted output produced on the screen. The result of face recognition is shown in Fig. 10 with the identified faces marked with red squares and the respective names under it. The screen displays Unknown in case the name is not present in the database. The results are much accurate compared to real time objects when tested with different images.



Fig. 7 Actual image of number plate

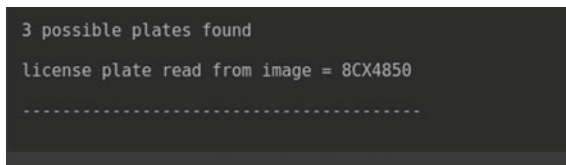


Fig. 8 Predicted output after recognition in python

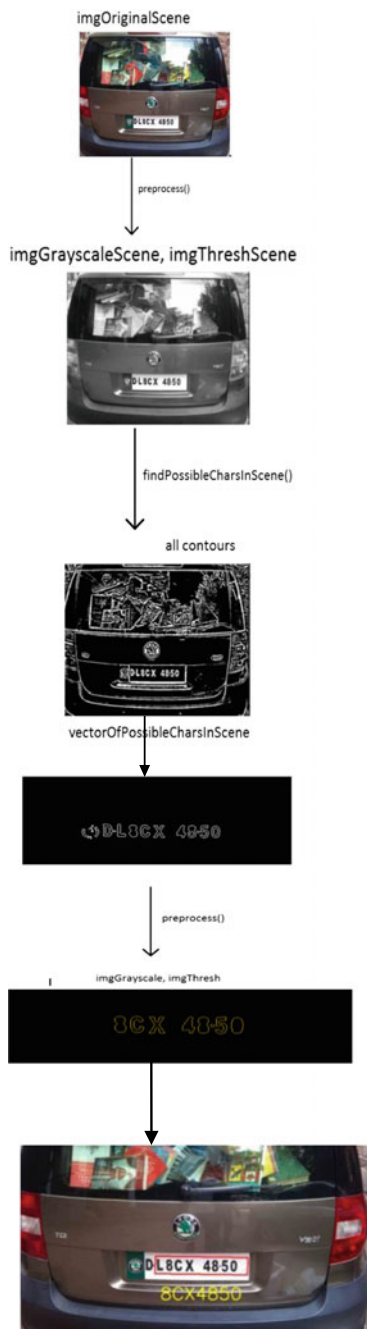


Fig. 9 Procedural steps with the final output

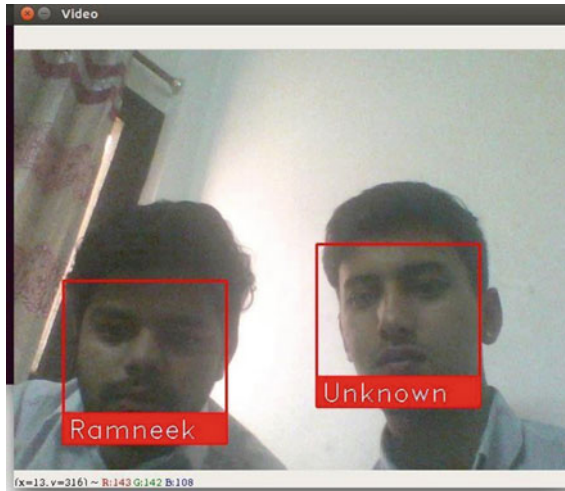


Fig. 10 Final output window of real time face recognition based on image dataset

5 Conclusion

The system for crime detected is improved with much more accurate results and higher efficiency. This system is complete and sophisticated in terms of crime detection. It uses the features of both text and face recognition. Segmenting the characters of the license plate and the English letters and numerals are trained as per the requirement of Indian number plate. Using Gaussian Blur filter in KNN algorithm makes it efficient to deal with all types of images for number plate detection irrespective of noise, intensity, and other factors. It generates multiple plate combinations and chooses one with the highest probability. Similarly, for face recognition, the self-built database is predicting much accurate results in terms of lighting conditions and the movement of persons. The localized facial features based on sensitivity and tolerance parameter in Haar Classifier can also be changed as per requirements. Also, the present values are fit enough to predict well in almost all conditions. The final implementation involves an effective tracking module in the form of a Python Email Script that sends an alert message to the concerned authorities and the nearby police stations and geo-coordinates of the place where the criminal activity took place including the number plate and criminal name detected after matching with their respective databases. The proposed system working with an accuracy of more than 85% successful detections in ambient light conditions.

6 Future Scope

For higher precision and accurate model, certain new features and techniques need to be incorporated to identify faces in all possible conditions of light, barriers in front of the face and aging of facial features with time. An additional phase of the exploration involves, developing a generic model to be installed at all traffic lights in the cities for operating it in dynamic moving traffic.

Acknowledgements This research project is a sole work of the authors mention in this paper and no organization has participated in its contribution. All experimentation performed involving human participants were in accordance with the ethical standards and their consent. There is no violation of copyrights in use of the images of the vehicle number plate. The authors declare that they have no conflict of interest [32, 33].

References

1. Du, S., et al.: Automatic license plate recognition (Alpr): a state- of-the-art review. *IEEE Trans. Circuits Syst. Video Technol.* **23**(2), 311 (2013)
2. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: *Proceedings of the IEEE*, pp. 586–591 (1991)
3. Bachhety, Shivam, Singhal, Ramneek, Rawat, Kshitiz, Joshi, Kunal, Jain, Rachna: Crime detection using text recognition and face recognition. *Int. J. Pure Appl. Math.* **119**(15), 2797–2807 (2018)
4. Do, H.N., Vo, M.T., Vuong, B.Q., Pham, H.T., Nguyen, A.H.: Automatic license plate recognition using mobile device. In: *2016 International Conference on Advanced Technologies for Communications (ATC)*, pp. 268–271. IEEE (2016)
5. Gou, C., Wang, K., Yao, Y., Li, Z.: Vehicle license plate recognition based on extremal regions and restricted Boltzmann machines. *IEEE Trans. Intell. Transp. Syst.* **17**(4), 1096–1107 (2016)
6. Radmanesh, A.: A real time vehicle’s license plate recognition system. In: *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS’03)*, vol 4, pp. 159–167 (2015)
7. Wei-gang, Z., et al.: A study of locating vehicle license plate based on color feature and mathematical morphology. In: *6th International Conference on Signal Processing*, vol 1, pp. 748–751 (2002)
8. Osuna, E., Freund, R., Girosi, F.: Training support vector machines: an application to face detection. In: *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, June, pp. 130–136, (1997)
9. Schneiderman, H., Kenade, T.: Probabilistic modeling of local appearance and spatial relationships for object recognition. In: *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, pp. 45–51 (1998)
10. Saraswathi, M., et al.: Evaluation of PCA and LDA techniques for face recognition using ORL face database. *Int. J. Comput. Sci. Inf. Technol.* **6**(1), 810–813 (2015)
11. GPU Based Face Recognition system for authentication, Bhumika Agrawal, et.al, © 2017 IJEDR|Volume 5, Issue 2|ISSN: 2321–9939
12. Yang, J., Luo, L., Qian, J., Tai, Y., Zhang, F., Xu, Y.: Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(1), 156–171 (2017)

13. Zhai, X., Bensaali, F., Sotudeh, R.: OCR-based neural network for ANPR. In: 2012 IEEE International Conference on Imaging Systems and Techniques Proceedings 2012 Jul 16, IEEE, Jul 16, pp. 393–397 (2012)
14. Chellappa, R., Ni, J., Patel, V.M.: Remote identification of faces: problems, prospects, and progress. *Pattern Recogn. Lett.* **33**, 1849–1859 (2012)
15. Yetirajam, M., Nayak, M.R., Chatopadhyay, S.: Recognition and classification of broken characters using feed forward neural network to enhance an OCR solution. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **1** (2012)
16. Almudhahka, N.Y., Nixon, M.S., Hare, J.S.: Automatic semantic face recognition. 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), Washington, DC, 2017, pp. 180–185
17. Lund, W.B., Kennard, D.J., Ringger, E.K.: Combining multiple thresholding binarization values to improve OCR output. Presented in document recognition and retrieval XX Conference 2013, California, USA, 2013. USA: SPIE (2013)
18. Satti, D.A.: Offline Urdu Nastaliq OCR for printed text using analytical approach. MS thesis report Quaid-i-Azam University: Islamabad, Pakistan. p. 141 (2013)
19. Bhansali, M., Kumar, P.: An alternative method for facilitating cheque clearance using smart phones application. *Int. J. Appl. Innov. Eng. Manag. (IJAIEM)* **2**(1), 211–217 (2013)
20. Roy, A., Ghoshal, D.P.: Number plate Recognition for use in different countries using an improved segmentation. In: 2nd National Conference on Emerging Trends and Applications
21. Zhang, K., Zhang, Z., Li, Z.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**(10), 1499–1503 (2016)
22. Anagnostopoulos, C.N.E., Anagnostopoulos, I.E., Loumos, V., Kayafas, E.: A license plate-recognition algorithm for intelligent transportation system applications. *IEEE Trans. Intell. Transp. Syst.* **7**(3), 377–392 (2006)
23. Erdinc Koer, H., Kursat Cevik, K.: Artificial neural networks-based vehicle license plate recognition. *Proced. Comput. Sci.* **3**, 1033–1037 (2011)
24. Bartlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face recognition by independent component analysis. *IEEE Trans. Neural Netw.* **13**(6), 1450–1466 (2002)
25. Mita, T., Kaneko, T., Hori, O.: Joint Haar-like features for face detection. In: IEEE International Conference on Computer Vision, pp. 1550–5499 (2005)
26. Almudhahka, N.Y., Nixon, M.S., Hare, J.S.: Automatic semantic face recognition. In: 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), Washington, DC, pp. 180–185 (2017)
27. Chen, Z., Huang, W., Lv, Z.: Towards a face recognition method based on uncorrelated discriminant sparse preserving projection. *Multimed. Tools Appl.* **76**, 17669 (2017). <https://doi.org/10.1007/s11042-015-2882-0>
28. Dillak, R.Y., Dana, S., Beily, M.: Face recognition using 3D GLCM and Elman Levenberg recurrent neural network. In: International Seminar on Application for Technology of Information and Communication (ISemantic), pp. 152–156 (2016)
29. Winarno, E., Hadikurniawati, W., Rosso, R.N.: Location based service for presence system using haversine method. In: 2017 International Conference on Innovative and Creative Information Technology (ICITech), pp. 1–4 (2017)
30. Qi, X., Liu, C., Schuckers, S.: CNN based key frame extraction for face in video recognition. In: 2018 IEEE 4th International Conference on Identity Security and Behavior Analysis (ISBA), pp. 1–8 (2018)
31. Deshpande, N.T.: Face detection and recognition using viola- jones algorithm and fusion of PCA and ANN. *Adv. Comput. Sci. Technol.* **10**(5), 1173–1189 (2017)
32. Zohra, F.T., Gavrilov, A.D., Duran, O.Z., Gavrilova, M.: A linear regression model for estimating facial image quality. In: 2017 IEEE 16th International Conference on Cognitive Informatics and Cognitive Computing (ICCI*CC), pp. 130–138 (2017)
33. Dhawanpatil, T., Joglekar, B.: Face spoofing detection using multiscale local binary pattern approach. In: 2017 International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–5 (2017)

Analysis of Classification Algorithms for Breast Cancer Prediction



S. P. Rajamohana, K. Umamaheswari, K. Karunya and R. Deepika

Abstract According to global statistics, breast cancer is the second of all the fatal diseases that cause death. It will cause an adverse effect when left unnoticed for a long time. However, its early diagnosis provides significant treatment, thus improving the prognosis and the chance of survival. Therefore, accurate classification of the benign tumor is necessary in order to improve the living of the people. Thus, precision in the diagnosis of breast cancer has been a significant topic of research. Even though several new methodologies and techniques are proposed machine learning algorithms and artificial intelligence concepts lead to accurate diagnosis, consequently improving the survival rate of women. The major intent of this research work is to summarize various researches done on predicting breast cancer and classifying them using data mining techniques.

Keywords Breast cancer prediction · Classification · Decision tree · Feature selection · K-NN · Random forest · SVM

1 Introduction

A malignant tumor developed from breast cells is termed as breast cancer, and it is the most commonly observed type of tumor in women. Almost 70% of deaths due to cancer arise in poverty-stricken and developing countries. Breast cancer affects about

S. P. Rajamohana (✉) · K. Umamaheswari · K. Karunya · R. Deepika
Department of Information Technology, PSG College of Technology,
Coimbatore 641004, India
e-mail: monamohanasp@gmail.com

K. Umamaheswari
e-mail: umakpg@gmail.com

K. Karunya
e-mail: karunyakandimuthu@gmail.com

R. Deepika
e-mail: deepika.rajendran97@gmail.com

10% of women at some point during their lives. As of late, the death rate continues to expand, with 88% of survival following 5 years from determination and 80% following 120 months from conclusion. There exists a popular method, called knowledge discovery in databases (KDD) which is used by medical researchers to predict the disease outcome by discovering patterns and relationships among the various number of variables stored in the database as historical data. In order to maintain the natural working mechanism of the body, a balance must be maintained between the growth and death rate of cells. But sometimes there is an abnormal and rapid growth of cells which can lead to cancer. Among numerous types of cancer, breast cancer is the major cause for the death of the women worldwide [1]. It is proved that there is a reduction of 38–48% in the mortality rate. There might be several reasons for breast cancer, including age, bosom density, obesity, and changing food habits like alcohol, causing undesirable effects. It is evident from the recent statistics tha, currently the scenario has become worse. The major symptom is a lump observed underarm or in the mammary gland that continues to exist even after the menstrual cycle of a woman. Out of the lump formed, most of them are painless, some may give prickly sensation. The other common symptoms are redness, swollen lymph nodes, or thickening or puckering off the skin. Milk secreting and duct cells are the main cells that cause tumors by draining the milk into the nipple from lobules. A proportion of growth developed from fibrous or fatty tissue can lead to breast cancer. This kind of gene mutation related to cancer is very popular among woman. The treatment for breast cancer includes chemotherapy, hormonal therapy, and radiation therapy. The efficiency result of treatment is comparatively low and requires more attention for preventive measures and control in the current research world. Many breast cancer charities are organizing campaigns such as National Breast Cancer Awareness Month (NBCAM) on the eighth month of every year to bring about awareness of breast cancer in society. This annual international health campaign assists in raising funds for research in breast cancer.

Breast Cancer is categorized as follows:

- (i) **Benign (Noncancerous)** The benign cases are those which are noncancerous and non-life threatening. But they could turn into cancerous cells easily. However, these cells could be easily separated.
- (ii) **Malignant (Cancerous)** Malignant case leads to abnormal cell growth invading nearby tissues. It is life-threatening in most cases.

Many researchers proposed various data mining algorithms that are deployed for the diagnosis of breast cancer. Among which, feature extraction and classification algorithms employed in it lead to the design of an efficient system. These techniques provide a significant process for extracting the key features which can lead to a proper diagnosis. It is experimentally proven that machine learning and deep learning algorithms are efficient when compared to conventional approaches [2].

2 Breast Cancer Overview

The twentieth century is also called as the cancer century because hundreds of cancer types were discovered in this century. After decades of hard work in analyzing various types of cancers, doctors are now able to identify the causes of these diseases, preventive measures to adopt, and type of treatment to be given. Among all types of cancers existing, breast cancer is rampant among women, very rarely in men. Heredity is a major factor among multiple factors that can cause breast cancer. Nearly 15–20% of women identified with breast cancer has had a recorded occurrence of the same through their generations. In extremely rare cases, a gene called p53 is responsible for breast cancer. The severity of breast cancer was 16 times more than average in families having this type of gene. The number of families with this gene is about 100 all over the world. Researchers have noticed the double risk of breast cancer in individuals producing wet wax of ear glands than in those producing dry wax. Some of the most common symptoms and types are listed in Tables 1 and 2.

Table 1 Breast cancer factors with symptoms

Breast cancer risk factors	Symptoms
Age	<ul style="list-style-type: none"> • Redness • Swollen lymph nodes • Thickening or puckering of the skin • Scaling of nipple • Dimpled skin • Change in the texture of the skin • Breast or nipple pain
Family history	
Genetics	
Breast cancer history	
Exposed to radiation in the chest or face below the age of 30	
Race/ethnicity	
Being overweight	
Menstrual history	
Drinking alcohol	

Table 2 Breast cancer types

Types of breast cancer	Description
Metastatic breast cancer	Cancer cells break and spread from the original tumor to other parts by means of the lymphatic system or through the blood
Phyllodes tumors of the breast	These are rare case tumors which require surgery to reduce the risk as they can grow rapidly fast. Phyllodes tumors are malignant and borderline benign
Lobular carcinoma in situ	In this type, the growth of cells in milk producing glands is abnormal and increases the person’s risk of developing invasive breast cancer
Inflammatory breast cancer	Instead of lump, this type of cancer is found with swelling and reddening in the area of the breast and spreads rapidly to other areas with symptoms worsening
Invasive lobular carcinoma	Also known as infiltrating lobular carcinoma. It starts in duct cells of milk carrying area and spreads beyond it

3 Related Works

There are several studies on breast cancer prediction based on machine learning algorithms.

A. A Novel Approach for Breast Cancer Detection Using Data Mining Techniques [2014]

Chaurasia and Pal [3] compared the performance of the supervised learning classifiers like Naive Bayes, decision tree, and SVM-RBF kernel and simple CART. The dataset that is used for classification is the Wisconsin breast cancer dataset that comprises 11 attributes. The attributes include sample code number, clump thickness, uniformity of cell size, cell shape, and marginal adhesion. The experiment's results proved that the SVM-RBF kernel has the maximum accuracy of about 96.84.

B. Artificial Neural Networks Applied to Survival Prediction in Breast Cancer [2000]

Lundin et al. [4] constructed the classification model using neural network and statistical analysis. The dataset that is used consists of the attributes like age, primary tumor size, axillary nodal status, mitotic count, and tumor necrosis. The dataset contains the details of 951 breast cancer patients, predicted 5-, 10-, and 15-year breast cancer by using ANN and logistic regression models. The accuracy obtained from the constructed model is about 94.35% which is a considerable accuracy.

C. Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis [2009]

Choi et al. [5] constructed the classification model using artificial Bayesian network. The dataset that is used consists of the attributes such as age at diagnosis, a clinical extension of the tumor, and the number of primary tumors. The dataset consists of the details of 2, 94,275 breast cancer patients. This dataset uses 15 attributes of which seven are primary attributes, seven derived, and one target attribute. The accuracy obtained from the constructed model using the artificial Bayesian network is 88.8%.

D. Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods [2004]

Delen et al. [6] refined a prediction model with the help of ANN, decision tree, and logistic regression to envision breast cancer and to determine the survival rate by analyzing the SEER cancer incidence database. The SEER breast cancer dataset consists of 4, 33,272 records and 72 variables. The accuracy obtained from the model is 89.2%.

E. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence [2013]

Ahmed et al. [7] analyzed the use of decision trees, SVM, and artificial neural network on breast cancer prediction. The dataset used contains population characteristics and includes 22 input variables. The sample dataset is the data collected from 1189 women. The experimental results found that the SVM produced the least error rate with greater accuracy of 95.7%.

F. Breast Cancer Prediction Using DT-SVM Hybrid Model [2015]

Sivakami [8] developed a model for the breast cancer prediction which uses the Wisconsin breast cancer dataset. The dataset encompasses 699 records, containing 458 belonging to the benign class and the remaining to malignant class. The data is collected from the needle extracts of patients' breasts. Before the prediction process, preprocessing must be done. Preprocessing fills up missing values with attribute's mean value in the dataset. The nine traits of the dataset include uniformity in size of the cell, the shape of cell, and thickness of clump. The implemented algorithm in this work is based on SVM and decision tree. The accuracy obtained from the classification model is 91%, and the error rate is 2.58. The number of precisely classified instances is 459 and imprecisely classified instances are 240.

G. Combining Bagging and Boosting [2007]

Kotsiant et al. [9] worked on various ensemble approaches such as bagging, boosting, and combination of both with a variety of base learners. They are C4.5, Naïve Bayes, OneR, and decision stump. These algorithms used datasets from the UCI repository dataset. The accuracy obtained is about 93.47%.

H. Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors [2017]

Islam et al. [10] used the most common dataset, Wisconsin breast cancer (WBC) dataset obtained from the repository of UCI machine learning. This dataset consists of 699 instances, where the cases are labeled as either benign or malignant. Classification is performed using SVM and K-NN. The accuracy obtained from the model using SVM and K-NN is 98.57% and 97.14%, respectively.

I. Heterogeneous Classifiers Fusion for Dynamic Breast Cancer Diagnosis Using Weighted Vote-Based Ensemble [2014]

This work proposed an ensemble approach by combining various classifiers including decision tree with Gini index and information gain, Naïve Bayes, SVMn and memory based learner. The classification is decided on the basis of weighted voting. Different datasets are collected from the public repository. Classification accuracy is enhanced using different preprocessing techniques and feature selections. The experimental results conclude that proposed approach contributed to major improvement than other existing classifiers. The accuracy obtained from the ensemble model is about 97.2%, the precision is 100%, and the recall value is 98.60%.

4 Proposed Methodology

4.1 Data Source

The publicly available breast cancer database is used. The database [11] constitutes about 570 records. The dataset consists of numeric attributes. Most of the research papers referred to the 32 attributes present in the dataset for breast cancer prediction.

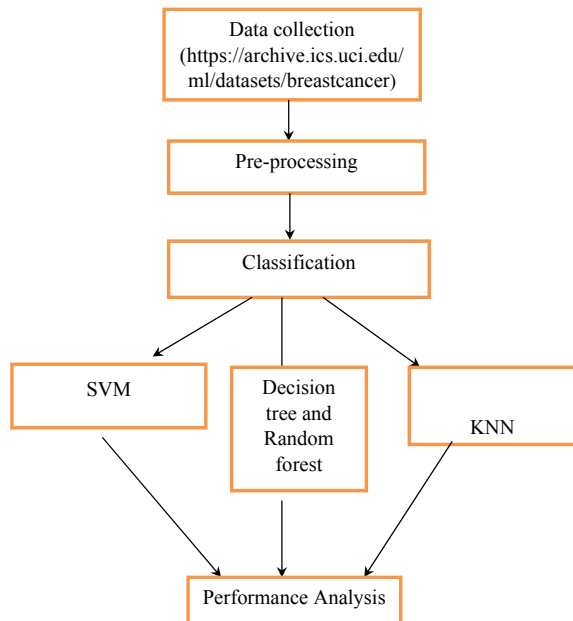
4.2 Data Preprocessing

The dataset applied in the proposed work is a clinical dataset which may contain many inconsistent, incomplete, or missing data. Such data reduces the accuracy of the model. Hence, preprocessing is an important step that needs to be carried out. To remove the inconsistencies in the dataset, several preprocessing techniques were applied. Also, normalization techniques were applied to handle the missing data. Thus, preprocessing techniques tend to increase the accuracy of the model constructed. The workflow of the proposed model is shown in Fig. 1.

4.3 Random Forest

For the classification of malignant and benign types of cancer, random forest algorithm can be employed. Based on various types of randomization, random forest has

Fig. 1 Workflow of the proposed model



been built as it is an ensemble of decision trees. As the random forests are very flexible rather it is widely used [12]. This supervised learning algorithm creates forests using many trees. The accuracy depends on the number of trees. One of the major advantages with random forest is that it could be functioned as both regression and classification. It can also handle the missing values and it will not overfit in case of a greater number of trees [13]. The random forest takes the test features and predicts the outcome of the randomly created trees based on rules and then stores the result. The votes for each predicted target is calculated, since each tree results in different prediction. Finally, the target receiving high vote will be considered as the final prediction. The random forest processes a huge amount of data at very high speed. The random forest has each tree in binary structure form, which is created based on top-down approach. Generally, the convergence in the random forest algorithm is very fast. The most important parameters are the depth and number of trees. Increasing depth led to an increase in the performance [14]. Thus, random forest is considered as the best classification algorithm on the basis of the processing time and accuracy. This algorithm is implemented on the dataset and the accuracy obtained is 97.34%. The overall view of the random forest algorithm is laid out in Fig. 2.

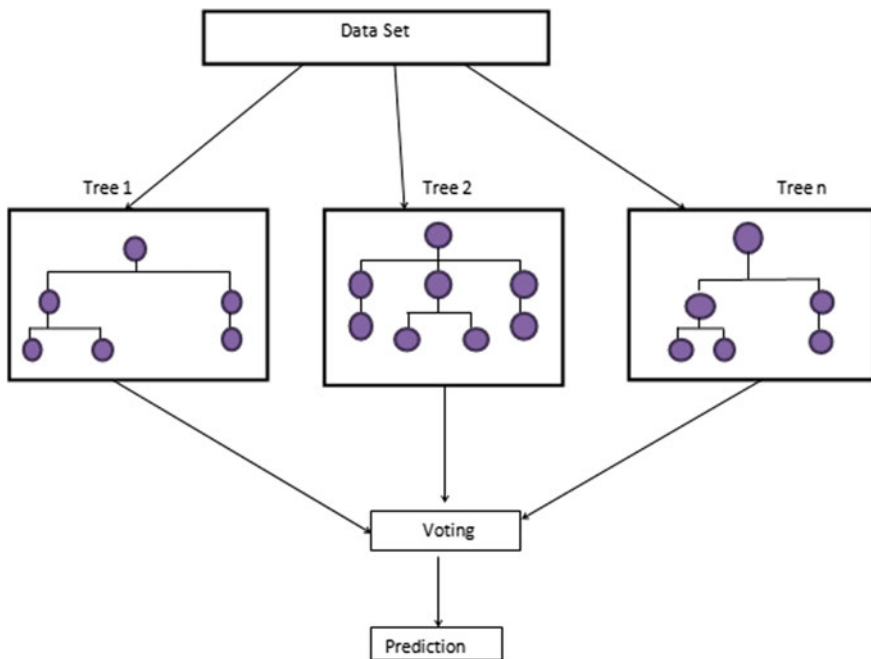


Fig. 2 Random forest

4.4 Support Vector Machine (SVM)

Another powerful supervised learning model is the support vector machine that examines the data for classification and regression analysis using the support of associated learning algorithms. SVM has been specialized for a certain range of problems and gained success in the field of pattern recognition specifically in bioinformatics and cancer diagnosis. In support vector machine, all the data points are drafted in an N-dimensional space [15]. SVM performs linear classification. In addition to this, it also performs nonlinear classification by mapping the inputs implicitly at the range of high-dimensional feature space. The main advantage of SVM is considered to be a unique technique called kernel trick, where lower dimensional space is converted to higher dimensional space and classified. In other words, a hyperplane or set of hyperplanes can be constructed with support vector machine constructs in a high- or infinite-dimensional space for the purpose of classification, regression, or outlier detection [16]. In general, the hyperplane with the largest functional margin achieves good separation as it lowers the generalization error of the classifier. Support vector machine gives an accuracy of 97% when employed on the dataset. The hyperplane used for classification is illustrated in Fig. 3.

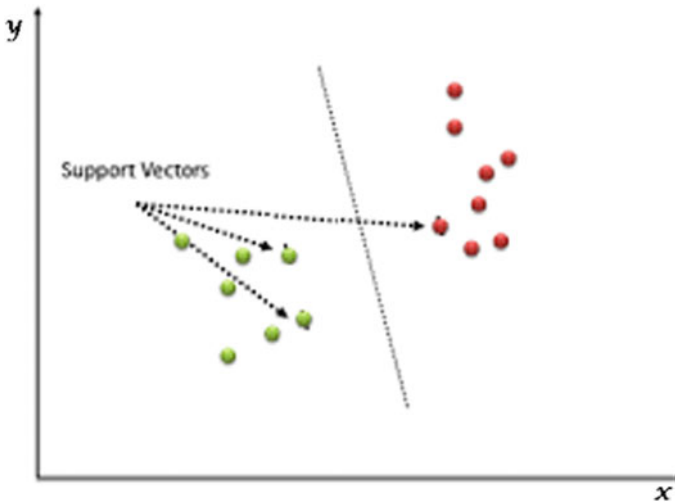


Fig. 3 Support vector machine

4.5 *K-Nearest Neighbor*

A nonparametric method, K-nearest neighbors algorithm (K-NN) is another algorithm to be utilized for various purposes like regression and classification. The output of the classification from k-NN is a class label which describes to which class or group it belongs to [17]. In K-NN, the membership is assigned based on the majority vote by the neighbors which is decided by the K value. In other words, each object has been assigned to one class that is most common among the neighbors. For example, if $K = 2$, then the query point is assigned to the class where two nearest neighbors belong to. Being the simplest machine learning algorithm, the explicit training step is not essential. In the training step, neighbors are chosen from the set of objects for which their corresponding classes are known. The algorithm is very sensitive to the local data [18]. Euclidean distance for continuous variables and Hamming distance for discrete variables are most widely used. However, the accuracy can be improved by using specialized algorithms like Large Margin Nearest Neighbor or Neighborhood Components Analysis. The K value is chosen based on the data. Choosing an appropriate K value is significant because that decides if the data is classified correctly. Heuristic techniques like hyperparameter optimization are used because the larger K value reduces the effect of noise but makes less distinct boundaries between classes [19]. The performance gets degraded as the noise in the data increases. The level of accuracy achieved from K-NN is about 95% with an appropriate K value. K-NN is shown in Fig. 4.

4.6 *Decision Trees*

These are classification algorithms where the attributes in the dataset are recursively partitioned. Decision trees contain many branches and leaf nodes. All the branches tell the conjunction of the attributes that leads to the target class or class labels [20].

Fig. 4 K-nearest neighbor

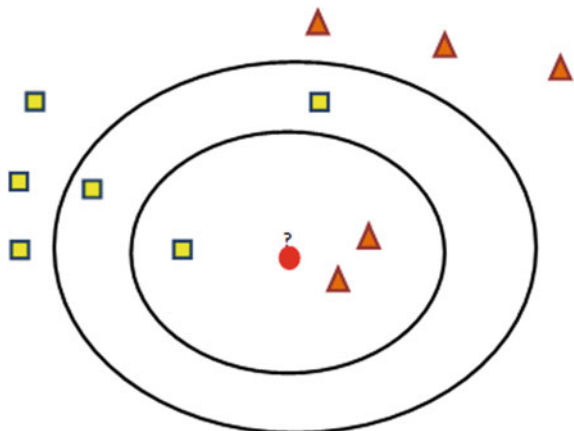
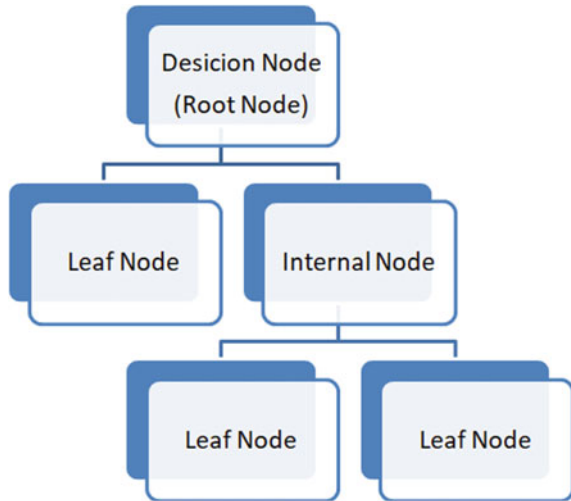


Fig. 5 Decision tree

The leaf nodes contain class labels or the target class that tells to which class tuple it belongs to [21]. There are various decision tree algorithms which can be used for classification of the data. Some algorithms include C4.5, C5, CHAID, ID3, J48, and CART [11, 22].

The Decision Tree can be built as

- The attribute splits decide the attribute to be selected.
- The decisions whether to continue for splitting or to represent as a terminal node is made.
- The assignment of the terminal node to a class.

Information gain, gain ratio, Gini index, etc. are the impurity measures to decide attribute splits done on the tree. After pruning, the tree is checked against noise and overfitting. As a result, the tree becomes an optimized tree. The main advantage of having a tree structure is that it is very easy to understand and interpret. The algorithm is also very robust to the outliers also. The structure of the decision tree is shown in Fig. 5.

5 Experimental Results

Initially, the entire dataset was preprocessed and normalized to handle all the missing values. The mandatory features are extracted using statistical methods. Similar to the experimental results shown in the table above, for the breast cancer prediction, random forests can give a better result that helps the people to get a systematic way of treatment and save their life and hence decreasing the mortality rate. Support vector machines are also equally good with an accuracy of about 91%.

Table 3 Performance analysis of various machine learning algorithms

Algorithm	Accuracy (%)	Precision (%)	Recall (%)
Support vector machine	91	82	81
Decision trees	90.6	83	79
K-nearest neighbor	87	78	76.8
Random forests	93.34	89	86.3

The decision tree and K-NN also give us a good result in predicting breast cancer with an accuracy of 90.6% and 87%, respectively. Hence, a good and accurate model can be built using the algorithms mentioned above. In conclusion, random forest has proven that it is very efficient for breast cancer prediction with an accuracy of 93.34% for diagnosis is exhibited in Table 3. It achieves the finest performance with respect to precision and recall.

6 Conclusion

Most of the Indian women die due to breast cancer [23]. Proper diagnosis of breast cancer is very important in the medical domain. Many models are built using machine learning and data mining methodologies to analyze huge voluminous data. But the challenge with the model is its accuracy and precision for the medical industry. The key for proper treatment and cure is significantly dependent on detecting breast cancer at its early stages [24]. This paper describes how several machine learning algorithms like random forests, decision trees, K-NN, and SVM are employed to model the exact diagnosis of breast cancer for an organized treatment which can save lots of life. The experimental results show the effectiveness of various machine learning algorithms and it proves to be efficient in breast cancer classification thus providing a major breakthrough in clinical diagnosis and treatment of the same. As mentioned earlier, the SVM, K-NN, decision tree, and random forest have their own advantages. Hence, the careful usage of the above will definitely lead to better results. Thus, the appropriate use of a proper algorithm eliminates the risk of death and alleviates the survival rate in women.

References

1. Ferlay, J., Héry, C., Autier, P., Sankaranarayanan, R.: Global burden of breast cancer. In: Li, C. (ed.) *Breast Cancer Epidemiology*, pp. 1–19. Springer, New York (2010)
2. Sharma, A., Kulshrestha, S., Daniel, S.: Machine learning approaches for breast cancer diagnosis and prognosis. In: *Proceedings of the International Conference on Soft Computing and Its Engineering Applications*, Changa, India, 1–2 December (2017)

3. Chaurasia, V., Pal, S.: Data mining techniques: to predict and resolve breast cancer survivability. *Int. J. Comput. Sci. Mobile Comput.* **3**(1), 10–22 (2014)
4. Lundin, M., Lundin, J., Burke, H.B., Toikkanen, S., Pylkkänen, L., Joensuu, H.: Artificial neural networks applied to survival prediction in breast cancer. *Oncology* **57**(4), 281–286 (1999)
5. Choi, J.P., Han, T.H., Park, R.W.: A hybrid Bayesian network model for predicting breast cancer prognosis. *J. Korean Soc. Med. Inform.* **15**(1), 49–57 (2009)
6. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.* **34**(2), 113–127 (2005)
7. Ahmad, L.G., Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M., Razavi, A.R.: Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med. Inform.* **4** (124), 3 (2013)
8. Sivakami, K., Saraswathi, N.: Mining big data: breast cancer prediction using DT-SVM hybrid model. *Int. J. Sci. Eng. Appl. Sci. (IJSEAS)* **1**(5), 418–429 (2015)
9. Kotsianti, S.B., Kanellopoulos, D.: Combining bagging, boosting and dagging for classification problems. In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 493–500. Springer, Berlin (2007)
10. Islam, M.M., Iqbal, H., Haque, M.R., Hasan, M.K.: Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 226–229. IEEE
11. Barghout, L.: Spatial-taxon information granules as used in iterative fuzzy-decision-making for image segmentation. In: *Pedrycz, W., Chen, S.M. (eds.) Granular Computing and Decision-Making*, pp. 285–318. Springer, Cham (2015)
12. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 37–44. IEEE (2006)
13. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
14. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*, pp. 1–8. IEEE
15. Marcano-Cedeño, A., Quintanilla-Domínguez, J., Andina, D.: WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Syst. Appl.* **38**(8), 9573–9579 (2011)
16. TP, L., Parthiban, L.: Abnormality detection using weighed particle swarm optimization and smooth support vector machine. *Biomed. Res. (0970–938X)*, **28**(11) (2017)
17. Mirkes, E.: *KNN and Potential Energy (Applet)*. University of Leicester, Leicester (2011)
18. Breast Cancer Organization: <http://www.breastcancer.org/symptoms/>
19. Han, J., Pei, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Elsevier, Amsterdam (2011)
20. Devi, R.D.H., Devi, M.I.: Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer. *Int. J. Adv. Eng. Technol.* **VII**(II), 98 (2016)
21. Cuingnet, R., Rosso, C., Chupin, M., Lehericy, S., Dormont, D., Benali, H., Colliot, O.: Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. *Med. Image Anal.* **15**(5), 729–737 (2011)
22. 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), 1114–1119 (2013)
23. Wolberg, W.H., Mangasarian, O.: *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine
24. Blockeel, H., Struyf, J.: Efficient algorithms for decision tree cross-validation. *J. Mach. Learn. Res.* **3**, 621–650 (2002)

Real-Time Footfall Prediction Using Weather Data: A Case on Retail Analytics



Garima Makkar

Abstract Be it a retailer, producer, or supplier, the weather has a substantial effect on each one of them. Climate variability and weather patterns have become critical success factors in retail these days. As a matter of fact, weather forecasting has become a \$3 billion business now. One of the main reason behind this surge is the capability of the forecasters to sell weather-related information to businesses who then strategize their various decisions regarding inventory, marketing, advertising, etc. accordingly. Hence only those retailers who stay “ahead of the game” will be able to enjoy huge sales while others who do not would face the consequences. Various studies regarding change in consumer behavior occurring due to the change in weather conditions have shown that even a degree change in temperature affects the store’s traffic and reflect the growing importance of predictive analytics in this domain. However, these studies incorporate only the historical weather statistics into account. In this paper, we will propose our methodology for footfall analytics to see how the changes in weather conditions will impact the retail store’s traffic and thereby retailing value chain, using real-time weather forecasts and footfall data. This analysis provides a platform for retailers to make evidence-driven decisions and strategize their business plan which would help them to deepen the customer involvement and to get efficiency in the planning process.

Keywords Footfall · Retail · Retailers · Real time · Weather

1 Introduction

Climate decides what shall we eat and wear, where shall we spend our holiday, how shall we commute, and even what shall we do every single day. All in all, weather affects four basic purchasing decisions: where, when, what, and in what quantity to purchase [1]. Be it a producer, supplier, or retailer, the weather has a substantial

G. Makkar (✉)
Tata Consultancy Services, Bangalore, India
e-mail: garimamakkar93@gmail.com

effect on each one of them. From the availability of products and logistics to purchaser demand patterns and sudden spikes, the groundwork for and mitigating the consequences of weather is a real trouble in all aspects of retail sector. Weather conditions affect whether customer would visit brick and mortar outlets or would shop online, making this another important reason for retailers to regulate their operations in order to avoid retail risks associated with different climate patterns. Hence, for retail success, meteorology is as important as geography is. All this have led to the growing importance of weather variability in the retail domain.

Traditionally, it was believed that the retail sector does not get affected much by climate, but a series of recent studies have concluded that retail sector is a weather-sensitive sector. In other words, weather plays an important role in sales of many product categories, store's footfall and store types. Many store managers often blame climate for poor store traffic/sales, but only some of them are able to manage the weather-related risk given that most of the suppliers offer diversified products thus alleviating the effect of weather on store's traffic. Incorporating the effect of weather on consumer behavior demand was considered as a real challenge in the past. For a long time, retailers have been aware of the fact that weather affects the footfall in the stores, but until now, nothing much was being done by most of the retailers to manage such weather-related risks. The rising variability in weather these days have stimulated new interest in studying the relationship between footfall and weather.

Footfall, being a bread and butter for any retail store manager, is directly linked to drive retail sales and improve conversion rates—the more people who come to the store, the more chance they will buy something. While the conditions like heavy rainfall, snowstorm coming, heat and dry conditions, etc., keep the tendency to impact the footfall drastically. So it becomes important to know the footfall peaks and troughs which could help retail managers to optimize in-store operations and manage their resources effectively. Thus we propose a methodology for predicting the footfall of customers in different retail stores given real-time weather forecasts and historical footfall data using a supervised analytical methodology. This paper allows retailers to find the threshold probabilities that will trigger warning and require necessary decisions to decrease inventory waste. Also, this methodology would predict favorable climatic conditions which could be taken care of beforehand in order to achieve retailer's motive of maximizing sale/footfall in real time. Thus, this analysis can be used by retailers to make evidence-driven decisions and strategize their business plan which would help them to deepen the customer involvement and to get efficiency in the planning process.

This paper is divided into the following sections: Sect. 2 highlights the literature survey and critical research gaps. Section 3 defines our problem statement. In Sect. 4, the methodology followed has been proposed. The results are highlighted in Sects. 5 and 6. The conclusion is explained in Sect. 7.

2 Literature Review

This section gives a brief about some of the works that have been done related to retail and weather so far. Generally speaking, these researches can be divided into two groups: (1) theories about the nature of the impact and (2) theories about the magnitude of the impact. Here, impact refers to the impact of weather on consumer activities in a retail store. So the former group explains why and how the change in climate brings the change in retail shopping behavior while the latter explains the extent of the weather change on activities such as consumer spending. The following is the description of some of these theories explaining the consequences which retail sector faces because of the change in the climatic conditions. First, we explain the theoretical work linking weather stimuli to shopper's behavior. Second, we review some of the empirical studies done in this context. Lastly, we review the analytical models predicting the footfall for retail firms during different climatic conditions.

2.1 Theoretical Work

There are different ways by which meteorological factors can affect both psychology and physiology of consumers. Overall, these researches have confirmed that weather has the tendency to influence an individual's mood. For instance, Sanders and Brizzolara [3], studied the effects of various climatic factors such as temperature, biometric pressure, relative humidity, precipitation, and wind speed on a one-dimensional mood rating scale. They concluded that sunshine and biometric pressure together have the strongest impact on mood. Following the same methodology but different mood scales, other researchers like Persinger and Levesque [2], Howarth and Hoffman [4] and Miranda-Moreno and Lahti [5], etc. have reported that rain tends to decrease the sense of comfort of pedestrians impacting consumer's behavior negatively. Similarly, studies by Cunningham [6] and Parrott and Sabini [7] have suggested that sunshine affects customer's mood to visit a retail outlet in a positive way. Experimenters like Sherman et al. [8] and Keller et al. [9], etc., have found that two major factors responsible for moderating psychological effects of weather are the season and the amount of time spent outside and thus pointing toward a positive association between mood and shopping intentions. Robert and John [10] and Gardner and Hill [11] etc., are some of the other contributors of theoretical work done in this context.

2.2 *Empirical Studies*

All these behavioral instances which are explained in above subsection generally comes from investigational and fieldwork. However, there is also empirical evidence demonstrating the impact of weather on different aspects of retail sector, but the number of works is limited. For example, Steele [12], has presented a paper about weather effects on sales of a departmental store and found that factors like snow, rain, and extreme temperature are responsible for making shopping at store less attractive and thus, affects sales and store traffic negatively. Agnew and Thornes [13], conducted a survey study on the weather sensitivity of food and beverages sector and concluded that in terms of number of customers, the bad weather affects the hypermarkets located in urban areas more adversely as compared to small retail stores located in rural stores. Kirk [1], concluded that about 35% of inquiries in his study would avoid shopping at low temperatures, 37% would avoid going shopping when it is raining, and 30% would not go shopping in excessive heat. Murray et al. [14], further investigated that high temperature is correlated positively to emotions as well as to greater propensity to consume.

2.3 *Analytical Work*

Various studies regarding change in consumer behavior occurring due to the change in weather conditions have shown that even a degree change in temperature affects the store's traffic, reflecting the growing importance of predictive analytics in this domain. Most of the work on footfall analytics based on weather data has been done using multiple regression analysis. Agnew and Palutikof [15], implemented separate regression models for each month of the year to examine the impact of temperature, precipitation, and sunlight on the total U.K. retail sales and sales of some products categories. Their approach concluded that the sign of the effects of weather variables and percentage of sales volatility is fluctuating between months and product categories. Another regression analysis by Starr [16], concentrating on effect of temperature on the U.K. retail sector, showed that temperature has dual effects, i.e., current and lagged effect in retail. They concluded that weather effect depends on sales assortment and thus, is not same for all retail types. After a year, Parsons [17], examined the correlation between shopping store attendance and weather on a daily basis using a multiple regression model. It was seen that temperature and rainfall are the most tangible weather variables having negative impact on number of visitors while variables like humidity, sunshine hours, etc., show insignificant effect on the store's traffic. However, this analysis was being done for the colder half of the year only. Also, Bahng and Kincade [18] carried out a simple regression analysis to uncover the associations between temperature and apparel retail sales in the U.S. and concluded that weather defines the start and the end of a

season, with summer and winter selling season lasting up to 20 weeks. And any unusual change in weather can delay the beginning of each of these selling seasons.

Thus all in all various researches have been conducted to give prominent result for the effect of weather change on a retail store but all of them links retail and consumer activities to historical weather events. And none of them forecast this relationship in real time. While this paper tends to analyze the correlation between weather and retail shopping behavior (i.e., footfall) in real time and in unique manner.

3 Problem Statement

Weather, being an unmanageable factor, tends to influence shopper's purchasing decisions and causes the revenue to move in any direction. In order to manage the considerable revenue swings caused by fluctuating weather patterns, retail distribution needs to understand the relationship between weather patterns and customer activity. Basically, there are two types of risk which can affect retail performance and retailer's abilities to fulfill customer demands: the risk of understocking and the risk of overstocking. If they miscalculate these, retailers end up losing millions through disappointed customers and incorrect stock levels. Thus, impacting the costs and efficiencies of whole retail business.

It is in this context that the present analysis aims to examine, quantify, and predict the impact of weather factors on retail traffic across various stores located at different regions of the U.S. The aim here is to create an early warning decision-making system relating the retail sector to real-time weather events, preparing both retailers and customers against any weather-related risk. It should be noted that this experiment focuses on weather as an important factor affecting footfall of these retail stores.

4 Analytical Methodology to Predict Footfall

Our experimentation contributes to the field of retail distribution system by giving significant instance of association between meteorological phenomena and retail business. This section provides a recommendation on how to make weather information useful to retailers, through expected footfall levels. The first subsection gives a detailed description about our dataset. Followed by data preparation and experiment analysis.

4.1 Data

For our analysis, the data has been taken from two different sources: (1) OpenWeatherMap, which is a global geographical platform providing various types of earth observation data. 2) Kaggle, also known as “Home for Data Science”, is a stage for analytics and predictive modeling competitions. To carry out our experiment we are considering three datasets namely, Footfall/Store data, Key dataset, and Real-time weather API data. The following is the description of each of these datasets:

- (a) *Footfall data*: This is a customer’s footfall data for 44 hypermarkets (all located in the United States) where footfall may be affected by the change in weather conditions. The footfall dataset provides us detailed information about each hypermarket/store, which can be used to find some interesting facts. For example, The column “station_nbr” corresponds to an id for one of the 10 weather stations covering these 44 locations. It should be noted that some of the stores are nearby and hence share the same weather station. Also, another column called “X3 h” measures the amount of rainfall (in inches) that occurred near each store on a daily basis. Similarly, there are other columns like “temp”, “snowfall”, “pressure”, etc. describing the type of weather conditions near each store location.
- (b) *Key data*: The key dataset comprises mainly of two columns: (1) “station_nbr” representing weather stations and (2) “store_nbr”, which is an id showing one of the store locations. These columns are then mapped together in such a way that stores nearby would come under the same weather station.
- (c) *Real-time weather API data*: This dataset gives us the 5-day weather forecast, i.e., how the weather conditions would be in the coming 5 days. Using an API key, the required dataset would be generated in JSON file format. And for our analysis, we converted this extracted dataset into csv format. The weather API data tells upcoming climate conditions helping retailers to plan their inventory stock accordingly. It should be noted that this data includes weather-related information every 3 h.

The entire dataset captures information about footfalls happening each day in each store for the complete 2 years. And with the count of shoppers, the retailers can easily find out the sales taking place in each store at the time of different climatic conditions. Retailers use various metrics to maximize their sales, with thinking behind making a purchase massively influencing where items are kept in a store, availability of customer’s background data and shopper’s every move observed as they move around stores. Amid all these factors, the most important factor from the perspective of retailers is “Footfall”, which is the count of shoppers who has visited a hypermarket during a given period of time. Thus, using the above-mentioned datasets, we applied real-time footfall analytics to forecast the number of people visiting each store at the time of different climatic conditions.

4.2 Data Exploration and Preparation

In today’s world, data comes from various industries, varied sources, and in various formats. The industries can be telecom, insurance, etc. while data sources can be open-source data providing sites, sensor data, social media data, etc. and formats can be .csv, .xls, .txt, etc. Each of these datasets would be having its own set of challenges and thus would be requiring a different series of steps for data preprocessing. Some of the problems which various analysts face while dealing with data include the presence of missing data, outliers, or operational challenges while capturing data and various such other challenges. Due to all these difficulties, it is believed that data preprocessing usually takes 70% of the total time devoted for building the entire model. Since it is the quality of inputs that decide the quality of output so once the hypothesis is built, it is recommended to spend a lot of time and effort performing this step (i.e., exploration, cleaning, and preparation of data).

Keeping this in mind, the second step we did is exploration and analyzing the data chosen for our analysis. Following are the steps performed for understanding, cleaning, and preparing the final dataset which was then used for building our predictive model:

- (a) *Variable Identification:* Under this, the first step is to identify independent (Input) and target (Output) variables and then identify the category and datatype of all the variables. The following Fig. 1 shows this variable identification for our dataset:
- (b) *Missing Value Treatment:* Missing value/data is defined as the value that is not present for the variable under interest. Being a symbol of messiness in the data, if missing values are not handled properly then it could lead to invalid conclusions about the data. The absence of data can cause several problems like

Variable Type	Type Of Data	Variable Category
<ul style="list-style-type: none"> • Target Variable Footfall • Independent Variable temp temp_min temp_max X3h(Precipitation) L3_rainfall L3_speed L3_rainfall Snowfall etc. 	<ul style="list-style-type: none"> • Numeric Footfall temp temp_min temp_max X3h L3_speed L3_rainfall L3_snowfall etc. • Character Date City 	<ul style="list-style-type: none"> • Continous Footfall temp temp_min temp_max L3_speed L3_rainfall L3_snowfall • Categorical X3h Snowfall

Fig. 1 Example of variable identification (modified and appended from Analytics Vidhya blog [19])

reduction in statistical power of the study, producing biased estimates, complicating the analysis of the study, etc. Thus in order to avoid these distortions, the treatment of missing values is considered as a necessary step in analyzing the data. The dataset which we considered also contained missing values for some of the columns like, “temp_max”, “temp_min”, “X3 h”, “pressure”, “snowfall”, etc. But different treatments were being done for each of them. For example, missing data in column “snowfall” was handled by the sum of snowfall happened in past 3 days. Similarly, the maximum temperature in the past 3 days was used to replace the missing values present in the column “temp_max.”

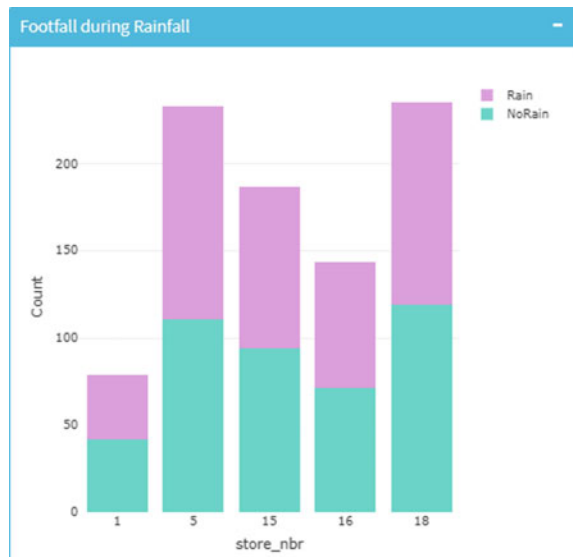
Hence, by following different treatment methods for each of the above-mentioned columns, we removed all the missing values from our dataset.

- (c) *Bivariate Analysis*: This analysis helps to understand the hidden insights present in the dataset. Basically, bivariate analysis is a way of finding the relationship between two or more variables present in the underlying dataset. And since visualization makes understanding of data easier so we used stacked bar charts for the demonstration purpose. The following figures (Figs. 2, 3, 4) show few relationships existed among variables in our dataset:

Due to the presence of detailed information about each store, the footfall data thus can be used to extract numerous useful statistics as shown above.

- (d) *Feature Engineering*: The art of extracting more statistics from existing data is known as feature engineering. So it is like making already existing data more useful from analysis point of view. For instance, consider the case of predicting

Fig. 2 Footfall during rain and no rain



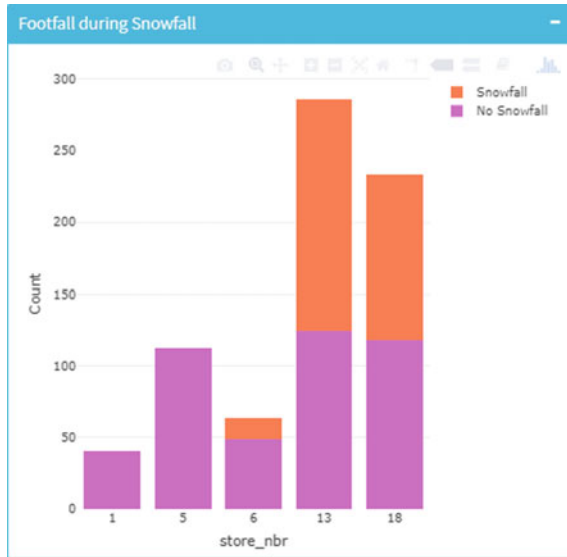


Fig. 3 Footfall during snow and no snow

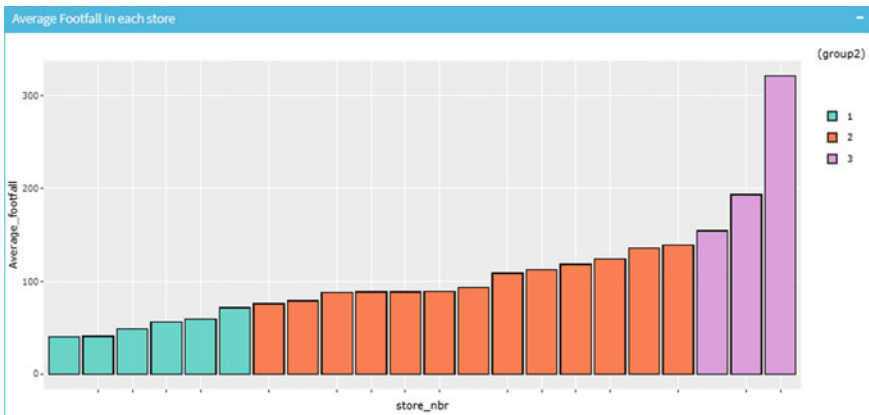


Fig. 4 Average footfall in each store

footfall in a shopping mart on the basis of dates. Here, using dates directly may not yield useful statistics from the data under consideration. This is because count of people visiting a shopping mart is likely to get less affected by the day of the month than the day of the week. And the latter information is present implicitly in the data which needs to be taken out for achieving reliable results. Thus, this activity of taking out information from data is known as “feature Engineering.” This step contains two substeps: (1) Variable Transformation and

(2) Feature Creation. In our analysis, we created some new variables from already existing variables. For example, “Date” variable was split into three other new variables—“Day”, “Month”, and “Year.” Also from input variables “temp_max” and “temp_min”, new columns called “L3_temp_max” and “L3_temp_min” were being generated. This step was being done thinking that these newly formed variables may have better relationship with the dependent variable and thus could impact the power of prediction remarkably.

In the next section, we will discuss how these steps are combined together so as to predict the footfall based on real-time weather API and store dataset.

4.3 Random Forest as a Supervised Detection Technique

With limited information about climate variability and weather patterns, we solve our problem statement using a supervised machine learning algorithm called random forest. Being a part of ensemble learning algorithm, random forest works by generating multiple models over training dataset where the output of each model is then combined to produce a stronger model.

In the present analysis, this algorithm is applied for regression purpose helping us to map both linear as well as nonlinear relationships existing between the variables. The concept here is to decorrelate the various trees that are produced from separate bootstrapped training data samples. After which, the mean of output of all the trees is used to reduce the overall variance which eventually would help to avoid overfitting. It should be noted that the tree-based algorithms are famous for building models with high stability, accuracy, and easy interpretability.

4.4 Footfall Prediction Based on Weather with Random Forest Technique

In the current study, random forest algorithm has been applied to forecast footfall in different hypermarkets during different climatic conditions. Because of continuous nature of our target variable, regression analysis was being done using random forest. Machine learning algorithms like Support Vector Machine (SVM) etc., could have also been used to solve this problem statement. But the reason for choosing random forest over these techniques is the ease with which it tells the important variables contributing toward regression (and classification) analysis and the relative importance of each of them based on their depth of location in the tree. The following steps explain how random forest for regression works in predicting store’s traffic given the weather data:

Each tree is generated as follows:

- (i) **Random Sample Selection:** Divide dataset into train and test. Using around 2/3rd of the training dataset, each tree is trained. This 2/3rd selection is done randomly with replacement and is the training set for that particular tree.
- (ii) **Random Variable Selection:** Each tree is trained using some number of predictor variables (say m) which are chosen randomly out of all the independent variables present in the dataset. For regression, the default value of m is total number of predictors divided by 3.
- (iii) **Splitting of node and Parameter Tuning:** Each split is done by examining all the variables one by one and then the best split is picked using methods like Gini Index, Chi-square, or Information Gain, etc. Also to get better accuracy important parameters like “mtry”, “ntree”, etc. are tuned accordingly.
- (iv) **Calculation of “Out of bag” (OOB) error rate:** Calculate the error rate of remaining 1/3rd data for each tree. Aggregate the errors for all the trees to get overall OOB error rate.
- (v) **Averaging the dependent variable:** On the remaining data, each tree is generated giving output equal to some value of dependent variable. Then the average of all these values is taken as the predicted probability.

In the next section, we have leveraged aforesaid approach to showcase some footfall prediction scenarios. There is lot of potentials to extend our methodology to different sectors, different scenarios related to weather and technical approaches.

