

Neha Sharma  
Amlan Chakrabarti  
Valentina Emilia Balas *Editors*

# Data Management, Analytics and Innovation

Proceedings of ICDMAI 2019, Volume 2

# **Advances in Intelligent Systems and Computing**

## **Volume 1016**

### **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 2



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 978-981-13-9363-1

ISBN 978-981-13-9364-8 (eBook)

<https://doi.org/10.1007/978-981-13-9364-8>

© 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 January 18 to 20, 2019. ICDMAI is a flagship conference of Society for Data Science, which is a nonprofit 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 toward 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 IBM industry leader. Other partners of 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) were selected for an oral presentation after a rigorous review process. Besides paper presentation, the conference also showcased workshop, tutorial talks, keynote sessions and plenary talk by the experts of the respective field.

The volumes cover a broad spectrum of computer science, information technology, computational engineering, electronics and telecommunication, electrical engineering, computer application and all the relevant disciplines. The conference papers included in these proceedings are published post-conference and are grouped into 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; 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 to ICDMAI 2019. The editors 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 Committee 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; and other eminent personalities like Mr. Eddy Liew, Cloud and Solutions Country Technical Leader for IBM Malaysia; Kranti Athalye, Sr. Manager in IBM India University Relations; Dr. Juergen Seitz, Head of Business Information Systems Department, Baden-Wurttemberg 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.

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 areas 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

<b>Sequence Mining-Based Support Vector Machine with Decision Tree Approach for Efficient Time Series Data Classification . . . . .</b>	<b>3</b>
D. Senthil and G. Suseendran	
<b>Identifying Phishing Through Web Content and Addressed Bar-Based Features . . . . .</b>	<b>19</b>
Nureni A. Azeez, Joseph Ade, Sanjay Misra, Adewole Adewumi, Charles Van der Vyver and Ravin Ahuja	
<b>A Unified Framework for Outfit Design and Advice . . . . .</b>	<b>31</b>
Adewole Adewumi, Adebola Taiwo, Sanjay Misra, Rytis Maskeliunas, Robertas Damasevicius, Ravin Ahuja and Foluso Ayeni	
<b>Contextual Relevance of ICT-CALL Modules and Its Role in Promoting Employability Skills Among Engineering Students—A Study . . . . .</b>	<b>43</b>
Yeddu Vijaya Babu	
<b>A Clinically Applicable Automated Risk Classification Model for Pulmonary Nodules . . . . .</b>	<b>55</b>
Triparna Poddar, Jhilam Mukherjee, Bhaswati Ganguli, Madhuchanda Kar and Amlan Chakrabarti	
<b>A Generalized Ensemble Machine Learning Approach for Landslide Susceptibility Modeling . . . . .</b>	<b>71</b>
Akila Bandara, Yashodha Hettiarachchi, Kusal Hettiarachchi, Sidath Munasinghe, Ishara Wijesinghe and Uthayasanker Thayasilvam	
<b>Comparative Evaluation of AVIRIS-NG and Hyperion Hyperspectral Image for Talc Mineral Identification . . . . .</b>	<b>95</b>
Himanshu Govil, Mahesh Kumar Tripathi, Prabhat Diwan and Monika	

<b>Real-Time Scheduling Approach for IoT-Based Home Automation System .....</b>	103
Rishab Bhattacharyya, Aditya Das, Atanu Majumdar and Pramit Ghosh	
<b>Floating Car Data Map-Matching Utilizing the Dijkstra's Algorithm.....</b>	115
Vít Ptošek, Lukáš Rapant and Jan Martinovič	
<b>Calculating AQI Using Secondary Pollutants for Smart Air Management System .....</b>	131
Gautam Jyoti, Malsa Nitima, Singhal Vikas and Malsa Komal	
<b>Enhanced Prediction of Heart Disease Using Particle Swarm Optimization and Rough Sets with Transductive Support Vector Machines Classifier.....</b>	141
M. Thiagaraj and G. Suseendran	
<b>DataCan: Robust Approach for Genome Cancer Data Analysis .....</b>	153
Varun Goel, Vishal Jangir and Venkatesh Gauri Shankar	
<b>DataAutism: An Early Detection Framework of Autism in Infants using Data Science .....</b>	167
Venkatesh Gauri Shankar, Dilip Singh Sisodia and Preeti Chandrakar	
<b>AnaBus: A Proposed Sampling Retrieval Model for Business and Historical Data Analytics .....</b>	179
Bali Devi, Venkatesh Gauri Shankar, Sumit Srivastava and Devesh K. Srivastava	
<b>Big Data Management</b>	
<b>Abrupt Scene Change Detection Using Block Based Local Directional Pattern.....</b>	191
T. Kar and P. Kanungo	
<b>A Framework for Web Archiving and Guaranteed Retrieval .....</b>	205
A. Devendran and K. Arunkumar	
<b>Business Intelligence Through Big Data Analytics, Data Mining and Machine Learning .....</b>	217
Wael M. S. Yafooz, Zainab Binti Abu Bakar, S. K. Ahammad Fahad and Ahmed M. Mithon	
<b>The Effect of Big Data on the Quality of Decision-Making in Abu Dhabi Government Organisations .....</b>	231
Yazeed Alkatheeri, Ali Ameen, Osama Isaac, Mohammed Nusari, Balaganesh Duraisamy and Gamal S. A. Khalifa	