5 Result and Experimentation

After preprocessing the raw data as explained in Sect. 4.2, we have our dataset ready which could now be used for building the model. With “footfall” as a target variable, we applied random forest algorithm steps (as explained in Sect. 4.4) to carry out regression analysis for the underlying problem statement. It should be noted that the optimal value of parameters such as “mtry” and “ntree” has been found using fivefold cross-validation technique. Following results are obtained from the application of random forest for prediction of footfall in a retail firm:

- (i) **Variable Importance Plot:** One of the advantage of random forest technique over other machine learning algorithms is that it tells which input parameters are important for predicting the target variable. Figure 5 shows the top five important parameters for predicting footfall in our model. This plot, called variable importance plot, tells which independent variables contribute more for the variation of target variable. Such information is used to choose the right set of features so as to make target predictions more accurate. It is important to note that these variables does not tell whether there would be more or less

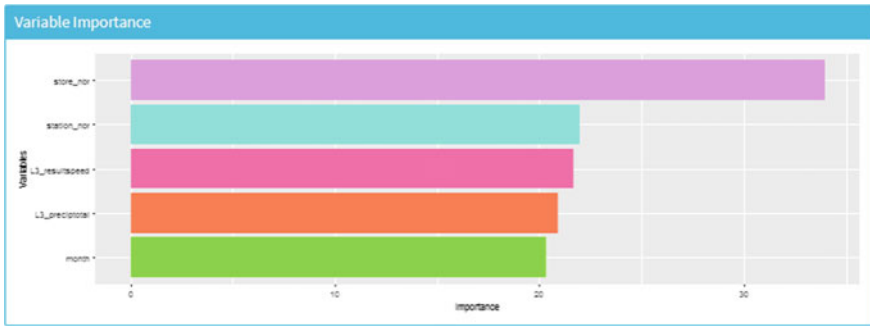


Fig. 5 Variable importance plot

footfall, but tell which all variables are important from prediction perspective. It can be seen that apart from store type, the weather variables such as precipitation and wind speed appear to be the most important parameters in our model.

- (ii) *Root Mean Square Error*: The Root Mean Square Error (RMSE) is a performance metric measuring the difference between actual and predicted values of the model. This measure tells us how well our model is able to predict the test outcomes. For our model, RMSE is equal to 11% indicating the good fit of our model, given the spread of our dataset.

Using these two performance metrics, we were able to predict the expected footfall given the weather conditions for all 44 retail stores for the test dataset.

6 Live Demo Model

The above section explains our study for predicting the footfall given the weather conditions near every 44 retail stores. It should be noted that this procedure is based on historical weather data of the U.S. Now to carry out the same experiment in real time, we will make use of 5-day weather forecast data obtained using an API key. Thus, by integrating this new information we are able to predict the expected number of people in each store for the upcoming 5 days over a google map. This live demo model showcase two things: (1) current weather conditions of all the 44 stores present at different locations in the U.S. on a google map and (2) expected footfall curve for the 5 days using which retailers can strategize their activities accordingly. Figure 6 shows this dashboard which we have built using RShiny. It should be noted that the output from this model would keep on changing as the inputs are subjected to real-time factors.

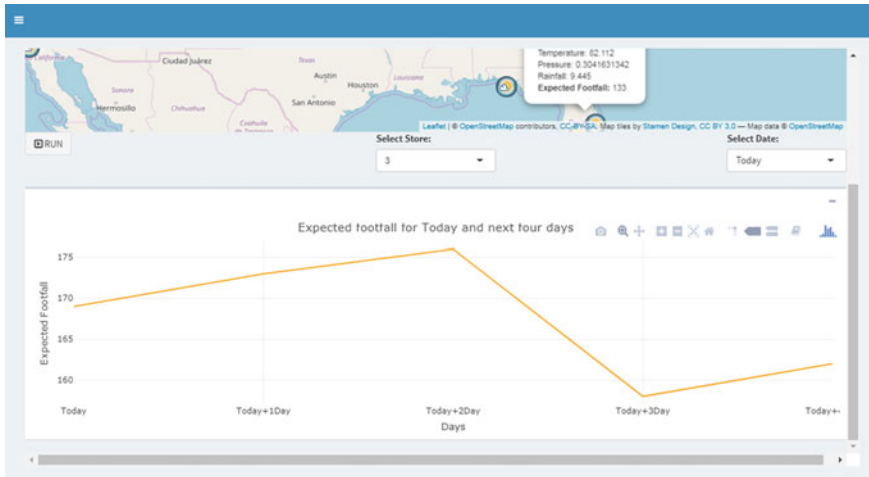


Fig. 6 Real-time demo model

7 Conclusion

The effect of climate variability and weather patterns on retail store’s traffic using a supervised machine learning algorithm has been investigated in this paper. Most of the retailers are unaware of the weather consequences on footfall or believe that climate exposure is unmanageable. But, even though weather cannot be managed, the revenue strategies can easily be managed which will secure retailers against any adverse weather conditions. Keeping this in mind, we developed an early warning decision-making system associating retail sector with changing weather events, preparing retailers as well as shoppers against any weather-related risk in this sector. Our methodology put weather in a business context and provides actionable insight into customer behavior, which results in increased footfall levels and improved profitability. Thus, this experimentation is applicable universally and enables comprehensive scenarios of daily footfall traffic to be explored using real-time weather API data, allowing retail strategies to be targeted and implemented effectively.

The major implications of this analysis are as follows:

1. This analysis provides a platform to researchers, practitioners, and decision-makers for sharing and examining how non-catastrophic weather events affect the retail’s footfall and the impact it will have in the future(real time).
2. It uses disaggregated data at a large scale and unlike past studies, the lagging effect of weather variables on store’s traffic has also been taken into account while building our predictive model.

3. Apart from providing valuable insights about the drivers of footfall, this application can be used as a tool for evaluating how successful the previous events in each of these stores were after taking account of various weather factors.
4. Also, measuring customer's foot traffic and engagement would enable retailers to ensure the right number of staffs as well as products that are required during the impetus of global climate variation.

References

1. Kirk, B.: Better business in any weather. *Res. Rev.* **12**(2), 28–34 (2005)
2. Persinger, M.A., Levesque, B.F.: Geophysical variables and behavior: XII. The weather matrix accommodates large portions of variance of measured daily mood. *Percept. Motor Skills* **57**(3), 868–870 (1983)
3. Sanders, J.L., Brizzolara, M.S.: Relationships between weather and mood. *J. Gen. Psychol.* **107**, 155 (1982)
4. Howarth, E., Hoffman, M.S.: A multidimensional approach to the relationship between mood and weather. *Br. J. Psychol.* **75**(1), 15–23 (1984)
5. Miranda-Moreno, L.F., Lahti, A.C.: Temporal trends and the effect of weather on pedestrian volumes: a case study of Montreal, Canada. *Transp. Res. Part D Transp. Environ.* **22**(2013), 54–59 (2013)
6. Cunningham, M.R.: Weather, mood, and helping behavior: quasi experiments with the sunshine samaritan. *J. Personal. Soc. Psychol.* **37**(11), 1947 (1979)
7. Parrott, W.G., Sabini, J.: Mood and memory under natural conditions: evidence for mood incongruent recall. *J. Personal. Soc. Psychol.* **59**(2), 321 (1990)
8. Sherman, E., Mathur, A., Smith, R.B.: Store environment and consumer purchase behavior: mediating role of consumer emotions. *Psychol. Mark.* **14**(4), 361–378 (1997)
9. Keller, M.C., et al.: A warm heart and a clear head: the contingent effects of weather on mood and cognition. *Psychol. Sci.* **16**(9), 724–731 (2005)
10. Robert, D., John, R.: Store atmosphere: an environmental psychology approach. *J. Retail.* **58**(1), 34–57 (1982)
11. Gardner, M.P., Hill, R.P.: Consumers' mood states: antecedents and consequences of experiential versus informational strategies for brand choice. *Psychol. Mark.* **5**(2), 169–182 (1988)
12. Steele, A.T.: Weather's effect on the sales of a department store. *J. Mark.* **15**(4), 436–443 (1951)
13. Agnew, Maureen D., Thornes, John E.: The weather sensitivity of the UK food retail and distribution industry. *Meteorol. Appl.* **2**(2), 137–147 (1995)
14. Murray, K.B., et al.: The effect of weather on consumer spending. *J. Retail. Consum. Serv.* **17**(6), 512–520 (2010)
15. Agnew, M.D., Palutikof, J.P.: The impacts of climate on retailing in the UK with particular reference to the anomalously hot summer of 1995. *Int. J. Climatol. J. R. Meteorol. Soci.* **19**(13), 1493–1507 (1999)
16. Starr, M.: The effects of weather on retail sales (2000). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=221728
17. Parsons, A.G.: The association between daily weather and daily shopping patterns. *Australas. Mark. J. (AMJ)* **9**(2), 78–84 (2001)
18. Bahng, Y., Kincade, D.H.: The relationship between temperature and sales: sales data analysis of a retailer of branded women's business wear. *Int. J. Retail Distrib. Manag.* **40**(6), 410–426 (2012)
19. <https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/>

Normal Pressure Hydrocephalus Detection Using Active Contour Coupled Ensemble Based Classifier



Pallavi Saha, Sankhadeep Chatterjee, Santanu Roy and Soumya Sen

Abstract The Brain plays an imperative role in the life of human being as it manages the communication between sensory organs and muscles. Consequently, any disease related to brain should be detected at an early stage. Abundant accumulation of cerebrospinal fluid in the ventricle results to a brain disorder termed as normal pressure hydrocephalus (NPH). The current study aims to segment the ventricular part from CT brain scans and then perform classification to differentiate between the normal brain and affected brain having NPH. In the proposed method, firstly few preprocessing steps have been carried out to enhance the quality of the input CT brain image and ventricle region is cropped out. Then active contour model is employed to perform segmentation of the ventricle. Features are extracted from the segmented region and Ensemble classifier is used to classify CT brain scan into two classes namely, normal and NPH. More than hundreds of CT brain scans were analyzed during this study; area of ventricle has been used as a measure of feature extraction. Experimental results disclosed a significant improvement in case of ensemble classifier in comparison to Support Vector Machine in terms of its performance.

Keywords NPH · Ensemble · ACM · CSF

P. Saha · S. Chatterjee

Department of Computer Science and Engineering, University of Engineering and Management, Kolkata, West Bengal, India

e-mail: pallavisaha321@gmail.com

S. Chatterjee

e-mail: chatterjeesankhadeep.cu@gmail.com

S. Roy

Department of MCA, Future Institute of Engineering and Management, Kolkata, West Bengal, India

e-mail: santanuroy84@gmail.com

S. Sen (✉)

A. K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India

e-mail: iamsoumyasen@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_38

543

1 Introduction

Recent studies have revealed biomedical image processing as an application in the detection of a wide range of diseases. It has a great contribution which allows physicians to perform an accurate diagnosis. The human brain is composed of cerebrum, brainstem, and cerebellum. Apart from these parts, there is a part called ventricle which is stuffed with a vivid and colorless fluid regarded as cerebrospinal fluid (CSF). Shape and size of the ventricle depend on the amount of CSF present in the ventricle. Accumulation of excessive amount of CSF in ventricle gives rise to a brain disorder known as NPH. This paper focuses on the detection of NPH brain disease by segmentation [4–6] of ventricle using ACM [2] and then estimates the area of ventricle which is then sent as input to the Ensemble classifier. Ensemble classifier performs classification of the CT brain scan and predicts as either normal brain or affected brain having NPH. This classifier comprises of three weak learners viz. Discriminant, Decision tree, and K -th nearest neighbor (KNN). The performance of Ensemble classifier is then compared with SVM and the best classifier is traced out. Features form the principal element in the training stage as well as in the testing stage. The efficiency of a classifier vastly depends on the extraction of features. At times, the best classifier fails to produce proper results if weak features are employed. In the field of image processing, a vast amount of work has already been undertaken by various researchers in the past and brought remarkable changes in the medical history. Classification of brain diseases has been achieved through several methods like Support Vector Machine (SVM), Random Forest, Artificial Neural Network skilled with Genetic Algorithm (GA) termed as ANN-GA, and ANN trained with particle swarm optimization (ANN-PSO). Ensemble classifier has also been applied in the detection of brain disease. Clangphukhieo et al. [1] designed an algorithm for the segmentation of the ventricular part from CT brain image. Initially, CT brain image is normalized and region of interest is estimated by applying gray level profile analysis followed by the Bayesian segmentation. This segmentation enables us to demarcate among the three intensities viz. white matter, gray matter, and CSF and at last the area of CSF is determined. Experimental results disclosed a minimum error of 3.14% and a standard deviation of 1.41 has been achieved by the proposed algorithm. Butman et al. [3] introduced an algorithm to calculate the size of the ventricle in the brain MRI examinations. This has been attained by firstly integrating the different serial images in order to enhance the SNR. Active Contour Model and fast marching techniques were applied to perform segmentation and later deformable registration was used to propagate the segmentation. Zhukov et al. [7] put forward a method based on the concept of a decision tree to estimate the credibility of power systems. Ensemble classifier is used for the classification of the system to determine whether it is in a safe state or not and aggregation is done based on boosting model and random forest model. Shenbagarajan et al. [12] proposed an efficient method to classify an MRI brain image into three types viz. normal, noncancerous, and cancerous. This method involves four phases namely preprocessing steps, segmentation, feature extraction

followed by the classification. Active Contour Model (ACM) is engaged for segmentation and then Artificial Neural Network (ANN) trained with Levenberg–Marquardt (LM) algorithm is used for classification. An enormous amount of research work has been performed in the past in the field of biomedical image processing. Still, it is difficult to meet the growing requirements. In the current study KNN, Decision tree, and Discriminant forming together the Ensemble classifier is implemented to perform classification between a normal brain and affected brain having NPH. The current paper is arranged as follows. Section 2 describes the contrast adjustment, segmentation, and classification model. Section 3 represents the proposed system laid out for the classification of the brain between a normal brain and affected brain having NPH. Section 4 reflects the depiction of experimental results in terms of its performance calculated from the confusion matrix.

2 Methodology

2.1 *Image Enhancement and Cropping*

Image enhancement forms one of the preprocessing steps that are extensively employed in image processing [15]. The main purpose behind image enhancement is to improve the intelligibility of data in images for viewers. Contrast is an essential factor in the calculation of the quality of an image.

In the current study, the contrast has been adjusted by adjusting the intensity values accordingly in the input CT brain scans to retrieve information efficiently from it. This, in turn, brightens up the image and then the ventricular part is cropped out from the input image.

2.2 *Active Contour Models*

Active Contour Model (ACM) also termed as snakes is well known in the field of computer vision and is widely applied in applications such as segmentation, edge detection, etc. ACM is classified into two types namely: Edge-based ACM and Region-based ACM [2, 16]. In order to hold back the contour during development for recognizing the boundary of the foreground object, edge-based model makes use of the gradient of the image.

The region-based model utilizes mathematical information regarding the regions lying both inside as well as outside the curve for contour development. In the proposed method, region-based model has been implemented for segmentation which focuses to encourage the curves to arrive at the boundaries of the input CT brain scans.

2.3 *Region-Based ACM for CT Brain Segmentation*

The region-based model has undertaken with the consideration that the pixel regions of the input image are mathematically similar. It works well on noisy, hazy images, and images having disconnected regions, etc. In the analysis of CT brain scans, region-based model believes the global properties like the length of contour and CT scan image pixel regions are opposite to the local properties like a gradient. The energy minimizing function can be written as:

$$\ln P(I_s|p) = \iint_A I_s(x, y) dA$$

$I_s(x, y)$ represents the intensity value at the pixel location (x, y) in the CT brain image and integral provides the sum total of area A confined by the curve p .

2.4 *Feature Extraction Based on Area*

The selection of the feature extraction technique plays a leading role in the performance of a classifier [8–11, 17]. In both the training and testing phase, the feature is the crucial element. In the current study, the area of the ventricle in the segmented region forms the measure of feature extraction. Extracted features become the dataset and thereafter enable us to classify between the normal and affected NPH brains.

2.5 *Ensemble Classifier*

Support Vector Machine (SVM) is a model that is widely applied for classification and regression analysis. The application of SVM in the field of image processing is an example of linear discrimination. The basic working principle of this model is that it divides the space into which the CT scan brain images are assigned into two classes by identifying a separating hyper plane. In 2D space, a line defines the boundary and hyperplane in higher-dimensional space. The inclination towards the application of SVM is that it employs the basis of fundamental risk reduction, which focuses to locate a hyperplane that decreases the area between training classes. The pixels that are nearest to the separating boundary describes the optimal separating hyperplane (OSH) produced by an SVM.

The SVM classifier is trained with the color features extracted from the pixels of several CT scan brain images to differentiate between two regions (classes) namely CSF and non CSF. In the testing phase, an unseen CT scan brain image is selected, then each and every pixel of the testing image is sent as input to the trained SVM

classifier and it assigns a class or a region to each of the pixels in the image as a response. This model is an age-old model and various researchers have adopted this classifier in the field of biomedical image processing to dig out some new innovations, here we have adopted this model and performed a comparison with the proposed model to determine the best classification model in terms of its performance.

2.6 Ensemble Classifier

The foundation of machine learning is the ensemble methods. The amount of diversification among the different learners that exists in the ensemble affects the efficiency of the ensemble to a huge extent. Ensemble learning employs various classification models together to achieve superior predictive performance during testing whereas this level of performance is never achieved by any classification model alone. This model works by integrating many poor learners namely Discriminant, Decision Tree, and KNN. Figure 1 describes the common workflow of ensemble classifier. Once the outcome from each of the weak learners is attained, an aggregation method such as AdaboostM1, Gentleboost, Logitboost, etc., is applied to forecast the final result for the test data.

3 Proposed System

Experiments have been carried out on more than 100 slices of CT brain scans that were gathered from the Apex Hospital situated at Kolkata in West Bengal. Initially, the contrast of the input CT brain image [13, 14] was enhanced and a part of the image was cropped out for later processing steps. Then segmentation of the ventricular part from CT brain scan was employed by using ACM. Features were extracted from the segmented region based on the area of that particular region. These extracted values of area for different CT brain scans together formed the dataset. These dataset were utilized in both the training and testing phase.

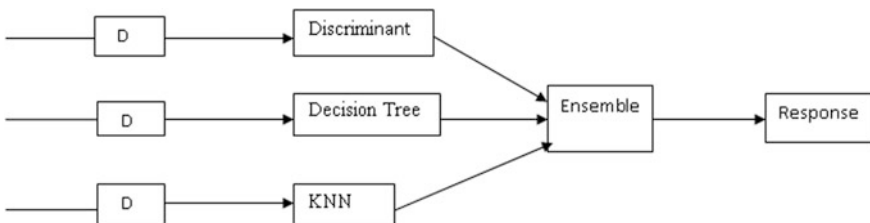


Fig. 1 Represents the workflow of Ensemble classifier

Further on, the extracted dataset is divided into two parts; former part is used for training the classification model and the later part is used during the training stage, depending on the value of data it is classified into either of the two classes namely: normal brain and affected brain having NPH.

In the present study, depending on the training data three weak learners such as Discriminant, Decision Tree, and KNN were trained and the predictions of each model were combined together by using few methods namely; AdaboostM1, Gentleboost, etc. Simultaneously, the training data is also sent to SVM model to train the model and prediction for the test data is noted. In the testing phase, test data was fed as an input to the three classification models and results are recorded and the efficiency of the ensemble classifier is determined and compared with SVM in terms of its performance. The different performance metrics which has been determined from the confusion metrics are as follows (Fig. 2).

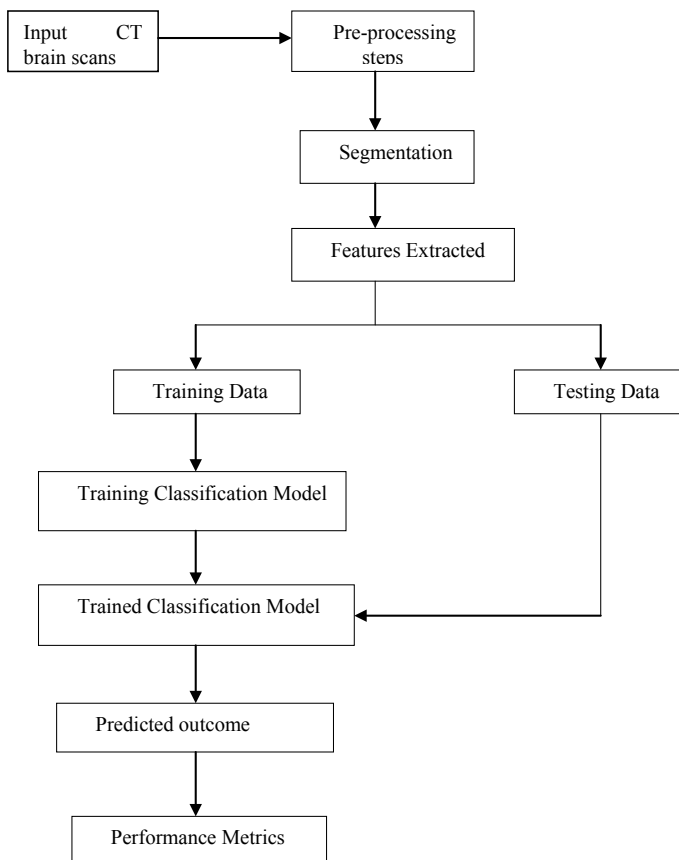


Fig. 2 Depicts the flowchart of the proposed system

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fn} + \text{tn}}$$

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

$$F\text{-measure} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

In the above equations, tp represents true positive, tn represents true negative, fp represents false positive and fn represents false negative for testing the efficiency of the classification model.

4 Experimental Results and Discussion

The simulations were carried out on more than 100 CT brain scans by using the proposed method. The resolution of each brain scan is 512×512 . The dataset comprises of 100 instances among which 60 are normal brain and 40 are affected brain having NPH.

Figures 3a and 4a depicts the input CT brain image, Figs. 3b and 4b represent the enhanced images of the input image. Figures 3c and 4c reflect the cropped part of the input enhanced image, Figs. 3d and 4d depict the segmented ventricle portion from the input CT brain image. Since CSF is stuffed in the ventricle, area of the ventricle can be used to determine the amount of CSF present in the brain which in turn helps to diagnose the condition of the patient. On further calculation, it has been found that the area of the ventricle in Fig. 3d to be around 268.687 mm and Fig. 4d to be 7764.5 mm. Radiologists have also manually calculated the area of ventricle for Fig. 3a which was 268.254 mm.

Moreover, Radiologists have also informed that the CT brain scan to be the healthy and normal brain. The estimated area of the ventricle in a CT brain scan of Fig. 4a is quite greater than the normal brain henceforth it is considered to be affected brain having NPH. Moreover, Radiologist has also determined manually the area of the ventricle in Fig. 4a to be 7764.21 mm. In addition, Radiologist has also intimated that the corresponding CT brain scan to be affected brain having NPH. Table 1 epitomizes the performance of Ensemble and SVM classifier in terms of accuracy, precision, recall, and *F*-measure.

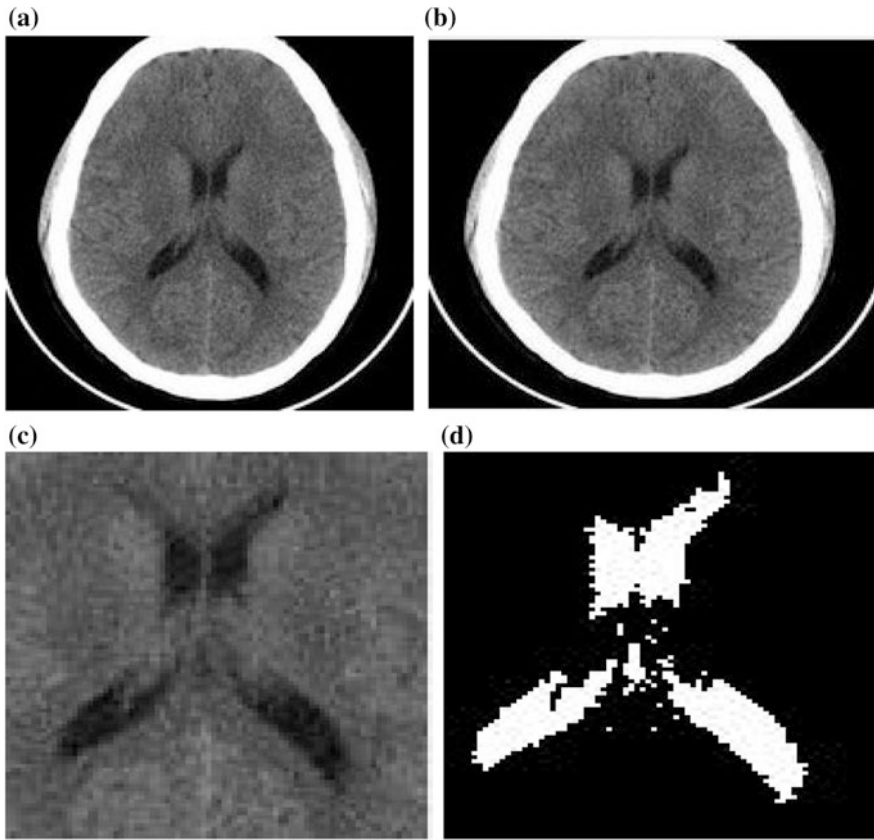


Fig. 3 a Source image. b Contrast adjustment. c Cropped image. d Segmented image

Table 1 reports that the accuracy obtained by the ensemble classifier is 93.06% in our proposed method while it achieves precision of 95.65%, recall of 94.28%, and F -measure of 94.96%. It has been observed that the performance of SVM is poor in comparison to the Ensemble classifier, achieved an accuracy of 87.12, precision of 88.40, recall of 91.42, and F -measure of 89.88.

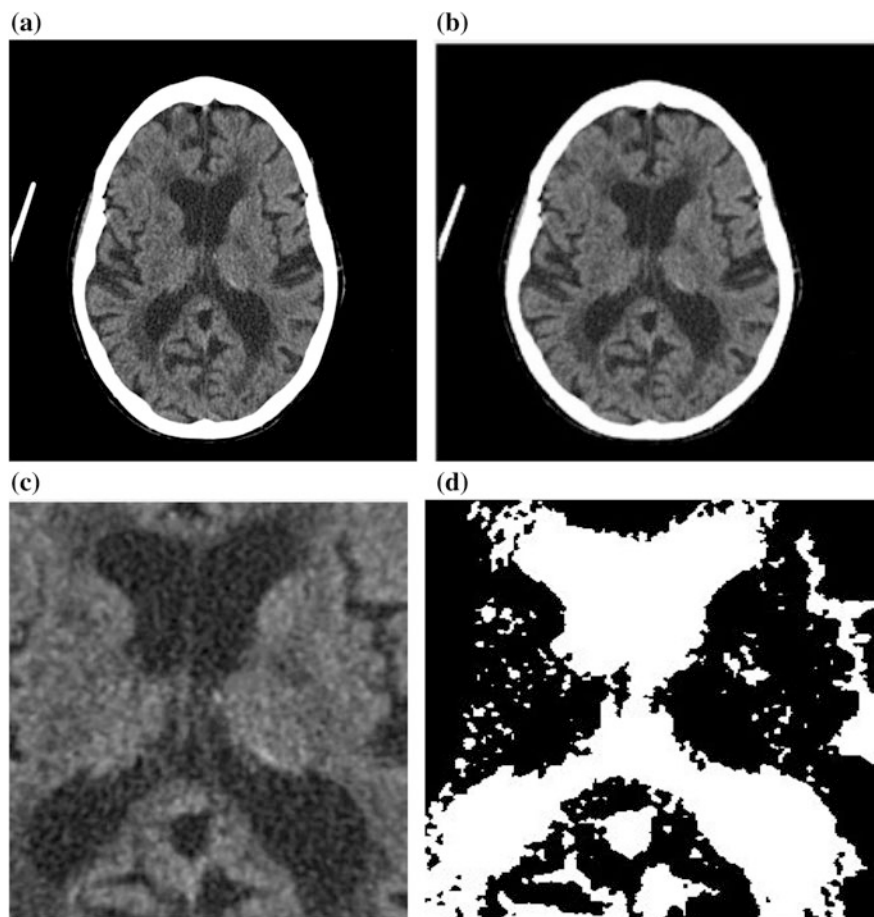


Fig. 4 a Source image. b Contrast adjustment. c Cropped image. d Segmented image

Table 1 Performance of the Ensemble classifier (in %)

Metrics	Ensemble	SVM
Accuracy	93.06	87.12
Precision	95.65	88.40
Recall	94.28	91.42
<i>F</i> -measure	94.96	89.88

5 Conclusion

Since the human brain controls many other parts of our body, hence proper functioning of the brain should be ensured primarily. Human brain forms the central organ of the nervous system, so any diseases related to the brain should be diagnosed at an early stage. In the current work, a powerful technique has been put forward for the detection of NPH in the CT brain scans. Experimental results disclosed that ensemble classifier produces better and satisfactory results in comparison to SVM by achieving an accuracy of 93.06%, precision of 95.65%, recall of 94.28%, and *F*-measure of 94.96%.

Acknowledgements We acknowledge Apex Hospital, Kolkata for their support to carry out this research work.

References

1. Clangphukhieo, B., Aimmanee, P., Uyyanonvara, B.: Automated segmentation of a ventricle boundary from CT brain image based on Naive Bayes Classifier. In: Proceeding of 7th International Conference on Computing and Convergence Technology (ICCT2012), Seoul, South Korea, pp. 1168–1173 (2012)
2. Mostaar, A., Houshyari, M., Badiyan, S.: A novel active contour model for MRI brain segmentation used in radiotherapy treatment planning. *Electron. Physician* **8**(5), 2443–2451 (2016)
3. Butman, J.A., Linguram, M.G.: Assessment of ventricle volume from serial MRI scans in communicating hydrocephalus. In: 5th IEEE International Symposium on Biomedical Imaging, Nano to Macro, pp. 49–52 (2008)
4. Chakraborty, S., Chatterjee, S., Dey, N., Ashour, A.S., Ashour, A.S., Shi, F., Mali, K.: Modified cuckoo search algorithm in microscopic image segmentation of hippocampus. *Microsc. Res. Tech.* **80**(10), 1051–1072 (2017)
5. Aung, P.T.T., Khaing, A.S., Tun, H.M.: MR brain image segmentation using region based active contour model. *Int. J. Sci. Technol. Res.* **5**(06), 92–97 (2016)
6. Karimia, H., Esfahanimehrb, A., Moslehb, M., Mohammadianjadvalghadamc, F., Salehpourc, S., Medhatia, O.: Persian handwritten digit recognition using ensemble classifiers. In: The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)
7. Zhukov, A., Tomin, N., Kurbatsky, V., Sidorov, D., Panasetsky, D., Foley, A.: Ensemble methods of classification for power systems security assessment. *Appl. Comput. Inform.* **15**, 45–53 (2019)
8. Chatterjee, S., Dey, N., Shi, F., Ashour, A.S., Fong, S.J., Sen, S.: Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data. *Med. Biol. Eng. Comput.* **56**(4), 709–720 (2018)
9. Bijalwana, A., Chand, N., Pilli, E.S., Rama Krishna, C.: Botnet analysis using ensemble classifier. *Recent Trends Eng. Mater. Sci.* **8**, 502–504 (2016)
10. Chakraborty, S., Chatterjee, S., Ashour, A.S., Mali, K., Dey, N.: Intelligent computing in medical imaging: a study. In: Dey, N. (ed.) *Advancements in Applied Metaheuristic Computing*, pp. 143–163. IGI Global, Hershey (2018)

11. Farhan, S., Fahiem, M.A., Tauseef, H.: An ensemble-of-classifiers based approach for early diagnosis of Alzheimer's disease: classification using structural features of brain images. *Comput. Math. Methods Med.* (2014). <https://doi.org/10.1155/2014/862307>
12. Shenbagarajan, A., Ramalingam, V., Balasubramanian, C., Palanivel, S.: Tumor diagnosis in MRI brain image using ACM segmentation and ANN-LM classification techniques. *Indian J. Sci. Technol.* **9**, 1–12 (2016)
13. Xia, Y., Ji, Z., Zhang, Y.: Brain MRI image segmentation based on learning local variational Gaussian mixture models. *J. Neurocomput.* **204**(C), 189–197 (2016)
14. Chakraborty, S., Chatterjee, S., Dey, N., Ashour, A.S., Shi, F.: Gradient approximation in retinal blood vessel segmentation. In: 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), pp. 618–623. IEEE (2017)
15. Kaur, S., Ritika, R.: Contrast enhancement techniques for images—a visual analysis. *Int. J. Comput. Appl.* **64**(17), 20–25 (2013)
16. François, L., Besson, S.J., Fadili, J., Aubert, G., Revenu, M., Saloux, E.: Region-based active contour with noise and shape priors. In: IEEE International Conference on Image Processing; Atlanta, GA, pp. 1649–1652 (2006)
17. Chatterjee, S., Dey, N., Sen, S.: Soil moisture quantity prediction using optimized neural supported model for sustainable agricultural applications. *Sustain. Comput. Inform. Syst.* (2018). <https://doi.org/10.1016/j.suscom.2018.09.002>

Question–Answer System on Episodic Data Using Recurrent Neural Networks (RNN)



Vineet Yadav, Vishnu Bharadwaj, Alok Bhatt and Ayush Rawal

Abstract Data comprehension is one of the key applications of question-answer systems. This involves a closed-domain answering system where a system can answer questions based on the given data. Previously people have used methods such as part of speech tagging and named entity recognition for such problems but those methods have struggled to produce accurate results since they have no information retention mechanisms. Deep learning and specifically recurrent neural networks based methods such as long short-term memory have been shown to be successful in creating accurate answering systems. This paper focuses on episodic memory where certain facts are aggregated in the form of a story and a question is asked related to a certain object in the story and a single fact present is given as answer. The paper compares the performance of these algorithms on benchmark dataset and provides guidelines on parameter tuning to obtain maximum accuracy. High accuracy (80% and above) was achieved on three tasks out of four.

Keywords Episodic memory · LSTM · Memory networks · Question answering system · Recurrent neural networks

V. Yadav · V. Bharadwaj · A. Bhatt (✉) · A. Rawal
Research and Innovation Lab, Tata Consultancy Services, Bangalore, India
e-mail: alok.bhatt@tcs.com

V. Yadav
e-mail: vineet.y@tcs.com

V. Bharadwaj
e-mail: vishnu.bharadwaj@tcs.com

A. Rawal
e-mail: ayush.rawal@tcs.com

1 Introduction

Question answering is a complex NLP task which involves understanding of the meaning of text and after understanding, ability to reason about important facts [1]. This becomes tricky as computer does not infer like humans given text data and a set of questions to answer. The problem is very specific and involves answering factual questions, for example, given a paragraph on the US Administration and being asked: “Who is Donald Trump?”.

A lot of interest is being shown by organizations in creating question–answer systems that can accurately answer simple questions without the use of a lot of training data. The focus is on obtaining the best possible approach to certain problems in the benchmark bAbI dataset using LSTM.

The bAbI dataset consists of total 20 tasks. These tasks are related to support facts about a question for which an algorithm should provide an answer. These supporting facts are single supporting fact, e.g., Mary traveled to office. Where is Mary? Two or three supporting facts, e.g., “John is in the playground” and “John picked up the football”. Where is the football?

This problem involves learning from a large data which is in episodic form and has facts which are related to the asked question. An attempt has been made to solve this using a modern neural network architecture which uses the facts to train and is able to produce the specific answer.

2 Related Work

We have seen the emergence of QA-based strategies in the field of language understanding. QA is easy to evaluate, especially for those questions which have true/false or yes/no answers. These questions should be unambiguous which a human can understand and answer. Memory network proposed by Weston et al. [2] is a new network architecture called memory network which consists of a memory component m and four components I [Input feature map], G [Generalization component], O [Output feature map], and R [Response component] which are potentially learned.

Another work on Large-scale Simple Question Answering with Memory Networks by Bordes et al. [3] proves that question answering systems can handle millions of data points very well and that training on two related datasets enhances overall accuracy. It involves the process of generating the most similar candidates through n -gram matching and getting the best candidate through finding cosine similarity.

A work on end-to-end memory network by Sainbayar Sukhbaatar et al. [4] which can be seen as an extension of RNN search having multiple computational steps known as hops per output symbol. This model can be considered as a

continuous form of memory network which can be trained end-to-end with less supervision on an input–output pair. It also performs well on various tasks from question–answer task to language modeling.

3 Scope of the Work

Four tasks are selected and are mentioned below to build question–answer models. The selection was done with the objective of having unique unrelated contexts.

Task1: Single supporting fact

Task4: Two logical arguments

Task7: Counting

Task17: Positional Reasoning

The detailed description of these tasks is provided alongside model results. We have also conducted error analysis on Task 1 to study the algorithm in depth.

4 Our Approach

A deep learning Recurrent Neural Networks (RNN) architecture is designed which uses the concept of LSTM. The data used to train the models is the bAbI 10,000 sample dataset.

RNN is the preferred architecture for sequential information. In a standard neural network, there is an assumption that all inputs and outputs are independent of each other. Practically this assumption may not always hold true. To predict the next word in a sentence it would help to know the words appearing previously. RNN are known as recurrent neural networks because they perform the same task on every element of a sequence, where output depends on computations of previous steps. RNN has memory which captures information about the computations of previous steps. In theory, RNN has the potential to use information present in long sequences, but practically it is only applicable for looking back at few steps (Fig. 1).

LSTM belongs to RNN family which has the capability of persisting information in longer sequence, which RNN lacks in implementation [5]. Key element of LSTM is known as cell state. This cell state keeps track of long-term sequential information [6]. LSTM neuron has three gates input, output, and forget. These gates regulate the interaction of information with RNN data flow (Fig. 2).

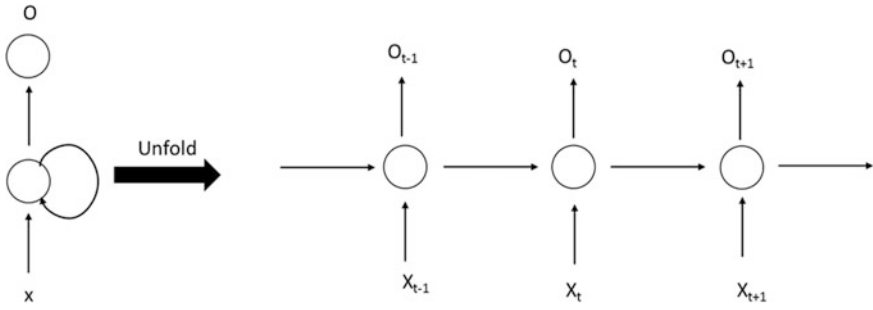


Fig. 1 Recurrent neural network architecture after unfolding the loop

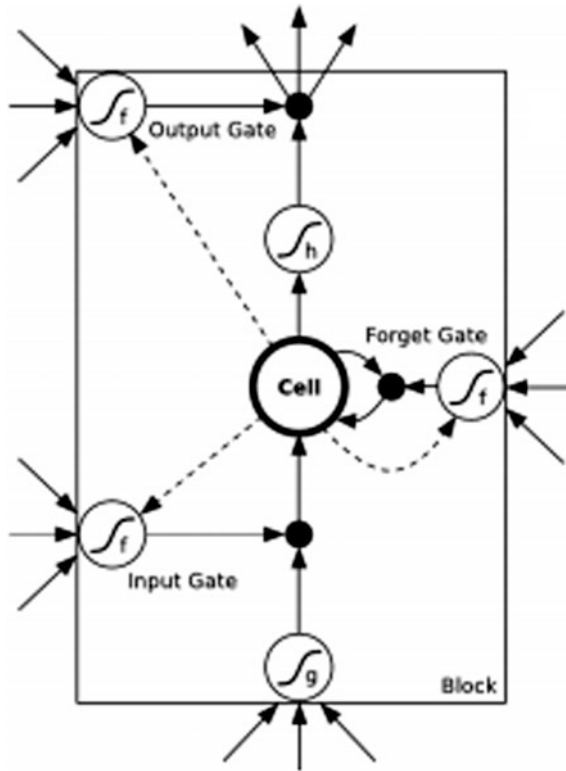


Fig. 2 LSTM cell with input, output, and forget gate which helps in passing specific information to the next level

4.1 Solution Architecture

The solution architecture used incorporates the memory network and has certain connections that help to emphasize the most appropriate words to pick the answer.

This consists of mainly four layers:

1. Embedding layer: It converts each word of dictionary to a vector.
2. LSTM: This is a type of RNN which is used for sequence data and overcomes short-term memory of regular RNN. Memory network is being used in this layer.
3. Dropout layer: It will shut off few neurons from a layer based on user-defined criterion and thus prevents model overfitting.
4. Dense layer: It is a layer with softmax activation [7], which will produce answer for a specific question by choosing the option with highest probability (Fig. 3).

First each question and story gets converted to embedding. After this, allow the network to follow the question based on story and give attention to certain most relevant words in the question; dot product between both the embedding was taken [8].

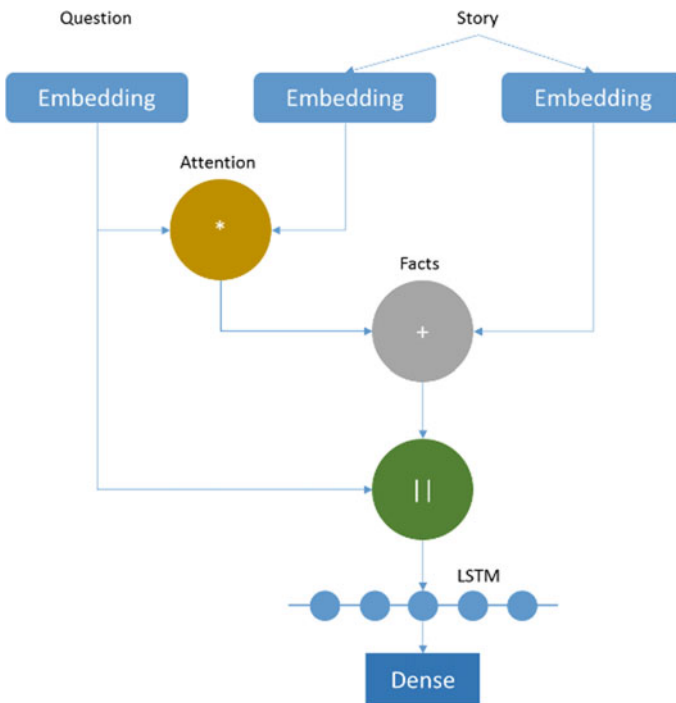


Fig. 3 Solution architecture involving embedding layers to dense layer that produces answer for a question given a story

This produces a vector which will have more weights to the words which are more relevant based on story and questions. This forms the network memory. This attention gets combined with the story embedding to find out facts about the question. These facts are then combined with questions and this information is fed into LSTM network to produce answer for a question.

4.2 Memory Network

This is the key component of the solution architecture and it controls information flow using gates. Memory networks [9] architecture enables them to store all the state from a given input, which means that queries can be asked and the memory is able to go many steps back to find the most suitable portions that need to be given greater “attention” as illustrated in the figure below (Fig. 4).

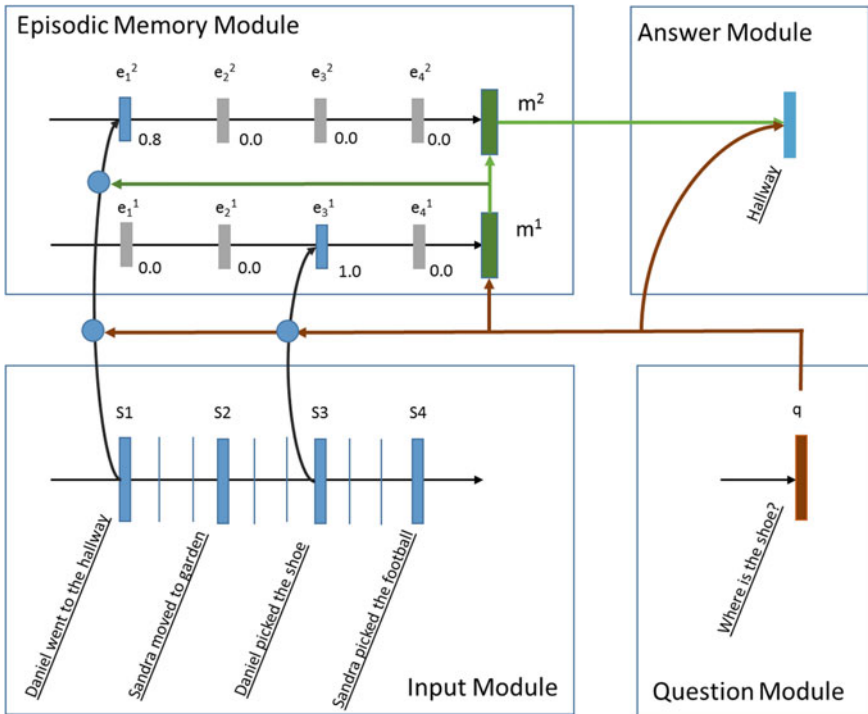


Fig. 4 Illustration showing memory network showing how relevant portions are being given more attention

4.3 Task Description and Model Results

This memory network was trained using different tuning parameters such as embedding size and LSTM size described below.

Embedding Size: This is the size of the Word2vec—word convert to a numerical vector.

Latent Size: Number of neurons in the LSTM layer.

Batch Size: Number of inputs processed together.

Epochs: Number of passes of forward and backward propagation in the deep network.

Changing dropout rate did not have any significant impact on model accuracy. Below are descriptions of each task and the results for different combinations and with 30% dropout for every model.

Task 1: This task consists of a story where answer is based on single supporting fact. For example,

Story: Dan went to the bedroom. Sunny moved to the bathroom. Q: Where is Dan?
A: bedroom.

Results for the task (Fig. 5, Table 1).

Epochs were not contributing to increased accuracy in a significant manner. Small latent size with less number of epochs is producing the best accuracy.

Task 4: Answer for this task will depend upon understanding two logical arguments. For example,

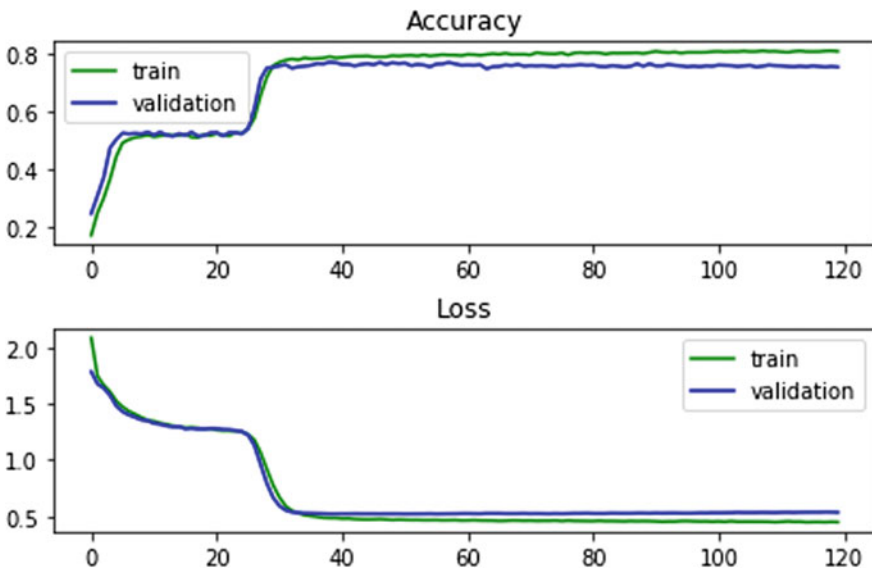


Fig. 5 Accuracy and loss for train and validation set w.r.t. epochs for Task 1

Table 1 Combinations of different parameters for Task 1 which shows changing architecture is having little impact on validation accuracy

Embedding size	Latent size	Batch size	Epochs	Validation accuracy (%)
32	32	64	120	77.98
32	64	64	300	74.83
32	128	64	120	76.2

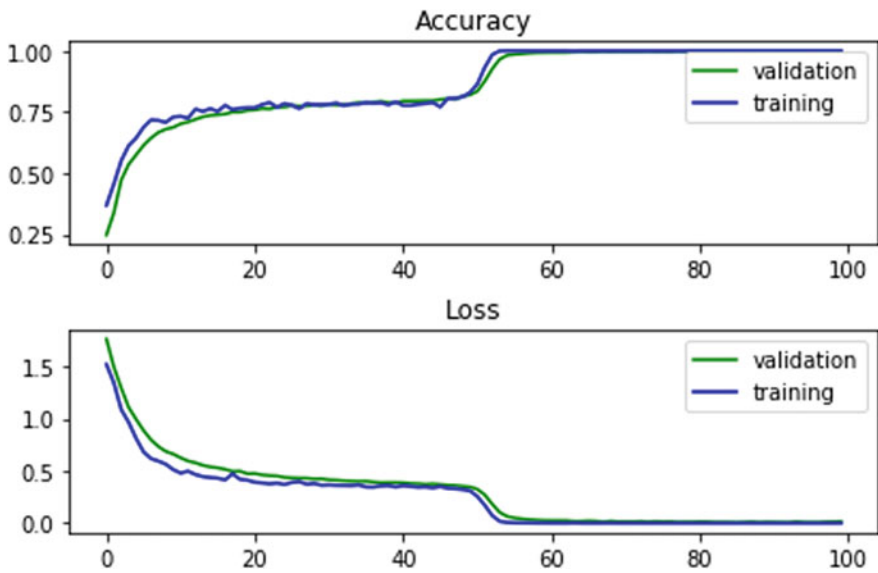


Fig. 6 Accuracy and loss for train and validation set w.r.t. epochs for Task 4

Table 2 Combinations of different parameters for Task 4

Embedding size	Latent size	Batch size	Epochs	Validation accuracy (%)
16	16	32	100	85.24
32	32	32	100	97.01
48	48	48	100	99.64

Story: The bathroom is west of the bedroom. The study room is east of the bedroom.

Q: Bedroom is west of what? A: study room.

Results for the task (Fig. 6, Table 2).

Increasing the embedding size is the key to obtain more accuracy on this task. Embedding is capturing more semantic relationship between words so increasing it is making model more robust. Even simple models give high accuracy on this task.

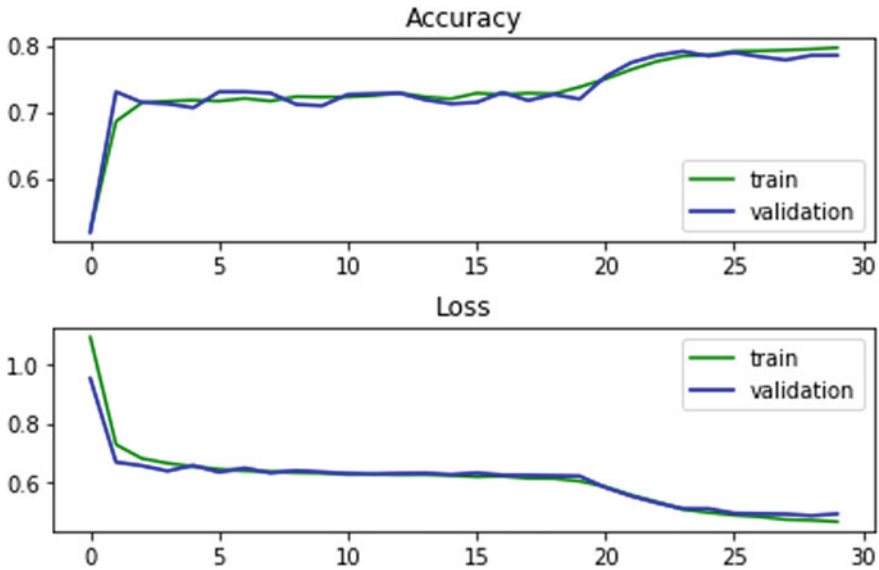


Fig. 7 Accuracy and loss for train and validation set w.r.t. epochs for Task 7

Table 3 Combinations of different parameters for Task 7

Embedding size	Latent size	Batch size	Epochs	Validation accuracy (%)
64	32	64	30	79.35
64	48	64	50	78.00
64	64	96	100	79.26

Task 7: This is a counting task, where counting of objects is to be done based on understanding the statements. For example,

Story: Jonny went to the garden. Daniel went to the bedroom. Sam picked up the apple there. Jonny got the milk there.

Q: Sandra is carrying how many objects? A: one.

Results for the task (Fig. 7, Table 3).

Other tests showed that embedding size was not contributing much to accuracy so the focus was on latent size and batch size. Overall we observe that models with low hyperparameter values are performing as well as those with high values.

Task 17: This is a task where the question relates to the position of an object based on given statements. For example,

Story: The red triangle is to the left of the yellow rectangle. The red triangle is below the pink square.

Q: Is the pink square to the right of the yellow rectangle? A: no

Q: Is the yellow rectangle to the right of the pink square? A: yes

Results for the task (Fig. 8, Table 4).

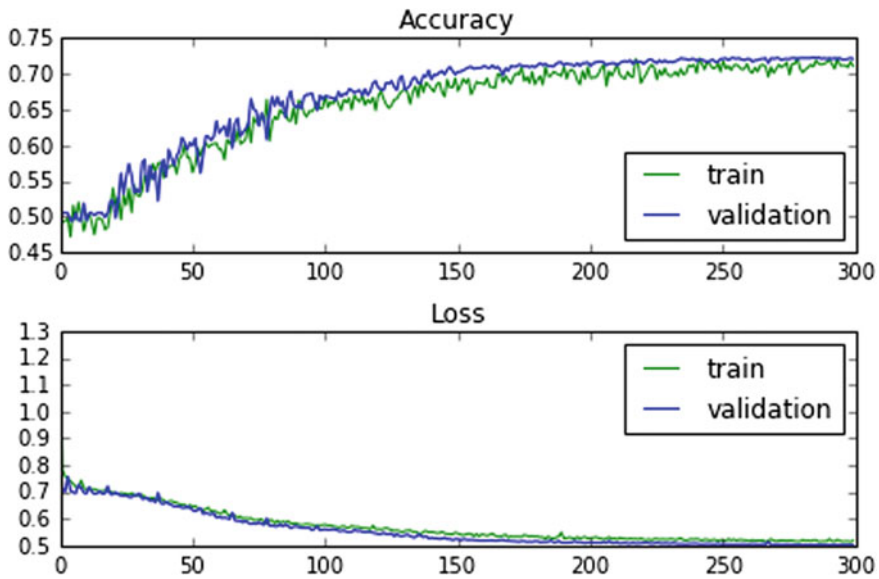


Fig. 8 Accuracy and loss for train and validation set w.r.t. epochs for Task 17

Table 4 Combinations of different parameters for Task 17 which shows increasing epochs and latent size does not improve accuracy much

Embedding size	Latent size	Batch size	Epochs	Validation accuracy (%)
32	64	64	300	70.70
32	128	64	300	72.10
32	300	64	300	72.30

Using different combinations of hyperparameters did not improve the accuracy of this task much with maximum accuracy of 72.3%. This is because in this task it becomes very hard for the memory network to exactly pinpoint the exact lines from which the solution gets extracted as there are several objects present in several positions.

5 Error Analysis

In this section, we analyze output on the basis of different criteria. This would help us in limitation and scope planning of output. The shortcomings which we have identified post-process will also help us to improve training models and algorithms.

Table 5 Accuracy rates by distance of relevant sentence

Distance	Accuracy (%)
1	100
2	100
4	65
5	55
7	20

Currently, this has only been done on the single supporting fact case. The following are a list of few experiments.

5.1 Position of Relevant Sentence from Answer

Our assumption is, if relevant sentence far from the question, then the algorithm has to disambiguate between lots of sentences (Table 5).

For example, in the case below, we see the question present at distance of 1 and 5, respectively, and therefore see a difference in accuracy.

Sandra went to the office

Daniel went to the hallway

Where is Daniel? *Prediction: Hallway Answer: Hallway*

Daniel went to the office

John moved to the hallway

Where is Sandra? *Prediction: Hallway Answer: Office*

We see that by the time the distance becomes seven sentences the accuracy reduces significantly since the question answering system has to look at much longer distances.

5.2 POS (Part of Speech) Tagging Cleanup

POS tagging will help in identification of content word which contains noun, verb, adverb, and adjectives. Also, it would help in ignoring the functional word and stop word like preposition and determiner (a, an, the, etc.). It would help in data cleanup process. The following sentences show input sentence and clean sentence. These extra words may increase the noise and reduce the discriminative power of the model.

Example: Daniel journeyed to the hallway gets converted to Daniel journeyed Hallway after excluding prepositions and determiners.

With POS tagging for 10 separate bootstrap tests, on an average the accuracy was **77.8%** with a while with ignoring unnecessary POS tagged keywords, we got average **82.8%** accuracy on test.

5.3 Number of Nearest Relevant Statement/Answer

In Babl dataset, the question is followed by two answers or statement. There are three different conditions.

- (a) Relevant sentence is not present in the previous two sentences.

John moved to the garden

Sandra moved to the bedroom

Where is **Mary**?

Here Mary word is not present in previous two relevant statements before the question “Where is Mary?”.

- (b) One relevant sentence is present in two different sentences.

Mary went to the bedroom

Daniel went to the garden

Where is **Mary**?

Here Mary is present only once in previous two statements before the question “Where is Mary?”

- (c) Two relevant sentences are present in the previous two sentences.

John moved to the bedroom

John journeyed to the hallway

Where is John?

Here John word is present in both previous two statements before the question, “Where is John?”.

Here, in Case a, algorithm has to look at relevant statement which is at longer distance from the question, so accuracy would be low.

In Case b, the algorithm has just identified answer from one relevant statement so this would give more accurate results.

Table 6 Accuracy rates for the three possible cases

Number of nearest relevant statement/answer	Accuracy
0	0.64
1	0.8
2	0.6

In Case c, the algorithm has to disambiguate probable answer from two relevant statements. So accuracy is expected to be lesser (Table 6).

6 Conclusion

In this paper, we have utilized recurrent neural network for typical question answering system. Generally, the question answering system explores feature engineering tasks like parsing, pos tagging, and tasking. We have explored how features like token sequences are important for question answering system. We have changed deep learning parameters like epoch, embedding size, latent size, and batch size and identified batch combination to generate the highest accuracy. We also have performed error analysis, where we identified how salient features like (1) presence of stop words, (2) number of relevant statement/answers for different questions, and (3) distance between relevant statement and questions can improve the algorithm in identification of current answers for respective question.

It has been shown that each task will have its own specific architecture with different parameters driving accuracy improvements in each context. So a one size fits all approach which cannot be used in the question and answering systems if we use this framework. Three of the four tasks described have good accuracy (75% plus) which shows we can create very accurate systems. Further steps would involve trying a radically different architecture on the tasks or performing NLP modifications on the inputs where accuracy obtained is low.