<b>The Impact of Technology Readiness on the Big Data Adoption Among UAE Organisations . . . . .</b>	249
Adel Haddad, Ali Ameen, Osama Isaac, Ibrahim Alrajawy, Ahmed Al-Shbami and Divya Midhun Chakkaravarthy	
<b>Artificial Intelligence and Data Analysis</b>	
<b>Sports Data Analytics: A Case Study of off-Field Behavior of Players . . . . .</b>	267
Malini Patil, Neha Sharma and B. R. Dinakar	
<b>Phrase Based Information Retrieval Analysis in Various Search Engines Using Machine Learning Algorithms . . . . .</b>	281
S. Amudha and I. Elizabeth Shanthi	
<b>The Politics of Artificial Intelligence Behaviour and Human Rights Violation Issues in the 2016 US Presidential Elections: An Appraisal . . . . .</b>	295
Patrick A. Assibong, Ikedinachi Ayodele Power Wogu, Muviwa Adeniyi Sholarin, Sanjay Misra, Robertast Damasevičius and Neha Sharma	
<b>Crop Prediction Using Artificial Neural Network and Support Vector Machine . . . . .</b>	311
Tanuja K. Fegade and B. V. Pawar	
<b>A Hybrid and Adaptive Approach for Classification of Indian Stock Market-Related Tweets . . . . .</b>	325
Sourav Malakar, Saptarsi Goswami, Amlan Chakrabarti and Basabi Chakraborty	
<b>Generative Adversarial Networks as an Advancement in 2D to 3D Reconstruction Techniques . . . . .</b>	343
Amol Dhondse, Siddhivinayak Kulkarni, Kunal Khadilkar, Indrajeet Kane, Sumit Chavan and Rahul Barhate	
<b>Impact of Artificial Intelligence on Human Resources . . . . .</b>	365
Sapna Khatri, Devendra Kumar Pandey, Daniel Penkar and Jaiprakash Ramani	
<b>Role of Activation Functions and Order of Input Sequences in Question Answering . . . . .</b>	377
B. S. Chenna Keshava, P. K. Sumukha, K. Chandrasekaran and D. Usha	
<b>GestTalk—Real-Time Gesture to Speech Conversion Glove . . . . .</b>	391
Varun Shanbhag, Ashish Prabhune, Sabyasachi Roy Choudhury and Harsh Jain	

<b>Deep Learning Algorithms for Accurate Prediction of Image Description for E-commerce Industry .....</b>	401
Indrajit Mandal and Ankit Dwivedi	
<b>Taj-Shanvi Framework for Image Fusion Using Guided Filters.....</b>	419
Uma N. Dulhare and Areej Mohammed Khaleed	
<b>Advances in Network Technologies</b>	
<b>Effective Classification and Handling of Incoming Data Packets in Mobile Ad Hoc Networks (MANETs) Using Random Forest Ensemble Technique (RF/ET) .....</b>	431
Anand Nayyar and Bandana Mahapatra	
<b>Community Structure Identification in Social Networks Inspired by Parliamentary Political Competitions .....</b>	445
Harish Kumar Shakya, Nazeer Shaik, Kuldeep Singh, G. R. Sinha and Bhaskar Biswas	
<b>An Integrated Technique to Ensure Confidentiality and Integrity in Data Transmission Through the Strongest and Authentic Hotspot Selection Mechanism .....</b>	459
Shiladitya Bhattacharjee, Divya Midhun Chakkavarthy, Midhun Chakkavarthy and Lukman Bin Ab. Rahim	
<b>Model to Improve Quality of Service in Wireless Sensor Network.....</b>	475
Vivek Deshpande and Vladimir Poulikov	
<b>Performance Analysis of the Mobile WSN for Efficient Localization .....</b>	485
Kailas Tambe and G. Krishna Mohan	
<b>Author Index.....</b>	499

# 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 Springer Journal of 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**

# Sequence Mining-Based Support Vector Machine with Decision Tree Approach for Efficient Time Series Data Classification



D. Senthil and G. Suseendran

**Abstract** The growing demand for an efficient approach to classify time series data is bringing forth numerous research efforts in data mining field. Popularly known applications like business, medical and meteorology and so on, typically involves majority of data type in the form of time series. Hence, it is crucial to identify and scope out the potential of time series data owing to its importance on understanding the past trend as well as predicting about what would occur in future. To efficiently analyze the time series data, a system design based on Sliding Window Technique-Improved Association Rule Mining (SWT-IARM) with Enhanced Support Vector Machine (ESVM) has been largely adopted in the recent past. However, it does not provide a high accuracy for larger size of the dataset along with huge number of attributes. To solve this problem the proposed system designed a Sequence Mining algorithm-based Support Vector Machine with Decision Tree algorithm (SM-SVM with DT) for efficient time series analysis. In this proposed work, the larger size of the dataset is considered along with huge number of attributes. The preprocessing is performed using Kalman filtering. The hybrid segmentation method is proposed by combining a clustering technique and Particle Swarm Optimization (PSO) algorithm. Based on the sequence mining algorithm, the rule discovery is performed to reduce the computational complexity prominently by extracting the most frequent and important rules. In order to provide better time series classification results, the Support Vector Machine with Decision Tree (SVM-DT) method is utilized. Finally, the Pattern matching-based modified Spearmen's rank correlation coefficient technique is introduced to provide more similarity and classification results for the given larger time series dataset accurately. The experimental results shows that the pro-

---

D. Senthil (✉)

Department of Computer Science, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

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

G. Suseendran

Department of Information and Technology, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

e-mail: [suseendar\\_1234@yahoo.co.in](mailto:suseendar_1234@yahoo.co.in)

posed system achieves better accuracy, time complexity and rule discovery compared with the existing system.

**Keywords** Classification · Time series · Hybrid segmentation · Pattern matching · Accuracy

## 1 Introduction

Time series data is considered vital among several time-based sequential data sets which could be acquired easily from applications involving both financial and scientific aspects [1]. Ideally, time series can be regarded as an accumulation of chronologically observed data sets or findings. Time series data comprises of large data size, high in dimensionality and may require updating by progressively. In addition to that, time series data has been defined [1] by means of continuous and numerical in nature and always assumed as either than an individual numerical attribute.

Time series data classification could begin by taking advantage of extensive research interests and so the developments along that data sets are put forward as an effort to establish advancements in data mining field [2]. At the same time, the complexity of the time series data imparts huge demand on the research efforts and hence targets the interest of the researchers since last decade. Recently, time series data mining framework possesses fundamental contribution toward analyzing the time series analysis [3] especially in the field of data mining.

Significant methods which are involved in time series data mining framework also possess the capability to successfully characterize [3] and also predict complex trend in time series that might be nonperiodic, irregular, and so on. Such methods involving concepts relevant to data mining would help to analyze time series data and also overcome the limitations along the conventional mechanisms adopted for time series analysis, especially of both stationary and linearity based requirements [4].

This proposed work deals with data set which is typically very large in data size having huge amount of the attributes. In order to overcome issues pertaining to low accuracy and high computational complexity, hybrid method comprising Sliding Window Technique with Improved Association Rule Mining (SWT-IARM) and Enhanced Support Vector Machine (ESVM) [5] is adopted for the time series evaluation. This technique also does not deal with large size dataset.

In the intent of dealing with abovementioned issues, the proposing work introduced sequence mining algorithm-based Support Vector Machine with Decision Tree algorithm shortly said to be as SM-SVM with DT for an efficient manner of time series analysis [6]. In this preprocessing has been carrying out as the first step. The preprocessing would take over in terms of using the Kalman filtering.

Second, the hybrid segmentation approach is taken place by means of combining the clustering technique [7] and the PSO algorithm. PSO stands for Particle Swarm Optimization algorithm. Depending on sequence mining algorithm, rule discovery

has performed in the aim of reducing computational complexity by means of extract the highly frequent and the important rules.

Then to bring out the hopeful time series of classification results, SVM-DT approach has been used in this review process. Whereas SVM-DT stands as Support Vector Machine with Decision Tree is a method used for obtaining better result in terms of classification process [8]. At last, Pattern matching-based modified Spearman's rank correlation coefficient mechanism has designed in the aim of providing high similar and the classification results for using large time series dataset in an accurate way [9]. The experimental results may also show up that the proposing review could achieve well-saying accuracy, the time complexity and also the rule discovery has been comparing with the existing system.

The paper is structured with the following sections: Sect. 2 discusses the reported literature on various models/approaches for time series classification. Section 3 overviews the proposed approach and the specific methodology being adopted. Section 4 presents the experimental results of the proposed scheme and finally Sect. 5 covers the concluding remarks.

## 2 Literature Survey

Win [1] dealt with the time domain statistical models and the approaches on time series analyze which used by applications. It brings out the brief view on basic concepts, nonstationary and stationary models, nonseasonal and the seasonal models, the intervention and also outlier models, the transfer function models, the regression time series model, the vector time series models, those applications. In this paper, author reviews the process in time series analysis which involves the model identification, the parameter estimation, the diagnostic checks, the forecasting, and the inference. We also discuss on autoregressive conditional heteroscedasticity model and more generalized in manner.

In the work reported by Lu et al. [2] an approach using ICA has been proposed for variables prediction helping the generation of components which are referred as independent components, called shortly as IC. Once finding and eliminating ICs containing noisy components, few of the remaining variables among them are used in reconstructing the forecast variables which already comprises fewer noise, thus serving as the variables to be used as input in the model termed as SVR forecasting. Towards understanding the performance evaluation of introduced approach, examples pertaining to opening index: Nikkei 225 and the closing index: TAIEX has been dealt in detail. The obtained output has shown that the proposed model would outperform SVR model constituting having non-filtered elements and also the random walk model.

Fu in [3] made a brief study and represented the review made comprehensively on the existing research involving data mining of time series date. The entire work could be divided into several sections comprising initial representation followed by indexing, and then follows further steps such as measuring similarity, segmentation,

visualization, and lastly the mining process. Additionally, research problems handled with advanced methodologies had been dealt significantly. The key importance is found to be that the review would help as a complete information to the potential researchers as it covers an up-to-date review of recent developments on time series data-based mining. Further, potential research gaps are also critically reviewed which would help the prospective researchers to understand and progress on the time series data investigation.

Varied techniques employed for the data mining of time series-based data has been reviewed chronologically in the work reported by Esling and Agon [4]. The review intends to give a complete overview of widely applied tasks that attracted the interest of several researches has been dealt. In most of the methods, assumptions related to classify the time series data have been found to be entirely dependent on the similar elements to implement the method specifically. Moreover, the literature survey has been classified into smaller sections based on some commonly dealt features including representative models, measure to understand distance followed by indexing approaches. Detailed review of literature corresponding to each single entity has been studied. Formalization of four robust methods and distance classification of any individual entity has also been dealt.

Senthil and Suseendran [5] has introduced an approach in the proposed research named Sliding Window Technique-based Improved ARM with Enhanced SVM (shortly represented as SWT-IARM with ESVM). In proposed system, preprocessing has performed in terms of using the Modified K-Means Clustering representing namely by MKMC. The indexing process could do with the help of R-tree which would provide a faster result. Segmentation is then made by SWT and it could minimize cost complexity by means of optimal segments. IARM has applied on the efficient rule discovery step by generating most frequent rules. By ESVM classification approach, the rules are been made to classified in more accurate way.

One of the approach holding core theoretical background with respect to the dynamic systems has been studied in the work reported by Povinelli et al. [6]. It has also included judicious selection of parameters, which gives the theory based on topology resulted due to signal reconstruction ensuring asymptotical mode of representing the entire fundamental system. This specific algorithm could calculate the appropriate parameters automatically with reference to spaces involving reconstructed phase, as only number of signals, mixtures, and class labels are only required to be given in the form of input. Among many, three individual data sets comprising motor current simulations, electrocardiogram recordings, and finally speech waveforms have been employed for data validation study. Robust output has resulted from the proposed method across various domains, from which it is evident that the method outperforms the baseline method, for instance, the neural network comprising time delay.

Methods to study data mining-based models for classifying time series datasets have also been reported by Keogh et al. [7], wherein commonly used representations having a linear piecewise approximation has given prime attention. Hence, this model has triggered studies related to support clustering, classification of time series data followed by specific indexing and mining. Also, several other algorithms

have been studied to derive the abovementioned representation, which eventually enabled to research every algorithm individually for several times. This method has been found to be undertaken for the first time to understand by empirically relating the proposed techniques. Authors also have shown that most of these represented algorithms typically contains issues in view of data mining.

To effectively forecast several temporal courses which combine to form temporal abstractions having case-based reasoning Schmidt and Gierl [8] has presented a different approach by introducing two prognostic applications which find huge importance in applications involving no consistent/particular standards that lack known periodicity and complete domain theory. The first model has dealt the prediction of time series data derived from kidney related function, especially from patients treated in intensive care units [10]. The underlying idea behind the model is to overcome the time delay, which eventually helps in cautioning the patients prone to kidney related failures. Yet, another application has worked in entirely different domain referred as geographical medicine and the basic aim is to compute the prompt cautioning so as to prevent from infectious disease characterized by means of time varied cyclic incidences.

In the methodology reported by Ghalwash and Obradovic [9], simple Multivariate Shapelets Detection (MSD) has been adopted which enables quick and user-specific classification, in this context classification of patient-relevant data sets which are typically in multivariate time series. This mechanism might extract the time series patterns and be called as multivariate shapelets, from all of dimensions of time series which is distinctly manifest target class in locally. The time series are made to classify by means of searching the most earliest in closest patterns.