References

1. Wang, Z., Yan, S., Wang, H., Huang, X.: An overview of Microsoft deep QA system on Stanford WebQuestions benchmark. Technical report, Microsoft Research (2014)
2. Weston, J., Bordes, A., Chopra, S., Rush, A.M. van Merriënboer, B., Joulin, A., Mikolov, T.: Towards Ai-Complete Question Answering: A Set of Prerequisite Toy Tasks (2014). <https://arxiv.org/pdf/1502.05698.pdf>
3. Bordes, A., Usunier, N., Chopra, S., Weston, J.: Large-Scale Simple Question Answering with Memory Networks. <https://arxiv.org/abs/1506.02075,2015>
4. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-To-End Memory Networks (2015). <https://arxiv.org/pdf/1503.08895.pdf>
5. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks (2015). <https://arxiv.org/pdf/1409.3215.pdf>
6. Murdock, J.W.: Decision Making in IBM Watson Question Answering. Ontology Summit (2015). Web presentations https://researcher.watson.ibm.com/researcher/view_person_pubs.php?person=us-murdockj&t=1

7. Su, V.: Solving the Prerequisites: Improving Question Answering on the bAbI Dataset (2015). http://cs229.stanford.edu/proj2015/333_report.pdf
8. Kumar, A., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask Me Anything: Dynamic Memory Networks for Natural Language Processing (2015). arXiv preprint [arXiv:1506.07285](https://arxiv.org/abs/1506.07285)
9. LSTMs and Dynamic Memory Networks for Human-Written Simple Question Answering. <https://cs224d.stanford.edu/reports/zack.pdf>

Convolved Cosmos: Classifying Galaxy Images Using Deep Learning



Diganta Misra, Sachi Nandan Mohanty, Mohit Agarwal and Suneet K. Gupta

Abstract In this paper, a deep learning-based approach has been developed to classify the images of galaxies into three major categories, namely, elliptical, spiral, and irregular. The classifier successfully classified the images with an accuracy of 97.3958%, which outperformed conventional classifiers like Support Vector Machine and Naive Bayes. The convolutional neural network architecture involves one input convolution layer having 16 filters, followed by 4 hidden layers, 1 penultimate dense layer, and an output Softmax layer. The model was trained on 4614 images for 200 epochs using NVIDIA-DGX-1 Tesla-V100 Supercomputer machine and was subsequently tested on new images to evaluate its robustness and accuracy.

Keywords Convolution neural network (CNN) · Softmax · Dropout · Galaxy type

1 Background and Introduction

The galaxy classification problem was given initially by Edwin P. Hubble in 1926 [1] and later extended by the scientist: de Vaucouleurs [2]. galaxy classification is a major task involved with all domains of analytical astronomy with major research agencies relying on automatic classification of galaxies into their corresponding correct classes to further understand the properties and possibilities. Galaxy shapes form the basis of classification of galaxies. Three major shapes of galaxies are noticed

D. Misra (✉) · S. N. Mohanty
Kalinga Institute of Industrial Technology, Bhubaneswar, India
e-mail: mishradiganta91@gmail.com

S. N. Mohanty
e-mail: sachinandan09@gmail.com

M. Agarwal · S. K. Gupta
Bennett University, Greater Noida, India
e-mail: ma8573@bennett.edu.in

S. K. Gupta
e-mail: suneet.banda@gmail.com

by scientists: elliptical, spiral, and irregular. As we know, the orbits of planets are elliptical due to gravitational force [3]. Similarly, galaxies which are a group of stars also form shapes similar to symmetric curves due to mutual gravitation between stars. Stars are in constant motion with respect to each other due to gravitational pull which causes galaxies to rotate [4]. Irregular galaxies are an exception from above circular shapes and do not have any symmetric shape. This is due to a lot of elemental hydrogen or dust in these galaxies. As we also see on earth dust particles do not form any regular shape, thus the reason behind the shape of irregular galaxies.

Generally, there are three major classes of galaxy, which is described as follows.

1.1 *Elliptical Galaxies*

A galaxy which has the ellipsoidal shape and a smooth nearly featureless image [5]. Moreover, the spherical form in $3D$ appears to us as a circular shape in $2D$. These are classified into $E0$ which are perfectly circular to $E7$, which are most flattened. Such type of galaxy is the brightest at center and brightness diminishes moving away from the center [6] (Fig. 1).

1.2 *Spiral Galaxies*

These have three major components: bulge, disk, and halo. Bulge is the central portion which consists of old stars [7]. Arm is the linear portion of stars which are in circular motion around bulge and is made of dust and younger stars. Halo is the spherical part around bulge and covers some part of the disk. Arms emerge directly from bulge (ordinary spiral) or from a bar of material around bulge (barred spiral) [8]. They are also further classified into lower case letters: a, b, c., on how tightly the arms are bound to bulge. The category has most tightly bound arms (Fig. 2).

Fig. 1 Sample elliptical galaxy image

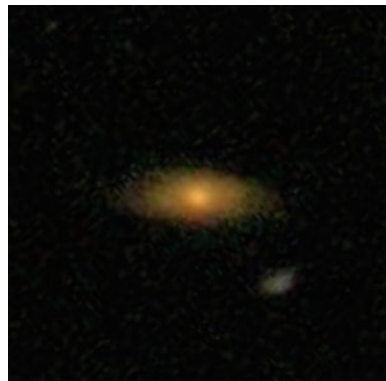


Fig. 2 Sample spiral galaxy image



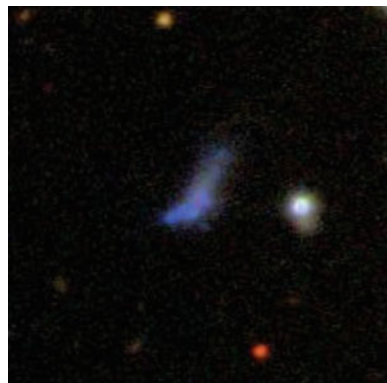
1.3 Irregular Galaxies

These are classified into Irregular I or Irregular II [9]. First category has a lot of elemental hydrogen and young stars. Irregular II has a lot of dust that makes the distinct stars not clearly visible [8] (Fig. 3).

We have made an effort to use deep learning and train the system using some existing labeled images. After this, if the machine is fed with new galaxy images, it will classify the new image into the correct type of galaxy with some degree of accuracy.

In this paper, we approached the problem of galaxy classification based on galaxy images into its three major classes elliptical, irregular, and spiral by applying a Deep Convolution Neural Network (CNN) Architecture. We applied the deep CNN comprising of four hidden layers, one flatten layer, and another dense layer using a Softmax Activation function. The network architecture used max pooling and dropouts wherever required and Tanh activation function was applied in every layer. The model was trained for 200 epochs on the dataset and then was subsequently tested on real-life images in batches of 64 for all the 3 categories. The model was

Fig. 3 Sample irregular galaxy image



highly accurate in its predictions and outperformed existing models and conventional machine learning models in this regard. The rest of the paper is organized in the following chronological order.

In Sect. 2, literature survey of related past work in this domain of galaxy classification is discussed. Followed by Sect. 3 which extensively elaborates about the proposed work and is divided into three corresponding subsections which are introduction to CNN, dataset description, and proposed architecture. Section 4 embeds the experimental setup and results obtained in forms of interactive visualizations and graphs. Concluding, Sect. 5 summarizes the whole paper and provides the conclusion of the research.

2 Related Work

Kormendy and Bender [10] explained about *S0* galaxies which are intermediate between *E7* (elliptical) and *Sa* (true spiral). They have a bulge and a disk, but no spiral so differs from both elliptical and spiral galaxies and are called lenticular galaxies. Authors make a parallel classification of *S0* galaxies to *Sa*, *Sb*, *Sc* and give them names *S0a*, *S0b*, *S0c*. This classification is done based on B/T ratio, i.e., bulge divided by total light. This value decreases from *a* to *c* for both spiral and lenticular galaxies. Buta et al. [11] proposed a new classification of galaxies by the name Comprehensive de Vaucouleurs revised Hubble-Sandage (CVHRS). In this, authors classified galaxies in notations of the form: *Sab-A Sab* galaxy that is closer to *Sa* than to *Sb*, *Sab-A Sab* galaxy that is closer to *Sb* than to *Sa* and so on for other galaxies. Shamir described the automatic classification of galaxy images into elliptical, spiral, or edge-on galaxies [12]. For the experimental purpose, authors used manually classified images to extract image features and discussed the Fisher score. Moreover, test images are classified using nearest weighted neighbor using the Fisher score as an important parameter. The author found automatic classification into elliptical, spiral, or edge with 90% accuracy [12].

In 2013, M. Martin et al. [9] have proposed galaxy classification algorithm based on Naive Bayes and random forest. The achieved accuracy was about 91% for the Naive Bayes and 79% for random forest classifier. In [13], authors have proposed a classification method based on supervised machine learning with non-negative matrix factorization for images of galaxies in the Zsolt Frei Catalog [14] and achieved the accuracy about 93%. In 2017 [15], authors proposed a new automated machine supervised learning astronomical classification scheme based on the non-negative matrix factorization algorithm with the accuracy of about 92%. Kim and Brunner [16] describe that most star-galaxy classifiers use reduced information from catalogs, this requires careful feature extraction and selection. With the latest advances in machine learning which use deep CNN allows the machine to automatically learn the features directly from the images and minimizes the need for human input. Authors present a star-galaxy classification framework using (ConvNets) directly on galaxy images pixel values.

In contrast to above mentioned approaches, we have used the deep convolution network to improve the accuracy. The details about the proposed model are discussed in the following sections.

3 Proposed Work

3.1 Dataset

In this research project, we classified galaxy Image data into its three corresponding major classes—elliptical type, spiral type, and irregular type using a Deep Convolutional Neural Network (CNN) architecture. The dataset containing the galaxy Images were downloaded from Kaggle [17] and NASA Hubble-Space Gallery Websites [18]. The dataset was categorized into three classes with images kept in two main folders: training and validation. The training folder is further subdivided into three subfolders for three classes: spiral, elliptical, and irregular. The validation folder is also similarly having three subfolders with same names as the classes. The number of images in these folders is listed in Table 1.

Initially, we had only 11 images for the Irregular class. So, we used the Augmentor Package of Python to perform Image Augmentation on those 11 images and generated 1615 augmented images.

Convolution Neural Network (CNN)—A deep learning framework used mostly for object detection and image classification.

Softmax—An activation function used for multiclass categorization.

Dropout—A regularization technique in neural networks to reduce overfitting.

Deep Learning—Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.

Galaxy—A cosmological cluster of stars systems and planets.

3.2 Brief Discussion About Convolutional Neural Network

Convolution Neural Networks (CNN) [19] is the state-of-the-art deep learning model for object recognition, image segmentation, classification, and analysis; but are not only restricted to this scope of problems and fare very well in many other aspects

Table 1 Different classes with the number of images for training, validation, and testing

Classes	Total images	Training set	Validation set	Testing set
Spiral	1464	1000	400	64
Elliptical	1464	1000	400	64
Irregular	1686	1232	390	64

including generating new images as used in Generative Adversarial Networks (GAN) [20]. CNNs are very robust and are very efficient in understanding the features in an image by applying simple convolution formula on the input features using random initialized weights.

The pictorial representation of the architecture of the convolutional neural network is given in Fig. 4.

We built our Convolutional Neural Network (CNN) model in Python using the Keras framework. The CNN architecture comprised of one Input Convolution 2D layer followed by four hidden layers, one penultimate dense layer, and finally one output layer. We used a filter size of 3×3 in each layer. All the images were resized

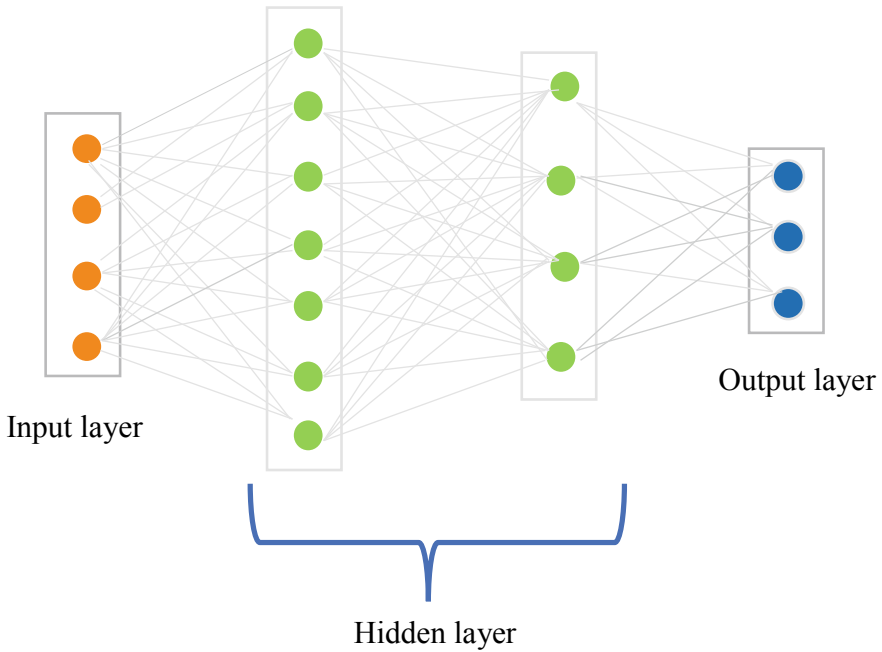


Fig. 4 CNN architecture Layman’s visualization

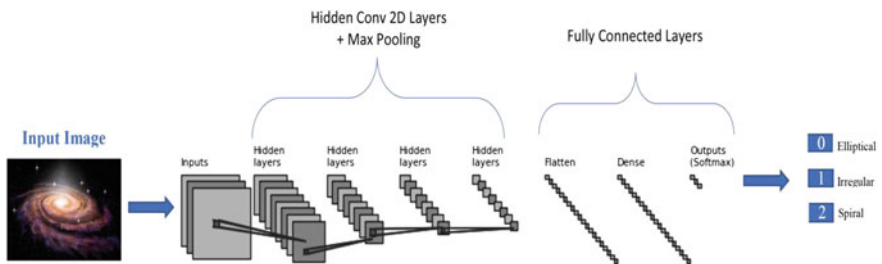


Fig. 5 The architecture of convolution neural network for galaxy classification

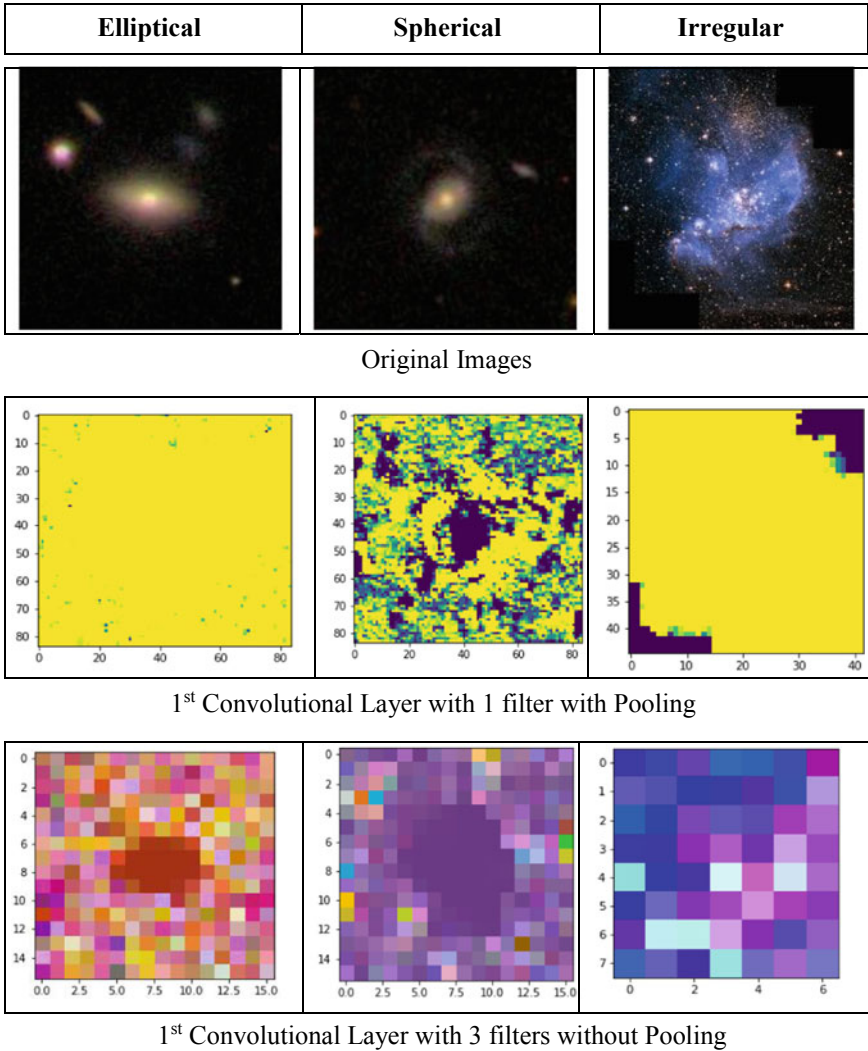


Fig. 6 Intermediate feature map representation with and without pooling

to 256×256 pixels. The batch size used was 64 and was trained for 200 epochs with 10 timestamps per epoch. We used dropout regularization after the hidden layers and before the output layer. The detailed architecture diagram is shown in Fig. 5.

After training the CNN model, we saved the weights of the model in weights .h5 file as we could now perform testing easily as many times we needed using the already trained model weights. This step was also performed as training took nearly 1.5 h to complete and it was not feasible to train the model every-time for testing. Thus, this time was saved by making the weights .h5 file for testing. The output images after the execution of intermediate operation of CNN is depicted in Fig. 6.

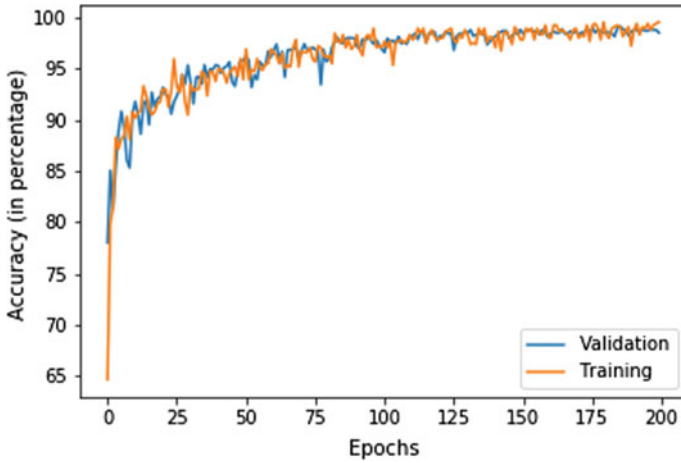


Fig. 8 Graphical representation of accuracy versus epochs

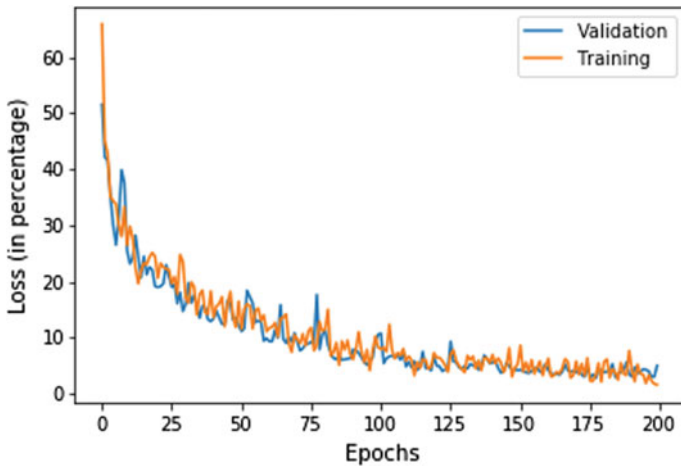


Fig. 9 Model loss curve

Elliptical 100%
Spiral 100%
Irregular 92.187%.

The testing accuracy of images of the various galaxy types is represented in the following Fig. 10 using bar chart representation.

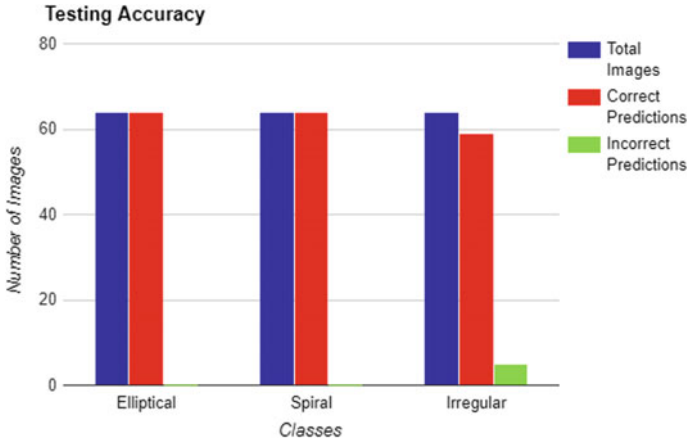


Fig. 10 The representation of testing accuracy using bar chart

5 Conclusion

In this research article, a galaxy classification algorithm has been proposed. The algorithm is based on deep convolution neural network with four hidden layers, max pooling layers after every convolution layer and one dense layer. The proposed research will be helpful for astronomical scientists and cosmologists. It will help to classify huge collection of galaxy images without manual effort of viewing each image individually. Research can be fine-tuned for further classification of galaxies into their subclasses as explained in the previous sections. The testing time was reduced to few seconds by saving the CNN weights file and thus it will be working on real time scenarios also.

References

1. Odewahn, S.C., Stockwell, E.B., Pennington, R.L., Humphreys, R.M., Zumach, W.A.: Automated star/galaxy discrimination with neural networks. In: *Digitised Optical Sky Surveys*, pp. 215–224. Springer (1992)
2. De Vaucouleurs, G.: Classification and morphology of external galaxies. In: *Astrophysik iv: Sternsysteme/Astrophysics iv: Stellar Systems*, pp. 275–310. Springer (1959)
3. Hupp, E., Roy, S., Watzke, M.: NASA finds direct proof of dark matter. Press Release (2006)
4. Basu, A., Mandal, A., Das, A.: Blueprint of the graphical galaxy of solar system segment: an algorithmic approach using CSS3.0
5. Liller, M.H.: The distribution of intensity in elliptical galaxies of the virgo cluster ii. *Astrophys. J.* **146**, 28 (1966)
6. Nieto, J.-L., Capaccioli, M., Held, E.V.: More isotropic oblate rotators in elliptical galaxies. *Astron. Astrophys.* **195**, L1–L4 (1988)
7. Mihalas, D., Binney, J.: *Galactic astronomy** wh freeman co. San Francisco (1968)

8. Benjamin, R.A., Churchwell, E., Babler, B.L., Indebetouw, R., Meade, M.R., Whitney, B.A., Watson, C., Wolfire, M.G., Wolff, M.J., Ignace, R., et al.: First glimpse results on the stellar structure of the galaxy. *Astrophys. J. Lett.* **630**(2), L149 (2005)
9. Grebel, E.K.: The evolutionary history of local group irregular galaxies. In: *Origin and Evolution of the Elements*, p. 234 (2004)
10. Kormendy, J., Bender, R.: A revised parallel-sequence morphological classification of galaxies: structure and formation of s0 and spheroidal galaxies. *Astrophys. J. Supp. Ser.* **198**(1), 2 (2011)
11. Buta, R.J., Sheth, K., Athanassoula, E., Bosma, A., Knapen, J.H., Laurikainen, E., Salo, H., Elmegreen, D., Ho, L.C., Zaritsky, D., et al.: A classical morphological analysis of galaxies in the spitzer survey of stellar structure in galaxies (S4G). *Astrophys. J. Supp. Ser.* **217**(2), 32 (2015)
12. Shamir, L.: Automatic morphological classification of galaxy images. *Mon. Not. R. Astron. Soc.* **399**(3), 1367–1372 (2009)
13. Selim, I.M., Keshk, A.E., El Shourbugy, B.M.: Galaxy image classification using non-negative matrix factorization. *Int. J. Comput. Appl.* **137**(5) (2016)
14. Astronomía, M., Cordero Garayar, J.P., Campusano Brown, L.E., Blanc Mendiberri, G., De Propris, R., Muñoz Vidal, R.: The dry merger rate and merger relic fraction in the coma cluster core
15. Selim, I.M., Abd El Aziz, M.: Automated morphological classification of galaxies based on projection gradient nonnegative matrix factorization algorithm. *Exp. Astron.* **43**(2), 131–144 (2017)
16. Kim, E.J., Brunner, R.J.: Star-galaxy classification using deep convolutional neural networks. *Mon. Not. R. Astron. Soc.* 2672 (2016)
17. <https://www.kaggle.com/c/galaxy-star-separation/data>
18. Van Dyk, S.D., Peng, C.Y., Barth, A.J., Filippenko, A.V.: The environments of supernovae in post-refurbishment hubble space telescope images. *Astron. J.* **118**(5), 2331 (1999)
19. Haykin, S.: *Network, neural: a comprehensive foundation*. *Neural Netw.* **2**(2004), 41 (2004)
20. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2015). [arXiv:1511.06434](https://arxiv.org/abs/1511.06434)
21. De La Calleja, J., Fuentes, O.: Machine learning and image analysis for morphological galaxy classification. *Mon. Not. R. Astron. Soc.* **349**(1), 87–93 (2004)
22. Angelica Marin, M., Enrique Suar, L., Gonzalez, J.A., Diaz, R.: A hierarchical model for morphological galaxy classification. In: *FLAIRS Conference* (2013)
23. Khalifa, N.E.M., Taha, M.H.N., Ella Hassanien, A., Selim, I.M.: Deep galaxy: classification of galaxies based on deep convolutional neural networks (2017). [arXiv:1709.02245](https://arxiv.org/abs/1709.02245)

Advances in Network Technologies

Energy-Based Improved MPR Selection in OLSR Routing Protocol



Rachna Jain and Indu Kashyap

Abstract Wireless Ad hoc networks are consisting of wireless nodes that communicate over wireless medium without any centralized controller, fixed infrastructure, base station, or access point. The networks should be established in a distributed and decentralized way. Performance of mobile Ad hoc network depends on the routing scheme chosen. Extensive research has been taken place in recent years to suggest many proactive and reactive protocols to make them energy efficient. In this work, table-driven routing protocol, i.e., optimized link-state routing protocol (OLSR) is tried to make more energy efficient which also helps in prolonging the network lifetime. OLSR is a proactive routing protocol in Mobile Ad hoc Networks (MANETS) which is driven by hop-by-hop routing. The conventional OLSR is hybrid multipath routing, in which link-state information is forwarded only by Multi-Point Relays (MPRs) selected among one-hop and two-hop neighbor sets of host. In this work, a novel mechanism is introduced to select MPR among nodes neighbor set to make it more energy efficient by considering will-iness of node. Proposed energy-aware MPR selection in MDOLSR is compared with conventional OLSR. Extensive simulations were performed using NS-2 simulator, and simulation results show improved network parameters such as higher throughput, more Packet Delivery Ratio (PDR) and lesser end-to-end delay as simulation time progresses.

Keywords Multi-point relay (MPR) · Optimized link-state routing (OLSR) · Packet delivery Ratio (PDR) · Residual energy · Throughput

R. Jain (✉) · I. Kashyap

Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India
e-mail: Rachna_19802000@yahoo.com

I. Kashyap

e-mail: Indu.fet@mriu.edu.in

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_41

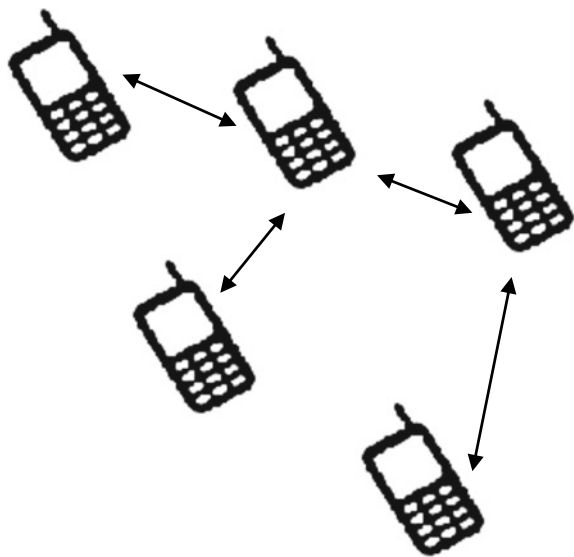
583

1 Introduction

Mobile Ad hoc Networks (MANETS) are general-purpose multi-hopping Ad hoc networks [1]. In these networks, mobile nodes can freely roam within the communication range. These nodes have limited battery capacity [2]. Due to the mobility of nodes topology of the network can change at any time. Routing of data, conservation of energy and bandwidth, and security of information have become major challenges in these battery operated nodes. Despite all these challenges, ad hoc networks have become popular especially in military applications, search and rescue operations, and vehicular communications. Based on network topology routing protocols are broadly categorized as proactive, reactive, hybrid as well as cluster based. In proactive routing protocols, each node maintains a routing table to carry forward the data, whereas in reactive routing route is searched on demand. Hybrid routing protocol combines the advantages of both. Figure 1 shows a scenario of mobile Ad hoc networks.

Although many protocols have been suggested for energy-efficient routing in MANETS but selecting a single path results in quick depletion of energy among nodes. Multipath routing ensures better utilization of the network through load balancing. Routing protocols depending on network topology can be divided into proactive, reactive, and hybrid routing protocol. In proactive routing protocol, every node maintains a routing table. Destination Sequence Distance Vector (DSDV) and Optimized Link-State Routing (OLSR) [3] come under this category. In case of reactive protocols, route is searched on demand. Ad hoc Distance Vector Routing (AODV) and Dynamic Source Routing (DSR) come under this category. Hybrid routing combines the advantages of both proactive and reactive routing.

Fig. 1 Mobile Ad hoc networks



Zone Routing protocol (ZRP) comes under this category. This paper is organized as follows: Sect. 2 discusses related work, Sect. 3 discusses OLSR routing protocol in detail, Sect. 4 discusses the concept of energy willingness in detail. Section 5 tells about performance parameters studied and Sect. 6 shows the simulation scenario and results obtained. In last, Sect. 7 discusses the conclusion and future work.

2 Related Work

A detailed literature survey of energy-efficient MANET routing protocols especially focusing on OLSR routing protocol is discussed in this section.

In [4] authors have proposed a modified OLSRM protocol which is based on conventional OLSR. Energy-aware metric used in this protocol is the number of nodes alive against varying simulation time of nodes. Authors have framed an analytical model to compute the correct behavior of the network.

In [5] authors have proposed EEOLSR energy-efficient routing protocol in MANETS which extends network lifetime without losing other Quality of Service (QoS) parameters. Multi-point Relay selection is done on the basis of node willingness to become MPR. This process of MPR selection is discussed in detail in Sect. 3.

In [6] a different approach is used to make energy-efficient OLSR with Autoregressive Integrated Moving Average timeseries model (ARIMA). In this method per interval energy consumption is computed. Composite energy cost model is developed with the help of the calculation of residual energy of the node and consumed transmission power of the node. Sakthivel et al. [7] gave a different metric for MPR selection using energy accuracy metric. Since MPR node is selected from one-hop neighbor set of the host which covers its two-hop neighbors also. Modified MPR is selected from an ordered set based on the value of highest residual energy along with considering timestamp values.

Jabbar et al. [8] gave a new metric, i.e., Multi-Criteria Node Rank metric (MCNR) metric comprising residual battery energy along with node speed to propose a new energy and mobility aware (EMA-OLSR) scheme for selection of MPR in OLSR. Jabbar et al. [9] have proposed Multipath Battery Aware OLSR (MBA-OLSR) which is based on OLSRv2 and its multipath links between source and destination. In this work, the author has considered remaining battery power of nodes to calculate the initial cost of links between sender and destination. Simulations are done using EXata simulator and results show better network lifetime.

Natarajan and Rajendran [10] proposed AOLSR which is hybrid advanced OLSR by modifying Dijkstra's algorithm to compute multiple paths in both dense and sparse networks from source to destination. Whereas Dhanalakshmi [11] proposed Energy Conserving Advanced Optimized Link-State Routing (ECAO) model by calculating energy costs of all nodes and compared results with OLSR and

AOLSR. The author has also implemented modified Dijkstra's algorithm to compute multipath and also checked for link failure for further analysis.

Moussaoui [12] devised stability of nodes (SND) and fidelity of nodes (FND) parameters to elect Multi-Point Relays (MPR) in case of OLSR. The author has proved via simulation that his selected method provides better Quality of Service (QoS) parameters than the traditional method in OLSR for MPR selection which considers Expected Transmission Time metric (ETX). The limitation of using ETX is overestimating link delivery ratio when packet size is too large.

Rango [13] gave a unique routing protocol which considered both link-stability and energy-aware parameters by proposing Link-stability and Energy-aware Routing protocols (LAER). The author has considered link stability along with minimum drain rate energy consumption. Same author Rango [14] has extended Dynamic Source Routing (DSR) protocol packet format to make it energy efficient. Modified DSR header packet included cost function both in route request and route reply packet. Both DSR and OLSR routing protocols have been thoroughly studied from an energetic point of view. The path according to minimum cost function is organized from best to worst path. Many node disjoint paths have been found, out of which most energy-efficient path is chosen. Since node disjoint provides higher fault tolerance. In non-disjoint routes link or node failure causes many routes to fail. Whereas in node or link disjoint routes link failure will cause only a single route to fail.

Another approach to improve routing efficiency is to use multipath in comparison with a single path. More than one path is used to increase routing efficiency. Multiple paths can be linked disjoint or node disjoint. In link disjoint paths, there is no common link, whereas in node disjoint path there is no common node. In link, disjoint paths goal is to reduce delay and increase efficiency. Multipath AODV or Multipath OLSR is proposed to prove it.

3 Optimized Link-State Routing Protocol

Routing algorithm for Ad hoc networks can be classified as routers obtain routing information along with the type of information used to calculate routing path. According to the way routers get information, protocols are classified as proactive, reactive, or hybrid as shown in Fig. 2. According to the type of information proactive protocols are further subdivided as distance-vector-based or link-state-based. Distance-vector protocol uses distance information to obtain path up to destination, whereas link-state-based protocols use topological information to obtain path up to the destination. Link-state algorithms are less inclined to routing loop formation as changes in network topology are known by nodes as compared to distance vector routing protocols. But link-state protocols are more expensive for implementation since they require more CPU power.

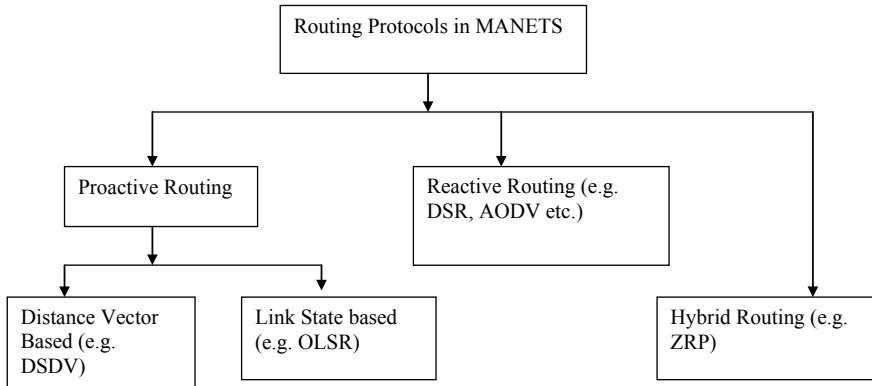


Fig. 2 Classification of routing protocols

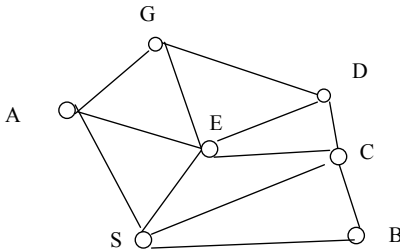
Optimized link-state routing protocol (OLSR) is a proactive, table-driven routing protocol which works on the concept of link sensing. Being a proactive routing protocol, routes are always available. But, if there is any change in network topology due to the mobility of nodes then the information should be flooded through the entire network. To control this flooding, Multi-Point Relays (MPR) are selected from one-hop neighbor set and two-hop neighbor set from the source node. Only selected MPR nodes are now responsible for the flooding of information, hence conserving the bandwidth by avoiding unnecessary flooding. Selecting the minimum number of one-hop neighbors which covers all two-hop neighbors is the goal of MPR selection. There are two types of control messages in OLSR.

- (i) HELLO Messages
- (ii) Topological Control Messages (TC Messages).

HELLO messages are periodically broadcasted to the neighbors that are only one-hop away. Hello messages obtain information about local links and its neighbors. Since links can be unidirectional or bidirectional, the host should know about all its neighbors. Links are subdivided as symmetric (bidirectional), asymmetric (unidirectional), or heard.

HELLO messages are periodically broadcasted after HELLO refresh time period and received by all one-hop neighbors. Each node attaches a list of its own neighbors to obtain the information about two-hop neighbors as shown by Fig. 3. Once the node has information about its one-hop and two-hop neighbors, it can select Multi-Point Relays (MPRs) which covers all two-hop neighbors as explained by Fig. 4.

Each node periodically broadcasts its HELLO messages to obtain the information about its neighbors and their link status. HELLO message contains a list of



Source Node	1-hop Neighbors	2-hop neighbors	Multipoint Relay
S	A,E,C,B	G,D	E

Fig. 3 Network example for MPR selection

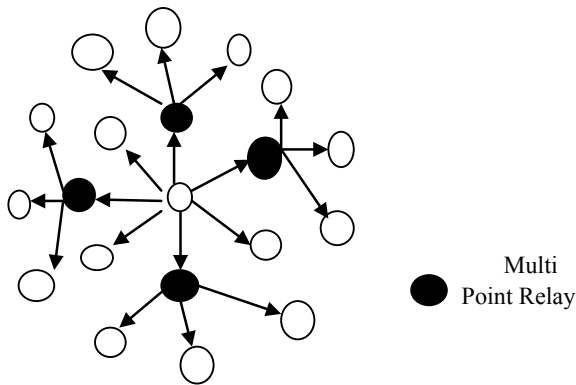


Fig. 4 Diffusion of broadcast message using MPR

addresses of neighbors to which valid bidirectional (symmetric) link exists. It also has a list of addresses of neighbors which have heard HELLO messages but the link is not bidirectional (asymmetric). Format of HELLO message is shown in Fig. 5.

Message Type	Vtime	Message Size	
Originator Address			
Time to Live	Hop Count	Message Sequence Number	
Reserved		Htime	Willingness
Link Code	Reserved	Link Message Size	
Neighbor Interface Address			
Neighbor Interface Address			

Fig. 5 HELLO message format in OLSR

Destination Address	Destination's MPR	MPR selector sequence number	Holding time
---------------------	-------------------	------------------------------	--------------

Fig. 6 TC message format in OLSR

HELLO message helps in link sensing as link code has the information about link type (Symmetric or Asymmetric) and neighbor type. Each node has knowledge about its two-hop neighbors so it helps in neighbor detection also. On the basis of this information, each node selects its MPR's which is responsible for the forwarding of control packets. On reception of HELLO message, each node constructs its MPR selector table. Multi-point relays are declared in the transmitted HELLO messages. If there is any change in the neighborhood up to two-hop neighbors, then multi-point relay set is recalculated.

Each node maintains a topology table on the basis of Topology Control (TC) messages. Routing tables are calculated on the basis of topology table. Format of TC message is shown in Fig. 6.

If there is entry to the same destination node with higher sequence number then the TC message is ignored. New topology entry is recorded if there is any entry with lower sequence number to the same destination node. Holding time of entry is refreshed in case entry is same as in TC message. New entry is recorded in case there is no corresponding entry.

Each node maintains a routing table to all known destinations in the network on the basis of topology table and neighbor table. Routing table is recalculated after a change in topology table. Destination address, next-hop address, and distance to every node are stored in the routing table.

4 Concept of EA: Willingness in OLSR

4.1 Energy Consumption Model

All the nodes in wireless networks are equipped with IEEE 802.11 g Network Interface Card. Energy required to transfer a packet p from any node is

$$E = i * v * tp \tag{2}$$

where i = Current (in Ampere), v = Voltage (in Volts), tp = Time taken to transmit the packet p (in Seconds)

Transmission time tp is given by

$$tp = (ph/6.10^6 + pd/54.10^6) \quad (3)$$

where ph = Packet Header in bits, pd = Size of payload

Assuming that energy consumption caused by packet overhearing is same as energy consumed in actually receiving a packet.

$$E(p, na) = Etx(p, na) + Erx(p, nb) + (N - 1)\{Eo(p, ni)\} \quad (4)$$

where $Etx(p, na)$ = Amount of energy spent in transmitting packet from node na , $Erx(p, nb)$ = Amount of energy spent in receiving a packet by node nb , $Eo(p, ni)$ = Amount of energy spent by overhearing nodes, N = Average number of neighboring nodes affected by transmission, $E(p, na)$ = Total energy spent in transmitting packet from node na to node nb .

From Eq. 4 it can be concluded that if the network is denser then overhearing causes much energy consumption. Modification is introduced in OLSR to include node willingness to be chosen as Multi-point Relay (MPR). In OLSR routing protocol, each node declares its willingness to be selected as MPR by default. In the proposed work, willingness is decided on the basis of remaining battery capacity of the node along with its lifetime which is dependent on drain rate.

4.2 Passive Hearing by Neighboring Nodes

As simulation time progresses, we can see from simulation results that energy of nodes that are not actively participating in routing is also becoming low. This is due to overhearing by neighboring nodes. Overhearing can be avoided by turning off device/node when unicast message is exchanged among neighboring nodes. It can be obtained by sending signaling messages (RTS/CTS) at MAC layer.

5 Performance Metrics

Performance parameters considered in this work are throughput, end-to-end delay, Packet Delivery Ratio (PDR), and energy required for transmission.

Throughput: It is the measure of rate of successful data transfer over the network. It is measured in bits per second (bps).

End-to-end delay: It is the total time taken in data transfer from source node to destination node. Latency in route discovery, delay in retransmissions at MAC, propagation delay as well as transfer time all contributes to end-to-end delay.

Packet Delivery Ratio (PDR): It is defined as a ratio of packet received by the destination node to the packets sent by source node. Higher the PDR, better is efficiency of routing protocol.

Consumed Energy: It is the sum of total energy spent by all the nodes.

Residual Energy: It is the initial energy minus consumed energy of the node.

Energy Cost per packet: It is the ratio of total consumed energy over number of successfully received packets at the destination.

6 Simulation Results

In this work, MDOLSR is built over UM-OLSR [19]. Simulation parameters are listed in Table 1.

Figure 7 shows that HELLO packet transmission is taking place while the simulation is in progress.

In this work, different performance parameters of modified OLSR, (MDOLSR) such as throughput of the network, packet delivery ratio, energy cost per packet are

Table 1 Network parameters

In this work following initial parameters are taken.	
Channel type	Wireless Channel
Radio propagation model	Two Ray Ground
Mobility Model	Random Way Point
MAC type	IEEE 802.11 a
Interface queue type	Drop Tail
Antenna model	Omni Antenna
Agent	UDP (User Data Protocol)
Application	CBR (Constant Bit Rate)
MANET topology	Dynamic
Speed of nodes	1 m/s...10 m/s
Max packet in ifq	50
Number of mobile nodes	20
Simulation area (X)	870
Simulation area (Y)	870
Simulation time	30 s....400 s
Pause time of nodes	50 s250 s
Initial energy	100 J per node



Fig. 7 Network scenario when the nodes have medium energy level

studied according to varying network conditions such as different values of pause times of nodes, varying node speed, and different values of simulation times (Table 2).

Simulation results obtained from modified OLSR are compared with conventional OLSR as shown in Fig. 8 which shows better results for PDR in case of MDOLSR, whereas Fig. 9 justifies the lesser time taken to search the route as the speed of node increases.

Now pause time of nodes is varied from 50 to 250 s and different performance parameters have been studied (Table 3).

The above results obtained are compared with conventional OLSR for same network parameters as depicted in the following graphs. Figure 10 shows there is a significant improvement in PDR of MDOLSR in comparison to traditional OLSR.

Now simulation time of nodes has been varied, and different performance parameters are captured. Table 4 clearly shows that as simulation time progresses residual energies of nodes deplete sharply.

Figure 11 shows the improvement in the throughput of the network of proposed MDOLSR in comparison to OLSR.

Table 2 Node's speed versus QoS parameters

Speed of node (m/s)	Total initial energy (J)	Consumed energy (J)	Residual energy (J)	Energy cost per packet (mWh)	Average energy (J)	Throughput (bytes/s)	PDR (%)	End-to-end delay (s)
1	2000	331.37	1668.62	9.36	16.57	54,464.2	50.38	1.24
4	2000	128.85	1871.15	1.97	6.44	93,571.6	93.08	0.006
6	2000	392.97	1607.03	8.34	19.65	73,470.6	67.05	0.908
7	2000	262.95	1737.04	4.22	13.15	92,629.7	88.56	0.013
8	2000	402.03	1597.97	8.49	20.10	68,018.2	67.36	0.887
9	2000	285.3	1714.7	29.36	14.27	11,263.8	13.83	3.35
10	2000	383.03	1616.96	6.37	19.15	98,112.5	85.61	0.0216

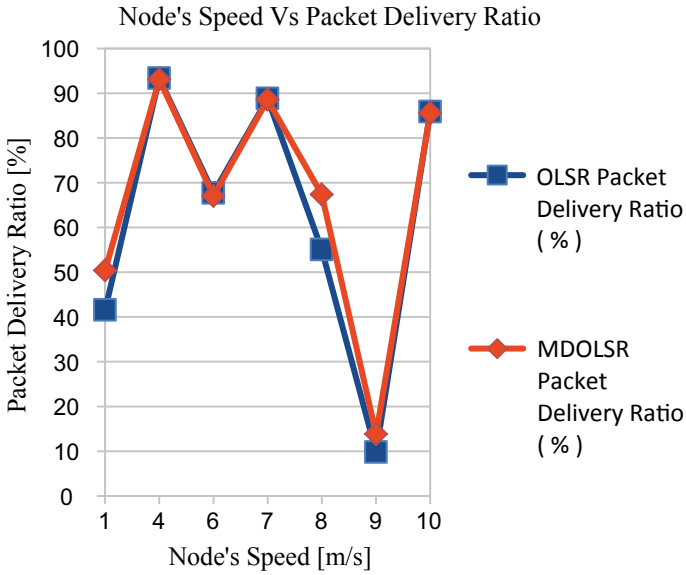


Fig. 8 Packet delivery ratio versus node's speed

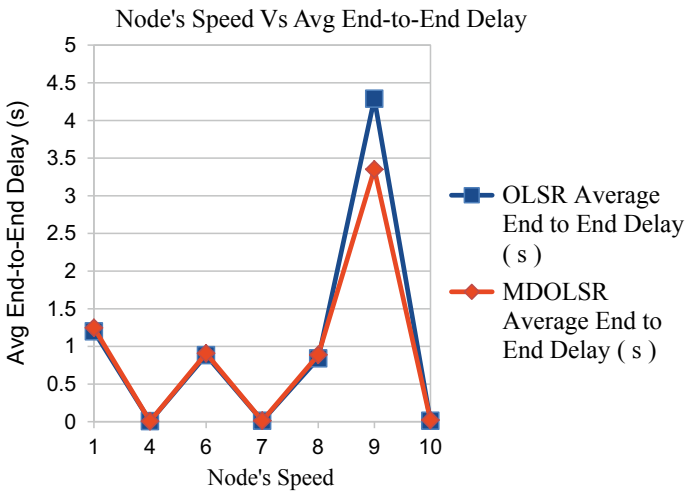


Fig. 9 End-to-end delay versus node's speed

Table 3 Pause time versus QoS parameters

Pause time of node (s)	Total Initial energy (J)	Consumed energy (J)	Residual energy (J)	Energy cost per packet (mWh)	Average energy (J)	Throughput (bytes/s)	PDR (%)	End-to-end delay (s)
50	2000	346.86	1653.13	9.7	17.34	52,355.5	50.89	1.24
100	2000	359.61	1640.38	7.54	17.98	71,637.2	67.81	0.88
150	2000	430.04	1569.96	12.14	21.50	53,490	50.43	1.23
250	2000	328.42	1671.58	5.38	16.42	91,143.7	86.88	0.014

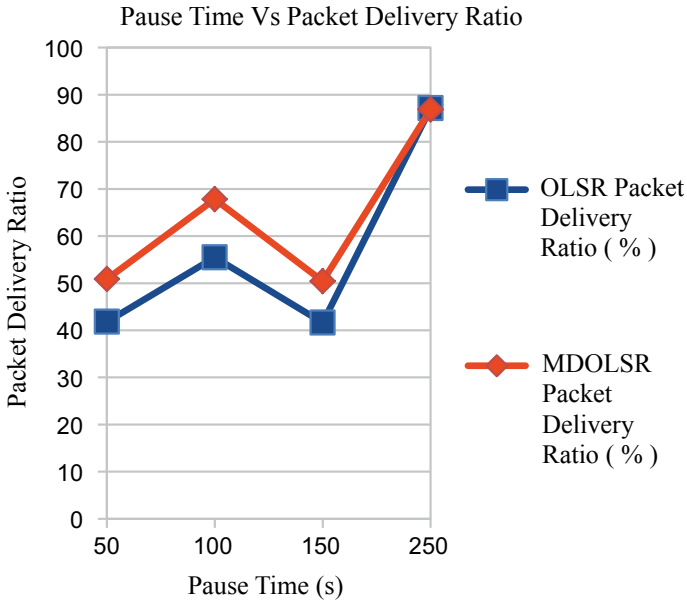


Fig. 10 Packet delivery ratio versus pause time

7 Conclusion and Future Work

Extensive simulations of OLSR routing protocol are chosen over other proactive protocols as it instantly knows about the status of the link and Quality of Service (QoS) information can be provided immediately. Another reason for considering OLSR is that reactive routing protocols result in large overhead as nodes are highly mobile but overhead in OLSR is independent of traffic as there is a fixed upper bound limit for overhead. This routing protocol is suitable for the large and dense network. Proposed methodology in modified MDOLSR results in better QoS parameters such as higher values of throughput and PDR whereas lesser values of energy cost per packet along with a smaller end-to-end delay. In future work energy-efficient OLSR can be further optimized using meta-heuristic techniques such as ant colony optimization, etc.

Table 4 Simulation time versus QoS parameters

Simulation time (s)	Total initial energy (J)	Consumed energy (J)	Residual energy (J)	Energy cost per packet (mWh)	Average energy (J)	Throughput (bytes/s)	PDR (%)	End-to-end delay (s)
30	2000	104.23	1895.77	1.59	5.21	98,170.5	93.08	0.0058
80	2000	689.14	1310.86	3.83	34.45	257,218	95.06	0.0119
130	2000	1609.8	390.19	6.45	80.49	349,408	90.24	0.0162
180	2000	1686.35	313.64	10.94	84.32	273,761	53.03	0.8454
200	2000	1601.37	398.63	11.33	80.06	197,894	55.10	1.6754
300	2000	1591.64	408.35	18.19	79.58	65,943.3	12.26	1.9375
400	2000	1685.6	314.4	6.74	84.28	280,896	73.32	0.0146

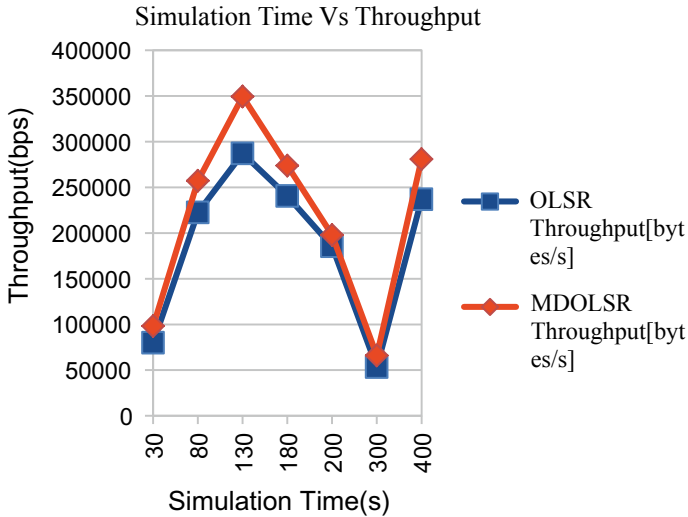


Fig. 11 Simulation time versus throughput

References

1. Chlamtac, I., Conti, M., Liu, J.J.N.: Mobile ad hoc networking: imperatives and challenges. *Ad Hoc Netw.* **1**(1), 13–64 (2003)
2. Conti, M., & Giordano, S.: Multihop ad hoc networking: the evolutionary path. In: *Mobile Ad Hoc Networking: Cutting Edge Directions*, 2nd edn., pp. 1–33 (2013)
3. Clausen, T., Dearlove, C., Jacquet, P., Herberg, U.: The optimized link state routing protocol version 2. In: *Internet Engineering Task Force (IETF)*, No. RFC 7181 (Standards Track) (2014)
4. Joshi, R.D., Rege, P.P.: Implementation and analytical modelling of modified optimised link state routing protocol for network lifetime improvement. *IET Commun.* **6**(10), 1270–1277 (2012)
5. De Rango, F., Fotino, M., Marano, S.: EE-OLSR: energy efficient OLSR routing protocol for mobile ad-hoc networks. In *Military Communications Conference, 2008. MILCOM 2008*. IEEE, pp. 1–7. IEEE (2008)
6. Guo, Z., Malakooti, S., Sheikh, S., Al-Najjar, C., Malakooti, B.: Multi-objective OLSR for proactive routing in MANET with delay, energy, and link lifetime predictions. *Appl. Math. Model.* **35**(3), 1413–1426 (2011)
7. Sakthivel, M., Palanisamy, V.G.: Enhancement of accuracy metrics for energy levels in MANETs. *Comput. Electr. Eng.* **48**, 100–108 (2015)
8. Jabbar, W.A., Ismail, M., Nordin, R.: Energy and mobility conscious multipath routing scheme for route stability and load balancing in MANETs. *Simul. Model. Pract. Theory* **77**, 245–271 (2017)
9. Jabbar, W.A., Ismail, M., Nordin, R.: Performance evaluation of MBA-OLSR routing protocol for MANETs. *J. Comput. Netw. Commun.* **2014**, 10 (2014)
10. Natarajan, D., Rajendran, A.P.: AOLSR: hybrid ad hoc routing protocol based on a modified Dijkstra's algorithm. *EURASIP J. Wirel. Commun. Netw.* **2014**(1), 90 (2014)

11. Dhanalakshmi, N., Alli, P.: Efficient energy conservation in MANET using energy conserving advanced optimised link state routing model. *Int. J. Parallel Emergent Distrib. Syst.* **31**(5), 469–480 (2016)
12. Moussaoui, A., Semchedine, F., Boukerram, A.: A link-state QoS routing protocol based on link stability for mobile ad hoc networks. *J. Netw. Comput. Appl.* **39**, 117–125 (2014)
13. De Rango, F., Guerriero, F., Fazio, P.: Link-stability and energy aware routing protocol in distributed wireless networks. *IEEE Trans. Parallel Distrib. Syst.* **23**(4), 713–726 (2012)
14. De Rango, F., Cano, J.C., Fotino, M., Calafate, C., Manzoni, P., Marano, S.: OLSR vs DSR: a comparative analysis of proactive and reactive mechanisms from an energetic point of view in wireless ad hoc networks. *Comput. Commun.* **31**(16), 3843–3854 (2008)

A Novel Approach for Better QoS in Cognitive Radio Ad Hoc Networks Using Cat Optimization



Lolita Singh and Nitul Dutta

Abstract Cognitive Radio is a Wi-Fi verbal exchange methodology that allows the user to engage except having a fixed preassigned radio spectrum. Cognitive Radio Networks (CRNs) are having the routing hassle that is one of the serious constraints. Ad hoc networks are non-centralized Wi-Fi networks that can be constructed and there is no need for any preexisting infrastructure for these networks. Here every point can work as a router. In this paper, the authors have explained the Cognitive Radio Networks (CRN) that are obtaining so a whole lot of recognition where the principal focus is on the dynamic undertaking of channels to wireless devices. In this paper, cognitive radio networks are primarily focused. Nowadays, almost all the networks rely on fixed allocated networks in an approved or unapproved frequency group. In this paper literature evaluates associated to CRN and an optimization algorithm to enhance the overall performance of TE under CRN has been discussed. Swarm intelligence technique is used in the paper. Swarm approach is clearly the combination of the decentralized attribute to gain excellent viable solutions. The motivation regularly creates from nature, more often than non natural outlines. One of the effective approaches known as Cat swarm has been used to acquire high price of accuracy and much low error rates which improves the lifespan of the network. The results are carried out by the use of CSO (Cat Swarm Optimization) algorithm and parameters like energy consumption, congestion, overhead consumption, and number of routing rules are used to analyze the overall performance of the algorithm.

Keywords Cognitive radio networks • CSO • Swarm approach • TE

L. Singh (✉)

Department of Computer Science Engineering, Marwadi University, Rajkot, India
e-mail: Lolita.singh101463@marwadiuniversity.ac.in

N. Dutta

Faculty of Technical Science, Jan Wzykowski University,
ul. Skalnikow 6 B, 59-101 Polkowice, Poland
e-mail: n.dutta@ujw.pl2

1 Introduction

Cognitive Radio is developed by popular person Joe Mitola in 1999 [1]. They allow wireless communication to come into a new method. The CR science provides a flexible result for the trouble of spectrum shortage by means of potential of dynamic spectrum allocation for community communication. It permits the coexistence of CR successful devices with licensed band customers [2] and allows the later to use the licensed spectrum opportunistically. The CR units can correctly feel on hand idle spectrum, reconfigure parameters to get admission to the quickly unused spectrum, and produce unbearable interference to licensed users, which make them successfully network among themselves [1].

Routing in traffic engineering using cognitive radio networks is one of the major topics. The routing hassle is one of those recent research topics which shows up in Cognitive Radio Networks (CRNs). The CRN has confirmed itself as a network to improve spectrum strength via self-organization and dynamic reconfiguration and thus, the Cognitive Radio Ad Hoc Network (CRAHN) [3] is introduced as a promising conversation technology. But, CR users need to stop their conversation or lower their transmission power to keep away from disturbances to authorized licensed customers (called Primary Users (PUs)). Because PUs are the owners of the channel, and the CR units (also referred to as Secondary Users (SUs)) opportunistically use the channel when PUs have no statistics to transmit [4]. That is why the unused spectrum (called spectrum hole) needs to be used successfully to make CRN successful. Besides that, if transmissions of CUs do no longer purpose dangerous interferences with PU transmissions, then CUs can communicate among themselves. CU and SU actually mean the same. From overall discussion CU, SU, and CR are the terms used interchangeably for a cognitive node (Fig. 1).

Network investigation is succeeding in all areas nowadays. Rare greatest capable expertise like 4G [1], facts-centric communication, and software-defined interaction [4] are receiving substantial importance. However additional idea named green

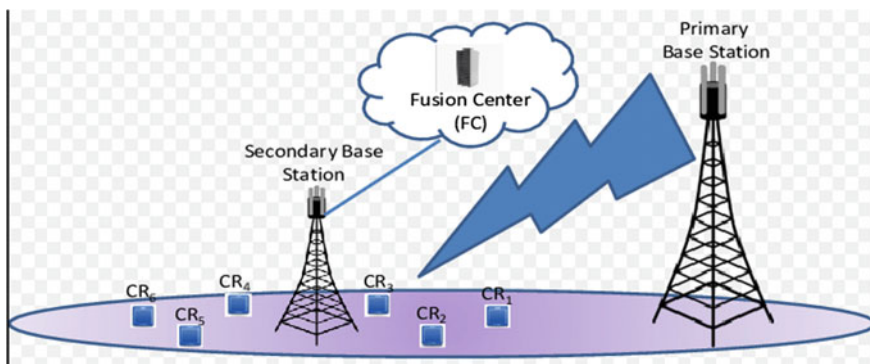


Fig. 1 CRN (cognitive radio network)

mobile computing is also attaining attention. In this, the procedures are intended to reduce power intake thus carbon emanation is minimized throughout the calculation. A review of such green computing constructed routing is established. Biology impressed algorithms are a unit extra step to green computing. Time serving packet forwarding and routing in Delay Tolerant Network (DTN) [9] is alternative space wherever several analysis work is established. These embrace transport network primarily based DTN and transport detector network. The routing in MANET with QoS improvement acquires an equivalent position in several analyses. Though, of these systems declared on top of depends on statically allotted channels either in authorized or unaccredited band. Such fixed channel job construct ends up in spectrum congestion and thus spectrum insufficiency in today’s thickly inhabited situation. That’s why, a replacement technology psychological feature i.e. Cognitive Radio Networks (CRN) attaining quality that works on vibrant task of channels to wireless devices.