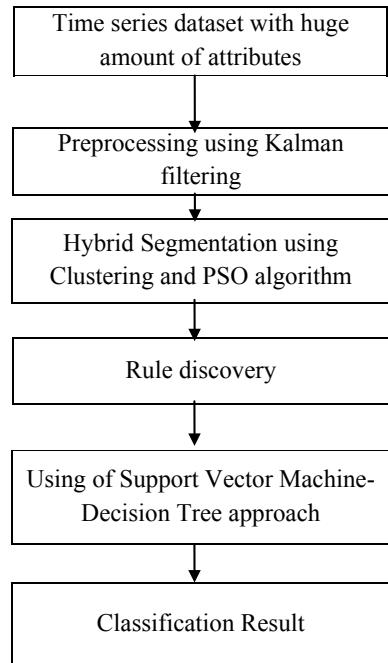
Thiyagaraj and Suseendran [11] performed role of data mining techniques in the process of heart disease prediction. This research discusses about the various data mining techniques, their working procedure during the prediction of heart disease has been provided. Rohini and Suseendran [12] introduced the new hybrid method for classifying the spirometry data. This work introduced the new technique namely Hybridized K means Clustering with decision tree algorithm. Thiyagaraj and Suseendran [13] introduced the modified Firefly Algorithm-Based Radial Basis Function With Support Vector Machine for the accurate prediction of heart disease. An input dataset encompasses three kinds of attributes such as Input, Key, and Prediction attributes. After the normalization, an attribute reduction and feature extraction are performed by using FA and Principal Component Analysis (PCA), respectively. Finally RBF-SVM is classified a features as normal or heart diseases.

### 3 Proposed Methodology

Time series dataset having high number of attributes has been used in this review [14]. First preprocessing has carry out with the help of Kalman filtering. Then hybrid segmentation is performed by means of combining the Particle Swarm Optimization with clustering approaches. The rule discovery process is carrying out to extract

the most significant rules and attributes in the intent of reducing the complexity of computation process. Finally SVM-DT is introduced to obtain a better classification result in terms of accuracy. These above said process which is proposing in this particular work has been illustrated in flow chart Fig. 1.

**Fig. 1** Architecture of proposed system



### 3.1 Time Series Dataset

A time series dataset refers to a progression of the data points series based on time order which is jacketed, enlisted, or may be recorded or graphed. Generally, a sequence of progressively equal spaced points in the time order approach denotes time series. Time series analysis could involve different strategies to examine the time series data in expectation of extracting the meaningful statistics and furthermore distort the characteristics of data.

### 3.2 Preprocessing

In this introducing work as first step preprocess is being carried out to obtain an efficient approach. Preprocess would be performed by means of using Kalman filtering. Kalman Filtering is also said to be LQE that stands for linear quadratic estimation. This algorithm utilizes different set of the measurements in series confronting some statistical noise and inaccuracies observed over time during the analysis. Furthermore, it creates the estimation of exhibiting unknown variables. These variables tend to be more accurate than single measurement by valuation of the joint probability distribution among the variables over the dispersion of factor every timeframe.

Kalman filters are relying upon different straight dynamical frameworks that are discrete in time domain. The state of system was being denoted by vector of the real numbers. The Kalman filter was used for the purpose of estimating internal state of a specific process governed by a sequence of the noisy observations. Thus a model based on the structure of Kalman filter framework has been developed to carry out the process analysis. This means that specifying the following matrices:

- $\mathbf{F}_k$ , denotes state transition model
- $\mathbf{H}_k$ , portrays the observation model
- $\mathbf{Q}_k$ , represented by covariance of process noise
- $\mathbf{R}_k$ , defines covariance of observation noise
- At times  $\mathbf{B}_k$ , denoting control-input model, for every time-step of  $k$  that states as follows.

The Kalman filter model would be speculated as the true state of time  $k$  has derived from state at  $k - 1$  in accordance with,

$$\mathbf{x}_k = \mathbf{F}_k \mathbf{x}_{k-1} + \mathbf{B}_k \mathbf{u}_k + \mathbf{w}_k$$

Thus,

$\mathbf{F}_k$  denoting a state transition model which is applied to a previous state namely  $\mathbf{x}_{k-1}$   
 $\mathbf{B}_k$  defining a control-input model that can be correlated to control vector  $\mathbf{u}_k$   
 $\mathbf{w}_k$  expressed as the one of process noise that presumed to draw from zero mean multivariate normal distribution,  $\mathcal{N}$  in terms of covariance,  $\mathbf{Q}_k$ . Where,  
 $\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{Q}_k)$ .

At time  $k$ , observation or measurement  $\mathbf{z}_k$  of true state  $\mathbf{x}_k$  postulated on the basis of equation follow

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k$$

Where,  $\mathbf{H}_k$  stated as an observation model mapping true state space to observed space and the  $\mathbf{v}_k$  as an observation noise considered as zero-mean Gaussian white; the noise having covariance  $\mathbf{R}_k$ : Thus,  $\mathbf{v}_k \sim \mathcal{N}(0, \mathbf{R}_k)$ . as initial state, the noise vectors at every step  $\{\mathbf{x}_0, \mathbf{w}_1, \dots, \mathbf{w}_k, \mathbf{v}_1 \dots \mathbf{v}_k\}$  are considered as mutually independent in nature.

Thus by using this Kalman Filtering model designed as per proposing work pre-processing could carry out. The result obtained from this preprocessing stage is given as input for segmentation process which is a hybrid approach using clustering and PSO-based approach.

### **3.3 Segmentation**

Segmentation has been performed with the help of combination of both Particle Swarm Optimization techniques to acquire better experience with time series analysis. Time series Segmentation is an approach among various methods of time series analysis. Here inputs are made to separate as various discrete segments in the aim of revealing properties of source. Most widely using algorithms in this field are based on change point detection such as sliding windows, bottom-up, top-down methods. Particle Swarm Optimization is a classification of optimization problem solving using an iterative population-based approach primarily repetitive approach originated from the flocking behavior of the birds. In the PSO, the composition of particles set constitutes as population, whereas each particle might really represent potential resolution to optimization problem. Each particle was predominantly composed of two properties which are unique to each other. Initial property as position defined as the particle's position in solution space and the later property known as velocity indicating the stratagem of particle's current new position in every of iteration. Position and Velocity of particle have defined as  $\mathbf{x}_i^{(t)}$  and  $\mathbf{v}_i^{(t)}$  respectively.

Where,

$$\mathbf{v}_i^{(t)} = \mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}.$$

In each clustering techniques depending Particle Swarm Optimization, three problems should get addressed namely particle representation, definition of similar measures, and definition of fitness function. In the following, each issue is portrayed for time series and data clustering based on the arrangement of PSO.

```

Initialize a swarm of  $P$  particles
 $t \leftarrow 1$ 
repeat
  for  $i = 1$  to  $P$  do
    Update the velocity and position of particle  $i$ 
     $\mathbf{v}_i^{(t+1)} = \omega \mathbf{v}_i^{(t)} + c_1 r_1 (\mathbf{x}_i^{pb(t)} - \mathbf{x}_i^{(t)}) + c_2 r_2 (\mathbf{x}_i^* - \mathbf{x}_i^{(t)})$ 
     $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \mathbf{v}_i^{(t+1)}$ 
    Update the personal best of particle  $i$  at iteration  $t + 1$ 
    as:
     $\mathbf{x}_i^{pb(t+1)} = \begin{cases} \mathbf{x}_i^{pb(t)} & \text{if } F(\mathbf{x}_i^{(t+1)}) \geq F(\mathbf{x}_i^{pb(t)}) \\ \mathbf{x}_i^{(t+1)} & \text{otherwise} \end{cases}$ 
  end for
  Update the global best as:
   $\mathbf{x}^{*(t+1)} = \operatorname{argmin} F(\mathbf{x}_i^{pb(t+1)}); \quad i \in \{1, \dots, P\}$ 
   $t \leftarrow t + 1$ 
until stopping criteria are satisfied

```

The distance measure had computed as follows:

$$S_{\text{dist}}(\mathbf{y}, \mathbf{z}) = 2 \times \frac{1}{\sqrt{2\pi}} \int_{\Phi_{y,z}}^{\inf} e^{-\frac{x^2}{2}} dx,$$

$$\Phi_{y,z} = \sqrt{(\underline{\mathbf{y}} - \underline{\mathbf{z}}) \Sigma^{-1} (\underline{\mathbf{y}} - \underline{\mathbf{z}})^T},$$

$y$  and  $z$  are being obtained by means of average over rows in time dimension of every time series matrix namely  $y$  and  $z$ , and  $\Sigma$  is covariance matrix. The hybrid similarity measure,  $S_H$  between the two multivariant time series had computed by combine two metrics which mentioned above,

$$S_H(\mathbf{y}, \mathbf{z}) = \alpha_1 S_{\text{dist}}(\mathbf{y}, \mathbf{z}) + \alpha_2 S_{\text{PCA}}(\mathbf{y}, \mathbf{z}),$$

where,  $\alpha_1$  and  $\alpha_2$  reflect contribution of every metric in final similarity measure and are chosen as  $\alpha_1 + \alpha_2 = 1$ .

This combination of above said clustering mechanisms and the above-discussed particle swarm optimization. This combination would bring out the best hybrid segmentation part in this review. Since segmentation has been carried out by using sliding window-based approach but in this discussing review, hybrid segmentation has been adopted to overcome the issues that are seen in existing work. Thus this hybrid approach will give out hopeful segmentation to obtain better accuracy result in terms of classification.

### 3.4 Rule Discovery

As third most step rule discovery would carry out. This rule discovery process could take place based on sequence mining algorithm the rule discovery. This usage of sequence mining approaches will pave a way to minimize the computational complexity in prominently by means of extracting highly more frequent and the important rules. This approach in data mining would concern with finding out patterns that are relevant among all in nature. Usually it deals with discrete values, thus time series mining is somehow closely relevant to this. At the end of the rule discovery stage, we could able to find out frequent rule generation which is highly most in nature. Thus it would obtain high confidence rules from frequent itemsets that are available.

### 3.5 Support Vector Machine-Decision Tree

In proposed work, hybrid SVM-based decision tree has been introduced to obtain best classification result and to speeding up the process. SVM make pattern recognition and could do data analysis as possible. Regression analysis and the classification are being carried out using Support Vector Machine. Thus the result got from applying SVM would act as a decision-making model. Support vector machine represented in short form as SVM is one among supervised learning mechanisms in computer science and the statistics. Support Vector Machine intent in analyzing the data and for recognizing the patterns. It may deal by individually with the classification and also regression analysis. Data would linearly that are separable which makes the researchers by means of identifying both hyperplanes in margin. This evaluation purely depends on the method in no points present in between and it may maximize distance among all. SVM might help in splitting the data having hyperplane and would also extend nonlinear boundaries by means of kernel trick. SVM would do classification method by correct in terms of classifying data present. It is also been described mathematically as following:

$$\begin{aligned} x_i \cdot w + b &\geq +1 \text{ for } y_i = +1 \\ x_i \cdot w + b &\leq -1 \text{ for } y_i = -1 \end{aligned}$$

Above equations may also combine in forming one set of the differences as shown below,

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad \forall i$$

Thus,

$x$  denotes vector point

$w$  denotes weight parameter as vector.

Splitting of data was based on the value got from the above equation. If output is founded more than 0, it meant to accept for split input data. SVM would identify the point which having distance as long from hyperplane among each of them. Hyperplane could likewise expand accessible edge margin availability and after that divides lines in the nearest focuses on the raised frame hull of datasets accessible. The discussed processes are taken place if hyperplane chosen is at as far point that possibly from the data. In addition to that, this approach has applied on points at another side. Summed Distance has defined in terms of splitting hyperplane to points which at nearest. At last, summed distance has been observed by solving and then subtracting currently available two distances.

SVMs in testing phase were predominantly used to test the tasks subjected underneath binary classification. Many methods exist that are addressing toward this task which intends to minimize number of the support vectors. It aims to reduce number of the presenting test data points which are in need of SVM's approach to be classified. In rule discovery, the decision boundary of SVM has to be aggregated within the decision trees. The obtained resulting tree, which is one among the hybrid tree containing univariate and the multivariate SVM nodes. The SVM nodes were used by the hybrid tree to identify and classify the crucial data points lying within the decision boundary; however, the remaining less crucial data points also been classified by univariate nodes. The hybrid tree classification accuracy was ensured as far as tuning the threshold parameter.

Thus this combination of decision tree approach with Support vector machine to obtain a beneficial rule discovery process. Finally with the help of Support Vector Machine-Decision Tree technique, rule discovery has been carried out and important rules are being identified.

### **3.6 Classification**

Thus by performing the above-discussed steps sequentially, hopeful classification could obtain. As a final step, Pattern matching depending modified Spearman's rank correlation coefficient technique has been applied in the intent of providing very most similarity and also the good classification results for the using larger time series dataset inaccurately. The experimental results obtained show up that the proposing system in this review would achieve time complexity, better accuracy, and rule discovery as compared to various systems that exist.