2 Literature Review on CRN

Author	Description
Cacciapuoti et al. (2010)	Presented a routine procedure known as CRN On-demand Distance Vector or CAODV. This suggested technique designed for MANET is an adjunct of Ad Hoc On-demand Distance Vector routing protocol. Like AODV, this procedure also makes a path once there is a route for a package to send. Numerous manage packages of AODV, e.g., path appeal; neighbor discovery packets are moreover castoff in CAODV
Karim Habak et al. (2013)	Proposed a technique of routing known as LAUNCH. This procedure has certain features like effective usage of the frequent manage network, negligible path arrangement postponement. However, this procedure doesn’t warranty the steadiness of the mounted path for the duration of conversation [6]
Sun et al. (2014)	Introduced a CRN-based totally vehicular network routing protocol. As mentioned in the paper, the protocol is appropriate for conditions like herbal disasters which cause extreme damage to communication infrastructure. Additionally, such herbal disasters are accompanied through sharp spikes in the utilization of commercially licensed spectrum. In such case, if pretentious victims try to transfer facts with excessive bandwidth requirement, wireless network, then the proposed protocol can suitably be used in CRN [11]
Ian F. Akyildiz et al. (2009)	Introduced CRAHNs current challenges and intrinsic features. The Software-Defined Networking (SDN) has been introduced and proposed. CRN deployments interference is managed by using the advantages of cloud computing and SDN in case of residual areas [5]
Anatolij Zubow et al. (2015)	A new CR SDN based structure has been introduced in which Spectrum Broker (SB), centralized controller primarily related cloud and it takes over spectrum task to CR Base Stations (CR-BS). The Wi-Fi information are combined to the SB beneath CR-BSs

(continued)

(continued)

Author	Description
	manipulate report, in the CRN community site visitors situation up to date records is done by way of SB controller with the aid of configuring acceptable policies in Open Flow-enabled CR-BSs [8]

3 CSO (Cat Swarm Optimization)

The Swarm Intelligence (SI) algorithms are optimization algorithms that were formulated for copying the intelligent behavior and nature of animals. In these technologies, a population of living things such as ants, bees, birds, and fish is communicating with each other and with their place via sharing information that results in the use of their place and things. Cat Swarm Optimization (CSO) algorithm is one of the more recent SI-based algorithms in which the nature of cats is simulated. The CSO algorithm is developed by Chu and Tsai (2007), and its things have been implemented for different optimization problems.

3.1 *Natural Process of the Cat Swarm Optimization Algorithm*

High alertness and curiosity about the environment is the nature of cats. They cannot spend most of their time resting; moving objects in their surroundings are their normal things. By this behavior, cats can cheat and seek their food. They spend very less time chasing prey to consume their energy as compared to the time dedicated to their resting. Chu and Tsai (2007) really got inspired by this behavior of cats and introduced CSO which is called as Cat Swarm Optimization. CSO has two modes: “seeking mode” it is the mode when cats are having rest and “tracing mode” it is the mode when cats are seeking their prey or food. CSO algorithm does the same thing. In CSO, a population of cats is first formulated and assigned in the M-dimensional solution space. Here each cat is representing a result. Then further two subgroups are made by the population. In the first group, the cats are rest mode and are very alert about their surroundings (i.e., seeking mode), while in the second group, the cats are trying to chase their preys (i.e., tracing mode). The hybridization of these two groups helps CSO to get the global solution in the M-dimensional solution space. The tracing subgroup should be small because the cats do not spend much time in the tracing model. The Mixture Ratio (MR) is defined for this which has a small value. New positions and fitness functions can be obtained after sorting the cats into these two modes. After this, the cat with an optimistic solution is

selected and saved in the memory. These steps have to be repeated until the stopping criteria are satisfied (Fig. 2).

The steps of CSO are given below:

Step 1: Initial population of cats has to be created and make them into the M-dimensional solution space ($X_{i,d}$) and provide each cat a velocity in the range of the maximum velocity value ($t_{i,d}$). Then further two subgroups are made by the population. In the first group, the cats are in rest mode and are very alert about their surroundings (i.e., seeking mode), while in the second group, the cats are trying to chase their preys (i.e., tracing mode).

Step 2: Give every cat a flag to arrange them into the tracing mode process or seeking mode process. This depends on the value of MR. The tracing subgroup should be small because the cats do not spend much time in the tracing model. The Mixture Ratio (MR) is defined for this which has a small value.

Step 3: Calculate the fitness value of every cat and the cat with the best fitness function has to be selected. The position of the best cat tells the best solution so far. X_{best} is the position of the best cat.

Step 4: The cats are applied to the seeking or tracing mode process and is based on their flags. New positions and fitness functions can be obtained after sorting the cats into these two modes. After this, the cat with an optimistic solution is selected and saved in the memory.

Step 5: If the termination process is accepted, end the process. Otherwise, repeat steps 2–5.

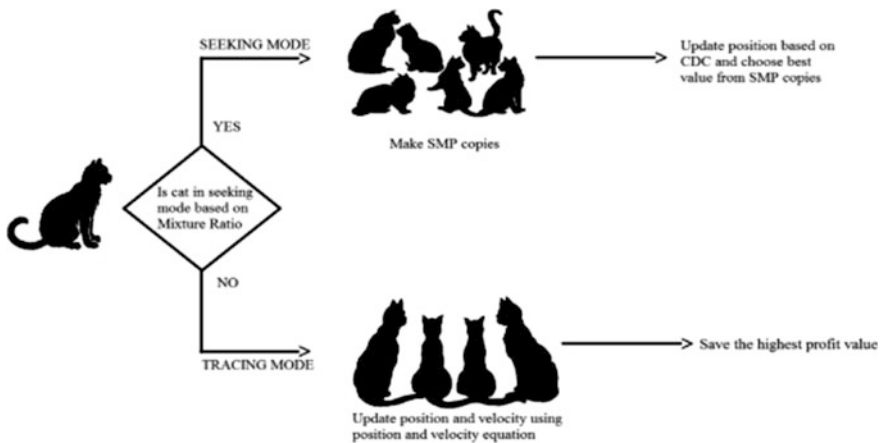


Fig. 2 Cat swarm optimization

3.2 CSO Algorithm

Algorithm for Cat swarm Optimization is given below:

```

Begin
Input parameters of the algorithm and the initial data
Initialize the cat population  $X_i$  ( $i = 1, 2, n$ ),  $v$  and  $w$ 
While (the stop criterion is not satisfied or  $I < I_{max}$ )
Calculate the fitness function values for all cats and sort them
 $X_g$  = cat with the best solution
For  $i = 1: N$ 
  If  $w = 1$ 
    Start seeking mode
  Else
    Start tracing mode
  End if
End for  $i$ 
End while
Post-processing the results and visualization
End

```

CSO algorithm normally uses the behavior of the cats and main focus is on two modes, i.e., “seeking mode” it is the mode when cats are having rest and “tracing mode” it is the mode when cats are seeking their prey or food.

4 Problem Formulations

Cognitive radio networks in routing in traffic engineering are one of the serious topics. The routing problem is one of those recent research topics which occur in Cognitive Radio Networks (CRNs). There is no need for preexisting infrastructure for Ad hoc networks (CRN). These are non-centralized wireless networks that can be easily formulated. An optimization algorithm is used to solve the problems in routing in traffic engineering using cognitive radio networks. Optimization algorithms are those that help in obtaining the best optimistic results. The optimization can be done using (CSO) Cat Swarm Optimization to optimize link utilization for the routing. Cat swarm is one of the effective approaches to get such optimize solutions to achieve a high rate of accuracy and fewer error rates which increase the lifespan of the network. The cat with an optimistic solution is selected and saved in the memory. The optimistic result would be the greatest and best position which is nothing but one of the cats to keep the best solution till it spreads the end of repetitions.

Objectives:

- To implement optimization algorithm called as CSO (Cat Swarm Optimization)
- To analyze the performance of the CSO for less overhead, congestion, and energy consumption to achieve a high quality of service.

The parameters are

- Energy consumption
- Congestion
- Overhead
- Variation of the routing rule.

Swarm intelligence approach has been used because it is a metaheuristic approach which is globally applied to any file to achieve high-end results according to the designed objective to achieve the best solution from the number of solutions.

5 Results and Discussions

The swarm intelligence approach is simply the combination of the decentralized characteristic to obtain the best optimistic results. The idea is utilized in an efficient method at computerized cognitive.

Swarm intelligence frameworks include regularly of a population of direct particles known as swarms interfacing with each other for the solution. The motivation frequently formulates from nature, mainly organic outlines. Cat Swarm optimization is used in this paper. The main aim of selecting this algorithm in the paper is to get optimistic solutions. In Cat Swarm Optimization, high alertness and curiosity about the environment is the nature of cats. They cannot spend most of their time resting; moving objects in their surroundings are their normal things. By this behavior, cats can cheat and seek their food. They spend very less time chasing prey to consume their energy as compared to the time dedicated to their resting. Chu and Tsai (2007) really got inspired by this behavior of cats and introduced CSO which is called as Cat Swarm Optimization. CSO has two modes: “seeking mode” it is the mode when cats are having rest and “tracing mode” it is the mode when cats are seeking their prey or food. CSO algorithm does the same thing. In CSO, a population of cats is first formulated and assigned in the M-dimensional solution space. Here each cat is representing a result. Then further two subgroups are made by the population. In the first group, the cats are rest mode and are very alert about their surroundings (i.e., seeking mode), while in the second group, the cats are trying to chase their preys (i.e., tracing mode). The hybridization of these two groups helps CSO to get the global solution in the M-dimensional solution space. The tracing subgroup should be small because the cats do not spend much time in the tracing model. The Mixture Ratio (MR) is defined for this which has a small value. New positions and fitness functions can be obtained after sorting the cats into these two modes. After this, the

cat with an optimistic solution is selected and saved in the memory. These steps have to be repeated until the stopping criteria are satisfied.

6 Result Explanation

The results are shown in the form of graphs and table.

Figure 3 shows the initialization of the network and shows that the cognitive radios are deployed and primary users are shown for the traffic communication and for the services and the common receiver is deployed to receive the packets and act as the destination.

Figure 4 shows the overhead consumption in which the overhead consumption is decreasing rapidly and is the desired output which is done using Cat swarm optimization approach and is able to achieve fewer collisions and intelligent transmissions (Fig. 5).

As we can see from the above figure that the energy consumption is decreasing in an efficient manner which is one of the main constraints in traffic engineering using the processing elements. The energy consumption is decreasing as the mobility increases which is also one of the desired outputs of the cognitive radios.

Figure 6 shows the congestion in the network which is one of the important parameters in the cognitive-based traffic engineering and shows that the congestion also increases as the mobility increases which increases the lifespan of the cognitive network (Fig. 7).

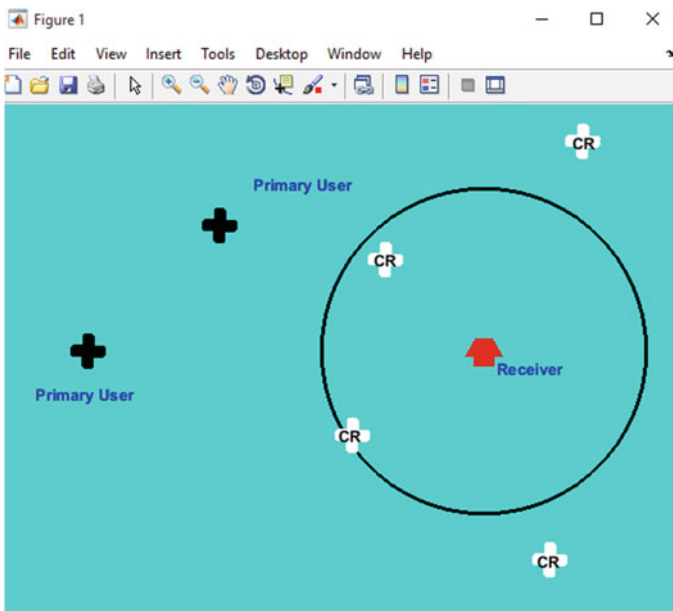


Fig. 3 Network initialization

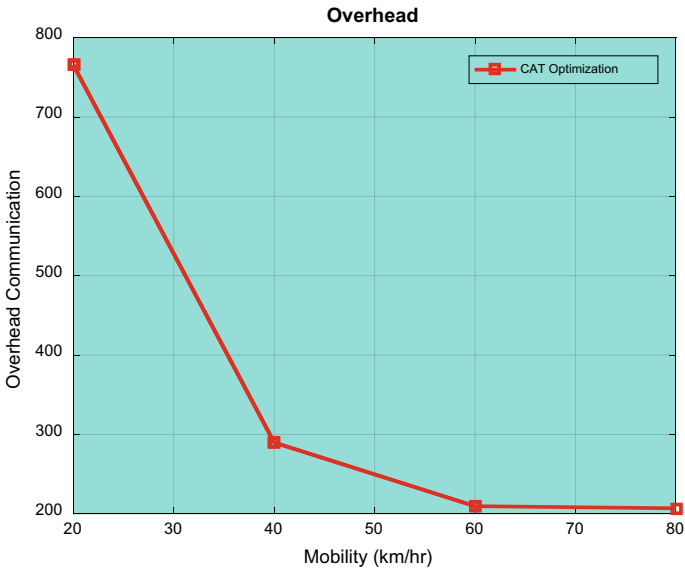


Fig. 4 Overhead consumption

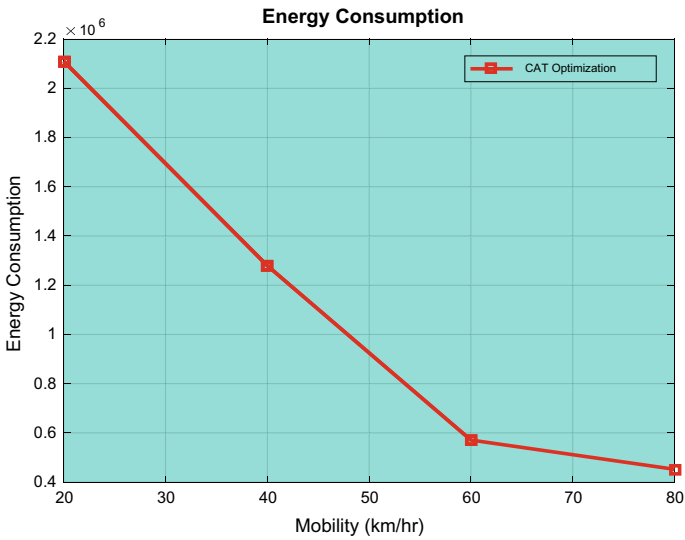


Fig. 5 Energy consumption

In the above figure, we can see the performance in terms of the routing rule which increase with an increase in the dynamic change in the topology using cognitive networks. Route rule is one of the significant parameters which divides

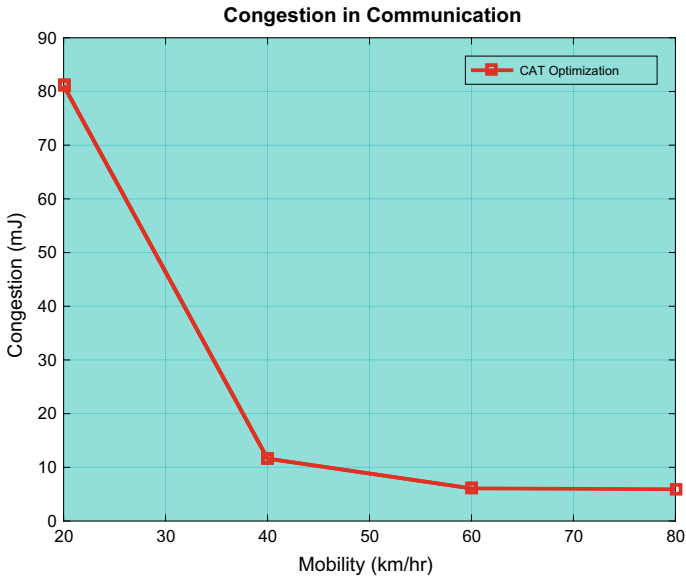


Fig. 6 Congestion

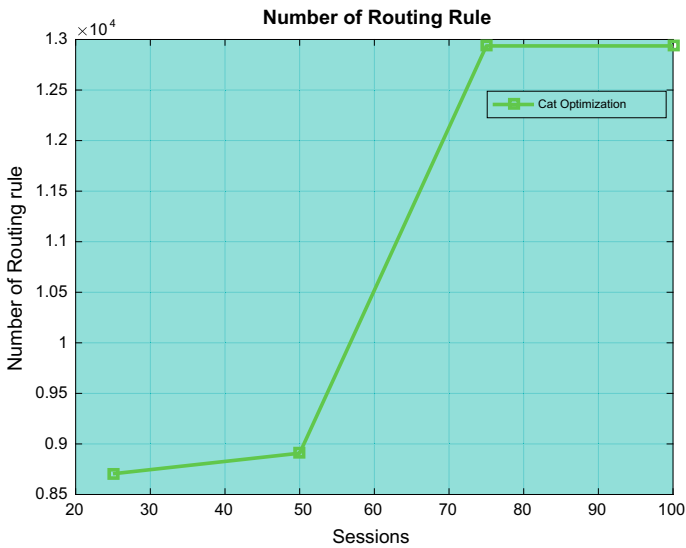


Fig. 7 Number of routing rule

the number of workflows into various threads and provides high hierarchy to the cognitive-based networks and is must be high.

Table 1 Performance analysis using CAT swarm optimization

Parameters	Proposed approach
Energy consumption	0.5×10^2 mJ
Congestion	5 mJ
Overhead consumption	200
Number of routing rules	1.3×10^4

Table 1 shows the performance analysis of CRN using Cat swarm optimization.

7 Conclusion and Future Scope

This paper explained the unique technique called Cognitive Radio Networks (CRN) that is becoming so much popular where the main aim is on active consignment of networks to wireless expedients. The routing in MANET with QoS improvement acquires an equivalent position in several analyses. Though, of these systems declared on top of depends on statically allotted channels either in authorized or unaccredited band. This static channel assignment proved spectrum congestion and thus spectrum shortage in a thickly occupied situation. Cognitive radio networks are introduced and a literature review related to CRN is discussed in the paper. An optimization algorithm to improve the performance of TE under CRN has been discussed. Swarm intelligence approach is used in the paper. Swarm approach is simply the combined part of the decentralized characteristic to obtain optimistic possible results. One of the optimistic approaches called Cat swarm has used to get a high rate of accuracy and fewer error rates which improves the lifespan of the network. The results are obtained using CSO (Cat Swarm Optimization) algorithm and parameters like energy consumption, congestion, overhead consumption, and number of routing rules are used to analyze the performance of the algorithm.

In future work, another Swarm optimization algorithm (BAT swarm) will be implemented to reduce congestion and overhead consumption. After getting results from BAT swarm hybrid model i.e. Cat Swarm +Bat Optimization will be achieved to get better results and comparison of this hybrid model will be done with the individual optimization techniques.

References

1. Cheng, G., Liu, W., Li, Y., Cheng, W.: Joint on-demand routing and spectrum assignment in cognitive radio networks. In: This full text paper was peer reviewed at the direction of IEEE Communications Society Subject Matter Experts for Publication in the ICC 2007 Proceedings, vol. 6501–6503 (2007)
2. Akyildiz, I.F., Lee, W.-Y., Chowdhury, K.R.: CRAHNs: Cognitive Radio Ad Hoc Networks. Elsevier Ad Hoc Networks, pp. 1–25 (2009)

3. Zubow, A., Döring, M., Chwalisz, M., Wolisz, A.: A SDN Approach to Spectrum Brokerage in Infrastructure-Based Cognitive Radio Networks. *IEEE DySPAN 2015*, pp. 213–221 (2015)
4. Alasadi, E., Al-Raweshidy, H.S.: SSED: servers under software-defined network architectures to eliminate discovery messages. *IEEE/ACM Trans. Netw.* **3**, 321–336 (2017)
5. Kaur, K., Garg, S., Aujla, G.S., Kumar, N., Rodrigues, J.J.P.C., Guizani, M.: Edge computing in the industrial internet of things environment: software-defined-networks-based edge-cloud interplay. *IEEE Communications Magazine* **5**, 44–51 (2018)
6. Tahaei, H., Salleh, R., Ab Razak, M.F., Ko, K., Anuar, N.B.: Cost effective network flow measurement for software defined networks: a distributed controller scenario. *IEEE Access* **12**, 501–514 (2017)
7. Hongli, Xu, Huang, He, Chen, Shigang, Zhao, Gongming, Huang, Liusheng: Achieving high scalability through hybrid switching in software-defined networking. *IEEE/ACM Trans. Netw.* **26**, 618–632 (2018)
8. Wang, Q., Zheng, H.: Route and Spectrum Selection in Dynamic Spectrum Networks. In: *IEEE CCNC 2006 Proceedings*, vol. 4, pp. 625–629 (2006)
9. Ali, A., Iqbal, M., Baig, A., Wang, X.: routing techniques in cognitive radio networks: a survey. *Int. J. Wirel. Mob. Netw.* **3**, 96–110 (2011)
10. Foerster, K.-T., Ludwig, A., Marcinkowski, J., Schmid, S.: Loop-free route updates for software-defined networks. *IEEE/ACM Trans. Netw.* **3**, 621–642 (2017)
11. Neama, G.N., Awad, M.K.: An energy efficient integral routing algorithm for software-defined networks. *IEEE* **5**, 401–406 (2017)
12. Asadollahi, S., Goswami, B., Raoufy, A.S.: Scalability of software defined network on floodlight controller using OFNet. In: *2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECOT)*, vol. 23, pp. 557–561 (2017)
13. Karakus, M., Durrezi, A.: Economic viability of QoS in software defined networks (SDNs). In: *2016 IEEE 30th International Conference on Advanced Information Networking and Applications*, vol. 23, pp. 139–146 (2016)
14. Kuang, L., Yang, L.T., Wang, X., Wang, P., Zhao, Y.: A tensor-based big data model for QoS improvement in software defined networks. *IEEE Netw.* **21**, 30–35 (2016)
15. Körner, M., Stanik, A., Kao, O.: Applying QoS in Software Defined Networks by Using WS-Agreement. In: *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, vol. 3, pp. 893–898 (2014)
16. Li, G., Wu, J., Li, J., Zhou, Z., Guo, L.: SLA-aware fine-grained QoS provisioning for multi-tenant software-defined networks. *IEEE* **7**, 121–134 (2017)
17. Ren, S., Dou, W., Wang, Y.: A deterministic network calculus enabled QoS routing on software defined network. In: *2017 9th IEEE International Conference on Communication Software and Networks*, vol. 24, pp. 181–186 (2017)
18. Oluwaseun, A., Twala, B.: QoS functionality in software defined network. In: *IEEE ICTC 2017*, vol. 26, pp. 693–699 (2017)
19. Even, S., Itai, A., Shamir, A.: On the complexity of timetable and multicommodity flow problems. *SIAM J. Comput.* **4**, 691–703 (1976)
20. Al-Fares, M., Loukissas, A., Vahdat, A.: A scalable, commodity data center network architecture. In: *ACM SIGCOMM Computer Communication Review*, vol. 38, pp. 63–74 (2008)

(T-ToCODE): A Framework for Trendy Topic Detection and Community Detection for Information Diffusion in Social Network



Reena Pagare, Akhil Khare and Shankar Chaudhary

Abstract The increased use of social network generates a huge amount of data. Extracting useful information from this huge data available is the need of today. Study and analysis of this data generated provide insight into the behavior of the customers or users and thus will be beneficial to increase the sales of products or understand customers. To achieve the same, we propose a novel framework which will extract trendy topics, identify communities related to these trendy, topics, and also identify influential or seed nodes in communities. The framework intends to find the list of topics which are popular, second, find trend-driven communities, and from these trend-driven communities find nodes which act as seed nodes and thus dominate the spread of information in the community. Analysis of real-world data is done and results are compared with baseline approaches.

Keywords Community detection · Information diffusion · Topic detection · Trend topics · Social network

1 Introduction

Over the past few decades, we can see rapid change in technology and the way technology is used by the world. Social network has become an inseparable part of every person's life today. As the number of user increases day by day there lies great potential in analyzing the social network for the kind of relationships and the actors in the network. The study of social network is not only limited to the behavior of people, but it can be a relationship between people, organization, cells, genes, etc.

R. Pagare (✉) · S. Chaudhary
PAHER, Udaipur, Rajasthan, India
e-mail: reenawp5@gmail.com

A. Khare
MVSr COE, Hyderabad, India
e-mail: khare_cse@mvsrec.edu.in

Social network application expands over a vast horizon from psychology, sociology, marketing, communication, and political science.

Formally a social network can be defined as a network which is a collection of objects in which some pairs of these objects are connected by links. When the structure of the network is considered, it is more in terms of “who is connected to whom” and the other connection at the “level of behavior”, one person can be friend of other, or he may be follower of the person, or he must have retweeted on a post or commented on a post, etc. This means that in addition to the language for discussing the structure of networks, there is also a need of a framework for reasoning the behavior and interaction in the network context.

Much research effort has been put into analyzing information diffusion, with most studies investigating which factors affect information diffusion, which information diffuses most quickly, and how information is disseminated. These questions are answered using information diffusion models and other methods, which play an important role in understanding the diffusion phenomenon. Information diffusion is about understanding who the important users are and which factors are influencing the information diffusion process. A good performance model is very important for understanding how to predict and influence information diffusion and has significant reference value to various applications, e.g., rumor controlling, behavior analysis, gaging public opinion, the study of psychological phenomena, and for resource allocation in public health care systems.

The social influence in a network can be identified through tracking communities and studying the behavior in the community. Social influence was studied in the context of influence maximization where the method was proposed to find the nodes in the network which will increase the spread of information in the community and hence in the network [1, 2]. Their work highlights the social diffusion model and marketing strategies for various applications [3, 4]. The work of [5] relates to the study of how one person influences the other person for buying of products. The author suggests a method to increase this influence to increase the spread of information in the network. Social influence is defined as the force exerted on a person due to her influencer and homophily is considered as the tendency that the other inactive node or person will follow the active node or influencer [6, 7, 8].

The work done in the above methods only exploit the topology of the network, and ignore other important properties, such as nodes' features and the way they process information. Most of the users in social network are in passive mode, based on this observation, researchers [9] suggested an approach similar to HITS algorithm which is a graph-based approach. The method assigns to every node in the network a value based on the influence and the rate at which the information is forwarded in the network. The users who work as a catalyst in spreading of information are called the seed nodes. These seed nodes are specific to a community. Therefore, Pal and Counts [10] develop a non-graph-based, topic sensitive method. The method uses structural as well as social attributes of the network to identify the most influential nodes. The set of identified users are clustered by using probabilistic technique and

therefore could find a list of nodes which are more in the network related to the topic. In paper [1], the author uses independent cascade and linear threshold model for influence maximization. The influence maximization problem asks for a parameter k , to find a k -node set of maximum influence in the network. Depending on the basis of number of nodes activated at the end of information cascade, the influence of a given set of nodes is identified. The author uses a greedy hill-climbing strategy for the optimization problem and provides appropriate solutions.

A **Social network** can be defined as a graph in which vertices represent people or groups of people and edges represent some kind of social relationship or social interaction between vertices. These social graphs can further be defined as the ones in which direction to the links or edges may or may not be known as directed or undirected graphs like Twitter and Facebook, respectively [11].

A **community** in the social network is a set of nodes or users which can be potentially grouped and examined together such that the interactions or the ties between the communities are quite dense [1]. In terms of the graph, we can say it is a giant connected component that is very dense. The users in social network exhibit a behavior where users ignore the information available with themselves and are influenced by the behavior of the neighbor's. This decision taken is transferred in the form of information to all the connected users in the networks. This phenomenon is known as **information diffusion** in social networks.

Information diffusion is a process in the social networks in which the information is transferred from one user to the other, who are connected with each other via social network relationship. Information diffusion includes (a) Detection of popular topics (2) Community Detection (b) Identification of Influential spreaders (Fig. 1).

Even though the framework associated with social networks can change with time, communities stay fairly steady. The primary problem is how one can identify individual's communities which have higher influence inside a social network, and several techniques are suggested for this; some techniques use links between users and some use network attributes. Information diffusion is a hot topic within social networks investigation recently. Even though there has been numerous revolutionary research in this area, there are still some less explored areas.

The remaining paper is arranged as follows: second part is related to state of art; third part describes our framework; fourth section presents result and experimentation; last section presents conclusion and future work.

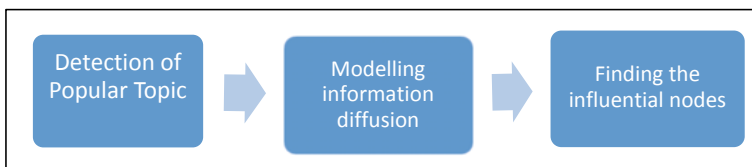


Fig. 1 Processes involved in information diffusion

2 Related Work

2.1 Information Diffusion

In paper [12], the author explores connections as well as designs developed by the aggregated interactions within Facebook webpages throughout catastrophe reactions. Evaluates social roles as well as crucial gamers utilizing social network evaluation. Research suggests measures to enhance information diffusion by considering social as well as network attributes. In Paper [13], the author suggests a differential equation model associated with information within the network. The model is associated with temporal–spatial information diffusion. This model is set up prior to the topological framework associated with the network, semantic content material as well as interactive actions associated with users within the network. Experimental calculations display the feasibility from the suggested model, conformity along with network topology with potential users associated with scalability with regard to big networks. The authors in paper [14] suggest parameters which evaluate topical importance as well as social. The work suggests that to study the information diffusion, it is important to study the behavior of user by taking into account different parameters like retweet count, followers information. In [15], through mixing users social as well as picture information, the model forecasts the actual diffusion route of the image much more precisely when compared with alternative baselines, which possibly encode just picture or even social functions, or even absence storage. Through mapping person users to person prototypes, the model generalizes for new users. The models are able to produce diffusion trees and shrubs, as well as display how the produced trees and shrubs carefully look like ground-truth trees and shrubs. Paper [16] detects communities based on not only structure properties of the network but also used ground-truth information. The author studied a collection of 180 real-world social, effort, and data networks exactly where nodes clearly condition their own community subscriptions. The suggested technique accomplishes 30% relative enhancement more than present nearby clustering techniques. In [17], the author proposes a CID model with regard to damaging information diffusion within competitors towards an optimistic information circulation. The model is an aggressive information diffusion model or even CID with regard to brief. This is the first model which talks about negative as well as positive information flow in the network and propagation of the same. The authors [18] proposes a grading opinion diffusion model. The opinion leaders play an important role in information diffusion in microblogs. The model is combined with communication power. The opinion leaders influence people who are unsteady concerning specific event. The work demonstrates how the brand new suggested models tend to be sensible as well as efficient for information spread within microblog information diffusion. The actual Information Systems (IS) community is constantly on the help to make inroads to improve using systems to aid catastrophe administration [19]. The actual community is constantly on the determine as well as lengthen the actual appropriate hypotheses, and also to create brand new paradigms

that may be delivered to keep about the ownership as well as diffusion associated with information techniques with regard to catastrophe administration. Research by authors in [20] handles the problem of advertising data diffusion, the degree to that data spreads, upon social media systems. This specific A–B–C framework (1) examines the textual options that come with Internet record articles making use of linguistic issue as well as phrase rely, (2) does apply the prior results in creating information suggestions, as well as (3) suppliers verified training supplies based on information tips to promote information diffusion among Internet record site visitors. The author suggests an evolutionary game theoretic framework [21] to model diffusion procedure within social networks. The framework is analyzed for degree and nonuniform degree network. Present social networks tend to be of very large scale producing huge data moves for each and every second. Exactly how data diffuses more than social networks offers has attracted a lot of interest. Author obtains diffusion characteristics within just complete networks, typical degree, without homogeneous degree networks, with identify connected with two distinctive networks. The authors in their work [22] depicted a person's behavior of online social networks opens new alternatives to systematically realize the data diffusion process after social networks. Random Recursive Tree (RRT) is used to help model with this advancement associated with Cascade trees. The use of stochastic sensitive look at forwards eliminated model to help demonstrate the particular potent individual behavior including producing, seeing, forwarding, and overlooking the information for the furnished social network.

The authors argued Twitter data is very loud [23] every twitter is brief, unstructured and along with casual vocabulary, challenging with regard to present subject modeling. About the additional hand, Twitter posts tend to be associated with additional data, for example, authorship, hashtags and the user-follower network. Taking advantage of this information, the author suggests the Twitter-Network (TN) subject model collectively model the written text and the social network inside a complete Bayesian nonparametric method. The TN subject model utilizes the hierarchical Poisson–Dirichlet Procedures (PDP) with regard to textual content modeling and the Gaussian procedure Random function model with regard to social network modeling. The author implies that the TN subject model considerably outperforms a number of current nonparametric models because of its versatility. Furthermore, the TN subject model allows extra educational inference, for example, authors' interests, hashtag evaluation, in addition to resulting in additional programs, for example, author suggestion, automated subject marking, and hashtag recommendation. Be aware the common inference framework may easily be reproduced to additional subject models along with inlaid PDP nodes. Another work [24] uses Louvain community detection algorithm and Girvan algorithm to find communities and centrality measure to find the nodes in the communities which help in the diffusion process. The author also uses the shortest path algorithm to find the communities formed using the seed nodes. The work considers the network attributes for community detection. The communities formed are not dense and are formed on the basis of only structure of the network. Does not consider the content in the network. As per study, in [25], twitter comprises a good obtainable system with

regard to learning and tinkering with the character of data dissemination. The author uses hashtag to find topics which disseminate most in the network. As per [26], it is vital to find out, monitor, review, as well as forecast well-liked subjects and occasions happening within the social networking within the space–time framework. Simultaneously, it's very helpful which a number of “what if” situations could be created to estimation the meme diffusion. The author has made use of location and time attribute to study the diffusion. The work [27] represents impact maximization issue in attempting to recognize some ‘K’ nodes through that distribute associated with impact, illnesses, or even data is maximized. The author suggests a ‘Greedy ad Community-Based’ formula.

2.2 Topic Detection

The paper [28] proposed LDA model which is the most popular and efficient model for topic modeling. But it does not consider the correlation among topics. It works only on content and is a probabilistic model that considers the distribution of words for identifying popular topics. An online LDA model is proposed [29], it considers the document in batches. Thus the memory requirement is reduced. It does not consider the social attributes of the network. Considers the distribution of word for topic definition. Takes only content into consideration while modeling. The authors [30] propose a model using wavelet analysis, detects event inferred using LDA gives a better description of the event and uses hashtag for topic inference. Topic popularity is found by using only content, the network attributes too can be considered to improve the accuracy of the model. In his work [8], the author proposes a model for unbounded topics. The hierarchical structure can help to identify popular topics. The model does not work for definite and bounded topics. A topic model based on time [31] is implemented, the model accuracy can be improved by consideration of other attributes of the network like retweet count and followers count. A parallel algorithm for topic modeling [32] gives a good speed up as compared to the sequential approach, but the resource requirement is huge. The model can be implemented on a small scale for less number of processor and it is important to study the behavior of the algorithm.

Most of the work done, find trendy topics irrespective of the social attributes and take content only into consideration, some works find communities and then find topics which are discussed mostly in these communities. In our proposed work, we make an attempt to identify trendy topics based communities that have played an important role in making the topic trendy. Given this real-time scenario, our objective is to (1) extract trendy topics (2) find communities related to these trending topics (3) identify the influential node in this topic-driven community.

To the best of our knowledge, our work is the first attempt toward presenting a framework which will identify trendy topics, find trend-driven communities, and then find the most influential nodes in the community. The topic-driven communities identified can be used to spread message or information related to the event or these set of users can be used for the marketing of products or to increase sales of

products. The work finds application in many areas related to identifying users for marketing, identifying users who spread wrong information in a closed network, or terrorist network. In order to detect communities related to trending topics, we focus on twitter, the widely used micro-blogging platform. The activity of twitter is depicted by tweets, retweets, comment, likes, and its structure of follower and followed unidirectional relationships.

To spread information in the network, it is needed that the users have information related to the topic. This information can be received by connecting with users having similar topic information. Thus forming of groups or communities of similar interest, users will be effective for spreading of information in the network. Accordingly, we present a method to identify these communities of topic-dependent users who created the post related to the topic U_c . Another important parameter of our work is followers users U_f information of these users U_c . These U_f users will be responsible for the spread of information in the network.

To summarize, given a trending topic, the objective is to detect community of users who (1) are related to the trending topic (2) are related to the same topic (3) are connected to each other (follow) (4) can be identified as U_c or U_f .

3 Proposed Work

3.1 Problem Formulation

Definition 1

Let $T = \{t_1, t_2, \dots, t_m\}$ be the set of tweets collected during a particular period.

Each tweet $t_i = \{w_1, w_2, w_3, \dots, w_i, \dots, w_n\}$ where w_i represents the sets of word in the tweet t_i .

Analyzing these tweets T for the word occurrences in each tweet and document as whole and finding which words w_j are maximum talked about is topic detection.

Where w_j is subset of w_i ; $w_j \in w_i$

These w_j words are the set of Trendy topics, Tr .

The first goal is to identify this Tr set

Definition 2

Let $Tr = \{Tr_0, Tr_1, Tr_2, \dots, Tr_9\}$ be the set of trendy topics. Here we assume that we identify top 10 topics only. For each topic Tr_i belonging to Tr (a) find set of users U who have tweeted, retweeted about the topic Tr_i (b) Generate a graph G , of these users u_i from U where u_i belongs to U . The graph G formed represents a community Ctr_i which is related to topic Tr_i . Each vertex in community Ctr_i is related to user u_i who in turn is related to topic Tr_i . There exists an edge between u_i and u_j , where $(u_i, u_j) \in U$ iff u_j is in followers list of u_i or vice versa or u_i retweeted on u_j . The users $u_z \in U$, who are not a part of the connected graph are discarded.

The second goal is to generate graph $G(V, E)$ of users related to topics. This graph represents a set of users who will play an important role in identification of communities in next phase.

Definition 3

For the graph G generated for topic Tr_i we run community detection algorithm to identify community. The community detection algorithm will identify communities associated with topic tr_i , which will be used to identify node with highest degree, $u_i \in U$

A degree here means the node which will have the highest number of connections. The node having the highest connection is the one who will spread maximum information in the community.

To summarize, our contributions

- (a) A model to find trend-driven communities and influential nodes.
- (b) A unique framework which combines topic detection and topic-driven community detection.

3.2 Overview

We propose a novel framework where both topic modeling and community detection are collaborated to study the information diffusion in the network. The topic modeling module is the first module which takes as input tweets dataset and identifies topics which are trendy. This module can also be called as trendy topic detection. The second submodule is related to forming communities which are related to a particular topic. The communities formed are called as topic-driven communities. We name them so because the communities identified are related to the topic. After identification of communities, the nodes with highest degree centrality are identified since these are the nodes which influence the propagation of information in the network. Thus the framework will not only detect trendy topics but will also help to identify users and communities which play a substantial role in spreading of information (Fig. 2).

3.2.1 Data Collection

The Twitter dataset is used as input to the system. The twitter data is extracted using twitter API and stored in CSV form. The tweets collected have fields like retweet count, comment count, user who tweeted or retweeted about post, followers. We also store separate information about the users or owner (the user who first posted the tweet) of the tweet. We maintain separate information about the user—follower relationship.

The twitter data was collected from March 2018 to August 2018. Due to restriction by twitter, only 15 lacs of tweets were collected. The data collected included fields like user id, username, tweet, retweet count, mention count, followers of users, and followers count. Figure 3 shows the distribution of messages over different domains. The diagram below shows that floods, health, and politics topics were mostly covered in the tweets captured.

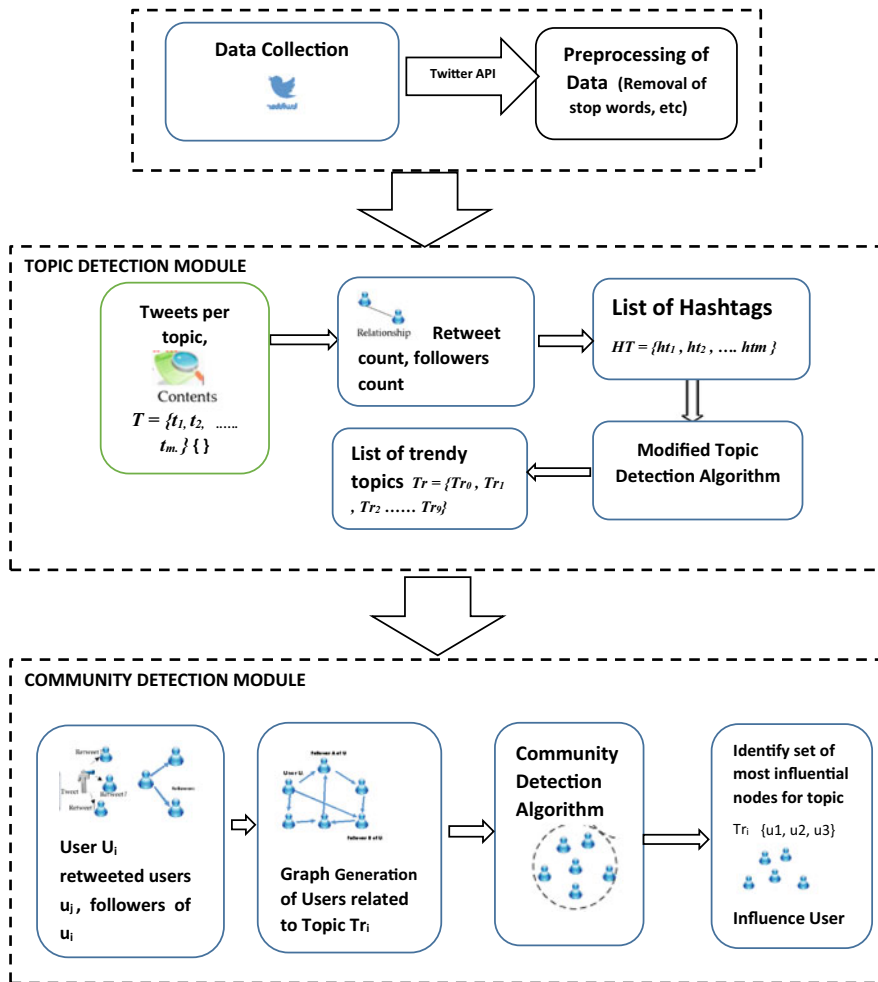


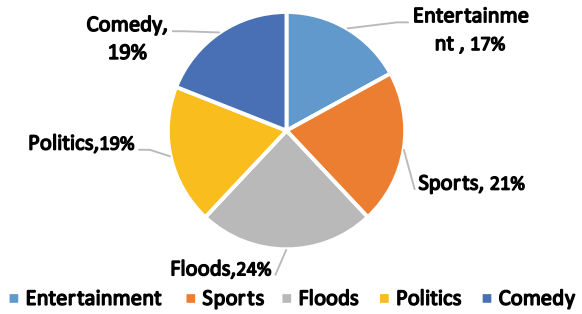
Fig. 2 Collaborative framework design

3.2.2 Data Preprocessing

In this step, the noise in the data is removed. The preprocessing deals with stop word removal, repeated data removal, filling of missing data by taking a mean of the values.

The next section provides details of our approach named T-ToCODE (Trendy Topic detection and trendy COMMUNITY DEtection) which is able to detect hot topics and second the communities associated with every popular topic.

Fig. 3 Distribution of messages in the dataset over different domain



3.2.3 Trendy Topic Detection

Latent Dirichlet Allocation (LDA) is a classic model for topic modeling of documents. LDA detect underlying topics in text documents. LDA is based on the assumption that documents with similar topics will use similar groups of words. Documents are probability distributions over latent topics. Topics are probability distribution over words. Thus LDA is an unsupervised, probabilistic model. We propose a novel approach for finding trendy topics, where we consider the hashtag, comment count, retweet count, and likes for a particular post to identify if the topic is trendy or not. We compare our modified Multiparameter LDA (MP-LDA) with Twitter LDA and baseline LDA.

The proposed algorithm is mentioned below.

Generative Process for MP-LDA:

1. Initialize: Time distribution value of L_w , vector δM , δV .
// word distribution for topic t
 2. For each topic $t = 1, 2, \dots, T$
 - a. Draw $\phi_t \sim \text{dir}(\beta)$
 3. For each user $u = 1, 2, \dots, U$
 - a. Draw $\theta_u \sim \text{dir}(\alpha)$ // topic distribution of user u
 - b. For each document/tweet $s = 1, 2, \dots, N_u$
 - i. Draw $Z_{u,s} \sim \text{Multi}(\theta_u)$ // Topic selection based on users topic distribution
 - ii. Draw $V_{n,s} \sim \text{dir}(\alpha)$
 - c. For each word $n = 1, 2, \dots, N_{u,s}$ // words are selected one by one based on value of P
 - i. compute $P = P_w || P_d$ according to L_w
 - if $P = 0$ then
 - draw $Y_{u,s,n} \sim \text{Multi}(\vartheta g, u)$
 - draw $W_{u,s,n} \sim \text{Multi}(\Phi_z, Y_{u,s,n})$
 - if $P = 1$ then
 - draw $Y_{u,s,n} \sim \text{Multi}(\vartheta h, m)$
 - draw $W_{u,s,n} \sim \text{Multi}(\Phi_{y,u,s}, \delta m, nu)$
 - end if
- end for
end for

Let θ_t be word distribution for topic t . P parameter is used to decide word belongs to a general topic or hot topic. The hashtag in the model is associated with a topic and word. δM is the hashtag associated with documents and δV is hashtag associated with word. The values that vectors can take is either 0 or occurrence count over the entire dataset. Latent variable learning is done by using Gibbs sampling. The MP-LDA is applied after preprocessing step to only those tweets which are above the threshold value of popularity count, comment count, and retweet count. All the tweets with parameter values below the threshold are not considered. Tweets with lesser retweet count or likes will not be popular and thus processing of these tweets for identifying popular topics is not required. The selection of threshold for retweet count, comment count and likes is based on the empirical analysis. It is seen from the statistics that most of the messages have less than 50 retweet and comment count. As the count reaches 1000 the distribution in number of messages is almost constant. Therefore we choose the threshold as 1000. After the application of MP-LDA to the dataset, we were able to find the topics which were most popular in the dataset. The topics like yoga for health, women's health, Asian games, US open championship and Kerala floods, Japan floods were found to be the trendiest topics.

3.2.4 Topic-Driven Community Detection

The steps involved in this submodule are given below.

Input: Topic Tr_i

STEP A: Identification of posts $P_t = \{P_1, P_2, P_3, \dots, P_1\}$ which are related to topic Tr_i

STEP B: Identification of user sets U_o who created the post, related to popular topic Tr_i

STEP C: Identification of user sets U_r who retweeted. The set identified in STEP B and STEP A together will be merged to form a set U_{rel} where $U_{rel} = U_r + U_o$

STEP D: Identification of followers of users U_f who created the post P_1 . The followers play a crucial role in the propagation of information and thus follower's information will be useful.

STEP E: Graph creation based on user- follower data, user— U_{rel} data

STEP F: Detection of communities related to a topic

Output: Community Ctr_i

STEP A:

Given a set of tweets $t = \{t_1, t_2, t_3, \dots, t_m\}$ related to a topic. Identify tweets t_i which are related to popular topic Tr_i . The data set of tweets will be searched and it will retry tweets related to these popular terms.

The output of this step will be a set of tweets $T_{Tr} = \{t_1, t_2, t_3, \dots, t_j\}$ where $j < m$ and $t_j \in t$

STEP B:

From stepA all the tweets from set T_{Tr} are parsed to identify the tweets or post which are source post. The tweets are parsed to get the user id of the users who have

tweeted about a topic and is the origin of the post. This step will give output a set of users.

$U_o = \{U_{o1}, U_{o2}, ..U_k\}$ such that $U_{oi} \in U$ and U_{oi} is the owner of the post and $r < m$

STEP C:

Given a set of tweets T_{Tr} . Identify users who retweeted Ur post in T_{Tr} , This third step gives a set of users who have retweeted a post.

At the end of step C, we get a set of users $U_{rc1} = \{U_o, U_r\}$ who may play an important role in the formation of communities

STEP D:

The belief that followers will play a substantial role in spreading message, the followers list is identified. Given a set of users U_o , find the follower of U_o and store it in the form (user id, user id follower[]). The set is identified as U_f

Also, a separate data related to User— U_{rc1} user data is stored. It contains those users who commented, retweeted a post by user.

STEP E:

Graph Generation:

Step D gives the vertices of the graph G . The set U_f (User - Follower) and the set U_m (user— U_{rc1}) will form the vertices of the graph.

Vertex set $V = U_f \cup U_m$; union of set U_f and set U_m

The edge between a set of vertices can be of two types. The edge set is represented by E . User $u1$ and user $u2$ there exists link between $u1, u2$ if and only if

Type 1: User $u2$ retweet on user $u1$ post

$$R = \{\text{retweet}(u1, u2) = \text{TRUE}; u1, u2 \in U_m\}$$

Type 2: User $u2$ is follower of user $u1$

$$F = \{\text{follow}(u1, u2) = \text{TRUE}; u1, u2 \in U_f\}$$

The set of edges $E = R \cup F$

On the basis of links and vertices a graph is formed. The belief that users who retweet or are followers of a particular user say $u1$ has more probability of spreading message which is posted by user $u1$ is used while formation of the graph $G(V,E)$.

STEP F:

On the graph generated in step E, we apply Louvain community detection algorithm [33].

The decision of applying community detection algorithm was done because it can handle in a very short time a huge network. Also, the algorithm reveals hierarchical structure which can be useful in understanding the overall functioning of the network.

The Algorithms working is explained below

1. A greedy assignment of nodes to communities, favoring local optimizations of modularity.
2. New—coarse-grained network in terms of communities found in the first step.

The above two steps are repeated until no further modularity increasing reassignments of communities are possible.

3.2.5 Identification of Influential Nodes

Once the communities are identified finding nodes which are having a higher degree in the community is obviously the node which is going to be more responsible for spreading of information in the network. Average degree is used to find nodes which are most influential in the network.

The list of influential users is given by $U_{In} = \{U_1, U_2, \dots, U_x\}$ where $U_i \in U$

4 Experimentation

4.1 Topic Detection

4.1.1 Dataset

We use twitter dataset for experimentation. It consists of 132,009 posts/re-posts as well as 73,257 person nodes. Table 1 gives statistics of dataset. The data collection is as mentioned in Sect. 3.2.1

4.2 Analysis

Table 2 gives the list of hot topics and top words in each topic. When the result was compared to LDA and T-LDA it was observed that the top words in MP-LDA were more accurate as compared to T-LDA and LDA. For example, the first topic has words like Japan under topic FLOOD, but LDA and T-LDA did not return this word as a top word in the list. If we consider the result and compare with trending topic during the period, there are tweets related to floods in Japan during the same time as in India in Kerala.

Table 1 Statistics of dataset

Total tweet	Original tweets	Retweet	Users
1,32,009	9600	1,22,409	73,257

Table 2 Top topics with top words

Topic id	Top words
0	Flood rescue stressed relief Kerala donation Japan Gordon
1	Sports Asian Games, Indians US open, football Cleveland baseman, Serena
2	Health, women's health, Rohingya, yoga, mental health, medicine, organic, weight loss

Table 3 Coverage rate

	Top 10	Top 20	Top 30
MP-LDA	0.66	0.75	0.78
T-LDA	0.45	0.50	0.53
LDA	0.23	0.35	0.40

We also compared the coverage rate of T-ToCoDE with the baseline LDA method and Twitter LDA. It was observed that the result obtained by our method which considered social attributes gave better accuracy as compared to traditional methods.

$$\text{Coverage rate} = (\text{Extracted Hot Topics} / \text{Actual Hot Topics}) * 100$$

The comparison was done for the top 10 topics, consecutively for 20, and then 30 topics. The proposed method showed consistent improvement in accuracy (Table 3; Fig. 4).

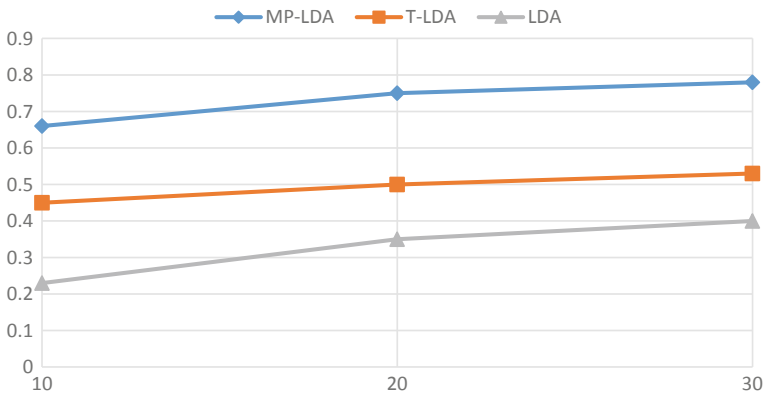


Fig. 4 Coverage rate versus no of topics for twitter data

4.3 *Community Detection*

4.3.1 **Dataset**

The trendy topic is given as input to the community detection module. The first topic which was trendy as per the topic detection was related to hashtag flood. First, we collected all tweets which were related to hashtag flood. This gave a list of 30,000 tweets related to floods. The captured tweets were then separated for the original tweet and retweet. It was found there were 2930 original tweets and remaining were retweets for flood dataset. From this set of original tweets identified, we considered only those users original tweets which had retweet count more than the threshold. After applying the threshold, we were left with 670 users who were original tweet users for flood dataset. We extracted the followers of these original tweet users. To this collected data then was Louvain Community detection method applied to find the communities related to Topic FLOOD.

The communities formed are evaluated on the basis of disconnected users, cohesiveness among users, and clustering coefficient. We compared our work in three different scenarios.

Scenario 1 (M1): The entire network is considered irrespective of tweet and retweet and an attempt was made to identify the communities. No threshold was applied for the selection of users. The graph was constructed on the basis of the followers and the followee network. Dataset considered for this method had all the 23,257 users included.

Scenario 2 (M2): The graph was constructed on the basis of retweet count and the retweet relationship and threshold considered was 1000. For the tweets collected for hashtag flood, out of 30,000 tweets, there were 2930 original tweet users, 7640 retweet users, and the remaining were duplicate users. There were a total of 10,570 users. When the threshold was applied to the original set of users, only 670 original users were considered for graph construction. The edge was considered between two users if user u_1 retweets about the post by user u_2 and vice versa.

Scenario 3 (T-ToCODE): The graph was constructed as per the proposed method in the framework. After the threshold applied, there were 670 users. For these 670 users, the followers list was considered and the graph was constructed on the basis of retweet as well as followers information. Due to the restriction by twitter, we could retrieve followers list for only 200 users and the total user were 41,235.

4.3.2 **Metrics**

Disconnected Users: This metrics will help to find users who are not part of the community

$$\text{ratio of disconnected users} = \frac{\text{users who are not connected to any other user}}{\text{total no of users}}$$

Cohesiveness among users: The second metric used is cohesion among users, this metrics allows to evaluate the quality of connection among users by using standard metrics such as modularity, the ratio between no. of communities and number of users, and density.

The ratio between the no. of community and no. of users allows to evaluate the ability of the approach to group the individual users into communities.

Lower the value, the better communities are formed, it signifies that communities are less in number with more number of users in each community.

The density is the ratio between the numbers of edges per node to the number of possible edges.

Clustering Coefficient: Clustering Coefficient quantifies the extent to which nodes in the graph tend to cluster together.

4.3.3 Analysis

Analysis of Disconnected Users

The result showed that 80% of users took part in the topic with our approach as compared to 76% with scenario 2 as compared to 72% of scenario 1. These results demonstrate that our graph construction approach includes more no of users.

Our approach considers the followers also thus increasing the number of users who can contribute toward spreading of information. In scenario 2, the graph is constructed using only retweet users thus leaving out a good percentage of users who may contribute toward information spread. Though in scenerio1 there are more numbers of users, due to consideration of the entire network a large number of users are disconnected users.

Analysis of Cohesiveness Among Users

Density is influenced by the fact that T-ToCODE framework share retweet, followers link and therefore has higher density compared to approach in scenario 2. The average number of communities found in Scenario 2 is 0.28 per user that exceeds the amount of community in T-ToCODE (Table 4).

Table 4 Cohesiveness among the users

Method	Modularity	Ratio of communities per node	Density
Scenario 1	0.420	0.321	0.010
Scenario 2	0.530	0.278	0.020
T-ToCODE	0.780	0.173	0.029

Table 5 Analysis of the structure of community (average)

Method	% of Post creators	% of followers	Clustering coefficient
Scenario 2	4.14	6.87	0.088
T-ToCODE	5.20	8.23	0.078

Analysis of Community Structure

The percentage of users who are the owners of post is more in scenario 3 (Proposed method) as compared to scenario 2 where only retweet relationship is considered. The increase in scenario 3 is obvious since we consider the retweet relationship along with follower’s relationship. T-ToCODE creates communities on the basis of retweet relationship and followers relationship; therefore, the followers percentage is observed more as compared to scenario 2. Since the communities formed are related to a topic and the communities are dense thus the clustering coefficient is also improved in T-ToCODE method (Table 5).

After identification of communities, the nodes with the highest degree were found in the network. For finding these nodes, we used GEPHI visualization tool. It was observed that the nodes which had the highest degree on an average were the nodes which were the most influential nodes in the network. The entire community could be reached through the users who were having a higher degree.