## **4 Experimental Results**

Evaluation of performance in respect to the time complexity, the accuracy and the rule generation of those namely SVM, ARM, SWT-IARM with ESVM, and SVM-DT has been conducted. The objectives were acquired by experimenting with the

real and synthetic data sets. It had tested about 35 time series datasets that acquired, and possess stock market values, the chemical and the physics phenomenon measurements.

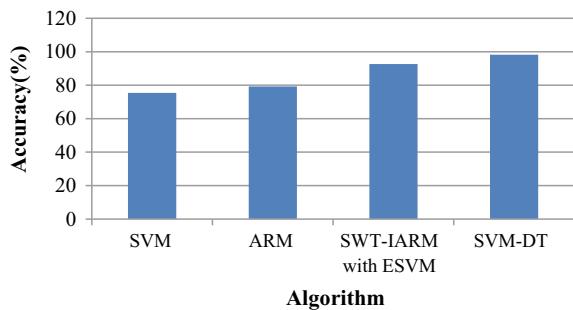
#### 4.1 Accuracy

Accuracy has been calculated as entire correctness of model designed and has computed as total actual classification parameters namely  $T_p + T_n$  which have segregated in terms of summing the classification parameters namely  $T_p + T_n + F_p + F_n$ . Thus accuracy is been computed by means of below formula represented:

$$\text{Accuracy} = \frac{T_p + T_n}{(T_p + T_n + F_p + F_n)}$$

From Fig. 2, it could be observed that the comparison metric has evaluated by

**Fig. 2** Accuracy comparison



means of using the existing and the proposed approach in terms of the accuracy. In  $x$ -axis, algorithms are represented and in the  $y$ -axis accuracy values are been plotted. The existing system may provide less accuracy, whereas proposed system could bring out the high accuracy for those inputted values. The proposed Support vector machine with Decision Tree simply representing as SVM-DT algorithm would select the best rules from time series data available. Finally these rules are made to apply for training and the testing phase to procure time series data which are in pertinent nature and closely connected with the given dataset. From the result obtained, we can conclude that the proposing system with SVM-DT would bring out better classification output. Thus proposed SVM-DT algorithm is considered as superior to previous one namely used in existing work.

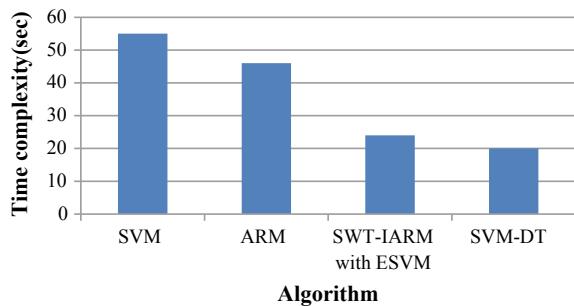
## 4.2 Time complexity

The system works well, thus algorithm would provide the lower complexity values and it is illustrated in Fig. 3.

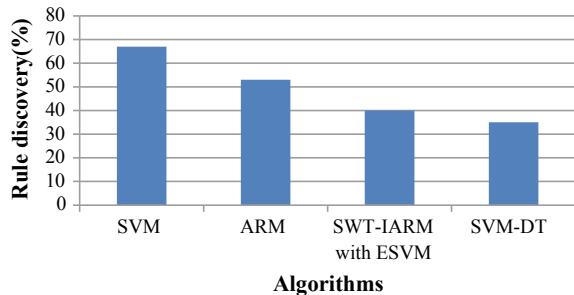
From Fig. 4, it has been noted that comparison metric is analyzed by the existing and the proposed method by means of the time complexity. In x-axis, algorithms are been taken and in y-axis time complexity value has been plotted. The existing method may provide high time complexity, while proposing system might provide low time complexity for inputting data. The proposing SVM-DT approach is used for selecting the good rules among all. At last, these rules are to be applied on train and test phase in the aim of producing highly more related data on the time series dataset. The result has proven that the introducing system would attain higher classification results with SVM-DT mechanism. Thus introduced SVM-DT is assumed as superior to previous one namely the SVM, the ARM and the SWT-IARM with ESVM algorithms (Fig. 4).

From the above draw chart, rules are generated by the existing and the proposed algorithms have been made to compared and showed. For x-axis, algorithms are been taken and in the y-axis, rule discovery value is placed. The proposing SVM-DT would provide very low number of the rules and thus it proven the superior time series classification.

**Fig. 3** Time complexity



**Fig. 4** Rule discovery



## 5 Conclusion

In this system, time series dataset is made to evaluate by using an efficient techniques. The indexing approach is focus on increasing the similarity and the faster access. The time required for constructing data series index which evolve to prohibitive as data grows, and they might consume less amount of time for the large sizing data series. In this preprocessing has been taken place as first step by means of Kalman filtering. Then it is applicable for hybrid segmentation process by means of combining the clustering approaches and particle swarm optimization methodologies. Finally SVM-DT stands for Support vector Machine-Decision Tree has been applied to carry out an effective sequence mining and thus obtains the better classification output.

In future work, a new system will develop by means of various data mining approached in terms of increasing the accuracy and reducing the time complexity as compared to this introduced system.

## References

1. Wei, W. W. (2006) Time series analysis. In *The Oxford handbook of quantitative methods in psychology*
2. Lu, C. J., Lee, T. S., & Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, 47(2), 115–125.
3. Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164–181.
4. Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1), 1–32.
5. Senthil, D., & Suseendran, G. (2018). Efficient time series data classification using sliding window technique based improved association rule mining with enhanced support vector machine. *Submitted to International Journal of Engineering & Technology*, 7(2), 218–223.
6. Povinelli, R. J., Johnson, M. T., Lindgren, A. C., & Ye, J. (2004). Time series classification using Gaussian mixture models of reconstructed phase spaces. *IEEE Transactions on Knowledge and Data Engineering*, 16(6), 779–783.
7. Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2001). An online algorithm for segmenting time series. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference*, pp. 289–296.
8. Schmidt, R., & Gierl, L. (2005). A prognostic model for temporal courses that combines temporal abstraction and case-based reasoning. *International Journal of Medical Informatics*, 74(2–4), 307–315.
9. Ghalwash, M. F., & Obradovic, Z. (2012). Early classification of multivariate temporal observations by extraction of interpretable shapelets. *BMC Bioinformatics*, 13(1), 1–12.
10. Thiagaraj, M., & Suseendran, G. (2017) Review of chronic kidney disease based on data mining proceedings of the 11th INDIACOM; INDIACOM–2017; IEEE Conference ID: 40353 2017 4th International Conference on “Computing for Sustainable Global Development”, 01st–03rd March, 2017 Bharati Vidyapeeth’s Institute of Computer Applications and Management (BVI-CAM), New Delhi (INDIA), pp. 2873–2878.
11. Thiagaraj, M., & Suseendran, G. (2017). Survey on heart disease prediction system based on data mining techniques. *Indian Journal of Innovations and Developments*, 6(1), pp. 1–9.
12. Rohini, K., & Suseendran, G. (2016). Aggregated K means clustering and decision tree algorithm for spirometry data. *Indian Journal of Science and Technology*, 9(44), 1–6.

13. Thiagaraj, M., & Suseendran, G. (2018). An efficient heart disease prediction system using modified firefly algorithm based radial basis function with support vector machine. *International Journal of Engineering & Technology*, 7(2), 1040–1045.
14. Senthil, D., & Suseendran, G. (2017). Data mining techniques using time series analysis. In Proceedings of the 11th INDIACom; INDIACom–2017; IEEE Conference ID: 40353 2017 4th International Conference on “Computing for Sustainable Global Development”, 01st–03rd March, 2017 Bharati Vidyapeeth’s Institute of Computer Applications and Management (BVI-CAM), New Delhi (INDIA), pp. 2864–2872.

# Identifying Phishing Through Web Content and Addressed Bar-Based Features



Nureni A. Azeez, Joseph Ade, Sanjay Misra, Adewole Adewumi, Charles Van der Vyver and Ravin Ahuja

**Abstract** Phishing which can also be called spoofing is mainly used to explain an approach being used by Internet scammers or cybercriminals to lure a genuine Internet user into revealing vital, confidential and classified information with the intention of using information gathered against them. Through this form of vulnerability, cyber-criminals use the information obtained to gain access into personal information to rob individual of valuables ranging. Since Internet users have increased gloabally, the number of people accessing email, social media and only transaction has increased accordingly. The upsurge in the number of Internet user has therefore enhanced the nefarious activities of the cybercriminals. Verification and checking of address bar features and content of web were adopted in handling phishing detection in this work. Efforts were made to properly study various features of websites considered as phishing as well as those that match the research work. To verify the efficiency of the system, a dataset comprising of phishing websites was downloaded from the popular phishing sites. To start with, a total number of 110 Universal Resource Locators (URLs) considered to be phishing were verified where the system was able to detect 88 websites considered to be phishing while only twenty two (22) URLs were

---

N. A. Azeez · J. Ade · C. V. Vyver  
Vaal Triangle Campus, North-West University, Potchefstroom, South Africa  
e-mail: [nurayhn1@gmail.com](mailto:nurayhn1@gmail.com)

J. Ade  
e-mail: [adebisijoe@gmail.com](mailto:adebisijoe@gmail.com)

C. V. Vyver  
e-mail: [Charles.VanDerVyver@nwu.ac.za](mailto:Charles.VanDerVyver@nwu.ac.za)

S. Misra (✉) · A. Adewumi  
Covenant University, Ota, Nigeria  
e-mail: [sanjay.misra@covenatuniversity.edu.ng](mailto:sanjay.misra@covenatuniversity.edu.ng)

A. Adewumi  
e-mail: [wole.adewumi@covenatuniversity.edu.ng](mailto:wole.adewumi@covenatuniversity.edu.ng)

R. Ahuja  
University of Delhi, Delhi, India  
e-mail: [ravinahujadce@gmail.com](mailto:ravinahujadce@gmail.com)

detected as non-phishing websites. With this result, the efficiency and accuracy level of the system is put at 80%.

**Keywords** Phishing · Address bar · Web · Features · Detection · Content

## 1 Introduction

The fact that Internet has paved ways for accessing information anywhere in the world cannot be over-emphasized. The ubiquitous nature of World Wide Web has offered a man to conveniently sit in a four-cornered room and be sourcing information in any part of the globe. Electronic transaction is very easy and can be done both in the day and night [1].

With the level of usage of Internet across the globe, there are various challenges currently associated with the full-scale utilization of Information and Communication Technology innovation. The prominent among these challenges are security, scalability and interoperability [2]. Of the three challenges mentioned, security has been identified as the main problem we have [3]. Through insecurity of data and information, many Internet users have become vulnerable. Valuable clandestine and classified information have been altered thereby making its originality and confidentiality questionable [4]. In fact, the level of insecurity has equally discouraged some individuals from making use of the Internet in carry out their day-to-day financial online transactions [5]. This decision could simply be attributed to various financial loss across and damaged reputation caused by insecurity through the handiwork of cybercriminals across the globe [6, 7].

In an attempt to find lasting solution to the challenges of insecurity, specifically, phishing and pharming researchers have conducted various researches with different techniques [8]. Some of these techniques have yielded results but not efficient and sustainable solutions. Against the backdrop of the phishing effect and challenges noticed in this cybercriminal act, thus study attempts to carry out phishing detection by checking address bar-based features and web content [9]. The results obtained so far have provided ample opportunity to Internet users to quickly recognize a typical phishing site and protect oneself from being a victim of nefarious activities of cybercriminals across the world [10, 3].

The paper is structured as follows. The literature review is given in the next Sect. 2. The component of the proposed framework is demonstrated in Sect. 3. The implementation of the result and demonstration are given in Sect. 4. Conclusion of the work is summarized in Sect. 5.

## 2 Literature Review

Cybercriminals, specifically Phishers are coming up with different strategies to lure and trick Internet users. Researchers are however, carrying out various studies to detect strategies being used by Phishers [11, 12].

A method that explains the inconsistencies between the structural features, website's identity and the HTTP transactions was developed in 2006 by Pan and Ding. They achieved this by proposing a novel technique that is based on DOM for anti-phishing. Through this technique, a typical phishing website will demonstrate a high degree of abnormality between HTTP transaction and DOM object [13].

The structural feature of the webpages will be inspected by the phishing detector. With this approach, there is no need for online transactions as users do not need to change their behaviours [8]. The authors are, however, of the opinion that this technique will perform better to give high positive rate if combined with other reliable and efficient techniques [14].

Fazelpour [15] used Poisson probabilistic theory, Bayesian probabilistic and K-Nearest Neighbor to classify emails to either fraudulent or non-fraudulent. They combined the results of various algorithms by using ensemble approach to attain reasonable and reliable accuracy level.