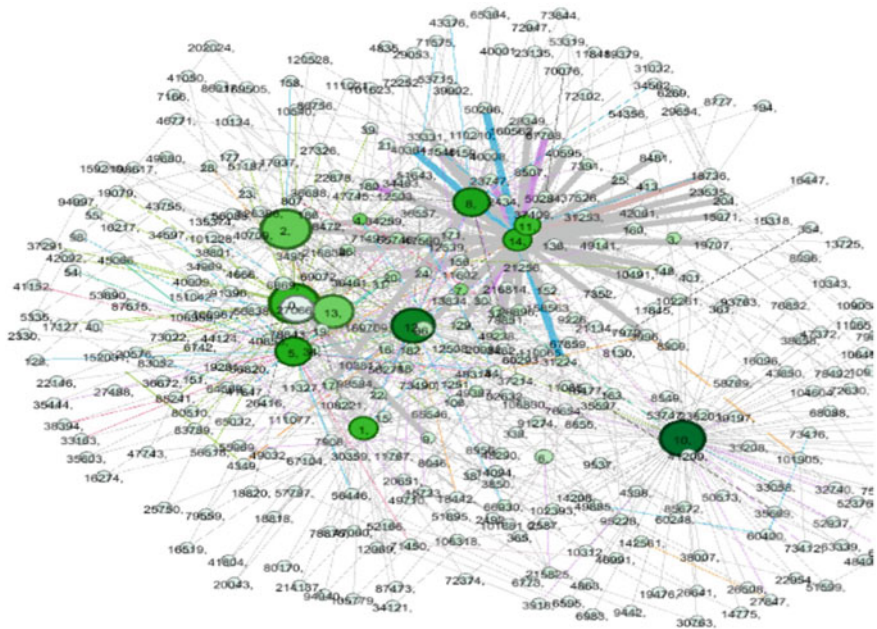


Fig. 5 Influential nodes for topic FLOOD: Scenario 3

The fig. 5 shows the influential nodes in the network ;nodes 8 , 11 and 13 are more dominant in the network. The edge thickness represents the ties among the nodes. From the graph, it can be observed that the nodes 8, 11, and 13 are the influential nodes in the community, along with a few other nodes in the network. There are communities identified for topic FLOOD, and there are nine influential nodes identified, the nodes which have the highest degree. The nodes at the center of each community are the node with the highest degree and thus they are the influential nodes in the network.

5 Conclusion

The novel framework proposed allows to identify hot topics and the communities which are influential in making those topics trendy. The analysis did prove that communities formed with relation to topics have more strength and lesser communities with more no of nodes are formed. The communities formed are denser and less overlapping. The topic detection model performs better as compared to the traditional model. The accuracy of topics identified is increased due to the inclusion of social attributes like retweet count, comment count and hashtag. Consideration of behavior of users in the network gives better results is proved.

5.1 Future Scope

The framework can be extended by considering the spatial and temporal attributes of the data. Thus adding these attributes can give a more personalized analysis, which can be used by an individual as well as businesses to target the customers.

Acknowledgements The authors declare that they have no conflict of interest. All procedures performed in studies involving human participants were in accordance with the ethical standards. Informed consent was obtained from all individual participants included in the study.

References

1. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03), pp. 137–146. ACM, New York, NY, USA (2003). <http://dx.doi.org/10.1145/956750.956769>
2. Bharathi, S., Kempe, D., Salek, M.: Competitive influence maximization in social networks. In: Deng, X., Graham, F.C. (eds.) Proceedings of the 3rd international conference on Internet and network economics (WINE'07), pp. 306–311. Springer, Berlin, Heidelberg (2007)

3. Bonchi, F.: Influence propagation in social networks: a data mining perspective. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 2(1) pp. 2–2 (2011)
4. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining, pp. 57–66 2001. <https://doi.org/10.1145/502512.502525>
5. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. *ACM Trans. Web (TWEB)* **1**(1), 1–39 (2007)
6. Goyal, A., Bonchi, F., Lakshmanan, L.V.S.: Learning influence probabilities in social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining—WSDM '10, p. 241 (2010). Available at: <http://portal.acm.org/citation.cfm?doid=1718487.1718518>
7. Saito, K., Nakano, R., Kimura, M.: Prediction of information diffusion probability ies for independent cascade model. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 67–75 (2008)
8. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10), 1029–1038. ACM, New York, NY, USA (2010). <https://doi.org/10.1145/1835804.1835934>
9. Liu, L., et al.: Modelling of information diffusion on social networks with applications to WeChat. *Physica A Stat. Mech. Appl.* **496**, 318–329 (2018)
10. Pal, A., Counts, S.: Identifying topical authorities in microblogs. In: WSDM '11, pp. 45–54 (2011)
11. Easley, D., Kleinberg, J.: *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York (2010)
12. Kim, Jooho, Hastak, Makarand: Social network analysis: characteristics of online social networks after a disaster. *Int. J. Inf. Manag.* **38**(1), 86–96 (2018)
13. Tu, H.T., Nguyen, K.P.: Differential information diffusion model in social network. In: Asian Conference on Intelligent Information and Database Systems. Springer, Cham (2018)
14. Shi, J., et al.: Determinants of users' information dissemination behavior on social networking sites: an elaboration likelihood model perspective. *Internet Res.* **28**(2), 393–418 (2018)
15. Hu, W., et al.: Who will share my image?: Predicting the content diffusion path in online social networks. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM (2018)
16. Yang, Jaewon, Leskovec, Jure: Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.* **42**(1), 181–213 (2015)
17. Tu, H.T., Nguyen, K.P.: Competitive information diffusion model in social network with negative information propagation. In: Asian Conference on Intelligent Information and Database Systems. Springer, Cham (2018)
18. Liu, X., Liu, C.: Information Diffusion and Opinion Leader Mathematical Modeling Based on Microblog. *IEEE Access* (2018)
19. Beydoun, G., et al.: Disaster management and information systems: insights to emerging challenges. *Inf. Syst. Front.* **20**, 1–4 (2018)
20. Liang, Y., Kee, K.F.: Developing and validating the ABC framework of information diffusion on social media. *New Media Soc.* **20**(1), 272–292 (2018)
21. Jiang, C., Chen, Y., Liu, K.J.R.: Evolutionary dynamics of information diffusion over social networks. *IEEE Trans. Signal Process.* **62**(17), 4573–4586 (2014)
22. Romero, Galuba, W., Asur, S., Huberman, B.: Influence and passivity in social media. In: ECML/PKDD '11, pp. 18–33 (2011)
23. Lim, K.W., Chen, C., Buntine, W.: Twitter-network topic model: a full Bayesian treatment for social network and text modeling (2016). arXiv preprint [arXiv:1609.06791](https://arxiv.org/abs/1609.06791)

24. Jain, S., Mohan, G., Sinha, A.: Network diffusion for information propagation in online social communities. In: 2017 Tenth International Conference on Contemporary Computing (IC3). IEEE (2017)
25. Stai, E., et al.: Temporal dynamics of information diffusion in twitter: modeling and experimentation. *IEEE Trans. Comput. Soc. Syst.* **5**(1), 256–264 (2018)
26. Ye, X., et al.: Open source social network simulator focusing on spatial meme diffusion. In: *Human Dynamics Research in Smart and Connected Communities*. pp. 203–222. Springer, Cham (2018)
27. Jalayer, M., Azheian, M., Kermani, M.A.M.A.: A hybrid algorithm based on community detection and multi attribute decision making for influence maximization.”. *Comput. Ind. Eng.* **120**, 234–250 (2018)
28. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
29. AlSumait, L., Barbará, D., Domeniconi, C.: On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: 2008 Eighth IEEE International Conference on Data Mining, Pisa, pp. 3–12 (2008). <https://doi.org/10.1109/icdm.2008.140>
30. Cordeiro, M.: Twitter event detection: Combining wavelet analysis and topic inference summarization. In: *Doctoral Symposium on Informatics Engineering, DSIE* (2012)
31. Liu, G., Xu, X., Zhu, Y., Li, L.: An improved latent dirichlet allocation model for hot topic extraction. In: 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, pp. 470–476. Sydney, NSW (2014). <https://doi.org/10.1109/bdcloud.2014.55>
32. Wu, C., Wu, B., Wang, B.: Event evolution model based on random walk model with hot topic extraction. In: Li, J., Li, X., Wang, S., Li, J., Sheng, Q. (eds.) *Advanced Data Mining and Applications. ADMA 2016. Lecture Notes in Computer Science*, vol. 10086. Springer, Cham (2016)
33. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **10**, P10008 (2008)

ns-3 Implementation of Network Mobility Basic Support (NEMO-BS) Protocol for Intelligent Transportation Systems



Prasanta Mandal, Manoj Kumar Rana, Punyasha Chatterjee
and Arpita Debnath

Abstract In an Intelligent Transportation System for a Smart City, seamless connectivity is essential for each user during mobility for efficient data communication. For a group of mobile users in a vehicle (bus/train/flight), due to high mobility, implementing a protocol in order to manage handoffs smoothly is a challenging task. Network Mobility Basic Support (NEMO-BS) protocol was proposed to comply with this requirement. It is an extended version of Mobile IPv6 (MIPv6). But, the MIPv6 implementation in ns-3, which is the most widely used open-source simulator, is still not extended so far, to support network mobility. In this work, we have implemented the functionality of the NEMO-BS protocol in ns-3.25 by modifying the existing MIPv6 module to enrich the ns-3 library.

Keywords Intelligent Transportation Systems · Mobile IPv6 · Network · Mobility · Smart City

1 Introduction

In the era of Smart City and Internet of Things (IoT), each and every user needs to be connected to the Internet seamlessly. Intelligent Transportation Systems is one of the main aspects of the Smart City. In case of group mobility, i.e., when a group of

P. Mandal · M. K. Rana · P. Chatterjee
School of Mobile Computing and Communication, Jadavpur University,
Kolkata, India
e-mail: prasanta.mtechdmc@gmail.com

M. K. Rana
e-mail: manoj24.rana@gmail.com

P. Chatterjee
e-mail: punyasha.chatterjee@gmail.com

A. Debnath (✉)
The Calcutta Technical School, Kolkata, India
e-mail: arpita.debnath03@gmail.com

users is moving in a vehicle (bus/train/flight), due to the mobility of the vehicle [1], the wireless interface of it sometimes switches the on-road access points. This event is called the IP handoff. This results in severe service degradation of the vehicular users. To handle it, the Internet Engineering Task Force (IETF) standardizes the Network Mobility Basic Support (NEMO-BS) protocol by extending the base Mobile IPv6 (MIPv6) scheme [2]. The objective of this is to provide uninterrupted Internet connectivity to the vehicular mobile users by maintaining ongoing sessions during handoff. UMIP (Usagi Patched MIPv6 stack) is an open-source implementation of MIPv6 and NEMO-BS protocols for Linux and can be used for network simulator ns-3. But it is only executed on ns-3 Direct Code Execution with the Linux native stack and that is why if any changes may be proposed in NEMO, then the users must have to trace the Linux stack. But, it is quite complicated to the naïve users to do any changes in Linux stack.

To fix the above issue, we have implemented the NEMO-BS protocol [3–6], using the ns-3 stack, by modifying the existing MIPv6 module [7, 8] in ns-3, that may be added to the ns-3 library which may help the ns-3 users a lot.

The paper is arranged as: Sect. 2 describes the Literature Survey, Sect. 3 describes the working principle of the protocol, Sect. 4 gives the description of the Class Diagram Implementation, Simulation and Results are given in Sect. 5 and finally, Concluding remarks emerge in Sect. 6.

2 Literature Survey

In [1], they proposed a group-based network mobility controlling technique to diminish the signaling problem. Besides this, they reduce the handover latency as well. But they did not investigate the finest mode for vehicles to be clustered together or they did not employ the neighboring association among the vehicles when alliance them.

In [8], they introduced a NEMO support protocol for Intelligent Transportation Systems that support mobility management as well as handover. But they did not support the guaranteed seamless connectivity during a handover.

Hager et al. [9], illustrate MINT- a Mobile Internet Router having enough computational capability to execute all essential communication protocol operations while enabling connections for the nodes. Transparency of the communication software is provided by the MINT router, and there is no requirement of any modifications with the basic software of mobility support while connecting via such a router to the Internet.

In the Request for Comments [10] which was on IP mobility support, clearly specifies the fact of using a Mobile Router for the mobile network.

By considering all the limitations of the above-related study, we try to implement NEMO-BS protocol using ns-3 network simulator that ensures the seamless network connectivity as well as the IP handoff management.

3 Working Principle of NEMO-BS Protocol

The working principle of the protocol is depicted in Fig. 1. The network entities for the handoff operation includes a Mobile Network Node (MNN), Mobile Router (MR), Home Agent of the MR (MR-HA), and Correspondent Node (CN).

Step 1: When MNN enters into a vehicle, it connects with the MR as MR is the default Internet Service Provider for the MNN inside a vehicle and acquires a Care-Of-Address (CoA), configured from the MRs prefix.

Step 2: As MNNs home prefix differs from that prefix, it binds the CoA with its HA (MNN-HA) through a Binding Update (BU) process.

Step 3: When the MR changes access point, it performs the BU process with the MR-HA. The BU processes result in two tunnels between entities (MR and MR-HA, MNN and MNN-HA).

A packet from CN to MNN follows the route: CN → MNN – HA → MR – HA → MR → MNN. It is encapsulated first by the MNN-HA such that it could reach the MR’s home. Reaching MR’s home, the packet is automatically redirected toward the MR-HA as the corresponding HA contains MR’s home links. MR-HA has the binding of the MR’s advertised prefixes to the MR’s current CoA. So, the MR-HA encapsulates the packet, setting MR’s CoA as the destination. The tunnel headers are decapsulated in the reverse order, i.e., in the MR first and then, in the MNN.

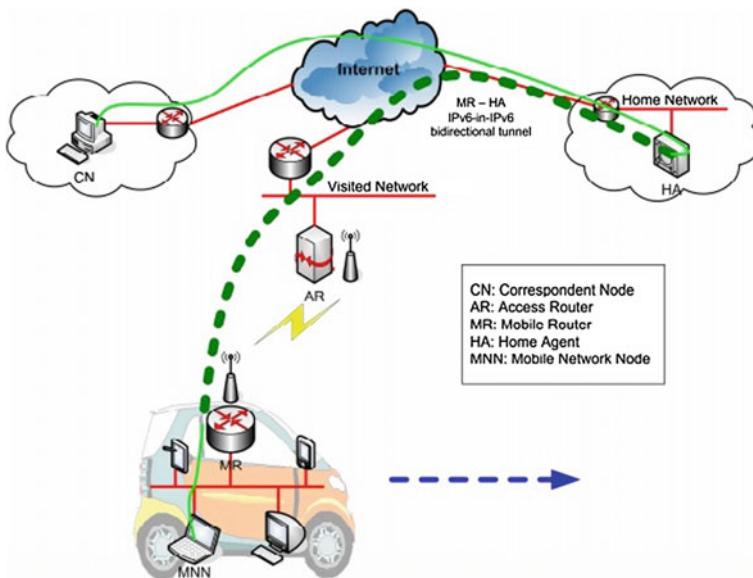


Fig. 1 NEMO-BS operation [4]

The working process of the NEMO-BS differs from the MIPv6 in the following two ways:

- The MR registers all its prefixes (which it advertises to the MNNs) during its BU process, instead of only the home address.
- Unlike MIPv6, the MR-HA must have to advertise all those MR’s prefixes to its neighbors such that the MR-HA can receive all the packets, destined to the MNNs CoAs, configured from those prefixes.

Here, we will describe the Binding Update (BU) process and data packet processing in NEMO-BS in detail.

3.1 Binding Update (BU) Process of NEMO-BS

In MIPv6, when a mobile host configures a new address in its interface, it sends single BU to its HA. But in NEMO-BS instead of sending single BU, each MH in the mobile network sends BU to its corresponding HA through MR when they configure a new address in the mobile network. In the same way, when MR configures a new IPv6 address in a new subnet from a new access router advertised prefixes, it also sends a BU to its HA. When a node configures an address on an interface, immediately Duplicate Address Detection (DAD) method is started to verify the delicacy of that address. After the time-out session of DAD, it calls the SetState() which is defined in ns3 core.

3.2 Packet Processing of NEMO-BS

After the successful completion of Binding Update process, a bidirectional tunnel is established between MR and MR-HA and another tunnel is established between mobile network MN and MNHA (Fig. 2). That is why when a mobile host in the

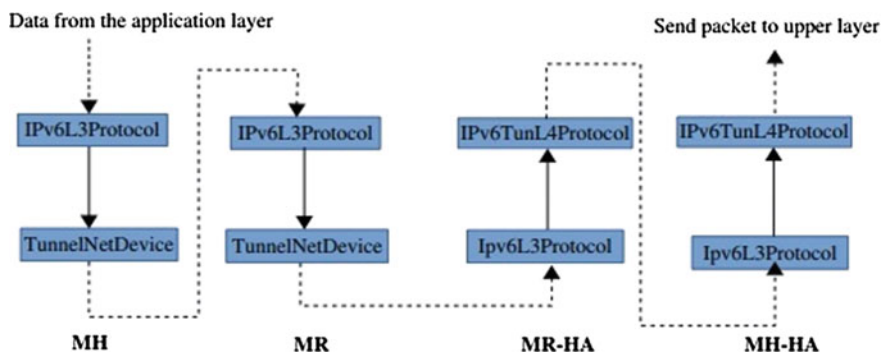


Fig. 2 Data packet processing from MH to MHHA

mobile network sends a data packet to a Correspondent Node (CN), the first level encapsulation is done by the tunnel between in MH and MHHA, and then is sent to MR. After that, second-level encapsulation is performed by MR because of the tunnel between MR and MR-HA. When the two-level encapsulated packet reaches the MR-HA, it decapsulates the outer encapsulation and forwards the remaining packet to the MHHA. When this packet reaches MHHA, it decapsulates the first level of encapsulation and sends it to the right destination that is the correspondent node. Besides this, the data packets are processed in reverse direction from the CN to MH.

4 Class Diagram Implementation

The existing classes of the implemented MIPv6 module [7] are modified to support the NEMO-BS functionalities as shown in the class diagram in Fig. 3. We divide our NEMO-BS classes into mainly four modules:

- Header
- Internet stack
- Net device
- Helper.

The important operations and functionalities of each module are detailed as follows:

4.1 Header

We have implemented the NEMO-BS functionality in ns3, by modifying some of the existing MIPv6 headers message formats in ns3 as defined in RFC3963. All the headers are inherited from the base class Header in ns3 system. In Fig. 3, we have depicted the main data members of NEMO-BS headers and relationship among them as defined in RFC3963. For NEMO-BS in ns3, we have added m flagR data field in Binding Update (BU) and Binding Acknowledgement (BA) header to denote the mobile node's status. If this flag value is set in BU, then HA assumes that MIPv6 mobile node acts as a NEMO-BS MR otherwise it considers it as MIPv6 Mobile Host (MH). A new mobility option header named as Mobile Network Prefix Option Header is added with existing headers shown in Fig. 3, which is also inherited from MIPv6OptionHeader class. It carries the one or more prefix information (128 bit or more) from MR to its HA within BU message and these prefixes are advertised by MR-HA to receive the packet correctly for the mobile host in the mobile network.

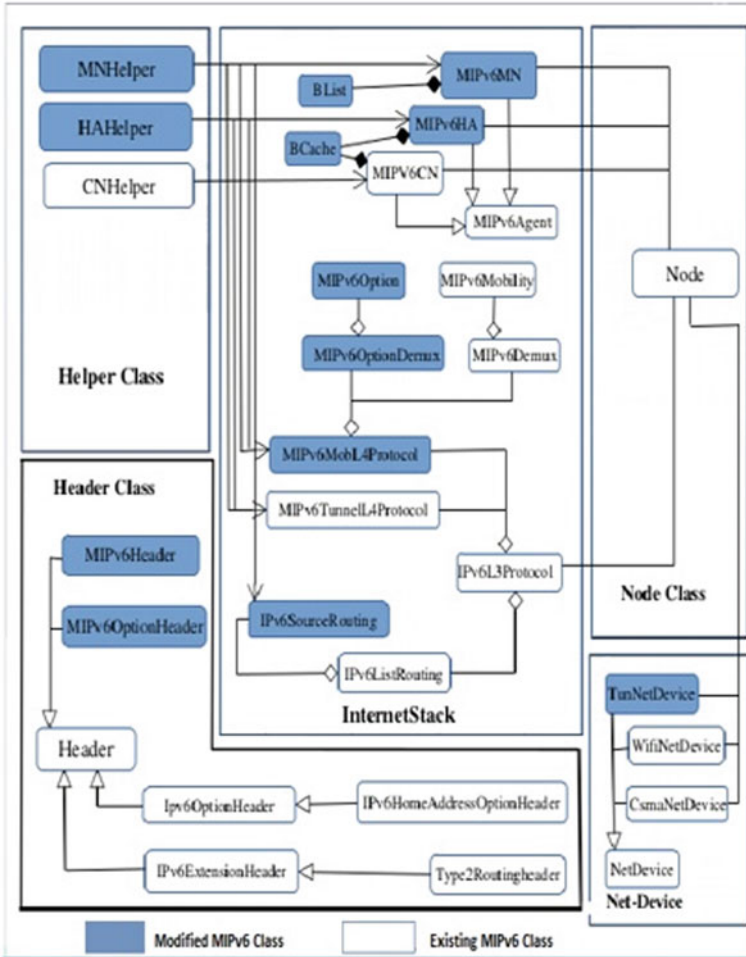


Fig. 3 NEMO-BS classes and their functional relationships

4.2 InternetStack

NEMO-BS implementation contains the two types of Demux classes named as MIPv6Demux and MIPv6OptionDemux. Details of the functionality of Demux classes are described in paper [7]. For NEMO-BS, a new mobility option is added to the MIPv6Option class. The MIPv6MobL4Protocol deals with mobility messages necessary for communication, whereas the MIPv6TunL4Protocol deals with data packets in the time of communication. When from lower layer, a packet is recognized by IPv6L3 Protocol and then sends it to the MIPv6-layer. The summarized data packet or a mobility message contains the MH type value in its header

which is dissimilar from normal data packet. The MH Type field value is checked by IPv6L3Protocol class, and depending on this value, appropriate upper layer L4 protocol class is called. If the packet is received by Receive() function of MIPv6TunL4Protocol class, then its Receive() function checks the summarized packet tunnel line which may use to de-capsulate the packet. It drops the packet if there is no matching tunnel interface found. The MIPv6TunL4Protocol class may be added, removed, or modified a tunnel.

4.3 *NetDevice*

A tunnel is made for virtual MAC layer by the TunNetDevice class. The TunNetDevice class reimplements the base class Send() function. After the reception of a packet from the upper layer, the Send() function accomplishes IPv6inIPv6 encapsulation by making a new IPv6 header. The main functionality of the MIPv6TunL4Protocol class is TunNetDevice class.

4.4 *Helper*

This is in the top of the above all implemented classes. When a user wants to use the NEMO-BS protocol, they don't have to bother about the complex internal structure rather they only call the Install() function of MIPv6Helper class to install the functionality on a particular node.

5 Simulation and Results

The simulation framework is shown in Fig. 4. We have used ns-3.25 for simulation. The MR has two interfaces.

- It connects with the Access Router (AR) through Wi-MAX interface and
- With the MNNs through Wi-Fi interface.

Simulation setup: The simulation snapshot is shown in Fig. 5. The Mobile Host (MH) and Correspondent Node (CN) perform as a host and remaining all others as a router. A middle router R1 connects the AR1 and AR2 connects with CSMA interface to MR-HA, another middle router R2 connects the CN and MHHA through CSMA interface to R1. The MR has two interfaces, it connects with access router through Wi-Fi interface, and connects with the mobile host in the mobile network through Wi-Max interface. IEEE 802.11 radio is used over here in AR1 and AR2. Beacons are generated in every 100 ms. Here, the data rate is set at

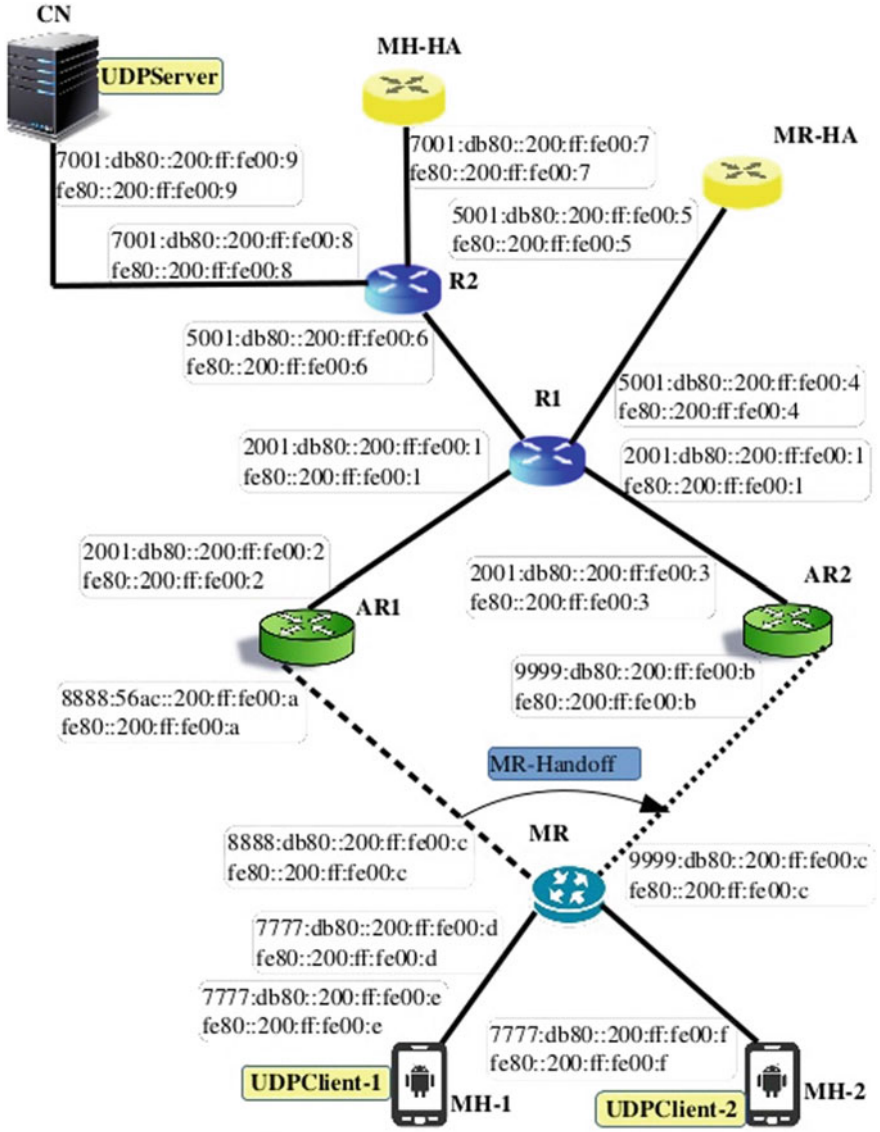


Fig. 4 Simulation Framework

2 Mbps and the link delay at 1.0 ms. On mobile hosts (MH1 and MH2), a UDP echo-server application is installed. On the CN, an echo-client application is running with 1024 packet size and maximum 100,000 packets. Here, the MR home network prefix is 5001:db80:/64, and MR's HoA of MR is expressed from this.

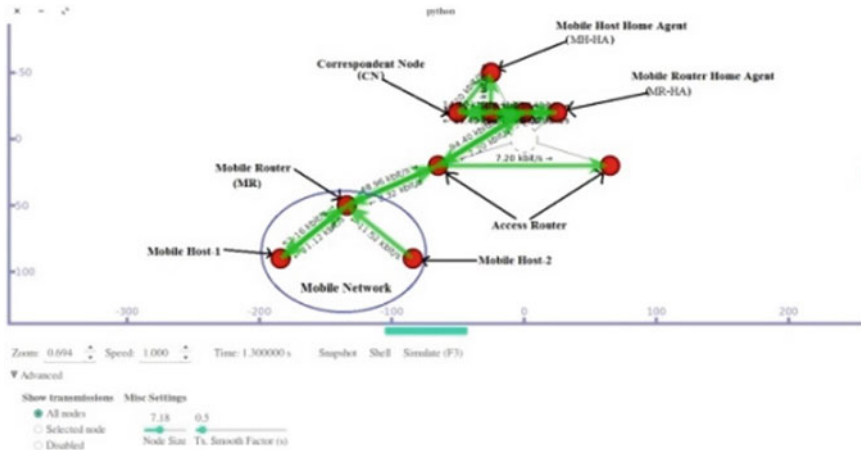


Fig. 5 Simulation snapshot of implemented NEMO-BS in NS3

The constant velocity model (20 m/s) is used for all the MR and moves from left to right fashion.

The simulation parameters are given below:

- Both the AR1 and the AR2 generate beacons in every 100 ms.
- All CSMA links use 2 Mbps data rate and 1.0 ms link delay.
- A UDP echo-server is fixed on the MNNs (MNN is called as MH) (MNN-1 and MNN-2) (say at port 9).
- The MR with all the MNNs in the mobile network uses the constant velocity mobility and travels from the left to right with 20 m/s velocity.

From the packet capture (PCAP) files, we can calculate the delay experienced by MNN1 in the mobile network due to the handoff of MR from AR1 to AR2. At $t = 7.240$ s, the MNN1 successfully receives the preceding packet from the CN. After that MR changes the point of attachment and after successful completion of the handoff of MR, MNN1 again successfully receives the next packet at $t = 12.245$ s. So, the delay experienced by MNN1 = $(12.245 - 7.240)$ s = 5.005 s.

6 Conclusion

NEMO-BS protocol provides seamless Internet connectivity by providing smooth and fast hand-off for a group of users moving from one place to another. In this paper, we have implemented this protocol as an extension of MIPv6 in ns3 environment, so that it can be appended as a new library file in ns-3 in future, which will help the naive users. The above implementation complies with the standard, defined by the IETF. But, a few future works are also required to make it fully usable by the

research community. As the open-source simulator like ns-3 presently mandates validation for any new implemented module, our implementation must be validated against some of the strong existing implementations. Above all, integrating NEMO into LTE would be very interesting as recent research trends mostly rely on LTE communication.

References

1. Kim, M.S., Lee, S.K.: Enhanced network mobility management for vehicular networks. *IEEE Trans. Intell. Transp. Syst.* **17**(5), 1329–1340 (2016)
2. Perera, E., Sivaraman, V., Seneviratne, A.: Survey on network mobility support. *SIGMOBILE Mob. Comput. Commun. Rev.* **8**(2), 7–19 (2004)
3. Network Mobility Basic Support Protocol. Retrieved July 25, 2018 from <https://github.com/prasanta2018/NetworkMobility>
4. Bernardos, C.J., De La Oliva, A., Caldern, M., von Hugo, D., Kahle, H.: Nemo: network mobility. Bringing ubiquity to the internet access. Demonstration at IEEE INFOCOM (2006)
5. Khan, M.Q., Andresen, S.H., Khan, K.N.: Pros and cons of route optimization schemes for network mobility (NEMO) and their effects on handovers. In: Proceedings of the 8th International Conference on Frontiers of Information Technology (FIT '10). ACM, New York, NY, USA, Article 24
6. Bauer, C., Ayaz, S., Ehammer, M., Grupl, T., Arnal, F.: Infrastructure-based route optimization for NEMO based on combined local and global mobility. In: Mobility Conference, p. 61 (2008)
7. Rana, M.K., Rana, B., Mandal, S., Saha, D.: Implementation and performance evaluation of a mobile IPv6(MIPv6) simulation model for ns-3. *Simul. Model. Pract. Theory* **72**, 122 (2017)
8. Lee, J.H., Ernst, T., Chilamkurti, N.: Performance analysis of PMIPv6Based network mobility for intelligent transportation systems. *IEEE Trans. Veh. Technol.* **61**(1), 74–85 (2012)
9. Hager, R., Klemets, A., Maguire, G.Q., Reichert, F., Smith, M.T.: MINT-AMobile internet router. In: 1st International Symposium on Global Data Networking, Cairo, Egypt, Dec. 13–15, 1993 (1993)
10. Perkins, C.: IP Mobility Support, IETF. RFC October 1996 (2002)

Modified DFA Minimization with Artificial Bee Colony Optimization in Vehicular Routing Problem with Time Windows



G. Niranjani and K. Umamaheswari

Abstract A NP-hard problem, vehicular routing is a combinatorial optimization problem. Vehicular routing problem with time windows indicates vehicular routing with specified start and end time. There will be “n” number of vehicles starting from the depot to cater to the needs of “m” customers. In this paper, Gehring and Homberger benchmark problems are considered wherein the size of customers is taken to be 1000. Artificial Bee Colony Optimization algorithm is executed on these 60 datasets and the number of vehicles along with total distance covered is recorded. The modified version of Deterministic Finite Automata is applied along with the Artificial Bee Colony Optimization and the results produce 25.55% efficient routes and 15.42% efficient distance compared to simple Artificial Bee Colony Optimization algorithm.

Keywords Artificial Bee Colony Optimization · Minimization · Number of routes · Total distance

1 Introduction

Vehicular routing problem with time windows is a NP-hard problem. It is a combinatorial optimization mechanism. Vehicular routing problem is similar to travelling salesman problem. In the travelling salesman problem, a sales person has to start from one city, travel to all the other cities only once and then return to the original city with the minimum cost possible. In VRPTW, the number of customers and the number of vehicles are specified. These vehicles start from the depot that is also specified in the problem, go to different customers that are distributed across

G. Niranjani (✉)

Department of CSE, PSG Institute of Technology and Applied Research, Coimbatore, India
e-mail: niragopal@gmail.com

K. Umamaheswari

Department of IT, PSG College of Technology, Coimbatore, India
e-mail: umakpg@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_45

643

and then return to the depot. The customers are specified as x and y coordinates of the point. When a customer is reached, the next customer is selected which has the minimum distance from the current point among all the points that are not yet visited. This paper considers the 60 datasets pertaining to Gehring and Homberger benchmark problems that service 1000 customers.

In the real world, these vehicle routing problems are being used in distribution and transportation logistics [1] and also specific applications such as solid waste, beverage, food, diary and newspaper industries [2].

The dataset, in general, contains the following details: vehicle count and the capacity of each vehicle. For each vehicle, the following details are specified: (1) Customer number: for depot, the customer number is zero. All the other vehicles have natural numbers sequentially. (2) x -coordinate of the customer. (3) y -coordinate of the customer. (4) Demand: this denotes the priority. (5) Start time: the customer needs to be serviced after the start time. (6) End time: the customer needs to be serviced before the end time. (7) Service time: the time taken to perform service for this particular customer. The Gehring and Homberger benchmark problems have six sets of problems: C1 type, C2 type, R1 type, R2 type, RC1 type and RC2 type. Each type has 10 problems. The types vary in terms of vehicle count and the capacity of each vehicle.

1.1 Reason for Considering the Gehring and Homberger's Problem for the Experiments

1. Solomon's benchmark problems have been considered in many algorithms associated with vehicular routing problem with time windows. In these problems, the maximum number of customers in a single instance is 100. Gehring and Homberger benchmark problems are a variation of Solomon's benchmark problems with more number of customers for each instance.
2. The number of customers represented in each of the instances is 1000. To demonstrate that the algorithm can handle a huge amount of customers comparative to the other algorithms such as branch and price, Tabu search, firefly algorithms and even some variations of ABC [Tournament selection, Vector Evaluated Technique and Roulette Wheel Selection technique], these problems are selected.

2 Artificial Bee Colony Optimization

Karaboga [3] started the Artificial Bee Colony algorithms after observing the behaviour of honey bees and applied them to real-time problems. He along with Karaboga and Basturk [4] further extended the usage of the algorithm for solving

various optimization problems. Karaboga and Basturk [4] and Singh [5] propose a solution that shows the better performance of bee colony algorithms in comparison with particle swarm optimization and genetic algorithms, particle swarm optimization [6, 7], Tabu search and ACO algorithms, respectively.

2.1 Working of Honey Bees

Artificial Bee Colony Optimization algorithms are based on the working of real-world honey bees. These algorithms come under swarm intelligence—wherein problems are solved based on the collective behaviour of social insect colonies and other animals. In this paper, the honey bees are considered. There are three types of honey bees: (1) Scout bees: these are used to search for new food sources randomly. (2) Onlookers: the onlooker bees calculate the nectar amounts available in the food sources. (3) Employed bees: these determine the nectar amount and the probability value with which the food source can be reached from the bee hive.

2.2 Essential Components of Forage Selection

(1) Food sources: a particular food source is selected based on its proximity to the hive, richness/concentration and the ease of extracting energy. (2) Employed forager: it carries information about the food source, how much distance it is from the hive and the direction of the food source from the hive, profitability and it shares information with a certain probability. (3) Unemployed forager: the scout bees and the onlooker bees come under this category.

2.3 Modes of Behaviour

There are two important modes of behaviour of the honey bees with respect to the food sources. (1) Recruitment of the food source, (2) Leaving behind the food source. The honey bees exchange information about the food sources among themselves by means of the “waggle dance”.

2.4 Basic Self-Organization Properties

(1) Positive feedback: when the nectar in the food source is high, the number of onlooker bees is high. (2) Negative feedback: when the nectar in the food source is poor, there are no bees nearer. (3) Fluctuations: probability value of the nectar in the

food source in comparison with others [8]. (4) Multiple Interactions: Communication between different types of bees with regard to the food source. Many tasks are performed at the same time leading to the division of labour [3].

2.5 Algorithm

The general Artificial Bee Colony Optimization algorithm proposed by Dervis Karaboga [4] is as follows:

Send the scout bees onto the initial food sources

REPEAT

Send the employed bees and find the nectar amounts in the food sources
The onlooker bees decide the selection of food source based on its probability value
They go to the respective food source and determine the nectar amount
The food source selected by the bees are exploited
Again, send the scout bees to search for new food sources
Remember the food sources exploited so far.

UNTIL (constraints are satisfied).

2.6 Parameter Setting for the ABC Algorithm

1. All the vehicles start from the depot.
2. It is assumed that the vehicles are not affected by traffic conditions.
3. All the customers will be serviced only after the start time but before their end time.
4. A particular vehicle should be available to service a customer within the service time of that particular customer.

3 VRPTW

The study of vehicular routing started in 1959 when G. B. Dantzig and J. H. Ramser [9] proposed the truck dispatching problem applicable to the delivery of gasoline to service stations. It was followed by Clarke [10] to provide an iterative procedure to select the optimum or near-optimum route in a better way. A lot of algorithms [11] have been proposed and some implemented to this problem which provide both exact and approximate solutions.

The vehicular routing problem with time windows has been solved by many algorithms. Some of the well known algorithms are presented in Table 1.

3.1 Objectives

The following are the objectives of the VRPTW:

1. Minimization of Total routes:
The total number of routes refers to the total number of vehicles needed to solve a particular problem. The number of vehicles varies depending on the placement of customers. If the customers are nearer, a single vehicle is enough to service all those customers, provided that particular vehicle is available within the start time and the end time of that customer. The algorithm is designed such that the total number of routes is minimized [12].
2. Minimization of the total distance travelled:
Distance of a single vehicle is measured as the sum of the distance between the depot and the first customer that the vehicle is servicing and the distance between the first and the second customer and then the second and the third and so on till the last customer. The distance between the last customer and the depot is added along with these values. The total distance of a particular customer is calculated as the summation of distances of individual vehicles. This algorithm takes care of the fact that the total distance is reduced.
3. Minimization of the total time taken:
Time taken for a single vehicle is calculated as the summation of the start time of the first customer and its service time and then subsequent service time of the corresponding customers. Total time is calculated as the summation of the time taken by individual vehicles which is to be reduced by the proposed algorithms.

It has to be taken care that the capacity constraints are not violated [13].

3.2 Procedure

There are different variations for solving VRPTW using Artificial Bee Colony Optimization algorithm. Some are presented in Table 2.

The vehicle count, the capacity of vehicles and the depot details—customer number: 0, x-coordinate and y-coordinate of the depot, demand, start time, end time and unload time—are got from the input dataset along with the details of the 1000 customers. The number of routes is initialized to 0.

The following procedures are executed until all the customers are serviced:

1. MemorizeBestSource (): This function is used to select the customer with the least start time and also that has not been serviced yet.

Table 1 Existing solutions for VRPTW

Author and year	Title	Strategy	Parameter	Dataset	Methodology	Benefits	Limitations
Bulhões and Uchoa [16]	A branch-and-price algorithm for the Minimum Latency Problem	Minimum Latency Problem	Minimize the sum of waiting time of customers	40 TSP Library instances with up to 200 vertices	In branch-and-price method, the arcs are considered instead of specific nodes	Roberti and Mingozzi (2014) did not solve nine instances. This method solved them. Also, branch-and-price could solve larger instances of size 159, 175, 180 and 195	When the vertices are greater than 150, it is difficult to solve
Michael Schneider, September 2015 [17]	The vehicle-routing problem with time windows and driver-specific times	Tabu search	Minimize travelling distance and working duration	Solomon's 100 customer instances	Relocation, Exchange of data and 2-Opt methods	10 runs are executed and the optimal ones are considered	Meta-heuristic method can be improved by means of splitting methods for solutions of higher quality
Huang et al. [18]	Multi-agent ACO for VRP with soft time windows and road condition	A mathematical model of VRP with soft time windows [penalty for lateness] and road factor	Transport cost, fuel consumption and customer satisfaction achieved by the arrival time of vehicles and the expecting time of customers	40 customer problem	Adaptive pheromone trail, state transition rule and pheromone updating, 3-opt solution	Multiagent ant colony optimization is considered better than ACO and AGA	For larger instances, yet to be verified

(continued)

Table 1 (continued)

Author and year	Title	Strategy	Parameter	Dataset	Methodology	Benefits	Limitations
Matthopoulos and Sofianopoulou [19]	A firefly Algorithm for the Heterogeneous Fixed Fleet VRP	HFVRP is a variation of VRPTW wherein the customers have different capacities and costs for goods distribution	Relative brightness, Degrees of attraction, Distance between fireflies	Augerat et al. (1995)	Relative brightness, Degrees of attraction, Distance between fireflies, Position update and Parameters values selection	Produces the best known solution for small datasets—up to 39 nodes	The next node is determined by means of a random walk of the firefly and hence, it is not much effective

Table 2 Existing solutions for ABC in VRPTW

Author and year	Title	Strategy	Parameter	Dataset	Methodology	Benefits	Limitations
Shi et al. [20]	A modified ABC algorithm for vehicle routing problems with time windows	Tournament selection	Minimize the total distance	Solomon's R102 problem	Fitness points are calculated between pairs of points and at the end, the one with the larger fitness value is selected	Travelled distance is better than the ant colony algorithm, iterated local search and genetic algorithms	Only one instance of 100 customer problem was considered
Nahum et al. [21]	Multi-Objective VRPTW: a Vector-Evaluated ABC Approach	Vector-Evaluated Technique	Count of vehicles and the total cost	Solomon's 25, 50 and 100 customer problem	(1) Minimize the number of vehicles, (2) The total path length is optimized based on the results calculated beforehand	Optimal solution can be obtained by running the algorithm again and again	The end result is based on previous runs. So, it takes lots of time for computation
Bhagade and Puranik [12]	Artificial Bee Colony (ABC) Algorithm for Vehicle Routing Optimization Problem	Effective path planning	Tour length and bee travel time	Results are simulated	Control parameters are set which help in determining the next node using the nearest neighbour method	Optimal distance is obtained by greedy selection strategy	Very few parameters are considered
Alzaqebah et al. [22]	Modified Artificial Bee Colony for the vehicle routing problems with time windows	Roulette wheel selection	Convergence speed increased	Solomon's 100 customer problem	The bees select the best node based on the random selection of nodes for the count of maximum trials	67% of the results are better than simple-ABC	The results need to be increased

2. **SendEmployedBees ()**: This function is used to execute the work of the employed bees. The customer selected by the **MemorizeBestSource ()** is to be processed. This function selects the route that this customer is to be placed in. A group of routes is selected based on the following conditions:
 - The distance from the depot should be greater than the distance from the routes latest point to this customer.
 - The time taken by the route until the latest point should be less than the customer's end time.
3. **CalculateProbabilities ()**: This function selects the route with the minimum distance among the set of routes selected in the **SendEmployedBees ()** function.
4. **SendOnlookerBees()**: If no such routes are selected by the **CalculateProbabilities ()** function, a new route is created with the following conditions:
 - The vehicle count of the new route is set to 1.
 - The customer is marked as serviced.
 - A new route is created with the present customer as the first vehicle.
 - Route time is equal to the sum of the starting time of a particular customer and the time taken to unload.
 - Distance of the route is calculated as the distance from the depot to the present customer.
 - The number of routes is incremented by 1.

When a sample route is selected to fit in this customer, the following procedures are carried out:

- The customer is marked as serviced.
- The present customer is added with the selected route as the last one.
- The number of vehicles processed by the current route is incremented by 1.
- If the start time of the present customer is before the route time, the route time is calculated as the sum of route time and the unload time for this particular customer. Else, the route time is calculated as the sum of the starting time of the present customer with its unload time.
- Distance of the route is calculated as the sum of the previous distance and the distance between the previous point and the present customer.
- The present customer is made as that particular route's last point.

All the distances are measured using the Euclidean distance—The distance between two points (x_1, y_1) and (x_2, y_2) is

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

5. **SendScoutBees ()**: The scout bees manage the customers that can be selected in the next iteration.
The route number, the number of vehicles, specific distance, number of customers serviced and time taken is displayed for each route. In the end, the total

number of routes, total distance—the summation of the distance of individual routes and the total number of serviced customers—are displayed. The number of vehicles and the total distance travelled is recorded to be compared.

4 Minimization of Deterministic Finite Automata

Deterministic Finite Automata refers to a machine that has a countable number of states and each input state has exactly one output state on a given input. The deterministic finite automata is indicated as $(Q, \Sigma, \delta, q_0, F)$ wherein Q refers to the set of states, Σ refers to the input alphabet, δ refers to the transition equation which contains the input state, input alphabet and the output state, q_0 refers to the initial state and F refers to the set of final states [14].

For minimizing a DFA, initially, the final states are considered as one component and all the other states are considered as another component. Each component is checked whether they all give the states belonging to the same component for each alphabet. If so, they are retained in the same component. Else, the component is split and the same procedure is followed again until there is no change. In the end, the states belonging to the same component are considered as one state and the Deterministic Finite Automata is reconstructed [15].

4.1 Algorithm for Modified DFA on ABC

On applying modified DFA minimization, the ABC algorithm changes as follows:

Send the scout bees onto the initial food sources

Group the food sources based on the distance between them

REPEAT

Send the employed bees and find the nectar amounts in the food sources

The onlooker bees decide the selection of food source based on their probability value

They go to the respective food source and determine the nectar amount

The food source selected by bees are exploited.

For each food source that is available in the group of the current food source

If the food source is currently available and is not yet exploited, it is exploited by the bees

Again, send the scout bees to search for new food sources

Remember the food sources exploited so far

UNTIL (constraints are satisfied).

4.2 *Parameter Setting for Modified DFA on ABC*

In addition to the above-mentioned parameters, a few are added to the modified DFA with the ABC algorithm:

1. The customers in a single group will be serviced by a single vehicle only when the start time and the end time coincide with the free time of the vehicle. Otherwise, different vehicles will be assigned to vehicles in a single group.

5 **Modified DFA Minimization with Artificial Bee Colony Optimization in Vehicular Routing Problem with Time Windows**

5.1 *Changes to Artificial Bee Colony Optimization in VRPTW*

In normal DFA minimization, the states that are grouped as a single component are considered as a single state. Here, it is modified such that the customers are grouped such that they belong to the same route if the capacity does not exceed. If the capacity exceeds the specified limit, the group can be split into different customers belonging to different routes.

The customers are considered in pairs having a minimum distance between them and they are grouped based on the following conditions:

- When both the customers are not in any groups, and they have a minimum distance between themselves compared to others; those two customers are combined together to form a new group.
- When any one customer of the pair is grouped and the other one is not paired, the customer not paired belongs to the group of the paired customer.
- When both the customers are paired, then two conditions are checked: (1) When they both belong to the same group, they are left as such. (2) When they both belong to different groups, any one of the customer groups is discarded and the customers belonging to the discarded group is added to the group of the other customer.

The functions of the Artificial Bee Colony Optimization for vehicular routing problem are executed as such except the functions:

- **SendOnlookerBees ()**: In this function, a small change is done. When a customer belonging to a group is selected to be attached to a route, the customers belonging to the same group are given a preference to be selected in the same route, provided the constraint pertaining to the capacity of the vehicle is not violated.

- `SendScoutBees ()`: This function is used to service the customers belonging to the group of the last serviced customer whose start time is less than the route time of the present route.

The route number, the number of vehicles, specific distance, number of customers serviced and time taken are again displayed for each route. In the end, the total number of routes, total distance—the summation of the distance of individual routes, the total number of serviced customers—are displayed. The number of routes and the total distance travelled is recorded to be compared with the previous one.

6 Results and Discussion

6.1 Comparison of Routes

The routes obtained from the Artificial Bee Colony algorithm is compared with the routes obtained in the Minimized DFA combined with ABC algorithm for the VRPTW. The comparison is done for each problem type separately.

`C1_Type`: Ten problems with count of vehicles: 200 and capacity: 250. There are 1000 customers in these problems each with service time of 90. The modified DFA with ABC algorithm gave better results for 90% of problems than the ABC algorithm as indicated in Fig. 1. The number of routes using minimized DFA is found to be 7.10% better than ordinary Artificial Bee Colony Optimization algorithm.

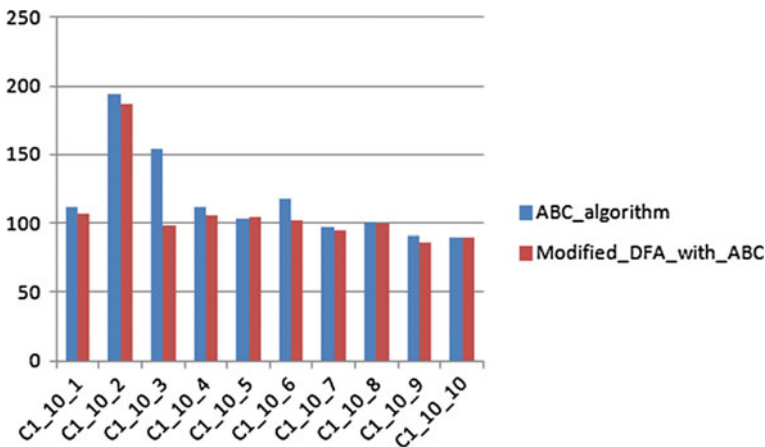


Fig. 1 Comparison of the number of routes for the ABC algorithm with the Modified DFA with ABC algorithm for C1 type problem

C2_Type: Ten problems with count of vehicles: 700 and capacity: 250. There are 1000 customers in these problems each with service time of 90. The modified DFA with ABC algorithm gave better results for 80% problems than the ABC algorithm as indicated in Fig. 2. The efficiency when minimized DFA is applied improves by 16.8% in terms of the number of routes.

R1_Type: Ten problems with count of vehicles: 200 and capacity: 250. There are 1000 customers in these problems each with service time of 10. The modified DFA with ABC algorithm gave better results for 90% problems than the ABC algorithm as indicated in Fig. 3. The number of routes is found to be 24.38% more efficient when minimized DFA is applied.

R2_Type: Ten problems with count of vehicles: 1000 and capacity: 250. There are 1000 customers in these problems each with service time of 10. The modified DFA with ABC algorithm gave better results for 100% problems than the ABC algorithm as indicated in Fig. 4. When minimized DFA is applied over Artificial Bee Colony Optimization, the results obtained are found to be 40.59% more efficient in terms of the number of routes.

RC1_Type: Ten problems with count of vehicles: 200 and capacity: 250. There are 1000 customers in these problems each with service time of 10. The starting time of customers in this data-set is far less compared to the previous problems. The modified DFA with ABC algorithm gave better results for 90% problems than the ABC algorithm as indicated in Fig. 5. These instances are found to be 10.43% more efficient when minimized DFA is applied for calculating the number of routes.

RC2_Type: Ten problems with count of vehicles: 1000 and capacity: 250. There are 1000 customers in these problems each with service time of 10. The starting time of customers in this dataset is far less compared to the R1, R2, C1 and C2 type problems. The modified DFA with ABC algorithm gave better results for 100%

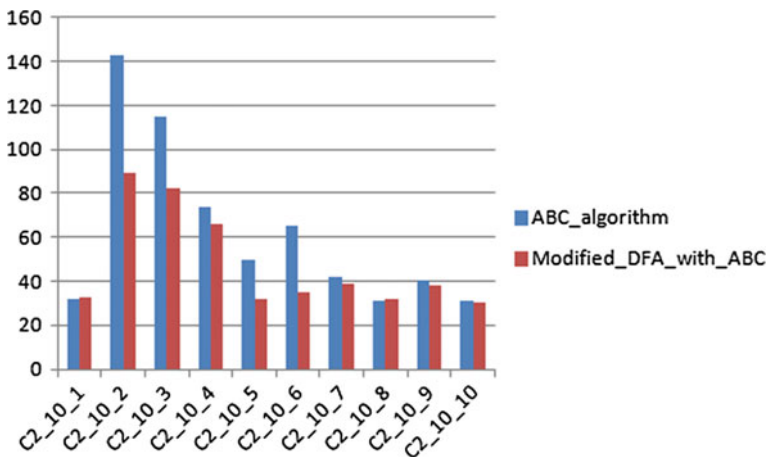


Fig. 2 Comparison of the number of routes for the ABC algorithm with the Modified DFA with ABC algorithm for C2 type problem

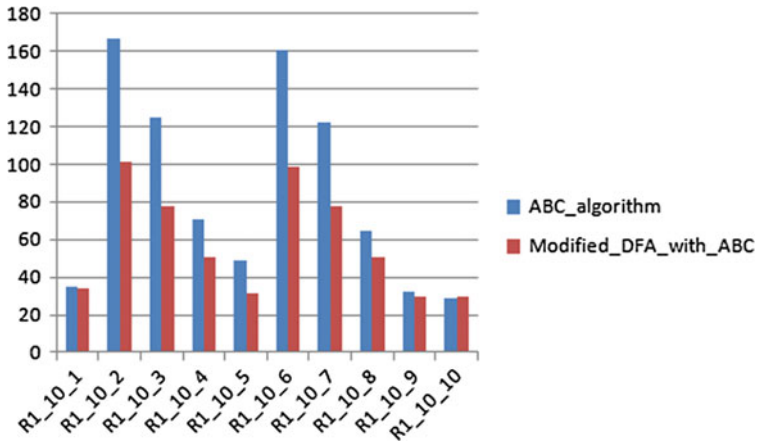


Fig. 3 Comparison of the number of routes for the ABC algorithm with the Modified DFA with ABC algorithm for R1 type problem

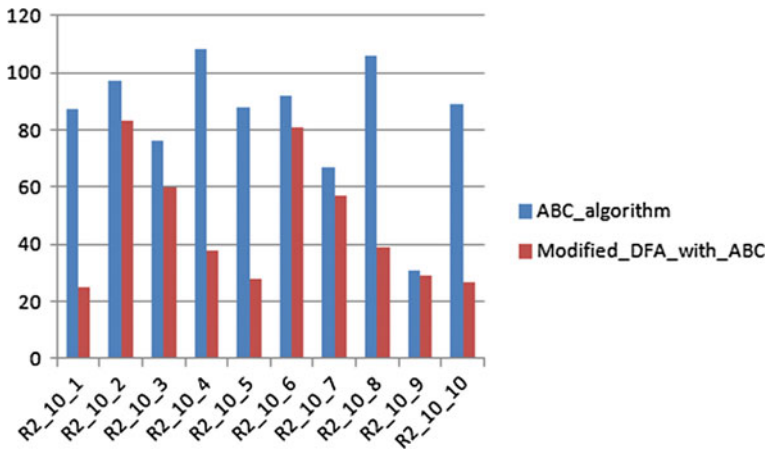


Fig. 4 Comparison of the number of routes for the ABC algorithm with the Modified DFA with ABC algorithm for R2 type problem

problems than the ABC algorithm as indicated in Fig. 6. 53.97% more efficiency is achieved for the RC2 type problems when minimized DFA is applied compared to Artificial Bee Colony Optimization algorithms for finding the number of routes.

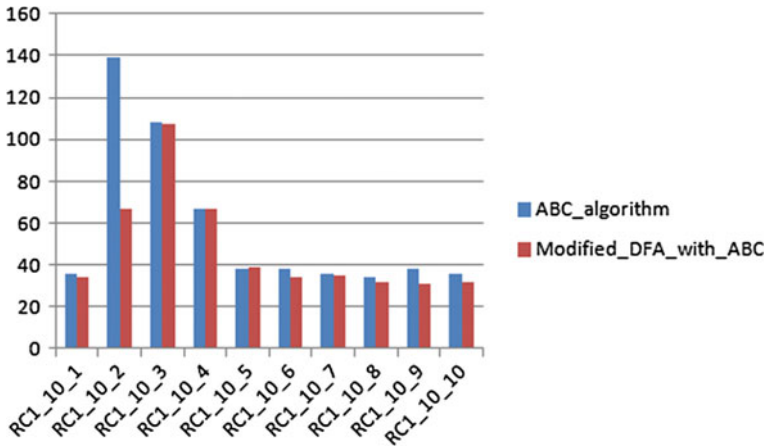


Fig. 5 Comparison of the number of routes for the ABC algorithm with the Modified DFA with ABC algorithm for RC1 type problem

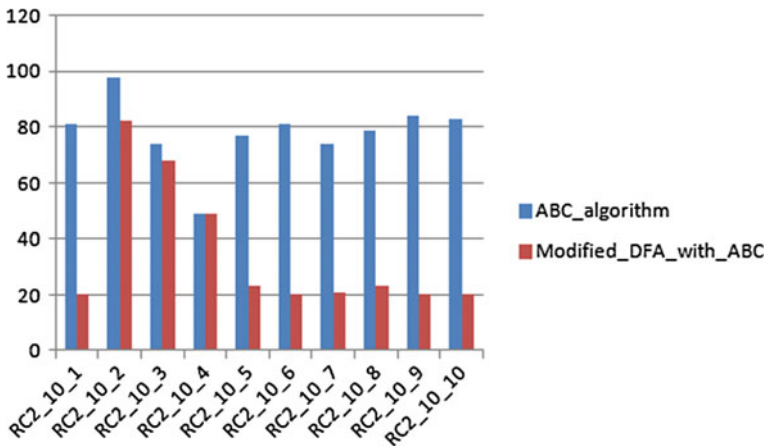


Fig. 6 Comparison of the number of routes for the ABC algorithm with the Modified DFA with ABC algorithm for RC2 type problem

6.2 Comparison of Distances

C1_Type: Ten problems with count of vehicles: 200 and capacity: 250. There are 1000 customers in these problems each with service time of 90. The modified DFA with ABC algorithm gave better results for 100% problems than the ABC algorithm as indicated in Fig. 7. The instances are found to be 7.29% more efficient when minimized DFA is applied on top of Artificial Bee Colony Optimization algorithms.

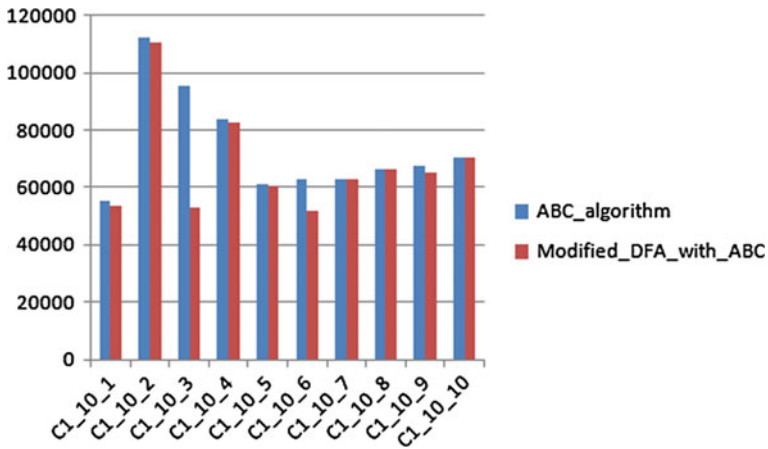


Fig. 7 Comparison of the total distance for the ABC algorithm with the Modified DFA with ABC algorithm for C1 type problem

C2_Type: Ten problems with count of vehicles: 700 and capacity: 250. There are 1000 customers in these problems each with service time of 90. The modified DFA with ABC algorithm gave better results for 90% problems than the ABC algorithm as indicated in Fig. 8. The efficiency when minimized DFA is applied improves by 18.02% in terms of the total distance.

R1_Type: Ten problems with count of vehicles: 200 and capacity: 250. There are 1000 customers in these problems each with service time of 10. The modified DFA with ABC algorithm gave better results for 80% problems than the ABC

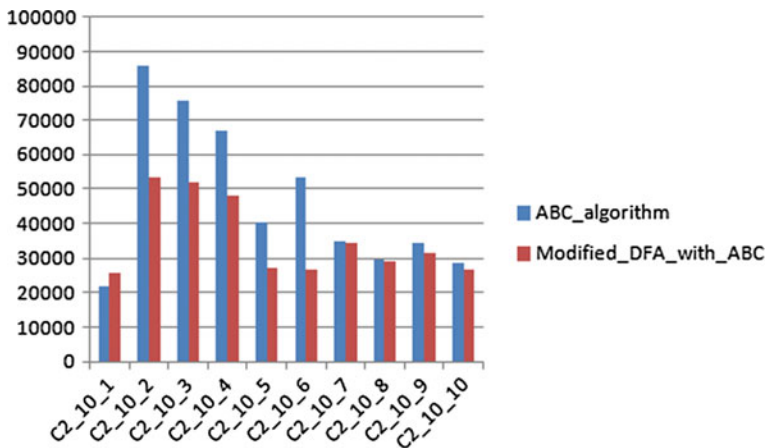


Fig. 8 Comparison of the total distance for the ABC algorithm with the Modified DFA with ABC algorithm for C2 type problem

algorithm as indicated in Fig. 9. The total distance is found to be 8.44% more efficient when minimized DFA is applied.

R2_Type: Ten problems with count of vehicles: 1000 and capacity: 250. There are 1000 customers in these problems each with service time of 10. The modified DFA with ABC algorithm gave better results for 100% problems than the ABC algorithm as indicated in Fig. 10. When minimized DFA is applied over Artificial Bee Colony Optimization, the results obtained are found to be 31.72% more efficient in terms of the total distance.

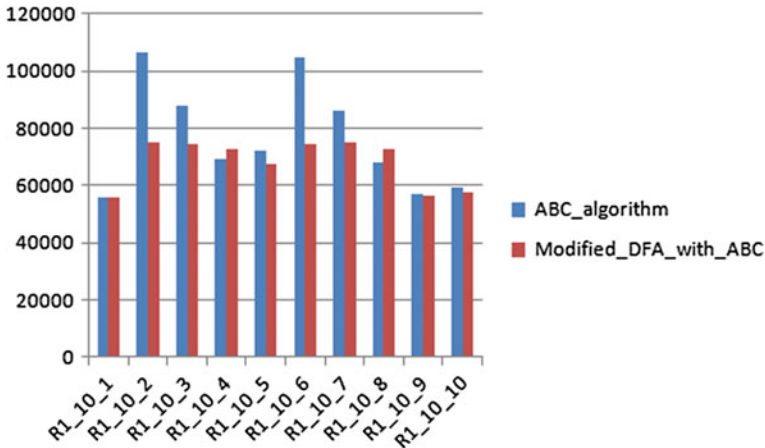


Fig. 9 Comparison of the total distance for the ABC algorithm with the Modified DFA with ABC algorithm for R1 type problem

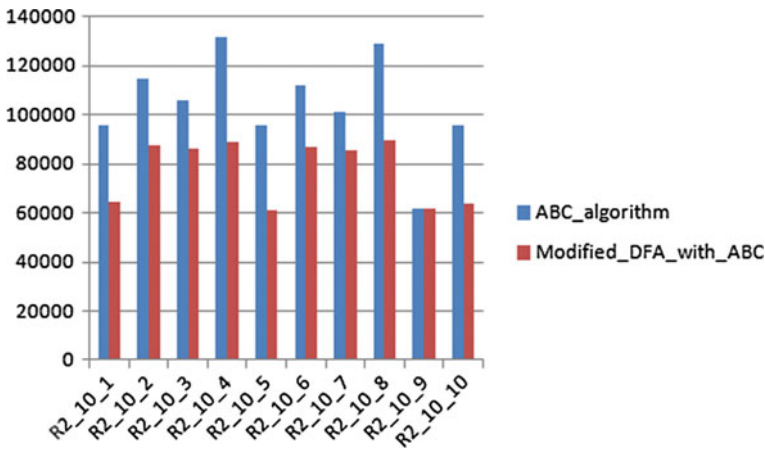


Fig. 10 Comparison of the total distance for the ABC algorithm with the Modified DFA with ABC algorithm for R2 type problem

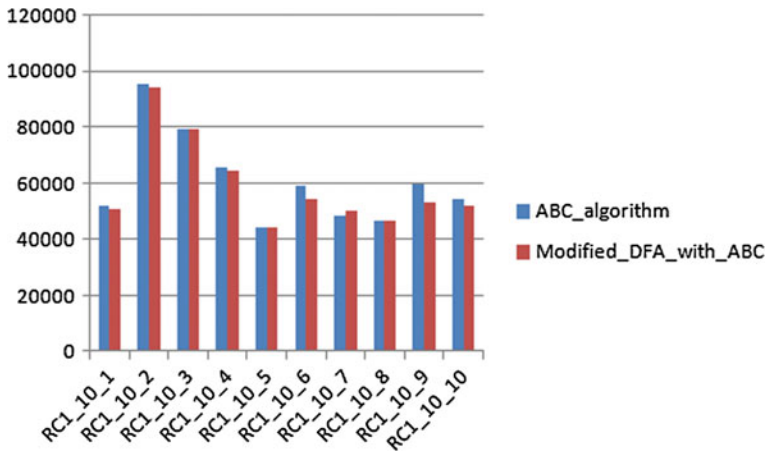


Fig. 11 Comparison of the total distance for the ABC algorithm with the Modified DFA with ABC algorithm for RC1 type problem

RC1_Type: Ten problems with count of vehicles: 200 and capacity: 250. There are 1000 customers in these problems each with service time of 10. The starting time of customers in this dataset is far less compared to the previous problems. The modified DFA with ABC algorithm gave better results for 90% problems than the ABC algorithm as indicated in Fig. 11. These instances are found to be 2.47% more efficient when minimized DFA is applied for calculating the total distance.

RC2_Type: Ten problems with count of vehicles: 1000 and capacity: 250. There are 1000 customers in these problems each with service time of 10. The starting

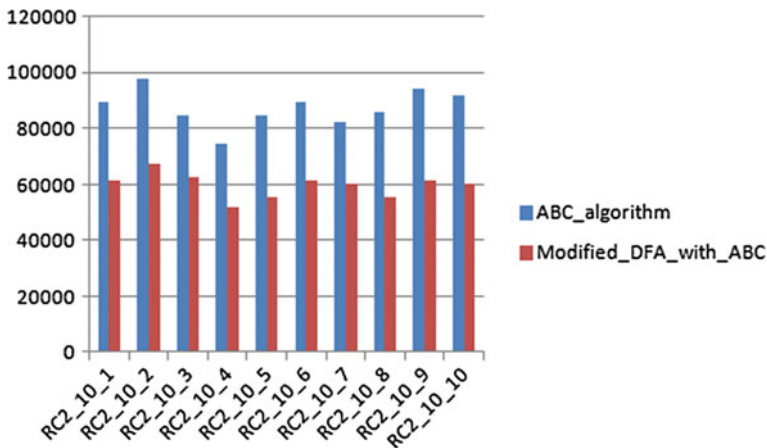


Fig. 12 Comparison of the total distance for the ABC algorithm with the Modified DFA with ABC algorithm for RC2 type problem

time of customers in this dataset is far less compared to the R1, R2, C1 and C2 type problems. The modified DFA with ABC algorithm gave better results for 100% problems than the ABC algorithm as indicated in Fig. 12. 53.97% more efficiency is achieved for the RC2 type problems when minimized DFA is applied compared to Artificial Bee Colony Optimization algorithms for finding the number of routes.

7 Conclusion and Future Work

The Modified DFA with Artificial Bee Colony algorithm has a higher performance when compared to the Artificial Bee Colony Optimization algorithm when applied to vehicular routing problem with time windows. Overall performance: 91.67% problems got better results, i.e. 25.55% more efficient in terms of the number of routes and 93.33% of the problems got better results, i.e. 15.42% more efficient for the total distance travelled by all the vehicles as a whole for the vehicular routing problem with time windows. The total time taken can also be added as a parameter along with these values. Even other algorithms such as genetic algorithms, particle swarm optimization and ant colony optimization can be optimized by applying DFA minimization along with those concepts for the VRPTW.

References

1. Toth, P., Vigo, D.: The vehicle routing problem. In: Toth, P., Vigo, D. (eds.) *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia (2001)
2. Golden, B.L., Assad, A.A., Wasil, E.A.: Routing vehicles in the real world: applications in the solid waste, beverage, food, dairy, and newspaper industries. In: Toth, P., Vigo, D. (eds.) *The Vehicle Routing Problem*, pp. 245–286. SIAM, Philadelphia (2002)
3. Karaboga, D.: An idea based honey bee swarm for numerical optimization. Technical Report-TR06, October (2005)
4. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J. Global Optim.* **39**(3), 459–471 (2007)
5. Singh, A.: An artificial bee colony algorithm for the leaf-constrained minimum spanning tree problem. *Appl. Soft Comput.* **9**(2), 625–631 (2009)
6. Umamaheswari, K., Sarathambekai, S.: Task scheduling in distributed systems using discrete particle swarm optimization. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **4**(2):510–522 (2014/2)
7. Sarathambekai, S., Umamaheswari, K.: Task scheduling in distributed systems using heap intelligent discrete particle swarm optimization. *J. Comput. Intell.* **33**(4):737–770 (2017/11)
8. Bonabeau, E., Dorigo, M., Theraulaz, G.: *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, New York (1999)
9. Dantzig, G.B., Ramser, J.H.: The truck dispatching problem. *Manag. Sci.* **6**(1), 80–91 (1959)
10. Clarke, G., Wright, J.W.: Scheduling of vehicles from a central depot to a number of delivery points. *J. Oper. Res. Soc. Am.* **12**(4), 568–581 (1964)
11. Desrochers, M., Desrosiers, J., Solomon, M.: A new optimization algorithm for the vehicle routing problem with time windows. *J. Oper. Res. Soc. Am.* **40**(2), 342–354 (1992). <https://doi.org/10.1287/opre.40.2.342>

12. Bhagade, A.S., Puranik, P.V.: Artificial bee colony (abc) algorithm for vehicle routing optimization problem. *Int. J. Soft Comput. Eng. (IJSCE)* **2**(2), 2231–2307 (2012)
13. <https://www.sintef.no/projectweb/top/vrptw/solomon-benchmark/>
14. Watson, B.W.: A taxonomy of finite automata minimization algorithms. Computing Science Report 93/44 Eindhoven, January 1995 (1995)
15. Perrin, D.: CHAPTER 1—Finite Automata Formal Models and Semantics, *Handbook of Theoretical Computer Science*, pp. 1–57. Elsevier, Amsterdam (1990)
16. Bulhões, S.R., Uchoa, E.: A branch-and-price algorithm for the minimum latency problem. *Comput. Oper. Res.* **93**, 66–78 (2017)
17. Schneider, M.: The vehicle-routing problem with time windows and driver-specific times. *Eur. J. Oper. Res.* **250**(1), 101–119 (2016). <https://doi.org/10.1016/j.ejor.2015.09.015>
18. Huang, G., Cai, Y., Cai, H.: Multi-agent ant colony optimization for vehicle routing problem with soft time windows and road conditions. In: MATEC Web Conference, Volume 173, 2018 International Conference on Smart Materials, Intelligent Manufacturing and Automation [SMIMA] (2018). <https://doi.org/10.1051/mateconf/201817302020>
19. Matthopoulos, P.-P., Sofianopoulou, S.: A firefly algorithm for the heterogeneous fixed fleet VRP. *Int. J. Ind. Syst. Eng.* **33**(1) (2018)
20. Shi, Y.-J., Meng, F.-W., Shen, G.-J.: A modified artificial bee colony algorithm for vehicle routing problems with time windows. *Inf. Technol. J.* **11**(10), 1490–1495 (2012)
21. Nahum, O.E., Hadas, Y., Spiegel, U.: Multi-objective vehicle routing problems with time windows: a vector evaluated artificial bee colony approach. *Int. J. Comput. Inf. Technol.* **3**(1), 41–47 (2014)
22. Alzaqebah, M., Abdullah, S., Jawarneh, S.: Modified artificial bee colony for the vehicle routing problems with time windows. *SpringerPlus* **5**(1), 1298 (2016)

Coverage-Aware Recharge Scheduling Scheme for Wireless Charging Vehicles in the Wireless Rechargeable Sensor Networks



Govind P. Gupta and Vrajesh Kumar Chawra

Abstract Recent advancement in the wireless power transfer technology has motivated the development of a wireless rechargeable sensor network (WRSN). In WRSNs, the formation of an optimal recharging schedule for each wireless charger vehicle is a well known NP-complete problem. To determine the optimal recharging schedule for each wireless charger vehicle, this paper presents a coverage-aware recharge scheduling scheme (CRS) where ACO-based metaheuristic algorithm is employed. In order to provide fast recharging in WRSN, the proposed scheme employs multiple wireless charger vehicles to perform the charging task. Performance analysis of the proposed scheme confirms its superiority in terms of charging latency.

Keywords Recharging scheduling · Wireless rechargeable sensor networks · ACO-based metaheuristic

1 Introduction

In WRSNs, a set of rechargeable sensor nodes are deployed over an area of interest where each sensor node carry a wireless rechargeable battery and one or more wireless charging vehicle (WCV) are deployed for efficient recharging task [1, 2]. Recharging of the sensor nodes before their battery power vanishes by using multiple WCVs is a very critical task. Hence, optimal scheduling of the WCV is a fundamental research issue in WRSNs [1–5].

In WRSNs, the formation of an optimal recharging schedule for each mobile recharger is a well known NP-complete problem [2–7]. There are many solutions

G. P. Gupta (✉) · V. K. Chawra
Department of Information Technology, National Institute of Technology,
Raipur, C.G. 492010, India
e-mail: gpgupta.it@nitrr.ac.in

V. K. Chawra
e-mail: vkchawra.phd2017.it@nitrr.ac.in

proposed in the literature for designing the recharge schedule for *WCV*. Most of the solutions consider only single *WCV* for completing the recharging task. However, these techniques suffer from charging latency problem [3–6].

In order to enhance charging efficiency and reducing the charging latency, this paper proposes a coverage-aware recharge scheduling (*CRS*) scheme. In the proposed scheme, *ACO*-based metaheuristic algorithm is used for deriving the optimal recharge schedule for each *WCV*. In *CRS*, multiple *WCVs* are used for delay efficient recharging of the deployed sensor nodes. The proposed scheme contains mainly two phases: task assignment process for the *WCV* and derivation of the optimal recharge schedule for each *WCV*. To ensure the coverage of *WCVs*, proposed scheme divides the monitoring area into a set of logical grids and a subset of grids are allocated to each *WCV* for power transfer task. In order to schedule the traversal path of each *WCV*, an *ACO*-based metaheuristic algorithm is used for deriving the recharge schedule for each *WCVs*.

The remaining parts of the paper are structured as follows. A brief overview of the related work on recharge scheduling methods for the *WCV* is described in Sect. 2. Section 3 presents the network model and various assumptions used in this work. Section 4 discussed the proposed coverage-aware recharge scheduling scheme for wireless charging vehicles. In Sect. 5, simulation result analysis and a detailed performance comparison of the proposed scheme with the existing scheme are discussed. Finally in Sect. 6, the paper is concluded.

2 Related Work

In the literature, several research works have been proposed for deriving the recharge schedule for wireless charging vehicles. In [3], the authors have described a mobile charging method where the first clustering algorithm is employed to cluster the nodes based on their energy consumption rate. After the formation of the cluster, a set of nested *TSP* tours are formed for traversing the nodes that require energy. In [4], authors have presented two different charging algorithms, keeping an objective to minimize charging time as well as travel distance.

In [5], Wang et al. have discussed a node deployment scheme where a data gathering vehicle and multiple charging vehicles are employed to balance the energy resources and maximize the network lifetime of the network. In this scheme, only a single mobile charger is used for recharging task. The authors in [6] have presented a multimode wireless charging method where the mobile charger can charge multiple nodes at a time. In this scheme, the network area is partitioned into hexagonal cells. In this scheme, it is assumed that charging point is always at the center of the cell.

In [7], the authors have considered charging problem as two subproblems such as tour formation and assignment problem. In this scheme, a greedy technique-based heuristic algorithm is proposed for solving these problems. In [8], Jia et al. have discussed an integrated solution for charging and routing problem in

WRSNs. In this scheme, a heuristic algorithm is employed to derive optimal charging tour using a predetermined routing tree. This scheme employed a *GA*-based metaheuristic optimizing technique for optimizing charging and the routing tasks. This scheme considered only a single mobile vehicle for charging and data gathering task. Thus it suffers from charging latency problem.

The authors in [9] have proposed a collaborative scheme for solving the recharging problem in *WRSNs*. In this scheme, the network area is divided into a set of concentric circles and authors have tried to optimize the number of charging vehicles required for completing the charging task. In the literature, there are many solutions proposed by different researchers to solve the recharge schedule problem in *WRSNs*. Most of the scheme only utilized a single mobile charger vehicle for completing the recharging task. Thus, suffers from charging latency problem. In this paper, multiple *WCVs* are employed for reducing the charging latency and enhancing total charging efficiency. In addition, the proposed work also considers the coverage issue of recharging so that all required sensor nodes are recharged by the dispatched *WCVs*.

3 Network Model and Assumptions

These research works consider a wireless rechargeable sensor network where network area is logically divided into a set of equal size grids. Each grid can contain a set of sensor nodes and one of the them is selected as cluster head node which collects the sensed data from its cluster members and handovers to mobile sink during its data collection process [11–13]. In this work, we assume that all sensor nodes are homogeneous and have equal sensing and communication range [13–16]. Figure 1 illustrates the network model used in this paper.

4 Coverage-Aware Recharge Scheduling Scheme for Wireless Charging Vehicles

This section presents a detail explanation of the working of the proposed coverage-aware recharge scheduling scheme (*CRS*). In the proposed scheme, *ACO*-based metaheuristic algorithm is used for deriving the optimal recharge schedule for each wireless charging vehicle (*WCV*). In *CRS*, multiple *WCVs* are used for delay efficient recharging of the deployed sensor nodes. The proposed scheme contains mainly two phases: task assignment process for the *WCV* and derivation of the optimal recharge schedule for each *WCV*.

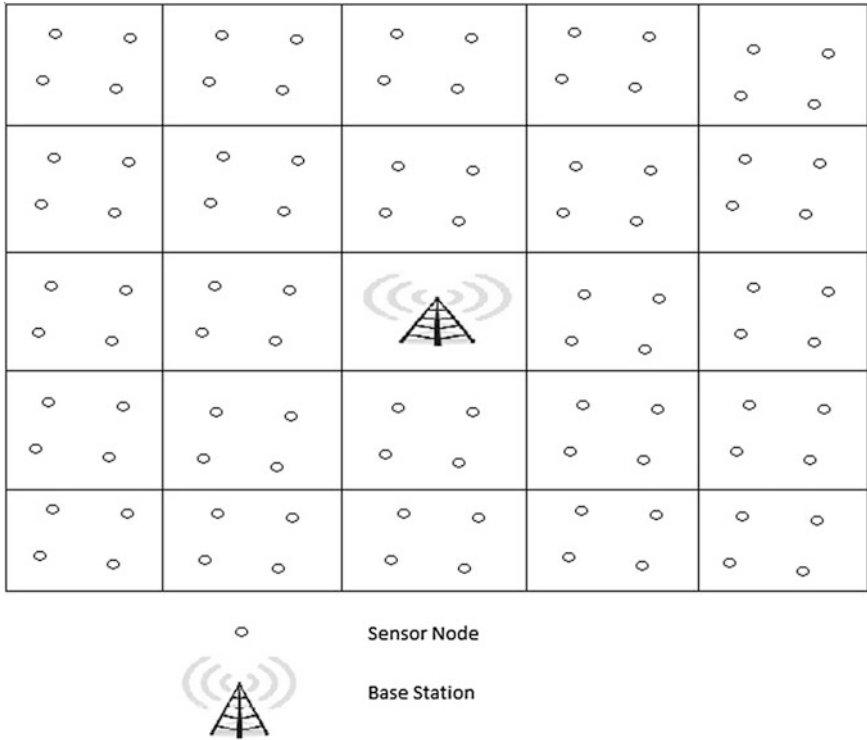


Fig. 1 Network model used in the proposed scheme

4.1 Task Assignment Process for the WCV

In order to assign the recharging task to each involved WCV, available area of interest is logically divided into equal size grids and a subset of grids is assigned to each WCV for traversing and recharging the sensor nodes located in that grid. In the experiments, the proposed scheme has taken 25×25 dimensions for each grid. In order to ensure coverage of recharging task by the WCVs, the monitoring area is divided into logical grids. For scheduling the traversal path for each WCV, an ACO-based metaheuristic algorithm is used which is described in Sect. 4.2.

4.2 Derivation of the optimal recharge schedule for each WCV

For the derivation of the optimal recharge schedule for each deployed WCV, the proposed scheme employed an ACO-based metaheuristic algorithm. The working of the ACO-based optimal recharge scheduling algorithm is described as follows:

- (i) **Initialization** In this phase, the proposed scheme selects a set of location points where WCV stay for a fixed charging time to wirelessly transmit power to recharge all sensor nodes located in the grid. Since each WCV stays at a particular point of each grid, a population set is randomly generated by selecting a random location point from each grid. Let a WCV needs to visit a set of M grid so a set of M stay points are randomly selected and need to devise a optimal visiting schedule for traversing these stay point such that total traveling cost of WCV will be minimum. Each WCV is dispatched from the base station to visit each grid and recharge all sensor nodes located in each grid. Base station (BS) and a set of stay points (S_i) form a connected graph G where source node BS connects all destination points S_i . Each link between (BS, S_i) is initialized with a variable called pheromone trail ($\tau_{ij} = \tau_0$). The value of the pheromone trail can be read and updated by ants.
- (ii) **Selection of visiting schedule for optimal recharging** In this phase, probability of each link (i, j) of the graph is calculated by using the following formula as given in [17, 18]. Here, N_i is set of neighboring stay points of the points i .

$$p_{ij} = \begin{cases} \frac{1}{d_{ij}} \times \frac{\tau_{ij}}{\sum_{j \in N_i} \tau_{ij}} & \text{if } j \in N_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

After calculation of probability (p_{ij}) of each link, a sequence of stay points is selected based on their computed probability value. This sequence of stay points considers as a recharge schedule which starts from the base station and ends at the base station. After getting the recharge schedule, calculate its cost (C_s) by calculating total traveling distance by WCV.

- (iii) **Pheromone Update** In this phase, after calculation of cost of the derived schedule, the value of pheromone trail (τ_{ij}) for each link is updated using the following formula which is given in [17, 18].

$$\tau_{ij}(t+1) = \rho \times \tau_{ij}(t) + \frac{q}{C_s(t)} \quad (2)$$

Here ρ is the evaporation rate whose value is between 0 and 1. Q is a constant and $C_s(t)$ is the cost of the derived schedule in t iteration. After updating the value of pheromone trail, Step (ii) and (iii) are repeated until we get an optimal cost schedule for WCV.

5 Result Analysis

This section presents the result analysis of the proposed scheme and compares its performance with the single *WCV-based* scheme. Implementation of the proposed scheme and the existing scheme are done using *MATLAB R2014*. In the experiments, this paper considers four *WCVs* for recharging tasks. Figure 2 illustrates traversing of each *WCV* for recharging the sensor nodes of each grid.

Figure 3 illustrates the result analysis of the proposed scheme (*CRS*) and compares its performance with the single *WCV-based* scheme in terms of charging latency by varying the speed of the *WCV*. It can be depicted from Fig. 3 that charging latency for the proposed scheme is significantly much lower than the existing scheme. This is due to the use of multiple *WCV* and derivation of optimal recharging schedule in the proposed scheme.

Figure 4 illustrates the result analysis of the proposed scheme (*CRS*) and its comparison with the single *WCV-based* scheme in terms of charging latency by varying the grid size from (10×10) to (60×60) within the network. It can be observed from Fig. 4 that charging latency for the proposed scheme is significantly much lower than the existing scheme. This is due to the use of multiple *WCV* and derivation of optimal recharging schedule in the proposed scheme.

Figure 5 illustrates the result analysis of the proposed scheme (*CRS*) and its comparison with the single *WCV-based* scheme in terms of charging latency by varying the number of stay points from 9 to 400 within the network. It can be

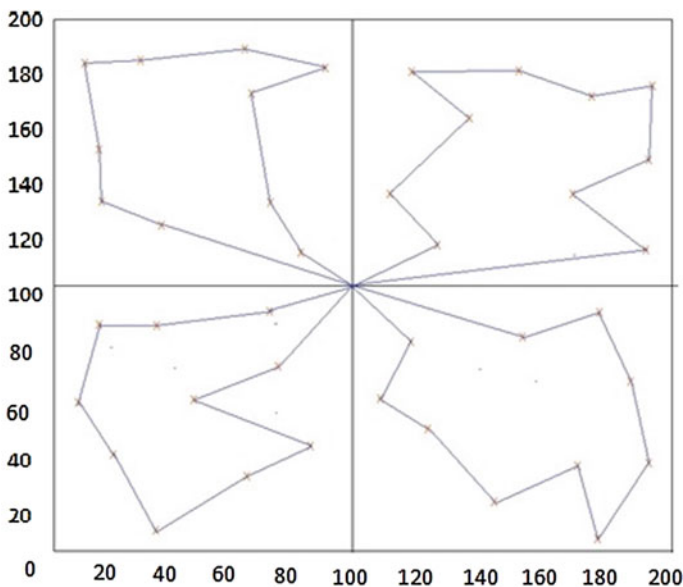


Fig. 2 Illustration of the traversal path of each *WCV* for recharging tasks

Fig. 3 Speed versus charging latency

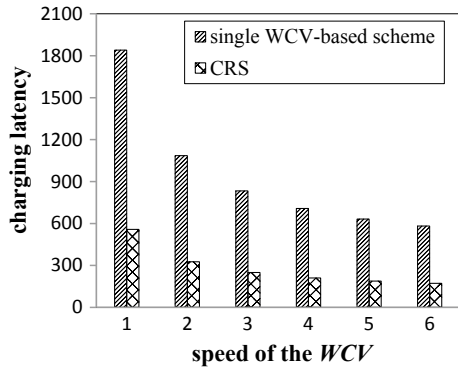
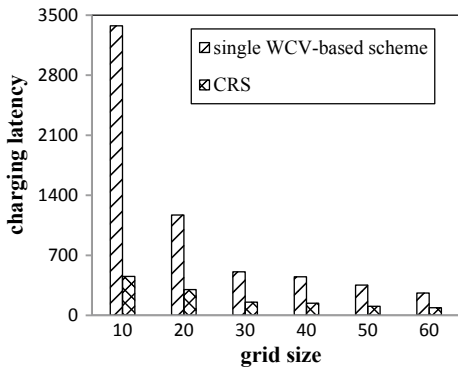


Fig. 4 Grid size versus charging latency



observed from Fig. 5 that as the number of stay points increases, the charging latency also increases. Charging latency for the proposed scheme is significantly much lower than the existing scheme. This is due to the use of multiple WCV and derivation of optimal recharging schedule in the proposed scheme.

Fig. 5 Number of stay points versus charging latency

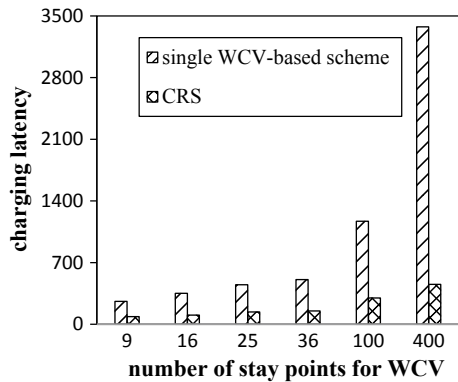


Fig. 6 Network area versus charging latency

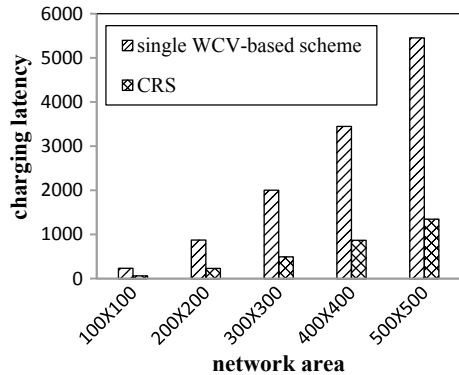


Figure 6 illustrates the result analysis of the proposed scheme (*CRS*) and its comparison with the single *WCV*-based scheme in terms of charging latency by varying the size of the network area (100×100) to (500×500). It can be observed from Fig. 6 that as the size of the network area increases from (100×100) to (500×500), charging latency also increases. This is due to the fact that the total number of grids visited by a *WCV* also increases. Charging latency for the proposed scheme is significantly much lower than the existing scheme. This is due to the use of multiple *WCV* and derivation of optimal recharging schedule in the proposed scheme.

6 Conclusion

This paper proposed a coverage-aware recharge scheduling scheme for wireless charging vehicle where ACO-based metaheuristic algorithm is employed to determine the optimal schedule for each wireless charging vehicle. In order to reduce charging latency, the proposed scheme employs multiple charging vehicles that are dispatched by the base station concurrently. Simulation results and its analysis confirm that the proposed scheme performs significantly better than the existing scheme in terms of charging latency.

References

1. He, S., Chen, J., Jiang, F., Yau, D.K.Y., Xing, G., Sun, Y.: Energy provisioning in wireless rechargeable sensor networks. *IEEE Trans. Mob. Comput.* **12**, 1931–1942 (2013)
2. Liao, J.-H., Hong, Jiang, J.-R.: An adaptive algorithm for charger deployment optimization in wireless rechargeable sensor networks. In: *ICS* (2014)

3. Fu, L.K., He, L., Cheng, P., Gu, Y., Pan, J.P., Chen, J.M.: ESynC: energy synchronized mobile charging in rechargeable wireless sensor networks. *IEEE Trans. Veh. Technol.* **65**(9), 7415–7431 (2016)
4. Lin, C., Wu, G.W., Obaidat, M.S., Yu, C.W.: Clustering and splitting charging algorithms for large scaled wireless rechargeable sensor networks. *J. Syst. Softw.* **113**(March), 381–394 (2016)
5. Wang, C., Li, J., Ye, F., Yang, Y.: A mobile data gathering framework for wireless rechargeable sensor networks with vehicle movement costs and capacity constraints. *IEEE Trans. Mob. Comput.* **65**(8), 2411–2427 (2016)
6. Xie, L., et al.: Multi-node wireless energy charging in sensor networks. *IEEE/ACM Trans. Netw.* **23**(2), 437–450 (2015)
7. Hu, C., Wang, Y.: Minimizing the number of mobile chargers to keep large-scale WRSNs working perpetually. *Int. J. Distrib. Sens. Netw.* **2015**, 1–16 (2015)
8. Jia, J., Chan, J., Deng, Y., Wang, X., Aghvami, A.H.: Joint power charging and routing in wireless rechargeable sensor networks. *Sensors* **17**(10), 2290 (2017)
9. Chen, Z., Chen, X., Zhang, D., Zeng, F.: Collaborative mobile charging policy for perpetual operation in large-scale wireless rechargeable sensor networks. *Neurocomputing* **270**(2017), 137–144 (2017)
10. Gupta, G.P.: Efficient coverage and connectivity aware data gathering protocol for wireless sensor networks. In: *Proceedings of the 3rd IEEE International Conference on Recent Advances in Information Technology (RAIT-2016)*, pp. 50–55 (2016)
11. Angelopoulos, C.M., Nikolettseas, S., Raptis, T.P.: Wireless energy transfer in sensor networks with adaptive, limited knowledge protocols. *Comput. Netw.* **70**, 113–141 (2014)
12. Gupta, G.P., Jha, S.: Integrated clustering and routing protocol for wireless sensor networks using Cuckoo and Harmony Search based metaheuristic techniques. *Eng. Appl. Artif. Intell.* **68**, 101–109 (2018)
13. Gupta, G.P.: Improved Cuckoo Search-based clustering protocol for wireless sensor networks. *Procedia Comput. Sci.* **125**, 234–240 (2018)
14. Gupta, G.P., Misra, M., Garg, K.: Towards scalable and load-balanced mobile agents-based data aggregation for wireless sensor networks. *Comput. Electr. Eng.* **64**, 262–276 (2017)
15. Kumari, S., Gupta, G.P.: Differential evolution-based sensor allocation for target tracking application in sensor-cloud. In: Bhateja, V., Coello Coello, C., Satapathy, S., Pattnaik, P. (eds.) *Intelligent Engineering Informatics. Advances in Intelligent Systems and Computing*, vol. 695. Springer, Singapore (2018)
16. Gupta, G.P., Misra, M., Garg, K.: Energy and trust aware mobile agent migration protocol for data aggregation in wireless sensor networks. *J. Netw. Comput. Appl.* **41**, 300–311 (2014)
17. Dorigo, M., Blum, C.: Ant colony optimization theory: a survey. *Theor. Comput. Sci.* **344**(2–3), 243–278 (2005)
18. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Trans. Evol. Comput.* **1**(1), 53–66 (1997)

A Transition Model from Web of Things to Speech of Intelligent Things in a Smart Education System



Ambrose A. Azeta, Victor I. Azeta, Sanjay Misra and M. Ananya

Abstract Several terms have been used to describe Internet of Things; Web of Things (WoT) is a term which can be used interchangeability and it is referred to as the capability of devices to interconnect to the World Wide Web and sharing the information and data to one another. WoT has been mentioned in the literature to improve interconnection between devices at all times. In WoT, two different modes of communication which are generally mentioned in previous studies include person-to-thing (or thing-to-person) and thing-to-thing. This paper presents an architecture for transiting from WoT to speech-enabled WoT known as Speech of Intelligent Things (SoIT). The system employs a combination of technologies such as system design, server-side scripting, speech-based system tools, and data management in developing the SoIT prototype system as a third mode of communication. This paper illustrates a scenario whereby remote monitoring and controlling of WoT devices within the university campus might be difficult to manage by only using the modes discussed in the literature. An evolution of WoT to SoIT was realized using speech technology to provide a prototype system. Technical implications involve using a telephone by connecting an object telephone number (OTN) and dial WoT objects and establish a control mechanism. The research limitation is mainly the cost of dialing an OTN number. The contribution of this paper is to favor and encourage the use of speech technology to enhance the convenience of communication between WoT devices within the school campus.

A. A. Azeta · S. Misra (✉)
Covenant University, Ota, Nigeria
e-mail: sanjay.misra@covenantuniversity.edu.ng

A. A. Azeta
e-mail: ambrose.azeta@covenantuniversity.edu.ng

V. I. Azeta
National Productivity Center, Kaduna, Nigeria
e-mail: victaazeta@gmail.com

M. Ananya
Technical University of Munich, Munich, Germany
e-mail: ge25daj@mytum.de

Keywords e-Campus · OTN · Speech interface · SoIT · WoT

1 Introduction

Among the web technologies that are presently gaining attention in the research community is the Web of Things (WoT). Web of Things transforms the connection between millions of devices and smart devices and interactions with humans and one another. In WoT, sending and receiving of information exists between cloud and one another, with the possible aim of collection and analysis of huge amount of data accurately [1] suggesting that distinct approaches exist across a variety of technologies such as vehicles which are able to sense fatigue of the user and execute maintenance which is self-scheduled and provide dynamic analysis, execute, and inform the forecasted time of arrival of waiting passengers.

The literature has failed to discover that the concepts of WoT have been applied to e-Learning and school environment to enhance teaching and learning results thereby providing greater achievements than in other fields. In other words, educational institutions are a good place in which WoT can be researched [2].

There are two different modes of the interface in WoT usually discussed in the studies. Thing-to-person or person-to-thing in which there is an interaction between people and things and thing-to-thing in which day-to-day objects communicate with each other. In this research, it is envisaged that control and monitoring of WoT objects remotely might be burdensome with the use of two modes only. Some uncertainties do exist on how the WoT will be controlled and monitored. Some years ago, Internet Protocol Version 6 (IPv6) was developed to facilitate a high number of systems with a distinct IP address that aided the transmission and receiving of information [3]. The entire world seems to approach a global village where the exchange and sensing of data are done through the connection to different types of networks (private and public). Data is collected, processed, and used from time to time from the connected device to provide more information about the devices. This process provides intelligent data for control, plan, management, and decision-making.

A method has been proposed on how to control the WoT objects involving mobile telephony. This method involves dialing the Object Telephone Number (OTN) through the use of a mobile device and connection to any of the WoT objects, which will begin this mechanism. In this environment, different SoIT devices are assigned a value for OTN which is used instead of an IP address for communication among the devices. The communication occurs when an OTN value is dialed via a mobile device. A lot of objects and entities such as traditional office systems, educational programmes, vehicles, mobile devices, and students and staff residents can be replaced by an SoIT-based campus. When universities are compared to this system, universities consist of, accommodations, clinics, athletic facilities, health institutions, hospitals, offices, parking areas, classrooms, restaurants, and libraries [4].

With Speech of Internet Things (SoIT), wearable computer can be used by students to explore the points of the university transportation which can be done by dialing an OTN number, attend their classes online, perform online bank transactions, be up-to-date with the library transactions, in other words, check the availability of a book, be aware of the prices of food items in canteens, and order items in the storage. In an educational environment, a large portion of information needs is shared by students and staffs which include information regarding classrooms, schedules, locations, seminars, lab equipment, student residences, assignments, presentations, sports events, etc. This information is highly relevant and related to people and objects found in the institutional environment. When wireless technology such as WLAN (Wireless Local Area Network) is implemented in conjunction with SoIT platform, technologies, and smart devices, it helps to detect devices or places, thereby making this feasible for fulfilling information needed by the individuals in a typical learning atmosphere in an efficient and effective way [5].

A discussion is made regarding the provision of accommodation for students and staffs of the higher institution in Nigeria whereas higher educational institutions which can provide this facility have not yet implemented the required technology in various other units such as transportation, storage, commercial institutions, cafeterias, classrooms, hospitals and clinics, dormitories, library, etc. In other words, manual processing of transactions is still done in many units. Supervision of a human is needed to control objects such as electrical appliances used in homes and offices. Devices have no interconnection and remote manipulation between them. Mostly, learning takes place 40% online and 60% involves the presence of an individual and this illustrates that ICT and the Internet are not being used with their full capabilities for administration, management, learning, and teaching.

Institutions which have been able to provide adequate accommodation for the staff and students still have the problem of providing exact locations of students to the facility as well as the visiting parents. There are several students' engagement activities on campus. Students can be found in places within the campus such as the library for the borrowing of books, storage, cafeterias, banks, and lecture halls for lectures. They sometimes tend to go home for personal reasons. The availability of transportation services for students and staffs not fully on ground. The wide use of SoIT devices should be encouraged for different systems of the campus to work together for successful management and easy access for everyone by an electronic device (remote control) and manipulation of OTN devices. A framework of SoIT based on the above is proposed in [27]. In this paper, we are extending the SoIT framework by developing a real system with a prototype and implementing it on a system.

This paper has six sections. In Sect. 2, the related work is provided. Section three presents the SoIT framework. The fourth section discusses the potential advantages and disadvantages of using WoT/SoIT. In section five, the SoIT prototype and evaluation is presented while section six concludes the paper.

2 Related Work

A research was carried out in [6] which gives a summary of the difficulties faced in the areas such as reliability, security, mobility, and connectivity of WoT with the use of IPv6 to achieve Internet of Everything (IoE). The main challenges and solutions of IPv6 were also addressed. Some of the future works which discuss the emergence of IoE to make applications mobile, distributed, secure, and powerful. An example of such application is smarte cities which were also highlighted. The study in [7] discussed the necessity of business organizations to develop WoT and show the power of digitalization to entities for the creation of more business opportunities. The design and evaluation of the model are presented that will give researchers and industrialists to record, see, and analyze the current and future commercial areas in WoT.

An attempt is made by studies to give an account of the architecture of WoT. The research that was carried out in [8] introduces key enablers of WoT and gives details of the main components of WoT such as the concept of machine-to-machine (M2M), portability and mobility of devices, protocols for communication that are appropriate for the WoT environment. It also discussed the difficulties faced in WoT. A prototype of navigation-based solutions for WoT and an architectural framework of WoT were also mentioned. In [9], an architecture and cloud-centric vision are given for WoT implementation. This study presented major technological development and applications which will make research in WoT feasible in the future. The authors in [10] showed the perception of the future by Rambus. As the number of devices is increasing dramatically and connection to the Internet is possible, the world is deviating from a PC-based model and going toward inter-connected systems.

3 The SoIT Architecture

The SoIT system signifies a vision by which the Internet interconnects on a daily basis using OTN. Physical entities are now connected to the virtual world which can serve as access points to Internet service and can be managed from a distance by phone calls. SoIT is a ubiquitous computing platform with a lot of possibilities for the advancement in society. Despite the numerous benefits, there are a lot of technical challenges and risks. SoIT can only continue to grow if the technological advances in the information and communication area are seen in the upcoming years. Due to the creation of smaller versions of technological devices, reduction in energy consumption and price devices can easily be adapted to daily lives. Smart devices play an important role in SoIT because information and communication technology are causing radical changes in the world [11].

According to authors of [12], smart campus comprises other smart components such as smart technology, lecturers, students, and environment. The framework by Abuelyaman needs a smart campus to contain a smart cafe, smart storage, smart residential areas, smart commercial banks, smart lecture rooms, smart hospitals and clinics, smart transportation, smart devices such as smartphones, tablets, laptops, etc. In SoIT, sensors interact with one another automatically and provide real-time response rather than interacting with humans only. Local area networks (LAN) will be used to mitigate physical devices. In WoT, each device will possess a unique IP address in IPv6 format only whereas in SoIT in addition to the IP address, the device will also have the OTN address.

The phone of the user would keep track of the physical and mental requirements of states through the interconnection of the components which contains the main system. Similarly, the complete set used at the higher institution may make efficient use of resources by sharing and distributing the resources with other components [13]. For example, during the arrival and departure from home, an automation system can control the front porch light according to the needs of the user [14]. Figure 1 shows the SoIT framework for a smart campus. Objects have been placed in a circular manner from 1 to N. Every object has a sensor attached and all of them (sensors) are interconnected to each other by wireless sensor network (WSN). This will aim to detect the transmission signals from the neighboring sensors which form the network. WSN which works for information and communication is also called revolutionary collection method. These sensors enhance the effectiveness and improve the efficiency of the architecture for a smart campus [15].

In SoIT, all the objects have an OTN value on every node in the network. Message transmission between users can be connected simply by calling an OTN for a particular system, for example, staff can call the OTN of a commercial bank to

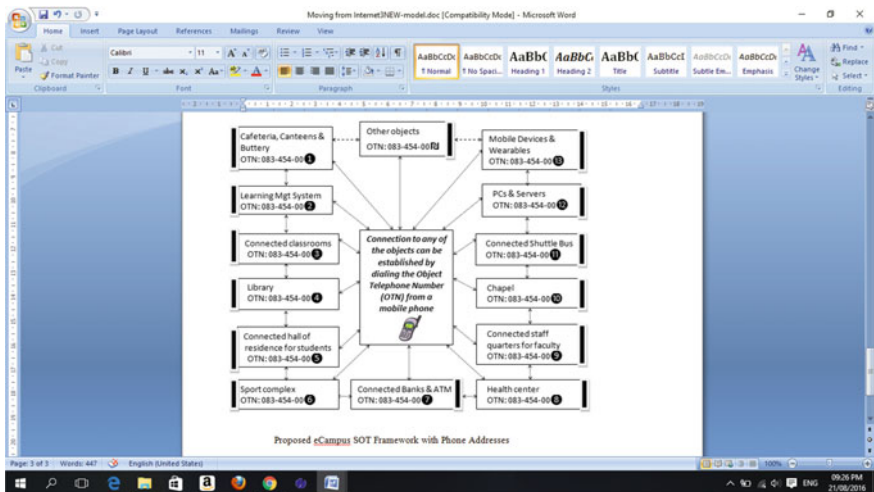


Fig. 1 The SoIT framework for smart campus

perform certain transactions such as checking the bank balance of the account. The student may want to check resources in the library or lecturers might want to connect with the students for giving lectures, answer questions, or make discussions on a particular topic. All these activities can be made possible when objects are interconnected to each other using the SoIT framework.

4 SoIT Prototype and Evaluation

In order to understand the SoIT concept, a simple prototype speech-based application was developed and deployed using VoiceXML [21] for speech interface, PHP for server side, and MySQL for the database. A database with a field named power status was created and assigned the value 1 to put off the computer system, and 0 to restart the computer system. This system allows a user to dial a phone number as an OTN address that is mapped to an object. The system responds through speech interface by saying say 1 to put off or 0 to restart the computer system. Once a 1 is selected, the computer will shut down. The web interface uses an auto-refresh command that constantly checks for the value in the database field of the power status button. The web interface of the system is shown in Fig. 2.

When OTN is dialed from a mobile phone, the system will say “Welcome to shutdown utility”, and request for authentication information from the user. After successful authentication, the user responds to the system request by saying 1 or 0. A 1 will shut down the system, and a 0 will restart the system. In experimenting

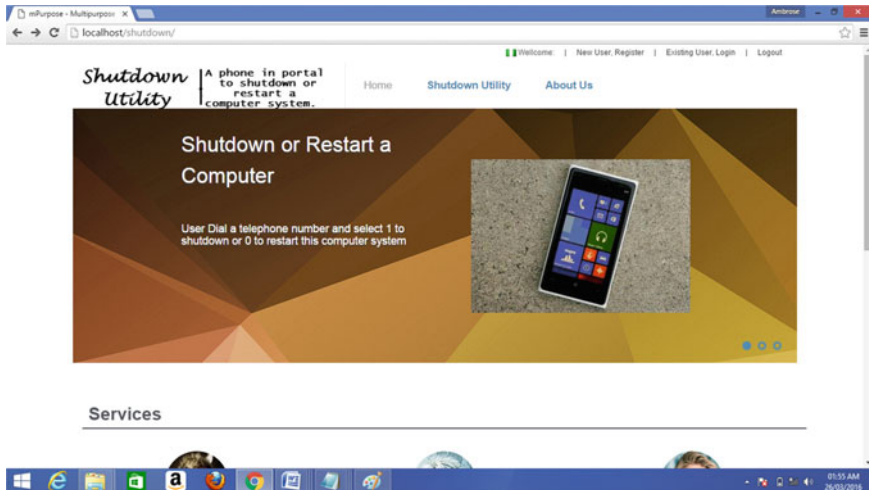


Fig. 2 The web interface. Source The researchers

with SoIT with respect to architectural considerations, hardware, software, and deployment requirements, the design and implementation testbed in [22] may be considered.

Typical Scenario

The call sequence in Fig. 3 exemplifies the interaction of the user with the system.

A prototype for an application called VoiceXML (speech user interface) is developed for the web application on a Voxeo voice server and can be retrieved from a mobile phone and as well as in fixed phone-line connections by adopting the following structure: <source country int. dial out #><destination country code><destination area code><generated voice network 7 digit #> [23].

A cognitive method called Cognitive Walkthrough Strategy was developed by Rieman and Redmiles [24]. Cognitive Walkthrough Strategy [24] makes use of one or a group of evaluators which are responsible for checking the user interface by following a certain number of tasks, making learning easy, and accessing the system's understandability.

A questionnaire was developed and used as a survey instrument. Twenty questionnaires were distributed to the users and we got responses from 16. The received responses were analyzed and reported, and the results were presented. We use a five-point Likert scale for our study and analysis (1 = strongly disagree and 5 = strongly agree). The findings show good usability of 4 as the mean rating on a 1–5 scale.

User : Dial OTN address 083-454-0012 to access a computer
System: Wellcome to Shutdown Utility.
 What is your userid?
User : My userid is kane.
System: What is your pin?
User : My pin is 4109
(System search the database to authenticate the userid and pin)
System: Say 1 to Shutdown your PC, or
 Say 0 to Restart your PC.
User : 1
(System process the input value 1 and initiates a shutdown process)
System: Please save your work and shutdown, the system will go off
 in 1 minute!!

Fig. 3 Sample call flow to shutdown and restart a computer. *Source* The researchers

5 Advantages and Disadvantages of Using SoIT and WoT

WoT and SoIT have numerous unlimited benefits. WoT systems can make our lives better by conservation of resources, the creation of new opportunities, innovation, and improvement of knowledge [16]. In [17], the authors have explained that WoT is developed to improve connections among devices by using WSN everywhere at a given time. WoT is presently being used positively by some business organizations. Users of business organizations can utilize the services of WoT for data analytics to generate more revenue, monitor the activities and assets, modify new business models, and develop an operational performance for innovation [18].

However, among several obstacles and hindrances of WoT security is one of the areas that prevents WoT from gaining its actual potential. The connection of hundreds of billions of devices gives a serious concern regarding the issue of security. Business owners and stakeholders need to take this issue into consideration for WoT to develop and overcome these obstacles [19]. Other major discoveries in [20] regarding WoT are given below.

- The connection between Secure Socket Layer (SSL) connections with cloud in 19% of mobile applications which are used for the control of WoT devices
- No provision of authentication between client and server in the devices
- Some devices lacked strong password and enforcement strategies
- Some WoT cloud interfaces were not compatible with two-factor authentication (2FA)
- Lack of lock-out or delaying measures against brute-force attacks in some WoT services which hindered the protection of user accounts
- Some devices were not protected from account harvesting mechanism
- Many WoT cloud platforms such as web applications were exposed
- No provision of encrypted firmware updates by majorly all WoT services
- Shortage of IP addresses.

SoIT and WoT and Smart Campus

The basic idea of IoT is the pervasive presence around us of a variety of things or objects connected via enabling technologies such as Radio-Frequency Identification (RFID) tags, sensors, actuators, and mobile phones, which through unique addressing schemes, can interact with each other and cooperate with their neighbors to reach common goals [33]. WoT includes the interconnection of devices and smart campus includes different smaller systems to interact with other systems to make the bigger system work.

Spoken dialog system has become very popular in the industry with the advent of continuous speaker-independent speech recognition technology over the last few years [28–30]. Speech technology is now combined with human spoken language capabilities to boost learning in students and to make communication between human and computers easier [31, 32]. This incorporation between spoken language and speech technology will make it possible for machines to take over the trivial tasks performed by humans. Intelligent things, on the other hand, will be able to

perform communication among humans and devices and choose the appropriate decision for any situation. Since smart campus consists of smart devices, addition of speech technology will make the device smarter. Current research claims that SoIT and WoT are being seen to work together for a better communication environment.

6 Conclusion

In this paper, a review of WoT and the ability to develop SoIT in a school environment was presented. The office or home appliances, and SoIT devices may be developed which will respond by calling on an OTN address when called by any telephone, cell phone, or mobile. This study has contributed to the area of transiting from WoT to SoIT. WoT is an emerging area of research and scientists are working worldwide by utilizing speech for monitoring systems in the WoT environment which will sharpen the technology and realize the expectation of a smart academic institution system.

With SoIT, students are able to dial an OTN address of a cafeteria and order for food, dial an OTN address of a classroom and receive lecture remotely while in the dormitory. Lecturers and students are able to dial the OTN address of any bank within the school premises and check balance, and also transfer money online. Campus shuttle bus service can be called by dialing the corresponding OTN number. The university library service and health centre can be linked by dialing the corresponding OTN address to ascertain the availability of a certain book and pharmaceutical drugs, respectively. In the case of home appliances, a user may forget to power off a television, fridge, or electric bulbs or any electrical home appliances when leaving home for work; all he needs do is dial the OTN address of the object and say shutdown to power off the system.

Dramatically, changes in information and communication technology also affected the style of education and the choice of students. Thus, a smart campus can be realized by making all the facilities in its jurisdiction smart. Educational institutions are some of the key centers in a smart city, which play an important role in preparing students for tomorrow [25]. In a smart academic environment, worldwide educational resources are shared to all faculty and students, and this will make the academic life of students and staffs easier, comfortable, and attractive. These potentials will be fully realized when IP for everyone in the case of IoT [26] changes to OTN for everyone in the case of SoIT.

Acknowledgements We acknowledge the support and sponsorship provided by Covenant University through the Centre for Research, Innovation, and Discovery (CUCRID).

References

1. David, N.D.: How the internet of things is revolutionizing healthcare. In: White Paper. Healthcare Segment Manager, Freescale Semiconductor (2013)
2. Adamkó, A., Kádek, T., Kollár, L., Kósa, M., Pánovics, J.: New challenges in smart campus applications. In: Recent Advances in Computer Science (2015)
3. Royer, M.: The Internet of Things (IoT). A trends white paper—August 2013. Bellevue College Economic & Workforce Development
4. Donald, S.: Parking on a Smart Campus: Lessons for Universities and Cities. Published by University of California Transportation Center (2005)
5. Rohs, M., Bohn, J.: Entry points into a smart campus environment—overview of the ETHOC system. This work was conducted as part the Entry Points project, which is funded by the ETH World Program (2003)
6. Jara, A.J., Ladid, L., Skarmeta, A.: The internet of everything through IPv6: an analysis of challenges, solutions and opportunities. *JoWua* **4**(3), 97–118 (2013)
7. Stefanie, T., Christoph, S.: A business model type for the internet of things. In: Research in Progress. 22nd European Conference on Information Systems, Tel Aviv (2014)
8. Katole, B., Sivapala, M., Suresh, V.: Principle elements and framework of internet of things. *Int. J. Eng. Sci.* **3**(5), 24–29 (2013)
9. Gubbia, J., Buyy, R., Marusic, S., Palaniswami, M.: Internet of things (IoT): a vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **29**, 1645–1660 (2013)
10. Rambus: The Internet of Things. How Rambus Sees the Future. 2014 Rambus Inc
11. Mattern, F., Floerkemeier, C.: From the internet of computers to the internet of things. In: Distributed Systems Group, Institute for Pervasive Computing, ETH Zurich (2010)
12. Abuelyaman, E.S.: Making a smart campus in Saudi Arabia. *EDUCASE Q.* **31**(2), 10–12 (2018)
13. Roman, R., Najera, P., Lopez, J.: Securing the internet of things. *Computer* **44**(9), 51–58 (2011)
14. StaReport: The Internet of Things: Privacy and Security in a Connected World. Staff Report January 2015. A Workshop Hosted by FTC
15. IEC: Internet of Things: Wireless Sensor Networks. White Paper (2014)
16. Ruggieri, M., Nikoogar, H.: Internet of Things: From Research and Innovation to Market Deployment. Rivers Publisher's Series in Communication (2014)
17. Ravikanti, S., Preeti, G.: Future's smart objects in IOT, based on BigData and cloud computing technologies. *Int. J. Innov. Res. Comput. Commun. Eng.* **3**(7), 6808–6817 (2015)
18. Azure, M.S.: Get Started with the Internet of Things in Your Organization. Introducing Microsoft Azure Internet of Things services. Executive Summary. © 2015 Microsoft Corporation
19. IBM: Saving the future of the IoT. IBM Institute of Business Value. Device Democracy (2005)
20. Barcena, M.B., Queest, C.: Insecurity in the IoT. *Candid Wueest*, Version 1.0 (2015)
21. Azeta, A.A.: Developing A computerized VoiceXML-based application for E-education: design, implementation and evaluation. Lambert Academic Publishing, Germany (2013)
22. Nati, M., Gluhak, A., Abangar, H., Headley, W.: Smartcampus: a user-centric testbed for internet of things experimentation. In: IEEE WPMC (2013)
23. Voxeo: Voice voice server, available online at: <http://community.voxeo.com> (2003)
24. Rieman, M.F.J., Redmiles, D.: Usability evaluation with the cognitive walkthrough. In: CHI '95 Proceedings, ACM (1995)
25. APKM: Smart Campus Guidelines-Draft. APKM—Smart Campus Draft Version 123/05/2015
26. Esaki, H.: Smart Campus Implementation Based on Internet-by-Design. Green University Tokyo Project. Internet Society (2015)

27. Azeta, A.A., Igbekere, E.O., Azeta, V.I.: Moving from Web-of-Things to Voice-of-Intelligent-Things in e-Campus. In: AFRICON, 2017 IEEE. IEEE (2017)
28. Nuance: <http://www.nuance.com>, 2002, as of 15 March 2002
29. SpeechWorks: <http://www.speechworks.com>, 2002, as of 15 March 2002
30. TellMe: <http://www.tellme.com>, 2001, as of 15 March 2002
31. Beck, J.E., Jia, P., Mostow, J.: Automatically assessing oral reading fluency in a computer tutor that listens. *Technol. Instr. Cognit. Learn.* **2**, 61–81 (2004)
32. Rickel, J., Johnson, W.L.: Task-oriented collaboration with embodied agents in virtual worlds. In: Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (eds.) *Embodied Conversational Agents*, pp. 95–122. MIT Press, Cambridge (2000)
33. Iera, A., Morabito, G., Atzori, L. (eds.): *The Internet of Things*. Springer, Berlin (2010). ISBN: 978-1-4419-1673-0

Intrusion Detection and Prevention Systems: An Updated Review



Nureni Ayofe Azeez, Taiwo Mayowa Bada, Sanjay Misra,
Adewole Adewumi, Charles Van der Vyver and Ravin Ahuja

Abstract The evolution of Information Technology (IT), cutting across several divides in our daily endeavors allows us to interact with all forms of data at different OSI model layers from application to physical. These data are susceptible to intrusion, aimed at compromising its integrity; thus, the need to protect these data, maintain its integrity, confidentiality, and availability cannot be overemphasized. Intrusion Detection and Prevention System (IDPS) is a device or software application designed to monitor a network or system. It detects vulnerabilities, reports malicious activities, and enacts preventive measures to keep up with the advancement of computer-related crimes using several response techniques. This paper presents an updated review on IDPSs given the fact that the most recent review found on the subject was done in 2016. It will also discuss the use of IDPSs to identify vulnerabilities in various channels through which data is accessed on a network or system and prevention mechanisms applied to mitigate against intrusion.

Keywords Vulnerabilities · Malicious activities · Prevention · Network · IDPS

N. A. Azeez · C. Van der Vyver
North-West University, Vaal Triangle Campus, Vanderbijlpark, South Africa
e-mail: nurayhn1@gmail.com

C. Van der Vyver
e-mail: charles.vandervyver@nwu.ac.za

T. M. Bada
University of Lagos, Lagos, Nigeria
e-mail: badataiwo@gmail.com

S. Misra (✉) · A. Adewumi
Covenant University, Ota, Nigeria
e-mail: Sanjaymisra@covenantuniversity.edu.ng

A. Adewumi
e-mail: wole.adewum@covenantuniversity.edu.ng

R. Ahuja
University of Delhi, New Delhi, India

1 Introduction

During 1984 and 1986, more research on intrusion detection system was done by several researchers. James P. Anderson's [1] presented a research on Intrusion Detection System (IDS). In the mid-1990s, IDS products were first commercialized by two companies, Internet Security System Inc (ISS) and Wheelgroup. They designed a network-based IDS called RealSecure and Netranger, respectively. ISS Inc. released the first version of RealSecure 1.0 for Windows NT 4.0. RealSecure used a knowledge base by matching signatures, however, it was ineffective for new attacks which became a major setback. Wheelgroup's Netranger was a known network-based IDS back in 1995; it functioned by scanning network traffic. Wheelgroup was acquired in by Cisco in February 1998; today, it forms an intrinsic part of Cisco's security.

Many researchers identified the setback in using the knowledge-based technique of matching signatures because it required continuous update of the database to recognize new attacks; more so, network and packet switching began to rise to a high speed from megabits to gigabits per sec. This was a major challenge as it became more difficult to scan through, analyze traffic, and detect attacks in real-time; thus, researchers were burdened with designing an IDS fit for high-speed networks. This led to the invention of host-based IDS, for example, TCP Wrappers, Tripwire, and Snort which provided analysis of system logs in real time. Snort is a free IDS tool, known for its multi-functionality as a network-based and host-based IDS. It was first released by Marty Roesch on December 22, 1998 for UNIX systems; later in 1999, a version of Snort (version 1.5) was released; it was effective in analyzing and logging packets in real-time; it was later modified for Windows system by Michael Davis in the year 2000 [2].

Today, as the functionality of IDS advances, attackers now explore means of detecting, bypassing, and disabling IDS before penetrating the infrastructure, resulting in denial of service (DoS). Security experts aim to curb these attacks by using Intrusion Detection and Prevention System (IDPS) architectures which are not visible to attackers by restricting communication permitted among various security components on a network. Due to the gradually increasing number of vulnerabilities, the identification of attack is essential. To this end, a number of reviews have been done on IDPSs in the literature [3] with the most recent one being [4] which was conducted in 2016. A lot has happened since that period that is worthy of reporting. For instance, it was in 2016 that the biggest DDoS attacks powered by a Botnet [5] were recorded. An example is Mirai, a Botnet primarily composed of infected routers and security cameras, low-powered and poorly secured devices which caused a lot of major DDoS attacks [6].

Internet attacks thus must be defined to measure security. Also, in recent times, infrastructure has evolved from a network of systems, private cloud infrastructure to the Internet of Things (IoT) offering several cloud-based services and solutions. While this has provided limitless opportunities on the choice of where to store data, the risk that accompanies these opportunities is also considered enormous because

these data can be compromised via several intrusion methods irrespective of the platform on which the data is stored [7].

2 Motivation

Intrusion is a criminal act committed against an information system, e.g., computer system, network or web infrastructure, such that security on the system is breached or compromised thereby putting it in an insecure state which allows for unauthorized access to the data being hosted by the system. This is done either by bypassing, disabling or exploring vulnerabilities on the system, typically leaving traces which can be discovered by the intrusion detection system.

The most recent review [4] focused majorly on network intrusion detection systems. The elements of security here are basically availability, accuracy, access control, confidentiality, integrity, and identification [8]. Intrusive attacks can be classified into passive [9] and active [5] attacks. An attack is classified as active when data is being altered with the intent to corrupt, destroy the data or the entire network hosting the data [5, 10]. An example of an active attack is interruption. This can include Denial of Service (DoS) [9], Distributed Denial of Service (DDoS) [6], SQL Injection [11], fabrication, replay attack [9], masquerading [12], and modification [5]. Some examples of passive attacks include traffic analysis [9], sniffing [5], and keyloggers [13]. The drawback of the study includes the fact that it mentioned but did not show the classification of intrusion detection systems. Also, network IDPS are known to identify abnormal behavior in network nodes only after the damage has been done to network resources [14]. Furthermore, with the increasing growth of the Web—a global network—network intrusion detection systems are limited in capacity at detecting anomalies on the web.

Newer models comprising a combination of machine learning techniques are being applied to combat attacks on the Web [15]. Some of the models also respond to security threats by detecting various malware intrusions and protocol authentication based on human behavior.

In addition, IDPSs are also being developed for cloud-based environments/systems; hence, the deployment of distributed IDPSs in cloud systems raises many challenges due to the diversity of its services and the complexity of its infrastructure.

3 Intrusion Detection and Prevention Systems

Security systems are designed in practice, to detect, identify and respond to malicious attacks against, a computing system, network or in general, information systems. These attacks are aimed at undermining the integrity of these systems,

steal information and in some cases cause damage to the systems thereby making the system unavailable.

IDS is either a software or hardware that automates intrusion detection, monitors network traffic for suspicious activities, and sends notifications to an administrator [7]. Intrusion Prevention System (IPS) is a software or hardware that prevents an intruder from gaining access to a network, let alone attack a network [16].

Today, security experts are trending with security appliance combining both intrusion detection and prevention capabilities which identify, log possible incidents, prevent attack, and send report to an administrator [16, 17]. Intrusion Detection and Prevention Systems (IDPS) ensures that the protection, availability, integrity, and confidentiality of information systems are guaranteed.

IDPS has become important when putting the security of information systems into context, preserving data, protecting data from unauthorized access or theft, and ensuring continuous availability of services that these information systems provide. Until recently, attackers' focus was on bank customers, where accounts were raided through fraudulent acquisition of personal details either by sending phishing emails or keyloggers, and credit cards were stolen [18].

3.1 Classification of Intrusion Detection and Prevention Systems

IDPS can be classified based on the following criteria [19]:

1. **Type of Intruder:** This is either external or internal. An external intruder is one who does not have any form of access to a network or service, while an internal intruder is one who has authorized access to a network but has restricted permissions on the network.
2. **Type of Intrusion:** There are various types of intrusions which are discussed in chapter two.
3. **Detection Technique:** Three different types of techniques are normally adopted for intrusion detection, misuse detection, anomaly detection, and stateful protocol analysis [20].
 - **Misuse Detection:** It is a signature-based detection method which matches intruder attack patterns, represented by signatures against a knowledge base of known exploit on software and system vulnerabilities. Signature-based detection analyzes and identifies specific patterns of events or behavior that portray an attack using either static, dynamic or hybrid approach. The static approach in misuse detection analyzes intrusion activities against a program and its code before execution. Dynamic approach analyzes attack patterns during or after the execution of a program while hybrid combines both static and dynamic approach to detect malicious attacks [21].

- **Anomaly-Based Detection:** It is a behavior-based detection technique which gets its input from audit logs generated from the operating system. This type of technique looks for variations in behavior which might indicate masquerading. Anomaly-based detection uses profiles created through monitoring the behavior of typical activity or models of intended behavior of users and applications over a period. It analyzes malicious attacks by using these profiles which represent the normal behavior of users, hosts, network connections, or applications against profiles of monitored activities; any deviation from the norm is triggered via an alert system [21].
- **Stateful Protocol Analysis:** Stateful [22] protocol analysis detects changes of protocol state. Unlike the anomaly detection method, this adopts predetermined universal profiles created based on accepted definitions of protocol behavior created by vendors and industry leaders [23].
- **Rule-based:** This involves making decisions based on rule sets which are defined by domain experts. They can detect known attacks but are incapable of detecting novel attacks. Also, with increase in network traffic, finding and coding rule sets is both difficult and time-consuming.
- **Supervised Machine Learning (ML):** It does not require model building as in the case of anomaly-based detection. Rather it is able to learn complex malicious and normal models.
- **Unsupervised Machine Learning:** An example is clustering-based IDPS. This approach to intrusion detection involves building models with unlabeled data; however, their performance is not as good as the supervised models (Table 1).

3.2 *Types of Intrusion Detection and Prevention Systems (IDPS)*

IDPS will be discussed based on the way they are deployed and the type of activities they monitor [16].

- Network-Based Intrusion Detection and Prevention System (NIDPS)
- Wireless Intrusion Detection and Prevention System (WIDPS)
- Network Behavior Analysis (NBA)
- Host-Based Intrusion Detection and Prevention System (HIDPS)

3.2.1 **Network-Based IDPS**

Network-Based IDPS (NIDPS) technology is designed to analyze packets at the network, transport, and application layer of the Open System Interconnection (OSI) model. NIDPS is most efficient when deployed within a network

Table 1 Advantages and disadvantages of intrusion detection techniques

Detection techniques	Advantages	Disadvantages
Signature-based detection	Effective and simple method of detecting known attacks since it uses signatures of known attacks	Cannot track unknown attacks and variants of known attacks
	Analyzes and identifies attacks by matching malicious signatures against known knowledge base	Attackers can make adjustment to attacks to avoid matching known attack signature
	Detection accuracy for known attacks is high	Requires continuous update of signatures or patterns
	Low computational cost	Newer attack signatures may not be in the signature database
	Rate of false alarm is very low	Detect only the attacks for which they are configured
Anomaly-based detection	Ability to detect and reduce the false alarm rate of unknown attacks	Detection accuracy is based on the amount of collected behavior or features
	Can detect new and unforeseen vulnerabilities	Well-known attacks may not be detected if they fit established a profile
	Dependency on the operating system is minimal and it is able to detect privilege abuse	Intruder can change profile slowly over a period
	Uses statistical test on collected behavior to identify intrusion	Configuring profiles is time-consuming
	No need for priori knowledge of security flaws	Less effective in the dynamic environment due to constant changes in monitored events
	System can also detect attacks from inside a network	
Stateful analysis	Adds stateful characteristics to regular protocol analysis	Resource intensive for protocol state tracing and analysis
	Distinguishes unexpected sequences of commands	Cannot detect attacks that do not violate the characteristics of generally accepted protocol behavior
	Identifies unexpected sequences of commands	
Rule-based	Can easily detect known attacks	Unable to detect unknown attacks
		Finding and coding rule sets is both difficult and time wasting
Supervised ML	Ability to learn complex and malicious models	They are hardly ever used in a real-world scenario owing to the fact that they require sufficient supply of labeled branding data
		Training data is labeled by domain experts which is both costly and time-consuming
Unsupervised ML	Works with unlabeled data on domain specialist may not be required	Performance is not as good as supervised ML

infrastructure with a specific design where it is able to monitor and analyze real-time packets for intrusion and take a decision on any suspicious activity. While NIDPS is effective in analyzing and detecting suspicious network packets in real time, it cannot analyze encrypted traffic, traffic over Virtual Private Network connection (VPN), SSH or HTTPS sessions, and traffic on mobile computing networks [17].

NIDPS has broad intrusion detection capabilities. An example of an NIDPS is KEMP Loadmaster which can detect intrusion and prevent intrusion by shutting down the device.

3.2.2 Wireless IDPS

WIDPS is a variant of NIDPS which monitors and analyzes packets and protocols on a wireless network. Despite its ability to analyze network traffic, WIDPS cannot detect abnormal activities within an application [17].

Advantages

- It is effective for monitoring and analyzing intrusion on a wireless network.
- WIDPS can identify various problematic issues like policy violations and mis-configurations at the WLAN protocol level.

Disadvantages

- It is vulnerable to DoS attacks.
- It cannot monitor and analyze packets on transport layer, network layer, and application layer.
- It is susceptible to evasion technique when an intruder attacks channels that are not currently monitored.

3.2.3 Network Behavior Analysis (NBA)

NBA is also a variant of NIDPS with the ability to monitor and analyze network traffic to detect unusual activities that may emanate from violation of policy, DDoS attacks or malware intrusion [17].

Advantages

- It is effective in detecting DoS attacks.
- It is effective for monitoring packets on transport, network application TCP/IP layer, etc.
- It can monitor and detect threats caused by malware, policy violation, and DDoS.

Disadvantages

- Packets are analyzed in batches, thus delaying the rate of intrusion detection.

3.2.4 Host-Based IDPS

Host-Based IDPS technology is designed for Application level and Operating System intrusion detection and prevention by monitoring the events on a single host on which it is installed. Aside from having the capability of monitoring and analyzing network traffic, HIDPS can analyze system-specific settings such as software calls, local security policy, and audits logs within the host for suspicious activities. HIDPS functionality can be divided into four categories [17]:

- **File System Monitoring:** Every system has a file system to detect and prevent intrusion; HIDPS monitors file systems regularly by checking variations in files size and file content against a known knowledge base. Whenever a system or user file shows a significant deviation, an alert is triggered which sends a notification to an administrator, indicating the detected intrusion and the action taken to prevent access or damage to the file [23].
- **Log File Analysis:** System events are generally logged in a file. These files (event logs) are analyzed constantly by HIDPS for changes or abnormal activities; a typical event log changes in the login information.
- **Connection Analysis:** HIDPS monitors and analyzes network packets (TCP/IP) for suspicious activities such as the ratio or sequence of TCP/IP connections on the host on which it is installed [24].
- **Kernel-Based HIDPS:** The kernel is provided with extra security capability which allows it to identify and prevent intruder activities itself.

Advantages

- It can detect intrusions on host applications, operating system, and network layer traffic.
- It can monitor and analyze suspicious activity on encrypted communication.
- It can detect intrusion on host systems by monitoring its file system, file access, system calls, etc.
- It does not require additional hardware since it is deployed on the host system.
- It can detect misuse of profile because it interacts with the user, as well as server installed application.
- It can prevent intrusion at the system level and detect attacks which NIDPS cannot detect.

Disadvantages

- It does not use a predefined database, therefore, detection accuracy is limited.
- Its uses more host resources, therefore, impacting on the system host performance.
- It must be deployed on each host which is expected to monitor.
- Its monitoring is restricted to the host on which it is deployed.
- There is a possibility of conflict with preexisting security configuration.

4 IDPS, Design, and Architecture

Information systems today have become a target for hackers whose only aim is to undermine the integrity, availability, and confidentiality of data. Therefore, proper design consideration must be put in place when designing an IDPS to increase its capacity to detect a threat and prevent it from gaining access to an information system. In designing an IDPS, the following must be considered [25].

4.1 *Speed and Accuracy*

These are highly desirable features. The sensitivity of an IDPS in terms of its speed and accuracy determines the rate of false negatives and false positives reported by the system. If the sensitivity of an IDPS is too low, it will have a high rate of false negative where intrusive activities are not detected, thus, no alert is triggered. Whereas if the sensitivity of an IDPS is too high, there is a high tendency of reporting false positives where an alarm is triggered for nonintrusive activities. False negatives and false positives can be triggered by several factors discussed below [26].

Causes of False-Negative Alerts:

- Improper spanning of switch ports which can cause network traffic to overwhelm the switch which can contribute to events with false-negative triggers.
- Flaws in the design of encrypted traffic which are usually not clear to the IDPS.
- A poorly written signature which does not have the capacity to detect an attack even though the attack is known.
- Improper communication of change management on network and server infrastructure to the information security team.
- Intrusive attacks caused by unpublicized or new attacks thus making it invisible to existing signatures.

Causes of False-Positive Alerts:

- A reactionary traffic alarm caused by equipment failure can trigger a false positive alert. For example, an ICMP flood caused by unreachable destination can trigger a false positive alert.
- An equipment-related alarm, e.g., a load balancer can trigger an alert generated from unrecognized packets from the equipment itself.
- A poorly written client software can trigger alerts of policy violation, e.g., alerts triggered by software bugs.
- Alerts triggered by unmalicious events.

4.2 Logging Capabilities

The logging capability of an IDPS is also very important because it facilitates its ability, identifies, detects, and reports malicious activities. The following are the logging capabilities to consider when designing an IDPS:

- IDPS must be able to log time stamps which include the date and time the malicious activity occurred.
- IDPS must be able to log connection ID, usually a unique number assigned to a session or a TCP connection.
- IDPS must effectively log an alarm type, set its severity rating, impact, and the priority of attack.
- IDPS must have the capacity to analyze protocols like TCP, UDP, ICMP at the network, application, and transport layer.
- IDPS must be able to identify the source and destination IP of connections and determine the number of bytes transmitted over the connection.
- IDPS must effectively understand the characteristics of application request and responses.

4.3 Information Gathering Capabilities

For an IDPS to be effective in detecting and preventing malicious activities on an information system, it must be able to gather information about the system upon which it is deployed.

- IDPS must be able to gather information on host profiles which include host IP and their corresponding MAC address.
- Ability to determine the OS version to enable it to determine the type of vulnerability it is susceptible to.
- Ability to identify network characteristic by gathering information on changes in network configuration.

4.4 Architecture of IDPS

Depending on the expected outcome, IDPS can be deployed using the following architecture [20]:

- Centralized: This architecture collects data centrally, sends it to a single location for analysis. Data collection is either from a single host or from several hosts.

- Hierarchical: This architecture collects data from several hosts which are analyzed according to the layers of the deployed IDPS.
- Distributed: This architecture collects data host by host and it is analyzed.

5 Conclusion

Security threats and incidents have evolved and pose a great challenge to information systems; thus, the importance of deploying an IDPS cannot be overemphasized and efforts to create more security techniques must continue to ensure that the integrity, originality, and confidentiality of information systems is guaranteed, thus making it accessible to everyone when the need arises.

IDPS in its various forms according to Chap. 3 is essentially beneficial in that it has the capacity to identify and detect vulnerabilities and prevent all forms of intrusion discussed in Chap. 2. However, consideration must be given to the type of deployment method to get the best of IDPS.

Furthermore, IDPS has extensive logging capacity which makes it effective in intrusion detection, most especially against signatures of various attacks against network systems; this has made a necessity for enterprise environment to protect data, as data sharing of all sorts has evolved to a global trend.

Lastly, when multiple IDPS technologies are combined into a single protection solution, it reduces management costs considerably because IDPS engages several techniques in intrusion detection and prevention; thus when developing a security strategy, it is important that it is comprehensive to stay ahead of the next threat.

References

1. Anderson, J.P.: Computer Security Planning Study. Washington (1972). Retrieved from <https://pdfs.semanticscholar.org/0735/6c5477c83773bd062b525f45c433e5b044e8.pdf>
2. Bruneau, G.: The History and Evolution of Intrusion Detection, vol. 7 (2001). Retrieved from <https://www.sans.org/reading-room/whitepapers/detection/history-evolution-intrusion-detection-344>
3. Patel, A., Taghavi, M., Bakhtiyari, K., Júnior, J.C.: An intrusion detection and prevention system in cloud computing: a systematic review. *J. Netw. Comput. Appl.* **36**, 25–41 (2013)
4. Amudhavel, J., Brindha, V., Anantharaj, B., Karthikeyan, P., Bhuvaneshwari, B., Vasanthi, M., Nivetha, D., Vinodha, D.: A survey on intrusion detection system: State of the art review. *Indian J. Sci. Technol.* **9**, 1–9 (2016)
5. Ahmad, K., Verma, S., Kumar, N., Shekhar, J.: Classification of internet security attacks. In: Proceedings of the 5th National Conference; INDIACom-2011. New Delhi (2011). Retrieved from https://www.researchgate.net/publication/262494946_Classification_of_Internet_Security_Atta
6. Symantec.: Internet Security Threat Report. Mountain View, CA 94043 (2017). Retrieved from <https://www.symantec.com/content/dam/symantec/docs/reports/istr-22-2017-en.pdf>
7. Kemmerer, R.A., Vigna, G.: Intrusion detection: a brief history and overview, pp. 27–29 (2002). Retrieved from <https://www.computer.org/csdl/mags/co/2002/04/r4s27.pdf>

8. Persa, S.: Network Security (2003). Retrieved from https://utcluj.ro/pub/docs/cursuri/prc_eng/stallings/securitatea.ppt
9. Pawar, M.V., Anuradha, J.: Network Security and Types of Attacks in Network, pp. 504–506 (2015). Retrieved from https://ac.els-cdn.com/S1877050915006353/1-s2.0-S1877050915006353-main.pdf?_tid=84c3d323-6ab4-4ca1-86f5-7eafc2cbfb30&acdnat=1530640171_af15fb42c5d503b379a7da902477f68d
10. Bloomberg, J.: Cybersecurity Lessons Learned From ‘Panama Papers’ Breach. The Little Black Book of Billionaire Secrets (2016). Retrieved from <https://www.forbes.com/sites/jasonbloomberg/2016/04/21/cybersecurity-lessons-learned-from-panama-papers-breach/#453217b12003>
11. Clarke, J.: SQL Injection Attacks and Defense, 2nd edn. Elsevier, Waltham (2012)
12. Salem, M.B., Stolfo, S.J.: Data collection and analysis for masquerade attack detection: challenges and lesson learned. Columbia University, Computer Science. New York: Department of Computer Science, Columbia University (2011). Retrieved from <https://doi.org/10.7916/D8D50VV1>
13. Wood, C.A., Raj, R.K.: Keyloggers in Cybersecurity Education. New York: Rochester Institute of Technology (2010). Retrieved from <https://pdfs.semanticscholar.org/d1d4/628a22e8d27c6cd202839d7bf0e3a7c7ea91.pdf>
14. Yerur, S.V., Natarajan, P., Rangaswamy, T.R.: Proactive hybrid intrusion prevention system for mobile adhoc networks. *Int. J. Intell. Eng. Syst.* **10**, 273–283 (2017)
15. Johnson Singh, K., Thongam, K., De, T.: Entropy-based application layer DDoS attack detection using artificial neural networks. *Entropy* **18**, 1–17 (2016)
16. Chee, J.: Host Intrusion Prevention Systems and Beyond, vol. 26 (2008). Retrieved from <https://www.sans.org/reading-room/whitepapers/intrusion/host-intrusion-prevention-systems-32824>
17. Letou, K., Devi, D., Singh, J.Y.: Host-based intrusion detection and prevention. *Int. J. Comput. Appl.* **0975–8887**(69), 27–32 (2013)
18. NSS Labs.: Security Value Map. Next Generation Firewall (NGFW), p. 1 (2017). Retrieved from <https://www.nsslabs.com/research-advisory/security-value-maps/2017/ngfw-svm-graphic/>
19. Santos, K.B., Chandra, S.T., Phani, R., Ratnakar, M., Baba, D.S., Sudhakar, N.: Intrusion detection system- types and prevention. *Int. J. Comput. Sci. Inf. Technol.* **77–82** (2013). Retrieved from <http://ijcsit.com/docs/Volume%204/Vol4Issue1/ijcsit2013040119.pdf>
20. Singh, A.P., Singh, M.D.: Analysis of host-based and network-based intrusion detection system. *Comput. Netw. Inf. Secur.* **41–47** (2014). Retrieved from <http://www.meecs-press.org/ijcnis/ijcnis-v6-n8/IJCNIS-V6-N8-6.pdf>
21. Ghafir, I., Husak, M., Prenosil, V.: A survey on intrusion detection and prevention (2014)
22. Williams, T., Shirley, M.: Next Generation Intrusion Prevention System (NGIPS) Test Report. NSS Labs (2017). Retrieved from <https://www.forcepoint.com/resources/reports/nss-labs-2017-next-gen-ips-report>
23. Masaryk University, Faculty of Informatics. Brno, Czech Republic: Researchgate. Retrieved from https://www.researchgate.net/profile/Ibrahim_Ghafir/publication/305957314_A_Survey_on_Intrusion_Detection_and_Prevention_Systems/links/57a75a2708aefe6167bc1de0/A-Survey-on-Intrusion-Detection-and-Prevention-Systems.pdf?origin=publication_detail
24. Scarfone, K., Mell, P.: Guide to Intrusion Detection and Prevention Systems (IDPS). National Institute of Standards and Technology Special Publication 800-94, 127 (2007). Retrieved from <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-94.pdf>
25. Sharifi, A., Zad, F.F., Noorollahi, A., Sharifi, J.: An overview of intrusion detection and prevention systems (IDPS) and security issues. *IOSR J. Comput. Eng. (IOSR-JCE)* **16**(1), 47–52 (2014)
26. Stawowski, M.: The Principles and Good Practices for Intrusion Prevention Systems Design, vol. 25 (2006). Retrieved from <https://pdfs.semanticscholar.org/8cac/01d90c44ae710719df662f858ca17ffc96d1.pdf>

Simulation-Based Performance Analysis of Location-Based Opportunistic Routing Protocols in Underwater Sensor Networks Having Communication Voids



Sonali John, Varun G. Menon and Anand Nayyar

Abstract Recently, Underwater Wireless Sensor Networks (UWSNs) have emerged as a prominent research area in the networking domain due to their wide range of applications in submarine tracking, disaster detection, oceanographic data collection, pollution detection, and underwater surveillance. With its unique characteristics like continuous movement of sensor nodes, limitations in bandwidth and high utilization of energy, efficient routing and data transfer in UWSNs have remained a challenging task for researchers. Almost all the protocols proposed for terrestrial sensor networks are inefficient and do not perform well in an underwater environment. Recently Location-Based Opportunistic Routing Protocols have been observed to perform well in UWSN environments. But it is also observed that these protocols suffer from performance degradation in UWSN networks with communication voids. The objective of this research paper is to discuss the working of major Location-Based Opportunistic Routing Protocols in UWSNs with communication voids and to highlight their issues and drawbacks. We analyzed the Quality of Service parameters, packet delivery ratio, end-to-end delay, throughput, and energy efficiency of two major Location-Based Opportunistic Routing Protocols, i.e., Vector-Based Forwarding (VBF) and Hop-by-Hop VBF (HH-VBF) in UWSNs with communication voids using NS-2 simulator with Aqua-Sim extension. Simulation results state that both VBF and HH-VBF protocols suffered from performance degradations in UWSNs with communication voids. In addition to this, the paper also highlights open issues for UWSN to assist researchers in designing efficient routing protocols for UWSNs having multiple communication voids.

Keywords Aqua-Sim · Communication void · Hop-by-Hop Vector-Based Forwarding (HH-VBF) · NS-2 · Opportunistic routing · Performance analysis · Quality of Service (QoS) · Underwater Wireless Sensor Networks (UWSNs) · Vector-Based Forwarding (VBF)

S. John · V. G. Menon (✉)
SCMS School of Engineering and Technology, Ernakulam, India
e-mail: varunmenon@scmsgroup.org

A. Nayyar
Graduate School, Duy Tan University, Da Nang, Vietnam

1 Introduction

Recently, Underwater Wireless Sensor Networks (UWSNs) [1, 2] have emerged as a prominent research area in the networking domain. The increasing interest in applying sensor networks into the underwater environment has found numerous applications in civilian and military fields such as coastline surveillance, pollution detection, and underwater surveillance. In UWSNs, a group of sensors is placed at different depths in the ocean to gather information. This collected information is then forwarded to the target devices positioned at the surface via a network of intermediate nodes. The data is then stored, processed, and passed to different applications for appropriate utilization. As radio signals have many limitations due to its mediocre propagation through water, the acoustic medium is used between the sensor nodes in UWSNs [3]. Most of the applications are highly sensitive and their success relies on the accuracy of the collected data. Many recent applications of underwater sensor network ranging from aquaculture to the oil industry, instrument monitoring, climate recording, natural disturbances prediction, and study of marine culture depend heavily on the accuracy of the collected data. Thus, efficient routing and data transfer among the nodes is vital in UWSNs.

Routing and data transfer from the sender to the sink stations, through the network, have been a tough task for researchers. Incessant movement of the nodes placed in the water due to the difference in the ocean environment and ocean currents is a major challenge. Limitations in bandwidth, frequent link interruptions, increased delay of data transfer, interference caused by marine mammals, etc., are some of the other elements influencing the efficiency of routing in UWSNs [4, 5]. Due to these unique characteristics, almost all the protocols drafted for conventional sensor networks do not perform well in UWSNs [6]. Recently, Opportunistic Routing Protocols (ORPs) [7–10] have found to give better performance in data transfer in UWSNs. The major advantage of this routing strategy is that it dynamically selects one best forwarder device from a group of candidate devices. This selection is based on the present scenario of the network, which leads to better performance of this latest category of protocols. Many ORPs have been designed for ad hoc, terrestrial sensor, and UWSNs [11–16]. Among all categories in opportunistic routing, Location-Based Opportunistic Routing Protocols (LBORPs) has found to give better performance. LBORPs make use of the knowledge on the present location of the devices to dynamically route message packets from the sender to the target. A few Location-Based ORPs have been efficiently used for data forwarding in UWSNs [17–23].

Communication holes or voids have been a major problem in most of the dynamic sensor networks. Often defined as the unreachability problem, the source node suffers from lack of adequate forwarder nodes in its transmission range. Frequent movement of sensor nodes in underwater increases this problem further. Limited research has been performed to determine the working and performance of LBORPs in UWSNs with communication voids [24–27]. In large UWSNs with a limited number of sensors, it is very significant to have a protocol that can handle

communication voids efficiently. In this article, we analyze the working and performance of two major LBORPs in UWSNs with communication holes. Vector-Based Forwarding (VBF) [21] and Hop-by-Hop VBF (HH-VBF) [20, 28] have been used by many applications for data transfer in UWSNs. VBF is an LBORP that constructs a virtual vector pipe between the sender and the sink node for routing. Only the devices within the virtual vector are selected for forwarding the data packet. Extending the VBF protocol, the HH-VBF protocol makes use of different virtual pipes for each node and the direction of the virtual pipe changes during the entire time of transmission.

Objectives of this research paper are

- To study the functioning of major LBORPs designed for UWSNs and to determine their issues and drawbacks.
- To analyze the simulation-based performance comparison of VBF and HH-VBF protocols in UWSNs with communication voids on parameters—PDR, throughput, end-to-end delay, and energy efficiency.
- And, to discuss the issues, challenges, and future directions in UWSNs routing research.

This research article is structured as follows. Functioning of the major Location-Based ORPs in UWSNs is discussed in detail in Sect. 2. The section also discusses the issues and drawbacks with these protocols. Section 3 describes the working of VBF and HH-VBF in a comprehensive manner. Section 4 presents the simulation-based analysis of VBF and HH-VBF protocols in UWSNs having communication voids using NS2 + Aqua-Sim simulator. Section 5 enlists open issues and challenges of existing ORPs in UWSNs. The paper concludes in Sect. 6 with future research directions.

2 Literature Survey

This section examines the functioning of some of the major LBORPs proposed for UWSNs. One of the most accepted protocols in UWSNs is Vector-Based Forwarding (VBF) [21]. VBF uses the details on the present position of the sensor devices for routing the information packets. In VBF, within a fixed virtual pipe the information packets are forwarded between the source and destination devices. These information packets are then transmitted along the redundant paths and thereby remain stable against packet loss. The major limitation with VBF is that the construction of a solitary virtual pipe will decrease the performance of routing in various node density areas and the productivity drops with communication holes.

Hop-by-Hop Vector-Based Forwarding (HH-VBF) [20] is a variant of the VBF protocol that uses virtual pipes from every intermediate device to the target device. Every device then dynamically makes packet forwarding choices with reference to its present location in the network.

Directional Flood Based Routing (DFR) [29] is a receiver-based, stateless, and LBORP in which each and every node is informed about the position of the destination device and one-hop neighbors. DFR utilizes a limited flooding method, in which every node passes its location details to all the other nodes and achieves more reliability in data transmission in the network. The flooding mechanism results in duplicate transmissions and energy loss. DFR also failed to address the void problem.

One of the earliest protocols that functions with reference to the pressure information proposed for UWSNs was Depth-Based Routing (DBR) [18]. DBR utilizes the knowledge on the depth of devices placed underwater to decide whether to transmit the packets or not. When a packet is received, each node will forward the packet to a smaller depth than that stored in the packet. Otherwise, the packet is discarded. Both communication holes and node mobility are the major problems influencing the limited performance of DBR.

GEographic and opportunistic routing with Depth Adjustment-based topology control for communication Recovery over void regions (GEDAR) [30] is another location-based anycast ORP that sends information packets from intermediate nodes to different target nodes. The protocol uses periodic beaconing to get the position information of every node in the network. The protocol suffers from increased duplicate message transmissions.

Hydraulic Pressure-Based Anycast Routing (HydroCast) [31] also utilizes the knowledge on the depth of the deployed device to choose the candidates for routing from the nearest nodes. HydroCast elects a subgroup of neighboring devices with the largest greedy progress to the target node. The performance of the protocol is not satisfactory with voids.

Multi-Path Routing (MPR) [32] helps to solve the data concussion issue at the sink nodes. The data collision is prevented at the receiving packets by creating a route that consists of various sub-paths between the sender and the sink. Geographic Partial Network Coding (GPNC) [33] is another geographic, partial network coding-based protocol proposed for UWSNs. Void is a major concern for this protocol too.

Focused Beam Routing (FBR) [34] was designed to minimize energy drainage during the routing process by controlled flooding of the packets. Although a preventive void handling technique was used, high delay was incurred. Void-Aware Pressure Routing (VAPR) [33] uses an opportunistic data forwarding technique coupled with information on the pressure of the deployed devices. VAPR selects a subgroup of candidate devices with the greatest progress to target. The holding time details of two-hop neighboring nodes helps to deal with the void areas but imposes a high overhead on the system. In Relative Distance-Based Forwarding (RDBF) [35], the packets are sent via the nodes that are nearest to the target node. RDBF uses a fitness value as an upper limit to supervise the number of transmitting devices to the sink. Moreover, the performance of RDBF also comes down with the presence of communication holes in the network.

In the next two sections, we discuss the working of the two LBORPs, VBF, and HH-VBF, used in UWSNs. VBF and HH-VBF utilize the knowledge of the location

of nodes in the network for routing. These protocols use greedy forwarding mechanism that chooses the node nearer to the target as the next forwarder for every data packet. We analyze the performances of these two protocols in UWSNs having multiple communication voids. We have selected these two protocols for the analysis because they are the two popular protocols used by various applications in UWSNs. Further, the design of these protocols is less complex and also incurs less overhead compared to other protocols.

3 Location-Based Opportunistic Protocols

3.1 *Vector-Based Forwarding (VBF)*

VBF [21, 36] was proposed to address the problem of energy limitation and efficient packet delivery. VBF utilizes the knowledge of the location of nodes in the network for routing. Each data packet in VBF contains the three position fields named as OP1, TP1, and FP1. This represents the coordinates of the source, the sink, and the forwarder node. The RANGE field in the packet deals with node mobility. As soon as the information packet advances to the region marked by its TPI, it is flooded in that region controlled by the RANGE field. The virtual routing vector from the source to the sink describes the transmitting path. The RADIUS field inside the packet is a precomputed value that is utilized by intermediate devices to check whether they are near to the vector pipe and suitable for packet transmission. In order to limit the number of intermediate transmitting devices and to conserve energy in the network, VBF makes use of a self-adaptive algorithm. Upon getting a data packet, each node first computes whether it is inside the virtual vector and whether it can act as a potential forwarder to the next node. Each potential member device awaits a limited period of time to check its desirability value in the pipe. It describes the closeness of the current node to the past forwarder node, and the virtual pipe between the sender and the sink. The waiting time depends on the desirability factor. If a node is more desirable, then it waits for less time. During this time, the device pays attention to the channel to observe the number of devices that are transmitting the same information packet as the present device. When the period expires, the node will transmit its packet if and only if the reduced desirability value of the other nodes is less than a precomputed level (Fig. 1).

3.2 *Hop-by-Hop Vector-Based Forwarding (HH-VBF)*

Hop-by-Hop Vector-Based Forwarding (HH-VBF) [20] is a variant of VBF that constructs virtual routing pipes from each hop in the network as packets travel from

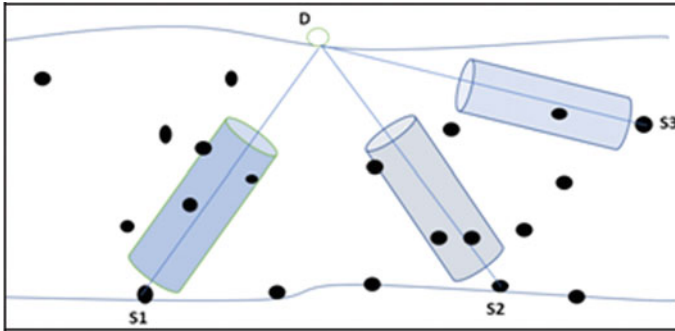


Fig. 1 Virtual vector pipe from sender to receiver node in VBF

the sender device to the receiver. This protocol creates a virtual routing pipe from each hop to the target device. Many studies have discussed the better working of HH-VBF in comparison with VBF. The pipeline radius is similar to the transmission count of the node. The protocol improves the direction of the transmitting pipe hop by hop, in the entire life period and in this way, all transmitting devices can generate a routing choice based on the present local topology details. By changing the direction of flooding pipeline dynamically, the performance can be enhanced. When a stable routing vector radius is set, the protocol shows a reduced performance in the sparse network compared with VBF. However, both the protocols are exposed to interferences caused by marine mammals as the data forwarding happens only inside the pipe. Here, the transmission can be interrupted when marine mammals block the pipe. There is an increased choice of finding a more acceptable forwarder within the hop-by-hop vector pipeline; HH-VBF gives better Packet Delivery Ratio (PDR). However, both the protocols do not succeed to yield energy fairness within the network. Also, they fail to effectively handle the communication hole problem.

The working of HH-VBF is described in Fig. 2. The sender node S2 wants to transmit an information packet to the destination device D2. Here S2 creates a virtual pipe to the destination D2. Once the packet reaches the intermediate sensor node M, it creates a virtual pipe to the destination. Now there are more nodes (N and P) included for sending the data packet to the target location. HH-VBF uses this forwarding strategy until the data reaches the target location. HH-VBF improves the performance in the network by dynamically moving the direction of the flooding pipe. Figure 3 illustrates the communication void problem in UWSNs. We can see that the source node S1 and S2 are trying to forward data packets to the destination D by creating virtual pipes. Data from S2 reaches the destination node located at the surface, but data from source S1 is unable to proceed due to the communication void near to the surface. Table 1 highlights the technical comparison of both the routing protocols.

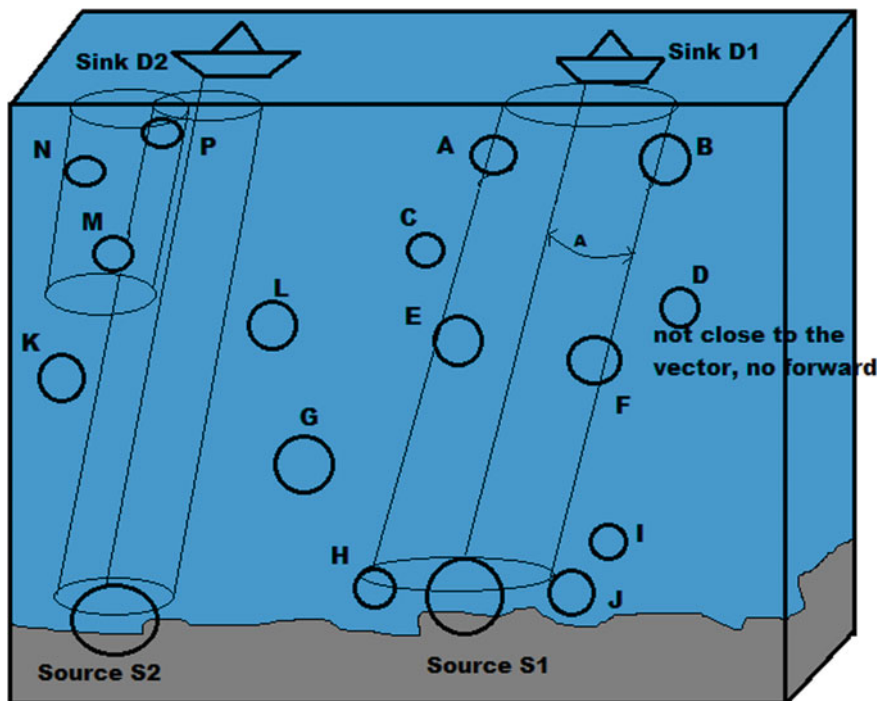


Fig. 2 Illustration of the functioning of VBF and HH-VBF

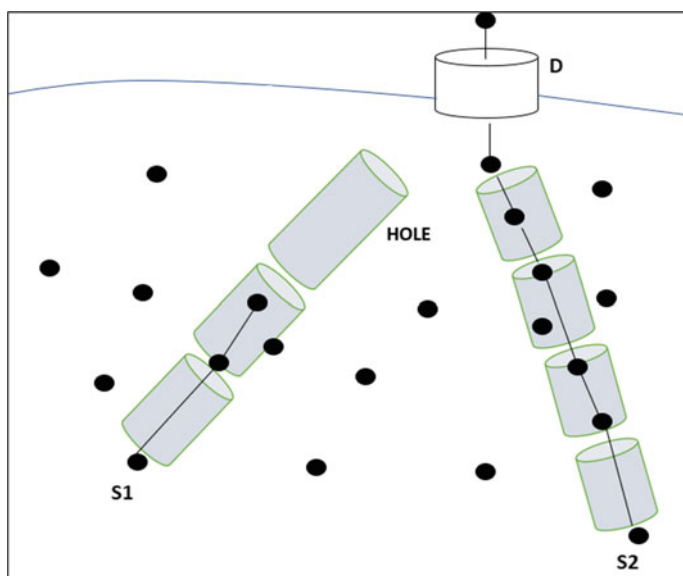


Fig. 3 Communication void problem with VBF

Table 1 Comparison of VBF and HH-VBF

Parameter	VBF	HHVBF
Metric used	Node location	Node location
Candidate coordination	Timer based	Timer based
Void handling	No	No
Routing	Single virtual pipe from the sender to the destination	Individual virtual pipes from each transmitting device to the target device
Advantages	Simple and flexible	Simple and flexible
Limitations	Duplicate messages performance degrades with communication holes	Performance degrades with communication holes

4 Simulation and Performance Analysis

4.1 Simulation Parameters

Performance of VBF and HH-VBF routing protocols are analyzed using simulations in Aqua-Sim [37]. Aqua-Sim is an extended version of NS-2 that efficiently simulates collision behaviors in large delay acoustic networks and also attenuation of underwater sensor networks. Aqua-Sim uses OTcl to model the protocol parameters implemented through C++ algorithms. The simulator creates a 3D environment of packet transmission in the network, which shows that the efficiency of both the protocols comes down with communication holes. Communication voids are created between the source and sink devices to measure QoS parameters in the overall network. Table 2 highlights the Simulation Parameters used to test the performance of the protocols in Aqua-Sim.

Table 2 Simulation parameters

Parameter name	Values
Simulator name	NS 2.35 + Aqua-Sim
Dimension of topology	1500 × 1500 × 1500 m
Transmission range	250 m
Antenna type	Omnidirectional
Data rate	50 kbps
Packet size	25–125 bytes
Number of nodes	100–600
Simulation time	300 s
Number of simulation runs	10
Protocols	VBF, HH-VBF

4.2 Performance Analysis

The performances of VBF and HH-VBF protocols were analyzed in a normal network scenario and in networks with communication void. Figure 4 shows the values of various parameters used in simulations. Four different QoS parameters were measured in the network.

- Packet Delivery Ratio (PDR):** PDR is the ratio of the overall packets arrived at the target device to the number of transmitted packets. Figure 4 shows the PDR in UWSN network scenario with varying number of nodes for the two protocols in normal network conditions and also with communication voids. Table 3 highlights the data values of VBF, VBF-Void, HH-VBF, and HH-VBF-Void. From the analysis, it is observed that both VBF and HH-VBF protocols give better performances in networks without communication voids. This proves that communication voids degrade the performance of even the latest LBORPs in UWSNs.
- Average End-to-End Delay:** It is the time incurred by the information packet in reaching the target device from the sender device. Figure 5 highlights the average delay incurred by VBF, HH-VBF protocols in transferring data packets across the network with varying number of devices in normal and void environments. Table 4 highlights the data analysis of delay observed by routing protocols with void and under normal operating conditions. the data values, it is

Fig. 4 PDR versus nodes

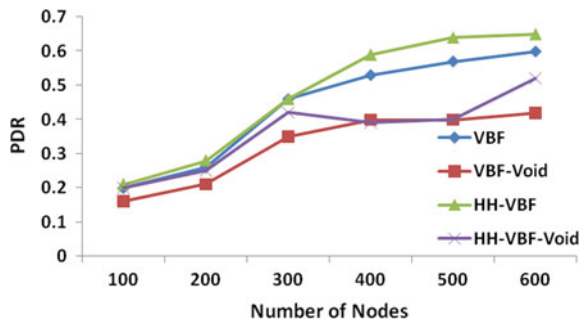


Table 3 PDR

No. of nodes	VBF	VBF-void	HH-VBF	HH-VBF-void
100	0.2	0.16	0.21	0.2
200	0.26	0.21	0.28	0.25
300	0.46	0.35	0.46	0.42
400	0.53	0.4	0.59	0.39
500	0.57	0.4	0.64	0.4
600	0.599	0.42	0.65	0.52

Fig. 5 Latency/End-To-End Delay versus nodes

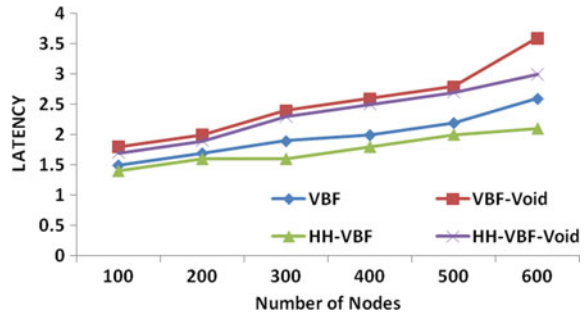


Table 4 Average end-to-end delay

No. of nodes	VBF	VBF-Void	HH-VBF	HH-VBF-void
100	1.5	1.8	1.4	1.7
200	1.7	2	1.6	1.9
300	1.9	2.4	1.6	2.3
400	2	2.6	1.8	2.5
500	2.2	2.8	2	2.7
600	2.6	3.6	2.1	3

evident that the delay is less VBF and HH-VBF in normal environments as compared to networks with voids.

- Throughput:** It is the rate or amount of data that is effectively transferred from the sender device to the target in a given time period. Figure 6 highlights the throughput of both VBF and HH-VBF protocols in normal and void environments. Table 5 enlists the data values of throughput. It is observed that HH-VBF protocol has the highest throughput compared to other routing protocols. Both the protocols suffer from degradation in throughput with voids in the network.
- Energy Consumption:** In order to test the validity of the performance of the routing protocols, energy consumption is regarded as a significant parameter. Energy consumption is regarded as how much energy is consumed by

Fig. 6 Throughput versus packet generation rate

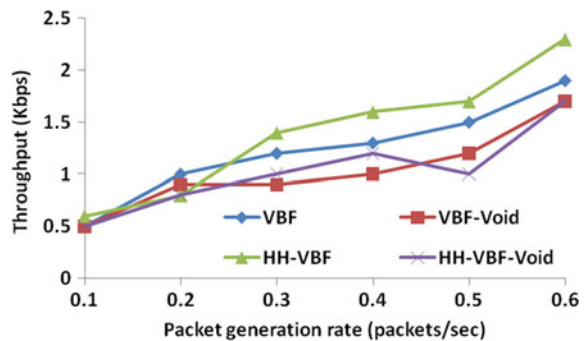


Table 5 Throughput

Packet generation rate	VBF	VBF-void	HH-VBF	HH-VBF-void
0.1	0.5	0.5	0.6	0.5
0.2	1	0.9	0.8	0.8
0.3	1.2	0.9	1.4	1
0.4	1.3	1	1.6	1.2
0.5	1.5	1.2	1.7	1
0.6	1.9	1.7	2.3	1.7

Fig. 7 Normalized energy consumption versus no. of nodes

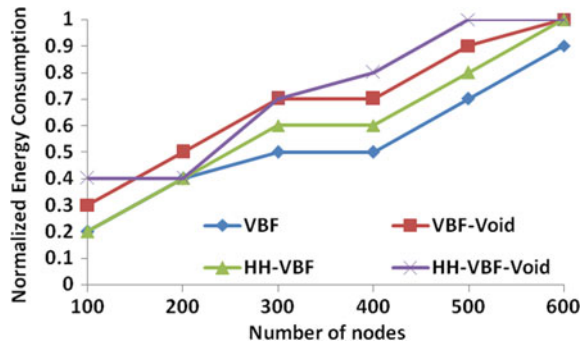


Table 6 Normalized energy consumption

No. of nodes	VBF	VBF-void	HH-VBF	HH-VBF-void
100	0.2	0.3	0.2	0.4
200	0.4	0.5	0.4	0.4
300	0.5	0.7	0.6	0.7
400	0.5	0.7	0.6	0.8
500	0.7	0.9	0.8	1
600	0.9	1	1	1

transmitting nodes and intermediary nodes to make sure that the packet reaches the destination. Figure 7 elaborates the graphical analysis of the energy consumption of energy nodes in different protocols under UWSN operating scenarios. Table 6 also highlights the data values of protocols and it is observed that both protocols consume more energy in void scenarios.

5 Open Research Areas and Recommendations

In this section, open research problems generated by communications holes in UWSNs are discussed. Researchers designing routing protocols need to address the following issues to come up with optimal routing solutions for UWSN scenarios.

- **Quality of Service (QoS):** Communication voids drastically degrade the QoS offered by the routing protocols in UWSNs. The data delivery rate is reduced with increased data loss due to communication holes and void nodes. Delay in data transmission increases with frequent loss of data packets.
- **Reliability:** Reliable data delivery is a vital factor in the success of any routing protocol. Void nodes are often unable to find the next forwarder node in the network and are forced to drop the data packet. Thus, reliability is a big concern in UWSNs with communication voids.
- **Scalability:** In many UWSNs, the problems of communication holes occur when the network size is increased. With a reduced number of sensor nodes in a large geographical area, the chances are high for the occurrence of communication holes. Solutions to avoid communication voids with a minimum number of sensor nodes are to be found out.
- **Mobility:** Frequent mobility of sensor nodes is another reason for the communication voids. Good protocols should handle the mobility of nodes effectively for better performance.
- **Energy:** Loss of energy due to communication voids is another area of concern. With constrained battery and inability in frequent recharges, underwater sensor nodes need to be protected from unwanted drainage of energy.

6 Conclusion and Future Work

Underwater Sensor Networks have come out as a prominent research area in the networking domain with a wide range of aquatic applications. Routing in UWSNs is an area of major concern due to its unique features such as continuous node mobility, frequent disruption of links, and interference caused by other underwater acoustic systems such as marine mammals. Recently in UWSNs, opportunistic routing has been accepted as an efficient routing approach. The paper discussed the functioning of various Location-Based Opportunistic Routing Protocols proposed for UWSNs and evaluated the performance of two major protocols, VBF and HH-VBF using simulations in Aqua-Sim. The performances of these protocols are evaluated in normal network scenarios and also in networks with communication voids. We could observe that the performance of these protocols comes down with communication voids in the network. Insights from this article would enable researchers to work toward designing more effective protocols for underwater environments having voids. In the near future, a novel routing protocol, especially

designed for UWSN operating scenarios would be proposed and compared with existing routing protocols in terms of PDR, Energy, Delay, and Throughput and above all will give good performance in networks with communication voids and that are highly secure from all sorts of intruders and man-in-middle attacks.

References

1. Chen, Y., Jin, X., Xu, X.: Energy-efficient mobile data collection adopting node cooperation in an underwater acoustic sensor network. *China Commun.* **14**(6), 32–42 (2017)
2. Wang, Z., Han, G., Qin, H., Zhang, S., Sui, Y.: An energy-aware and void-avoidable routing protocol for underwater sensor networks. *IEEE Access* **6**, 7792–7801 (2018)
3. Akyildiz, I.F., Pompili, D., Melodia, T.: Underwater acoustic sensor networks: research challenges. *Ad Hoc Netw.* **3**(3), 257–279 (2005)
4. Akyildiz, I.F., Pompili, D., Melodia, T.: State-of-the-art in protocol research for underwater acoustic sensor networks. In: *Proceedings of the 1st ACM International Workshop on Underwater networks—WUWNet '06* (2006)
5. Açar, G., Adams, A.: ACMENet: an underwater acoustic sensor network protocol for real-time environmental monitoring in coastal areas. *IEE Proc. Radar Sonar Navig.* **153**(4), 365 (2006)
6. Partan, J., Kurose, J., Levine, B.N.: A survey of practical issues in underwater networks. In: *Proceedings of the 1st ACM International Workshop on Underwater Networks—WUWNet '06* (2006)
7. Biswas, S., Morris, R.: ExOR. *ACM SIGCOMM Comput. Commun. Rev.* **35**(4), 133 (2005)
8. Bruno, R., Conti, M., Nurchis, M.: Opportunistic packet scheduling and routing in wireless mesh networks. In: *2010 IFIP Wireless Days* (2010)
9. Chakchouk, N.: A survey on opportunistic routing in wireless communication networks. *IEEE Commun. Surv. Tutor.* **17**(4), 2214–2241 (2015)
10. Nayyar, A., Bath, R.S., Ha, D.B., Sussendran, G.: Opportunistic networks: present scenario—a mirror review. *Int. J. Commun. Netw. Inf. Secur. (IJCNIS)* **10**(1), 223–241 (2018)
11. Menon, V.G., Prathap, P.M.: Comparative analysis of opportunistic routing protocols for underwater acoustic sensor networks. In: *2016 International Conference on Emerging Technological Trends (ICETT)* (2016)
12. Menon, V.G.: *Opportunistic Routing Protocols in Underwater Acoustic Sensor Networks: Issues, Challenges, and Future Directions*. *Magnetic Communications: From Theory to Practice*, pp. 127–148. CRC Press, Boca Raton (2018)
13. Menon, V.G., Prathap, P.M.: *Moving From Topology-Dependent to Opportunistic Routing Protocols in Dynamic Wireless Ad Hoc Networks: Challenges and Future Directions. Algorithms, Methods, and Applications in Mobile Computing and Communications*, pp. 1–23. IGI Global, Hershey (2017)
14. Menon, V.G.: Analyzing the performance of random mobility models with opportunistic routing. *Adv. Wirel. Mob. Commun.* **10**(5), 1221–1226 (2017)
15. Han, M.K., Bhartia, A., Qiu, L., Rozner, E.: O3. In: *Proceedings of the Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing—MobiHoc '11* (2011)
16. Menon, V.G., Prathap, P.M.: Survey on latest energy based routing protocols for underwater wireless sensor networks. *Int. J. Comput. Netw. Wirel. Commun.* **6**(6), 52–55 (2017)
17. Ayaz, M., Abdullah, A., Faye, I., Batira, Y.: An efficient dynamic addressing based routing protocol for underwater wireless sensor networks. *Comput. Commun.* **35**(4), 475–486 (2012)

18. Yan, H., Shi, Z. J., Cui, J.: DBR: depth-based routing for underwater sensor networks. In: NETWORKING 2008 Ad Hoc and Sensor Networks, Wireless Networks, Next Generation Internet, pp. 72–86 (2008)
19. Jafri, M.R., Sandhu, M.M., Latif, K., Khan, Z.A., Yasar, A.U., Javaid, N.: Towards delay-sensitive routing in underwater wireless sensor networks. *Proc. Comput. Sci.* **37**, 228–235 (2014)
20. Wang, C., Zhang, G., Zhang, L., Shao, Y.: Improvement research of underwater sensor network routing protocol HHVBF. In: 11th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2015) (2015)
21. Xie, P., Cui, J., Lao, L.: VBF: vector-based forwarding protocol for underwater sensor networks. In: NETWORKING 2006. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems (2006)
22. Ghoreyshi, S., Shahrabi, A., Boutaleb, T.: A novel cooperative opportunistic routing scheme for underwater sensor networks. *Sensors* **16**(3), 297 (2016)
23. Nicolaou, N., See, A., Xie, P., Cui, J., Maggiorini, D.: Improving the robustness of location-based routing for underwater sensor networks. In: OCEANS 2007—Europe, pp. 1–6. Aberdeen (2007)
24. Ghoreyshi, S.M., Shahrabi, A., Boutaleb, T.: Void-handling techniques for routing protocols in underwater sensor networks: survey and challenges. *IEEE Commun. Surv. Tutor.* **19**(2), 800–827 (2017)
25. Menon, V.G., Joe Prathap, P.M.: Opportunistic routing with virtual coordinates to handle communication voids in mobile ad hoc networks. In: Advances in Intelligent Systems and Computing, pp. 323–334 (2015)
26. Ghoreyshi, S.M., Shahrabi, A., Boutaleb, T.: An opportunistic void avoidance routing protocol for underwater sensor networks. In: 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA) (2016)
27. Menon, V.G., Prathap, P.M.: A review on efficient opportunistic forwarding techniques used to handle communication voids in underwater wireless sensor networks. *Adv. Wirel. Mob. Commun.* **10**(5), 1059–1066 (2017)
28. Darehshoorzadeh, A., Boukerche, A.: Underwater sensor networks: a new challenge for opportunistic routing protocols. *IEEE Commun. Mag.* **53**(11), 98–107 (2015)
29. Hwang, D., Kim, D.: DFR: directional flooding-based routing protocol for underwater sensor networks. In: OCEANS 2008 (2008)
30. Coutinho, R.W., Boukerche, A., Vieira, L.F., Loureiro, A.A.: GEDAR: geographic and opportunistic routing protocol with depth adjustment for mobile underwater sensor networks. In: 2014 IEEE International Conference on Communications (ICC) (2014)
31. Chen, Y.S., Juang, T.Y., Lin, Y.W., Tsai, I.C.: A low propagation delay multi-path routing protocol for underwater sensor networks. *J. Internet Technol.* **11**, 153–165 (2010)
32. Hao, K., Jin, Z., Shen, H., Wang, Y.: An efficient and reliable geographic routing protocol based on partial network coding for underwater sensor networks. *Sensors* **15**(6), 12720–12735 (2015)
33. Noh, Y., Lee, U., Wang, P., Choi, B.S., Gerla, M.: VAPR: void-aware pressure routing for underwater sensor networks. *IEEE Trans. Mob. Comput.* **12**(5), 895–908 (2013)
34. Jornet, J.M., Stojanovic, M., Zorzi, M.: Focused beam routing protocol for underwater acoustic networks. In: Proceedings of the Third ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization—WuWNeT '08 (2008)
35. Li, Z.L., Yao, N.M., Gao, Q.: Relative distance-based forwarding protocol for underwater wireless sensor networks. *Appl. Mech. Mater.* **437**, 655–658 (2013). <https://doi.org/10.4028/www.scientific.net/amm.437.655>

36. Nayyar, A., Puri, V., Le, D.N.: Comprehensive analysis of routing protocols surrounding underwater sensor networks (UWSNs). In: *Data Management, Analytics and Innovation* (pp. 435–450). Springer, Singapore (2019)
37. Nayyar, A., Singh, R.: A comprehensive review of simulation tools for wireless sensor networks (WSNs). *J. Wirel. Netw. Commun.* **5**(1), 19–47 (2015)

A Hybrid Optimization Algorithm for Pathfinding in Grid Environment



B. Booba, A. Prema and R. Renugadevi

Abstract Grid computing has been highly effective in the area of life sciences, financial analysis, research collaboration, and engineering. This paper is a study of existing algorithms like Swarm Intelligence (SI) algorithms such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC-PSO), and Parallel Particle Swarm Optimization (PPSO) to opt for the optimal path in a grid computing environment. These algorithms were used to solve the complex optimization problems in finding the path between source node to destination node effectively. Nature computing techniques based on the study of the collective behavior of ants, particle swarms, and bees are used to find the optimal path, improve the optimization methods and scalability in a set of representative problems. The hybridization of a grid computing environment and nature-inspired computing algorithms such as ACO, PSO, ABC-PSO, and PPSO has resulted in a class of solutions that differ in structure and design from the peer-to-peer network algorithms and the evaluated results showed the effectiveness of the pathfinding problem. ACO is implemented on a dynamic grid computing environment to demonstrate scalability and a solution for pathfinding. A class of four algorithms is used to find an optimal path and improve the optimization methods and shorten the computational time in a grid computing environment.

Keywords ACO · PSO · ABC-PSO · PPSO · Grid computing · Nature-inspired computing · Pathfinding · Swarm intelligence

B. Booba (✉) · A. Prema

Department of Information Technology, School of Computing Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, India
e-mail: boobarajashekar@gmail.com

A. Prema

e-mail: unjanai@gmail.com

R. Renugadevi

Department of Computer Science, School of Computing Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, India
e-mail: nicrdevi@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_50

713

1 Introduction

1.1 Overview of Grid Computing

Grid computing is a platform for solving computational problems in science, engineering, commerce, and other fields. Numerous solutions to grid computing are being developed. Grid computing has been proven to be highly effective in the area of life sciences, financial analysis, research collaboration, and engineering [1].

Grid computing is the accumulation of computer resources to achieve a common goal. The grid provides location-transparent scheduling that is a measure of registering resources required at a determined time interval for computation. Grid capacity has to make more cost-effective utilization of a given measure of computer resources as an approach to tackle problems that cannot be approached without a huge measure of computing power. The fact that it utilizes the resources of numerous workstations might be helpful synergistically to harness and manage as cooperation to a typical destination [2].

1.2 Problem Definition

The objectives are to find the grid computing based-pathfinding, identifying the position of the obstacles being developed on the grid computing environment and comparing, analyzing the performance results of the class of four algorithms in nature-inspired computing. This paper focuses on motion planning for “pathfinding”. It addresses the optimal pathfinding problem in a grid computing environment [3]. This paper focuses on the use of SI algorithms such as ACO, PSO, PPSO, and hybrid ABC–PSO to find the optimization problems for selecting the shortest path in a grid computing environment. These algorithms utilize the swarm intelligence in grid environment to obtain an optimal solution over an iterative process. The optimal solutions are evaluated using different kinds of aspects, the hybrid ABC–PSO has been evaluated against other SI algorithms. A conclusion has been shown that the hybrid ABC–PSO algorithm performs better in terms of intensification and diversification concepts. This paper is to demonstrate robustness, scalability of the nature-inspired pathfinding algorithms in a customized grid computing environment.

1.3 Proposed Work

The pathfinding algorithms are analyzed to stumble on the shortest path in grid computing milieu from source node to destination node even if there are any

obstacles. This paper focuses on the use of ACO, ABC-PSO, and PPSO for selecting the shortest path in a grid computing environment [4].

The following approaches are used in the proposed work:

- Election of shortest path using ACO algorithm.
- Selection of shortest path by means of Parallel Particle Swarm Optimization (PPSO) algorithm.

The above mentioned approaches were used to solve the complex optimization problem in pathfinding between source nodes to destination node effectively.

2 Related Works

ACO methods are useful in reducing the problems of finding paths to a goal state. I showed the behavior of real ants, and to provide heuristic solutions for optimization problem. Cellular-DPSABC algorithm is able to apply dynamic optimization problems. The principle of this algorithm was implemented in PSO to fitness value population in Artificial Bee Colony Algorithm (ABC). This projected algorithm can be applied for solving optimization problems. The new algorithm is simple and flexible swarm-based one. It is also robust, it was established on a limited set of test problems [5].

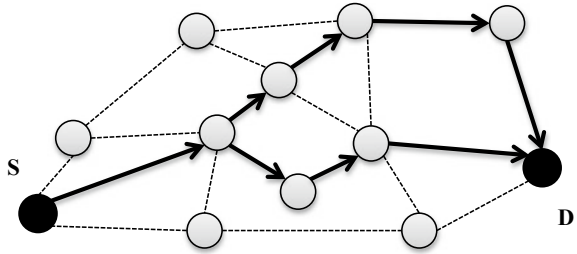
An abstraction mechanism which maintains memory competence using a large-scale grid. It is also effective in reducing planning costs. This mechanism called a Minimal-Memory (MM) abstraction distinguishes the abstraction representation of the techniques which use the abstraction for pathfinding. Memory-efficient way of precomputing subgoals reduces the main obstacle in applying real time search systems in video games. Bee Colony Optimization (BCO) algorithm is focused on obstacle prevention and shortest safe path from the source to the target. Field D^* can be used to approximate a distance find on the nodes of a complex, and the distance field can be used to weight the complex [6].

3 Contribution of the Paper

3.1 Design of ACO Algorithm for Pathfinding

ACO is a streamlining algorithm in light of the activities of strategies which are to diminishing the issues of finding the paths to a goal state. ACO basically is an artificial system that takes motivation from the conduct of genuine states. It locates the most efficient routes from their nests to food sources [7]. The attribute of ACO algorithms is the permutation of the structure of a solution. ACO algorithms are

Fig. 1 Grid pathfinding



capable of solving minimum cost path problems on more complicated graphs as shown in Fig. 1.

ACO is capable of solving the large computational problem in grid computing [8]. When all the jobs are assigned, the resources are allocated by the pheromone trails. The pheromone trail can be proposed from the source node while traveling through several nodes to the target node.

The trails are updated by

$$\tau_{ab} \leftarrow (1 - p)\tau_{ab} + \sum_m \Delta_{\tau_{ab}}^m \tag{1}$$

Equation (1) is the amount of pheromone deposited for a state transition, ab is the pheromone evaporation coefficient and τ_{ab} is the amount of pheromone deposited by m th system [9].

3.2 Design of ABC-PSO Algorithm

The benefit of ABC over other swarm intelligence approaches is that in the ABC algorithm the imaginable solutions indicate food sources. A total number of employed bees are identical to the number of nourishment sources, with each working honey bee speaking to a sustenance source [10]. Nourishment source signifies an answer for the issue that is fundamental to be upgraded. Every sustenance source is produced as pursues

$$X_{ij} = X_{\min j} + \text{rand}[0, 1](X_{\min j} - X_{\min j}) \tag{2}$$

ABC-PSO, two algorithms' components smoothly with the end goal to pick up profit by their specific qualities Fig. 2. The incorporating PSO with ABC is to build goal of the pbests of the particles. The algorithm with segments from both ABC and PSO is shaped with the end goal to have a calculation that easily tackles distinguishable issues as while having iteratively invariant conduct as PSO in the meantime [11].

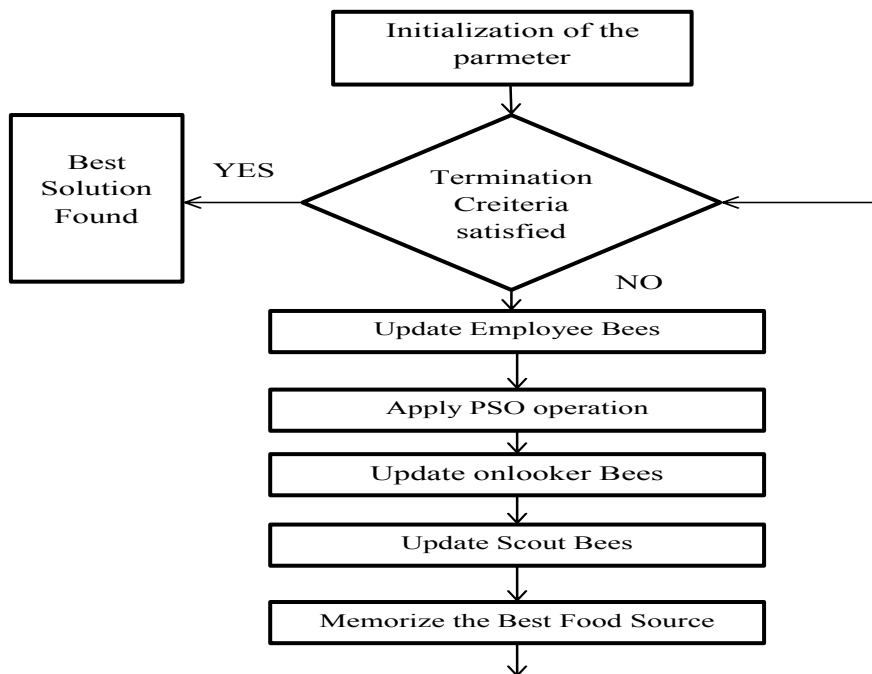


Fig. 2 ABC-PSO algorithm

For every molecule i in the swarm, the ABC calculation method refresh condition is down to business to the individual best $pbest_i$. This is done after iteratively choosing another molecule k , arbitrary issue variable j , hence $cpbest_i$ is refreshed as pursues

$$pbest_{ij} = pbest_{ij} + \phi_{ij} X(pbest_{ij} - pbest_{kj}) \quad (3)$$

3.3 Design of PPSO Algorithm

Parallel processing is to generate the same results and feasibly uses multiple processors to reduce the run time. Each time, all particles are independent it can be easily analyzed in parallel [12]. The PPSO algorithm solves complex functional optimization problems. The particles in PPSO algorithm are attracted to the best locations in the search space [13]. The process of PPSO group is shown in Fig. 3.

The above technique gives the worldwide best and the neighborhood best arrangements [14]. It builds the shot of particles to find a superior arrangement and to expand the nearby ideal arrangements.

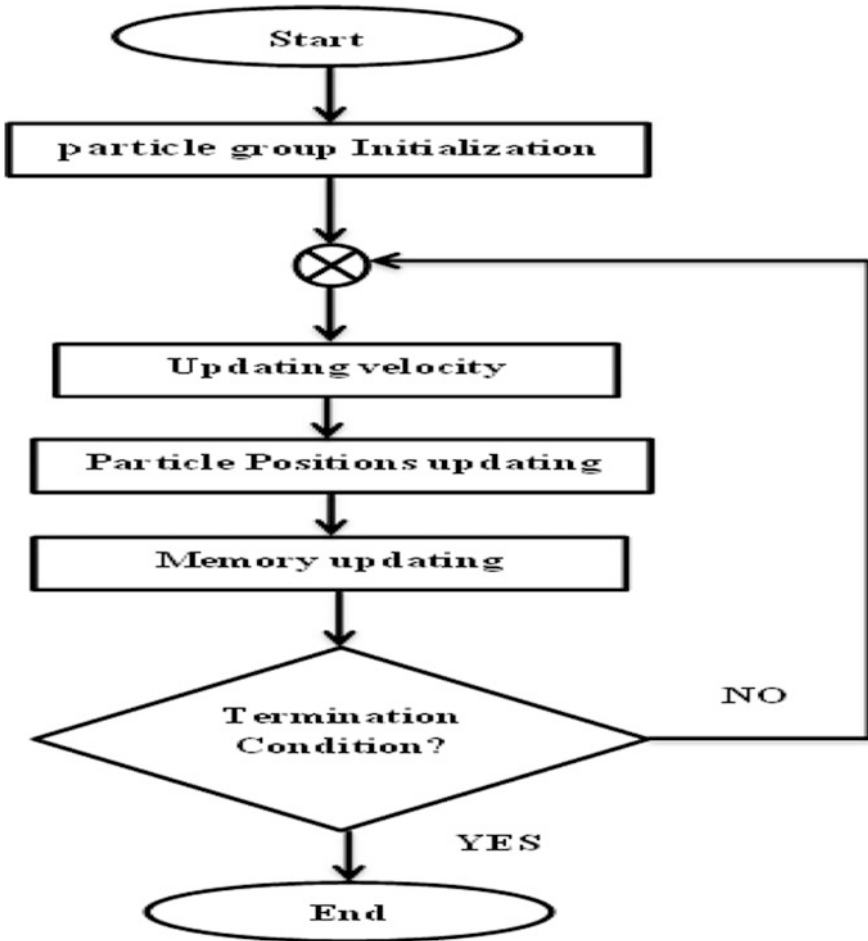


Fig. 3 The process of parallel particle swarm optimization/group

4 Simulation Results

In this work, MATLAB tool is used to build the simulation environment and the proposed techniques are simulated to evaluate the performance of Make span and to find an optimal path [15].

$$\text{Make span} = \text{Completion time of last job} - \text{Starting time of first job}$$

To evaluate the effectiveness of these algorithms, the average deviation of the present best result from the optimum in all iterations is determined. The yield of the dynamic condition gives the premise on which most limited ways can be figured between any two hubs utilizing pheromone esteems at every one of their neighbor hubs [16, 17]. The system show utilized in unique is made out of the hubs and connections that are viewed as bidirectional. It very well may be finished by dissecting the likelihood make length esteems [18].

5 Performance Analysis

The Proposed technique is replicated for various combinations of grid sizes such as 128×8 , 256×16 , 512×32 , and 1024×64 . The results are shown in Table 1. It demonstrates the performance of the ACO, PSO, ABC-PSO, and PPSO algorithms. This performance measure produces a fast rate of better optimization make span [19].

Figure 4 defines a method to find the solution for the problem of resource with total execution times which equal to the make span. According to this, the optimal solution for each algorithm is found based on the global pheromone.

Table 1 Probability make span of ACO, PSO, ABC-PSO, and PPSO

Probabilistic make span of ACO		Probabilistic make span of PSO		Probabilistic make span of ABC-PSO		Probabilistic make span of PPSO	
No. of iteration in matrice	Value	No. of iteration in matrice	Value	No. of iteration in matrice	Value	No. of iteration in matrice	Value
[0][0]	818	[0][0]	799	[0][0]	470	[0][0]	580
[0][1]	652	[0][1]	550	[0][1]	320	[0][1]	520
[0][2]	563	[0][2]	490	[0][2]	290	[0][2]	450
[0][3]	506	[0][3]	476	[0][3]	281	[0][3]	410
[1][0]	491	[1][0]	359	[1][0]	260	[1][0]	310
[1][1]	404	[1][1]	350	[1][1]	251	[1][1]	280
[1][2]	348	[1][2]	332	[1][2]	234	[1][2]	290
[1][3]	343	[1][3]	312	[1][3]	219	[1][3]	260
[2][0]	309	[2][0]	275	[2][0]	170	[2][0]	230
[2][1]	280	[2][1]	230	[2][1]	140	[2][1]	190
[2][2]	176	[2][2]	141	[2][2]	121	[2][2]	160
[2][3]	163	[2][3]	132	[2][3]	76	[2][3]	99

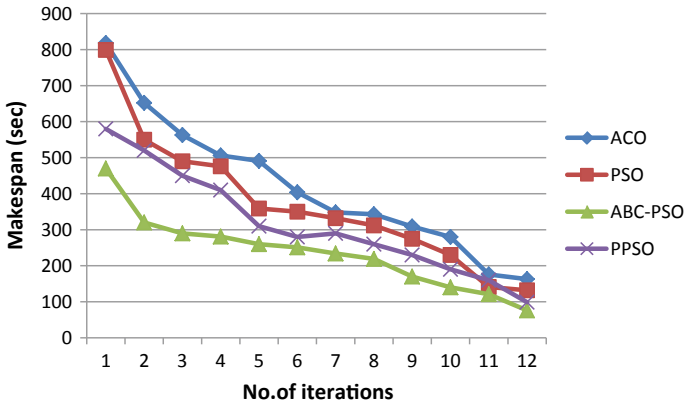


Fig. 4 ACO, PSO, ABC-PSO, and PPSO performance

6 Conclusion

This work reports the application of ACO for pathfinding in a simulated grid computing environment which is suitable for comparison with PSO, ABC-PSO, and PPSO algorithms. The hybridization of grid computing environment and swarm intelligence algorithms such as ACO, PSO, ABC-PSO, and PPSO has resulted in a class of solutions that differ in structure and design from the peer-to-peer network algorithms. The distinction between ant colony and bee colony is also studied through simulation studies in MATLAB. ACO is implemented on a dynamic grid computing environment to demonstrate the scalability of the proposed solution for pathfinding. For the future, this can be implemented in a cloud environment to avoid obstacles freely and improve the scalability.

References

1. Dorigo, M., Stutzle, T.: Ant colony optimization: overview and recent advances. In: Gendreau, M., Potvin, Y. (eds.) *Handbook of Metaheuristics*, 2nd edition, vol. 146 in International Series in Operations Research & Management Science, pp. 227–263. Springer, Verlag, New York (2010)
2. Reshamwala, A., Vinchurkar, D.P.: Robot path planning using an ant colony optimization approach: a survey. *Int. J. Adv. Res. Artif. Intell.* **2**, 65–71 (2013)
3. Zafarani-Moattar, E., Haj-Seyed-Javadi, H., Feizi-Derakhshi, M.R.: Parallel reverse niching PSO for multimodal optimization. In: *International Conference on Machine Learning, Electrical and Mechanical Engineering (ICMLEME'2014)* (2014)
4. Zhang, J.-R., Zhang, J., Lok, T.-M., Lyu, M.R.: A hybrid particle swarm optimization-back-propagation algorithm for feed forward neural network training. *Appl. Math. Comput.* (2007)

5. Li, M., Duan, H.: Hybrid artificial bee colony and particle swarm optimization approach to protein secondary structure prediction. In: Proceeding of the 10th World Congress on Intelligent Control and Automation (2012)
6. Sood, M., Kaur, M.: Shortest path finding in country using hybrid approach of BBO and BCO. *Int. J. Comput. Appl.* **40**(6), 0975–8887 (2012)
7. Mundra, P.S.: Ant colony optimization: a technique used for finding shortest path. *Int. J. Eng. Innov. Technol.* **1**(5) (2012)
8. Baktash, N., Mahmoudi, F., Meybodi, M.R.: Cellular PSO-ABC: a new hybrid model for dynamic environment. *Int. J. Comput. Theory Eng.* **4**(3) (2012)
9. Kumar, S., Sharma, V.K., Kumari, R.: Comparative study of hybrids of artificial bee colony algorithm. *Int. J. Inf. Commun. Comput. Technol.* **1**(2) (2013)
10. Uras, T., Koenig, S., Hernandez, C.: Subgoal graphs for optimal pathfinding in eight-neighbor grids. Association for the Advancement of Artificial Intelligence (2014)
11. Bulitko, V., Björnsson, Y., Lawrence, R.: Case-based subgoaling in real-time heuristic search for video game pathfinding. *J. Artif. Intell. Res.* **39**, 269–300 (2010)
12. Dorigo, M.: The ant colony optimization metaheuristic: algorithms, applications, and advances. International Series in Operations Research & Management Science (2003)
13. Ameer, M.S.B., Sakly, A., Mtibaa, A.: Implementation of real coded PSO algorithms using FPGA technology. In: 2014 15th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA) (2014)
14. Sadhasivam, G.S., Meenakshi, D.K.: Load balanced, efficient scheduling with parallel job submission in computational grids using parallel particle swarm optimization. In: 2009 World Congress on Nature & Biologically Inspired Computing (2009)
15. Bentley, P.J.: Perceptive particle swarm optimization: an investigation. In: Proceedings 2005 IEEE Swarm Intelligence Symposium 2005 SIS 2005, 2005 Publication Chu Shu-Chuan. Parallel Particle Swarm Optimization Algorithms with Adaptive Simulated Annealing, Studies in Computational Intelligence (2006)
16. Chu, S.-C.: Parallel Particle swarm optimization algorithms with adaptive simulated annealing. Studies in Computational Intelligence (2006)
17. Booba, B., Gopal, T.V.: Comparison of ant colony optimization & particle swarm optimization in grid scheduling. *Aust. J. Basic Appl. Sci.* **8**(7) (2014)
18. Booba, B., Gopal, T.V.: Efficient scheduling of packets in wireless sensor networks using priority based scheduling approach. *J. Comput. Sci.* (2014)
19. Booba, B., Gopal, T.V.: An efficient distributed computing technique for job scheduling is accepted by the journal. *Int. J. Appl. Environ. Sci. (IJAES)*

Dynamic Hashtag Interactions and Recommendations: An Implementation Using Apache Spark Streaming and GraphX



Sonam Sharma

Abstract Hashtag, started with Twitter is a keyword with prefix “#” and now being used mostly for all communication on social media. It has been identified as very powerful and effective in organizing communications according to the topic and trend. Hashtag can further help on various analysis, as it links users with their topic of interests. Hashtag aids in building communities of similar interests. With hashtags, we can follow current trend and interest on twitter which can help us in analyzing multiple factors, e.g., sensitivity of the ongoing trend, its spread, people getting affected, its effect on business and so on. Traditionally available approaches help us in analyzing batch data and finding interests and trends on it. Now with the advancements in the field of technology helps us in analyzing a large amount of online data within seconds. In this paper, we will be exploring dynamic hashtag interactions to find correlations among them and propose a methodology which can successfully find relevant hashtags based on the interest in focus. We will propose our methodology of analyzing and exploring tweets in real time with the extent of converting information; we are getting from twitter to knowledge.

Keywords Co-occurrence graph · Community networks · Hashtags · Graph mining · Real-time analytics · Streaming graph · Social media

1 Introduction

GRAPH structures have many application areas named cybersecurity, social network, e-commerce websites, etc. With the abundant data getting generated in wide variety, with different sets of volumes and rates. The increased usage and connectedness of social media have given rise to trends and interests which can be triggered for various reasons. Analyzing social media data whether in batch or in real time with this freely available interesting public information could yield

S. Sharma (✉)
Tata Consultancy Services, Noida, India
e-mail: sonam.sharma2@tcs.com

interesting results. People on social media have been expressing their feelings freely, which makes Twitter an ideal source for accumulating wide amount of data for real-time on-time analysis on the vast amount of opinions and on a wide range of topics. Although this information can be easily collected, getting the information from it and representing it in good form has its challenges. These challenges withstand even with the advancement of tools and technologies. Additionally, analyzing twitter data is indeed a challenging problem due to the nature, diversity, and volume of the data.

Whenever we talk about social media, the combination of graphical data and big data is inevitable and is the reason it is popularly known as Social Network Analysis (SNA). There are various types of graph structures that are possible where the most popular ones are user-followers graph, user-mentions graph, etc. But here we are basically building and analyzing the graphs based on hashtags for a particular trend, which can further help us in analyzing the trending hashtags and other related hashtags. We are focusing on open-source software technologies named Spark which provides a graphical layer on top of it.

Data Streams have gain popularity in data mining techniques, where time is the most relevant factor since such data helps in providing *on-time* analysis and focus on *now*. But it is a very challenging task in various aspects [1]:

1. Dynamic nature of the data streams can cause data distribution changes with time (called concept drift) [2]
2. It not feasible to store the data since the data is in infinite volume
3. The analysis of data streams must happen as quickly as possible so as to get quick insights and operate in real time.

We can also see that when it comes to real-time analytics, many canonical analytics approaches may not be suitable for handling real-time data and providing *on-time* analytics. Here, we require a knowledge-based system for social network analysis which can handle streamed data and find interesting patterns on it and deliver the results within seconds [3, 4].

It is expected that the new approaches developed for extracting insights from real-time data will help in advancing the development of more sophisticated analytics allows better decision-making. This will help in gaining insights from your data as you receive with improved decision-making. This will help customers in quick and immediate decisions based on the scenario that is happening now, not what happened weeks, or months ago. Most of the big data that we receive are in real time and it is most valuable at the time of arrival (Fig. 1).

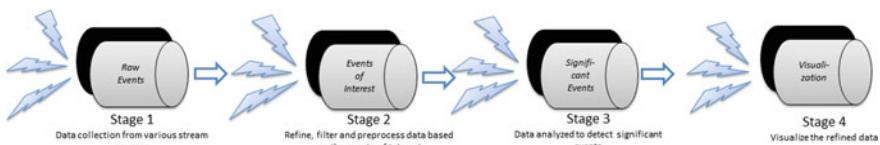


Fig. 1 A canonical real-time event processing pipeline

In any canonical real-time event processing pipeline at every stage, we polish for further step and at every stage we try to refine the data in a way to get new insights from it which can further help end users for decision-making. This streaming data can help in building final products which provide *actionable on-time insights*. The initial stage is to collect data from the various stream sources which is the first stage of any data processing pipeline. The main challenge with stream data preprocessing is that new data continuously arrives and we need to apply fully automated preprocessing methods. Also, we need to be able to update the methods with evolving data and optimize model parameters to deal with rapidly changing data. Streaming data needs to be fully automated with the required preprocessed methods.

In this paper, we present various methodologies on a linguistic process for building knowledge-based social media framework based on hashtag interactions. The framework works seamless with upcoming big data technologies and software. The state-of-the-art methodologies in the data analytic techniques on social media data are provided.

This paper is divided into the following sections: Sect. 2 discusses literature review and critical research gaps highlighted in this paper. In Sect. 3, the detailed problem statement for our analysis is defined. The proposed methodology is stated in Sect. 4. The results and conclusions are explained in Sects. 5 and 6 [5].

2 Literature Review

Activity on social media platform has become a part of daily activity of users because of which it has become one of the most powerful mediums for many researchers and enterprises to perform extensive analysis and derive interesting results. In this section a detailed study about various techniques on knowledge discovery using hashtags on social media. Although the problem of knowledge mining from twitter has been studied extensively during recent years and some we discuss below. Wei Wu designed a system for automatically generating personalized annotation tags to label Twitter user's interests and concerns [6]. However, user-level tags have its limitations as it only focuses on the particular user and their content and Twitter being a very large community, where we can focus on communities or on a large set of people which can help us in building strong knowledge-based system. Other such kinds of implementations have been performed on key-phrase extraction which determines important phrases from single documents or an entire collection of text documents [7, 8]. These approaches are not immediately applicable on Twitter because even if Twitter offers textual information but canonical NLP techniques cannot work on it as people on twitter tend not to follow basic norms of grammar on which we can do language processing. Apart from this, if we focus on single tweet for analysis we will most of the information which is being generated on Twitter.

However, current explorations being dined on social media are more matured than ever before and still it is one of the most chosen topics for researchers because

of its freely available data and whenever there is twitter, hashtag plays the most significant role in tweets and now people have started to understand this and follow it across various social media platforms. It has been identified that hashtag inclusion in the posts are increasing nowadays, because of which hashtags act as the fundamental information facet associated with social network messages [9]. Made a significant contribution by observing that hashtag recommendations aim to encourage the user to use hashtags more often and it is better to use appropriate hashtags and also avoiding synonymous in hashtags. All these observations indicate that the hashtag recommendation is not only interesting but an important problem statement.

Also, most of these existing solutions suffer from certain limitations. One problem is that the majority of approaches are limited to centralized environments. The proposed techniques are very time-consuming and take many computational resources. Most of the available solutions are neither sufficient nor suitable for building total knowledge analytics system on Twitter, since there is a huge mismatch between their processing capabilities and when we get the output from them, we lose the value (the fifth V of big data) of the data since that is not *on-time*.

Thus, in all, it can be stated that they are solving the purpose of providing knowledge to the user but all the knowledge that is provided is either insufficient for the business users to go with or not on-time. Most of the methodologies that are developed on Twitter are on batch data. With the advent of the technologies, we can modify or build these methods on the top of stream data to help the business in providing on-time and social media presence using this.

3 Problem Statement

(a) *Proliferation of trends on social media*

Trends in social networks are elements of communication which helps in analyzing the popularity of a particular event (news) on social media. The more popular the event is, the more the occurrence of interactions on social media is. On trending events, people express their opinions which further helps them to build community and helps in the spread or proliferate the trend.

Most often hashtags help in the spread of trend and there are various combinations of hashtags which help people in expressing their thoughts clearly and be able to connect to the like-minded people. Sometimes trends can be misleading and biased too as it is basically the opinions which people are expressing what they feel and their personal thoughts. Understanding the dynamics of trend spreading could be of the great value of information for social media research areas.

(b) *Objective*

The objective of this paper is to propose a methodology to analyze hashtag interactions dynamically and graphically detect current ongoing trends on Twitter by building real-time hashtag co-occurrence graph on the top of real-time streaming engine. Real-time co-occurrence graph can help in analyzing popular hashtags. For this study, we have used big data technologies like *Spark Streaming* and *GraphX*. Graph visualization libraries like *graphstream* are used to visualize the graph in real time. Only the activities of Twitter after the connection are made. We analyze the streaming pipeline for each sliding window and perform the analysis. Under such circumstances, it becomes essential to understand each and every aspect of the data beforehand and constantly update the algorithm so as to deal with a future discrepancy.

4 Analytical Methodology for *On-Time* Twitter Analysis Hashtag Recommendations

How should we decide on managing data across an enterprise or an organization which is streamed from multiple sources like various kind of sensors, security data in the form of video from webcam, data from IDS/IPS, telematics data, and so on. Apart from management of data which is getting generated in real time, the other most important task is to analyze and find insights from it the moment it gets generated. Properly managing the data what is the best possible way to find insights from it as soon as it gets generated, so that there is no loss of data value and take necessary “*on-time*” actions on it [10].

Whenever a tweet is being twitted on twitter which contains keyword that starts with ‘#’ called hashtags. Hashtags mostly contain current trends or hot topics in discussion. A single hashtag has the capability to describe the whole tweet. And if there are multiple hashtags available in the same tweet it can help in analyzing and getting interesting results. Multiple hashtags can further help us in analyzing user of the tweet. Multiple hashtags are getting created on Twitter and Trending Hashtags getting retweeted which helps in spread of the news faster not just on Twitter platform, but people use these hashtags on other social media sites to increase awareness. Hashtags refer to specific topics related to particular ongoing event or trend, but only a few of them become too exclusive and popular. People use hashtags to show their interest in a particular topic. Hence, a hashtag can provide valuable information about the topic, its evolution, and what user community of the analyzed hashtag is interested in through time.

Experiments and their results with real-time twitter data and analysis being done here provide real-time recommendations on what hashtag to follow next based on the topics which can help in building the related hashtags network and organization of trend. Finally, we discuss techniques that can describe the way hashtags, related to the same event, evolve through time and what are the most recommended

Table 1 Sample keywords to be crawled from twitter

Tradewar	Tradewars	Trump	trade
----------	-----------	-------	-------

Table 2 Combination of hashtags considered

US AND trade	US AND China	Trump AND tradewar	China AND tradewar
--------------	--------------	--------------------	--------------------

hashtags. Keeping this in mind, we now explain the procedure which we followed for our experiment.

(A) *Data*

Tables 1 and 2 have listed the keywords used to crawl the data from Twitter. Where Table 1 shows specific keyword whereas Table 2 shows combination of these keywords to further enhance our analysis.

We have built our analysis from July 16 to Sept 12 2018. And we will be showcasing the results for the 6-min window in spark streaming. We have decided this window based on the amount of data and cluster size so that we can get seamless and fruitful results. These hashtags will be the *Focus Point* for our analysis and will further be enhanced by building the real-time graph and getting related Hash-tags on it.

(B) *Real-time Tweet Statistical Analysis and Data-set Preparation*

The next steps we perform will further provide ongoing real-time analysis on the data. Since our analysis majorly focused on hashtags, so we have filtered hashtags and build our analysis on top of it with some other information like users who have used this hashtag and number of tweets, this hashtag has been used shown in Fig. 2, which gives us an understanding that how important the analyzed topic used by the



Fig. 2 A Map() collection displaying the Hashtag -> (#unique tweets, #unique users) for some specific window

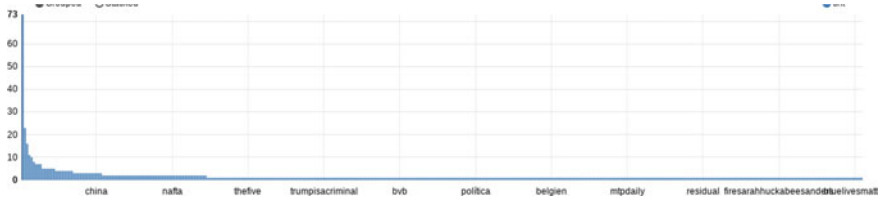


Fig. 3 Bar chart representation of hashtag distribution on some *window()* in Spark Streaming



Fig. 4 Classes of hashtags at a particular time window

hashtag is for this particular window of stream. In this section, our main goal is to get an aware of the streaming twitter data from different angles to get the full view (Figs. 3, 4).

For further analysis, we have to focus on major top hashtags and their co-occurrences on tweet. And for this analysis, we have filtered the tweets which contain no hashtags even when they can help us in analyzing the content more but these tweets do not fit our use case.

5 Hashtag Interaction and Recommendation

Here, we will look at the major top hashtags and their co-occurrences. For this analysis, we have filtered the tweets which have no hashtags available with them (Fig. 5) [11, 12].

While analyzing the tweets for multiple *window()*, we have identified that some hashtags which are particularly in trend and popularly occur together or individually for a specific period of time. This makes them most relevant among all. We have also added pairs count in the Edge property of the graph for its use in further analysis.

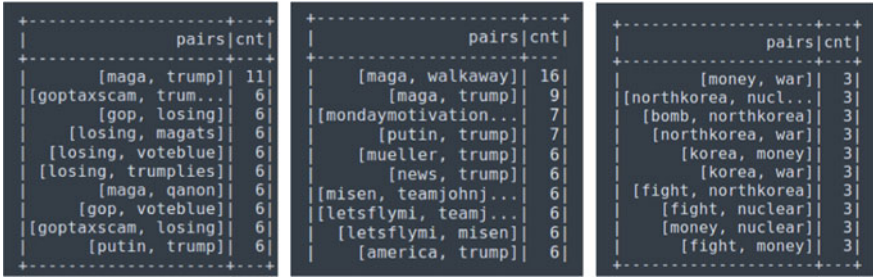


Fig. 5 Shows the top hashtag co-occurrences at different time window()

6 Architecture Flow of the Analysis

The streaming data we have used for our analysis and provided use case is taken from Twitter using Twitter Streaming API. Figure 6 shows the components we have used for this analysis. Specific to this use case the architecture is kept understandable with least and default configurations. The graph visualization uses *graphstream* library, which is both easy to integrate and visualize the graph in Spark and can help in streaming too if we will be using bigger batched in spark streaming window, and the graph which is getting visualized is small enough to fit the window. With multiple tweaks between the window sizes in *graphstream*, we were able to successfully visualize the graph [13, 14].

Kafka which is a distributed stream processing engine helps us in getting the streaming data and can also act as buffering the data for specified configuration and all the analyses of the data have been done on *Spark Streaming*.

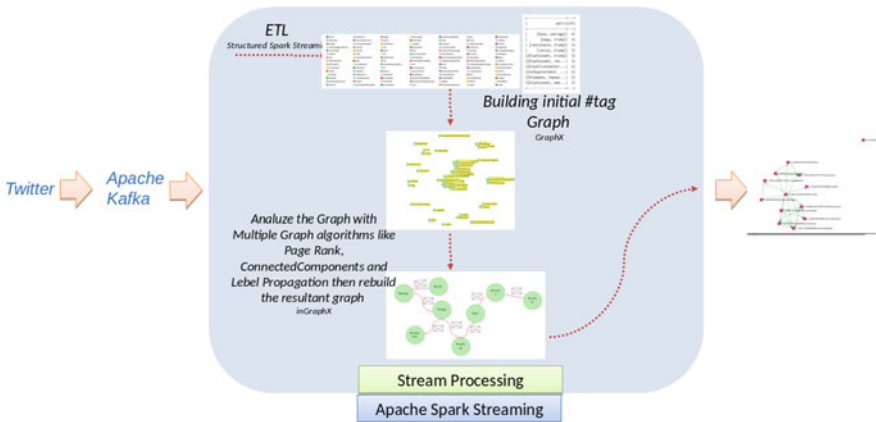


Fig. 6 Architecture defining the flow of analysis

7 Algorithms and Results

In this section, we describe the graph algorithms that we have used to build final hashtag co-occurrence graph for the specific batch window in Spark Streaming.

(a) *Preliminaries*

- (i) For any graph to be built, we have to preprocess and make dataset ready in the desired format for GraphX. For GraphX vertices, we will find *#unique* hashtags and provide them the *id* with *Long* datatype, since GraphX only understands long as a datatype for ID. For this, we have encoded the hashtag string using *MD5 hashing* algorithm so that we get unique id for each hashtag string. We also have to verify that if all hashtags have got different ids since no hashing algorithm can perfectly encode all the strings uniquely. This will make out Vertex RDD in the form of

```
vertexRDD : RDD[(Long, (String))]
```

- (ii) Another preliminary component to build a graph is edges. For this, we will create hashtag *combinations* for each tweet and filter out empty tweets. After creating the *combinations*, we will *flatMap* them and create edge RDD by encoding the hashtags in a similar manner as vertices.

```
edgeRDD : RDD[Edge[Int]]
```

With above vertex and edge RDD's, we will build a graph (Fig. 7)

```
Graph[(String), Int]
```

(b) *Applying Algorithms*

After building the graph we will perform the following analysis to get the final hashtag co-occurrence graph.

- (i) To understand structural property of a graph we apply GraphX *Graph* class provides a number of methods to explore the graph.

(a) **Connected Components**

This is the method which can help us in understanding the graph on its connect-
edness. To describe this, we can say that a graph is called connected when any
vertex in that graph can reach another vertex by following a sequence of edges from
one vertex to another. If not so, then a graph can have multiple connected com-
ponents which can act as a connected subgraph (Fig. 8).

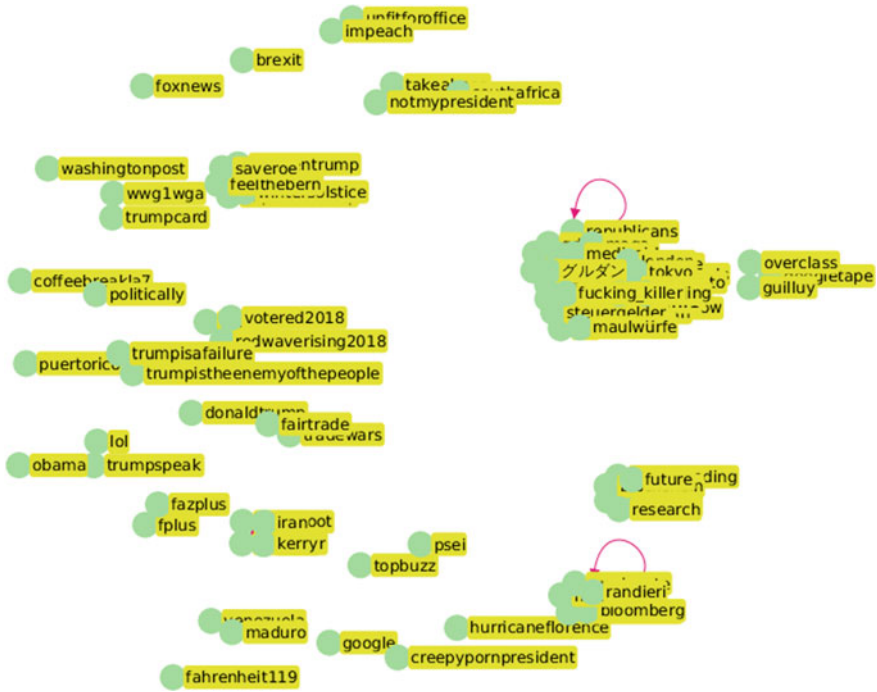


Fig. 7 Result of initial graph created and displayed using graphstream in real-time sliding window

(b) Degree Distribution

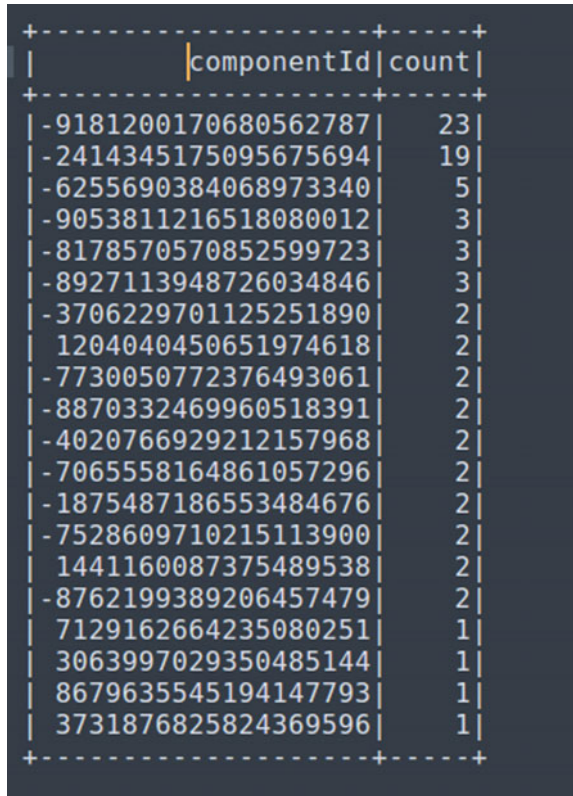
A connected graph can have many different structures. For example, there can be one vertex that is connecting the whole graph. Eliminating the total graph becomes disconnected. So, for all these types of insights degree distribution that is *degree* of each vertex, where degree is defined as a number of edges connected to the vertex (Fig. 9).

(c) Community Detection in GraphX

Community detection algorithms usually focus on determining the graph which can either be overlapping or non-overlapping communities. When the communities are disjoint, there are no common elements we find between the communities. But when communities are overlapping or non-disjoint the elements among the nodes have common elements which they share. In this paper, we are focusing on one of the algorithms available on GraphX, i.e., Label Propagation and we are trying to visualize these using graphstream library (Fig. 10).

How Label Detection works is that it propagates the labels of node throughout the network which helps it to form the communities. The propagation of label is called as label flooding. In label flooding at a point of time single label becomes

Fig. 8 Result of realtime connected components in a GraphX. - 9181200170680562787 is the most connected component with a node count of 23. We have identified that #trump helps in building the component with maximum node count



```
+-----+-----+
| componentId | count |
+-----+-----+
|-9181200170680562787 | 23 |
|-2414345175095675694 | 19 |
|-6255690384068973340 | 5 |
|-9053811216518080012 | 3 |
|-8178570570852599723 | 3 |
|-8927113948726034846 | 3 |
|-3706229701125251890 | 2 |
| 1204040450651974618 | 2 |
|-7730050772376493061 | 2 |
|-8870332469960518391 | 2 |
|-4020766929212157968 | 2 |
|-7065558164861057296 | 2 |
|-1875487186553484676 | 2 |
|-7528609710215113900 | 2 |
| 1441160087375489538 | 2 |
|-8762199389206457479 | 2 |
| 7129162664235080251 | 1 |
| 3063997029350485144 | 1 |
| 8679635545194147793 | 1 |
| 3731876825824369596 | 1 |
+-----+-----+
```

dominant when it is inside the densely populated community which is very well connected where we say that labels are trapped inside. Labels are supposed to be trapped inside whenever it is inside the densely populated network. Labels of each node are treated as opinion. Whenever the opinion is same, therefore the community is same.

In our case whenever the hashtags and their co-occurrence are closely related to trump the form of a community. Apart from trump, there is very less number of communities that are formed which are subsequent across windows of the given timeframe. So for our analysis, we will consider the community containing trump as the most important community and choosing other hashtags connected to it we will move to PageRank which further helps us in ranking the nodes in the graph and within the community.

In the community detection Label propagation algorithm using pregel in spark GraphX we have used a 5 step algorithm since waiting till convergence can hamper the streaming flow of the engine. And we have identified 5 to be the most optimal choice among all.

Fig. 9 Result of degree distribution for a streaming graph. Hashtags *trump* and *maga* are among the highest degree vertices subsequently among various stream windows

A terminal window with a dark background and light-colored text. It displays a table with two columns: 'hashtag' and 'degree'. The table is enclosed in a border of dashed lines and vertical bars. The data is as follows:

hashtag	degree
trump	362
maga	149
afp	91
bbc	89
news	87
ft	85
泛亚	85
中國	85
metoo	85
china	85
abc	85
foxnewsus	85
russia	85
岳昕	85
董瑶琼	85
xijinping	85
cnn	85
putin	77
soccer	73
huawei	71

(d) PageRank

Though the original PageRank algorithm was originally developed by Larry Page to rank Google search results. It can be applied majorly to all graphs where we are required to find the influence of the vertices in any graph.

Similarly for our problem statement, we have a graph where we have hashtags as vertices and an edge between the hashtag if two hashtags co-occur together in a tweet. So, our main motive here to apply the PageRank algorithm is to build the subgraph by removing all irrelevant nodes.

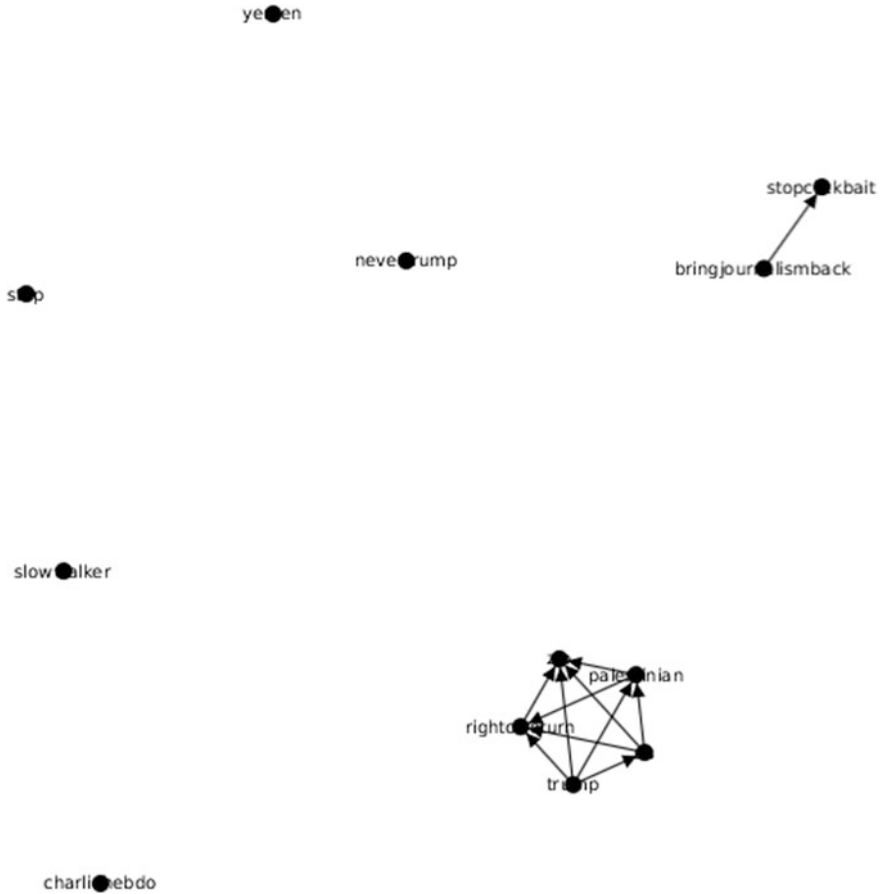


Fig. 10 Result of communities generated by the graph in one of the stream time window. Here, we can see various communities getting created with #trump is the largest community of in a particular sliding window

GraphX property graphs are directed multigraphs, which mean each edge in the graph specifies a unidirectional relationship, i.e., if there is an edge between two vertices in hashtag co-occurrence graph then hashtag1 and hashtag2 occur together in a similar fashion than hashtag2 and hashtag1. So we need to have two directions among two nodes for the PageRank to work correctly. *GraphX* offers static and dynamic implementations of PageRank where static PageRank runs for fixed no. of iterations and dynamic PageRank runs until convergence.

For our analysis, we will be using dynamic PageRank algorithm with a tolerance value of 0.01. We have done our analysis of multiple tolerance values like 0.01,

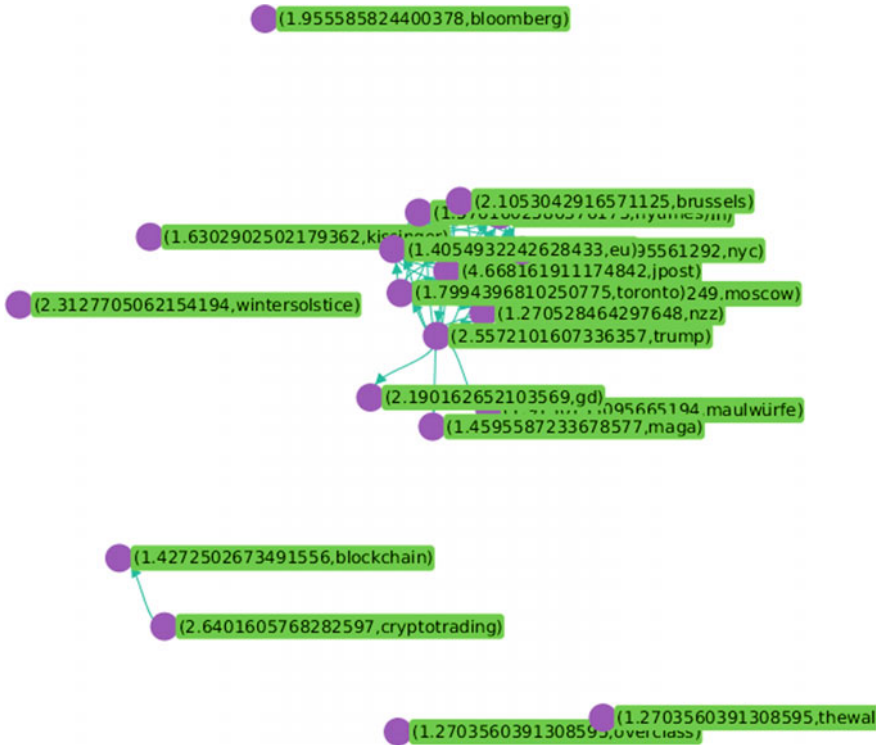


Fig. 11 Final graph visualization using graphstream. A result after community detection and PageRank in spark

0.0001, and 0.001, and found 0.01 to be the best. Since we are using real-time Spark Streaming and other tolerance values have taken longer than expected to converge and which might hamper our results for other subsequent windows of analysis in Spark Streaming [3, 15].

After we have got the PageRank, we will find top 20% of nodes in the graph and discard all irrelevant nodes. A node is called irrelevant, if it is not in topmost 20% of ranks in PageRank and create the new subgraph from this. This final graph will be our output graph for graphstream to visualize (Fig. 11).

PageRank helps in ranking across the communities in the graph. And after fetching the top 20% of the nodes for every community of Hashtag we can see that PageRank can help us in getting the most important hashtags that can be recommended to the user for further analysis.

8 Conclusion

The practicality of building real-time graphs using Spark Streaming and identifying the related hashtags which help in hashtag recommendation is being investigated in this paper. Twitter analytics has considered to be the top priority for most to get public opinions and views and hashtags helps us in understanding it in a better way. From the real-time hashtag co-occurrence graphs, we have seen the relevance of hashtag pairs and how to be so important for a specific period and how their importance fade with time. But some always remain relevant for a specified long period which played an important role in our analysis. Also, we have seen that we can seamlessly build real-time engine with graph functionality on top of it.

This can be proved to be a crucial tool for both social media and real-time applications. In this paper, we have also identified that if we will be able to combine the generated communities by Label Propagation with the PageRank algorithm, we get better results than waiting for both the algorithms to converge since in real-time analytics the one more thing followed is that how soon we are generating the results.

Further steps are required to improve this.

Work on enhancing PageRank and community detection algorithm and improvement in tolerance of PageRank for better convergence this can further help us in recommending the graphs better.

Improvement in graph creation is required and to enhance we will build the graph by saving previous window data on creating a bigger and better graph which can help in giving better results.

Also, instead of using hashtags we will focus on creating graph with additional properties from the data.

References

1. Sadik, S., Gruenwald, L.: Research issues in outlier detection for data streams. *ACM SIGKDD Explor. Newsl.* **15**(1), 33–40 (2014)
2. Bifet, A., Holmes, G., Pfahringer, B.: Moa-tweetreader: real-time analysis in twitter streaming data. In: *International Conference on Discovery Science*. Springer, Berlin (2011)
3. Bifet, A., Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In: *International Conference on Discovery Science*. Springer, Berlin
4. Duan, Y., et al.: Twitter topic summarization by ranking tweets using social influence and content quality. In: *Proceedings of COLING 2012*, pp. 763–780 (2012)
5. Ferragina, P., Piccinno, F., Santoro, R.: On analyzing hashtags in Twitter. In: *International Conference on Web and Social Media (ICWSM)*. AAAI Press (2015)
6. Wu, W., Zhang, B., Ostendorf, M.: Automatic generation of personalized annotation tags for twitter users. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (2010)
7. Turney, P.D.: *Information retrieval* **2**, 303 (2000). <https://doi.org/10.1023/A:1009976227802>

8. Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction (2003). <https://doi.org/10.3115/1119282.1119287>
9. Zangerle, E., Gassler, W., Specht, G.: Recommending#-tags in twitter. In: Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings, vol. 730, pp. 67–78 (2011)
10. Silberstein, A., Ashwin M., Raghu R.: Feed following: the big data challenge in social applications. In: Databases and Social Networks. ACM (2011)
11. Wang, X., et al.: Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM (2011)
12. Wang, M., Mizuho, I.: Hashtag sense induction based on co-occurrence graphs. In: Asia-Pacific Web Conference. Springer, Cham (2015)
13. Ramamonjison, R.: Apache Spark Graph Processing. Packt Publishing Ltd, Birmingham (2015)
14. Ryza, S., et al.: Advanced Analytics with Spark: Patterns for Learning from Data at Scale. O'Reilly Media Inc, Sebastopol (2017)
15. Pavlyshenko, B.: Forecasting of events by Tweet data Mining (2013). arXiv preprint [arXiv:1310.3499](https://arxiv.org/abs/1310.3499)

Author Index

A

Abayomi-Zannu, Temidayo, 339
Adepu, Divya, 93
Adewumi, Adewole, 685
Adnan, Kiran, 139, 301
Agarwal, Mohit, 569
Ahuja, Ravin, 685
Akbar, Rehan, 139, 301
Akila, D., 197
Alfred, Rayner, 151
Ali, Adnan Bin Amanat, 301
Alzahrani, Ahmad, 269
Ananya, M., 673
Andrade, Melvita, 93
Arora, Arushi, 3
Arunkumar, K., 325
Aswale, Neerja, 57
Awirothananon, Thatphong, 255
Azeez, Nureni Ayofe, 685
Azeta, Ambrose A., 673
Azeta, Victor I., 673

B

Bachhety, Shivam, 503
Bada, Taiwo Mayowa, 685
Badgujar, Aishwarya, 93
Baital, Kalyan, 111
Banodha, Umesh, 29
Bharadwaj, Vishnu, 555
Bhattacharya, Sonia, 365
Bhatt, Alok, 555
Bhosale, Shivjeet, 475
Biswal, Monalisa, 41
Booba, B., 713

C

Chakrabarti, Amlan, 111
Chakrabarty, Himadri Bhattacharyya, 365
Chandollikar, Neelam, 475
Chandrakar, Ruchi, 41
Chandrasekaran, E., 197
Chatterjee, Punyasha, 633
Chatterjee, Sankhadeep, 543
Chaudhary, Shankar, 613
Chawra, Vrajesh Kumar, 663
Chutatape, Opas, 125

D

Debnath, Arpita, 633
Deepika, R., 517
Devendran, A., 325
Dey, Anindita, 161, 171
Diwan, Prabhat, 161, 171
Dutta, Nitul, 601

E

Elavarasi, S., 185

G

Garg, Vikas, 69
Gill, Neetu, 161, 171
Goddy-Worlu, Rowland, 339
Govil, Himanshu, 161, 171
Grant, Emanuel, 339
Guha, Subhanil, 161, 171
Gupta, Govind P., 663
Gupta, Nirmal Kumar, 285
Gupta, Suneet K., 569

H

Haq, Rana Azhar Ul, 229
Hegde, Sarika, 463

J

Jain, Rachna, 503, 583
Jalnekar, Rajesh, 475
Janghel, Rekh Ram, 425
Joglekar, Pushkar, 475
John, Sonali, 697

K

Kamaruddin, Sk., 215
Karthika Renuka, D., 241
Karunya, K., 517
Kashyap, Indu, 583
Kashyap, Smiti, 69
Kaur, Gurkarandesh, 451
Kaur, Varinderjit, 439
Kavita, Choudhary, 15
Khan, Asif Riaz, 139
Khare, Akhil, 15, 613
Khor, Siak Wang, 301
Krishna, Ragini, 489
Kumar, Akshi, 3

M

Mahi, Gurjot Singh, 383
Majeed, Fiaz, 229
Makkar, Garima, 529
Mandal, Prasanta, 633
Menon, Varun G., 697
Misra, Sanjay, 673, 685
Misra, Diganta, 569
Mohanty, Sachi Nandan, 569
Mukul, Kavya, 57

N

Nale, Ruchita, 41
Naonueng, Kitti, 125
Nayak, Jyoti Ranjan, 401
Nayyar, Anand, 3, 503, 697
Niranjani, G., 643

O

Oberoi, Ashish, 439, 451
Odun-Ayo, Isaac, 339

P

Pagare, Reena, 613
Pawade, Dipti, 93
Peddawad, Dipali, 475

Phoophuangpairoj, Rong, 125
Prashanth, C.M., 489
Prema, A., 713

R

Rajamohana, S.P., 517
Rana, Manoj Kumar, 633
Rathore, Yogesh Kumar, 425
Ravi, Vadlamani, 215
Rawal, Ayush, 555
Renugadevi, R., 713
Rohil, Mukesh Kumar, 285
Roy, Santanu, 543

S

Sadaoui, Samira, 269
Saha, Pallavi, 543
Sahu, Anil Kumar, 413
Sahu, Rajkumar, 401
Sakhapara, Avani, 93
Sasi Kumar, A., 197
Saxena, Kanak, 29
Sen, Soumya, 543
Sharma, Sonam, 723
Shaw, Binod, 401
Shetty, Surendra, 463
Shilaskar, Swati, 475
Singh, Gurinder, 69
Singh, Lolita, 601
Sinha, G.R., 413
Soni, Sapna, 413
Sri Vinitha, V., 241
Suseendran, G., 185, 197

T

Thanjunpong, Sathaya, 255
Tomar, Kanika Singh, 69

U

Umamaheswari, K., 517, 643
Upasani, Shubhangi, 3

V

Van der Vyver, Charles, 685
Verma, Amandeep, 383
Verma, Archana, 425

Y

Yadav, Sanjeev Kumar, 15
Yadav, Vineet, 555
Yu, Hin Fuk, 151