Kan and Thi [16] conducted a research by classifying webpages without necessarily taking into consideration their contents but rather by using their URLs. The latter is noted to be very fast since there is no delay when fetching the page content and parsing the text [17]. The various features adopted in this work modelled different sequential dependencies among various tokens. It was concluded in their work that addition of URL segmentation and feature extraction improved the classification rate when compared to other existing techniques. A very similar research was conducted by Kan et al. [16]. They trained various classifiers and were able to improve the result of f-measure.

Justin et al. [18] adopted host-based features and the lexical to identify malicious URLs. The approach used could recognize the components and metadata of important websites without necessarily requesting for any domain expertise. Through this research, they evaluated close to 30,000 URLs with reliable and excellent results. They were able to realise a very high classification rate of 96% with a very low false positive rate [18].

Justin et al. [19] used one online algorithm to handle and manage some URLs whose characteristics evolve over time. Their system was able to gather updated URL features that were paired and shared “with a real-time feed of assigned and labelled URLs from a pool of mail provider” [20]. Ninety-nine percent (99%) “classification rate” was reported adopting confidence-weighted learning on a balanced dataset [19].

### 3 Conceptual Background and Proposed Framework

Some basic components are needed to guarantee absolute performance of the new proposed phishing detection system. The components are: white list of domain names, contents of the page, phases used in the system and URL features.

#### 3.1 First Phase (*Pre-filtering Phase*)

In a study, Chen et al. [21] considered some of the websites as lawful, used the domain name 2nd level as their registered brand name. In order to verify the identities' consistency, a “Consistency checking Identities (Cld) is established in the first phase; the second phase witnesses the examination of domain name of the input string against the Cld list” [22]. If there is a match of any identity on the Cld with the whole second level domain then the page is categorized as a lawful website. The checking will be carried out on the next phase to determine the status further [23].

#### 3.2 Classification Stage

This section classifies URLs based on particular characteristics itemized below using Naïve Bayes Classifier.

##### 3.2.1 Features Based on Address Bar

Various URL features were adopted to identify a typical phishing website. Those considered are explained hereunder:

###### (1) Randomness of URL (ru)

As stated and established by Chen et al. [21], the contents of URL, as well as lawful webpages, are mostly related. However, the URL contents of a typical phishing website are insignificant hence there is a greater chance of having the feature of long random strings. We can take an example of a web site which is phishing: “<http://signin.ebay.com.87ab3540af65fa59167f076ea075f9f7.ustsecurity.info/>” as well as a lawful website: <http://music.shop.ebay.com>.

As shown in the example given, in the lawful and genuine website, the domain token is concerned with “ebay”. On the other hand, in a phishing website, “ust” is useless and relevant with “ebay”, they (also) consist of long string. These situations may be as a result of the need by the phisher to quickly create phishing webpages by making use of randomly generated strings [24, 25].

Before the randomness of any URL could be determined, all available URLs are divided by the reserved symbols with the motive of obtaining URL tokens. In addition

to that, we also need to divide the domain name. Once produced and generated the required tokens, then RU is evaluated by using  $H = \text{Llog}_2 N$ . We can explain the formula in our case as the number of segments of a given URL token. Token is denoted as  $t$  as depicted in Eq. 1.

The parameters are defined as follows:

$$R(t) = \max(sl\_d(t), sl\_s(t)) * \log_2(as(t)) \quad (1)$$

where

$R(t)$  represents the randomness of token  $t$ ;

$sl\_d(t)$  the number of segments formed by splitting token  $t$  into digits;

$as(t)$  = the number of letters in token  $t$

## (2) Address of IP

Assuming the address of an IP is adopted as an option to the domain name in the URL, for example “<http://146.73.5.155/fake.html>” an Internet user should quickly suspect that someone is attempting to steal his classified and clandestine information. Anytime an IP address is seen as a domain name such a website is regarded as a phishing site.

## (3) URL's having “@” Symbol

Whenever “@” symbol is used in the URL, other characters and symbols preceding it are truncated and completely ignored. Therefore, anytime “@” sign is noted in a given URL, the website is noted a phishing site.

## (4) Redirection with “//”

The presence of “//” in the URL path indicates that an Internet user will be sent to a different website. In this study, the position of “//” within the URL path is critically examined. “//” is expected to appear in the 6th position anytime a URL begins with “HTTP”. If, however, a URL starts with “HTTPS” then the “//” will appear in the 7th position.

## (5) Including Suffix or Prefix Separated by (–) to the Domain

The (–) is not used in any lawful URL. Whenever this is identified in any URL, caution must be taken. It is a common sign that the URL is not a genuine or lawful one.

## (6) Sub Domain and Multi Sub Domains

Whenever a domain name includes country-code top-level domains (ccTLD), a second-level domain (SLD) and the actual name of the domain, producing a rule for extracting these features, there is need to exclude “www” from the given URL which is considered, “a subdomain”. Thereafter, the need to remove ccTLD and later counting the dots is available in the URL. Anytime, the number of dots is more than one, the website is tagged “suspicious”. Conversely, if the dot is more than two, it is regarded as “Phishing” because it will consist of many subdomains. Finally, if there is no subdomain in the URL, the system is classified as “Lawful” [26].

### 3.3 Content Features

In some cases, “the features and contents of a typical website can be vividly looked into to verify whether the website is phishing one or not. Document Object Model (DOM) tree, which is for interacting and representing objects in XHTML, HTML and XML is usually used for parsing the webpage after the page has been downloaded”. In this work, after parsing feature vector obtained would be used along with the features of address bar to classify if the website is phishing or non-phishing by using Naive Bayes Classifier [27].

### 3.4 Naive Bayes Classifier

This is a popular technique for carrying out classification with a postulation of absolute independence among various predictors. Naive Bayes is known to perform excellently well when compared with other prominent classification types. Bayes theorem gives a means of determining the  $P(\text{clx})$  from  $P(x)$ ,  $P(c)$  and  $P(x|c)$ -as depicted in Eq. 2.  $P(x|c)$  can be defined as the possibility which is considered as the probability of predictor given class.  $P(c)$  can also be defined as prior probability of class.  $P(x)$  is defined as prior probability of predictor.

Figure 1 provides the step by step way of carrying the experimentation before arriving at the results obtained. The moment the URL is entered, a comparative assessment is observed with trusted domain list to ascertain if it is phishing or not. If there is no concrete assurance regarding the status of the URL, features are extracted for detection purpose after which training and classification will be carried out on the suspicious URL. If URL is considered as phishing, the whole process will stop.

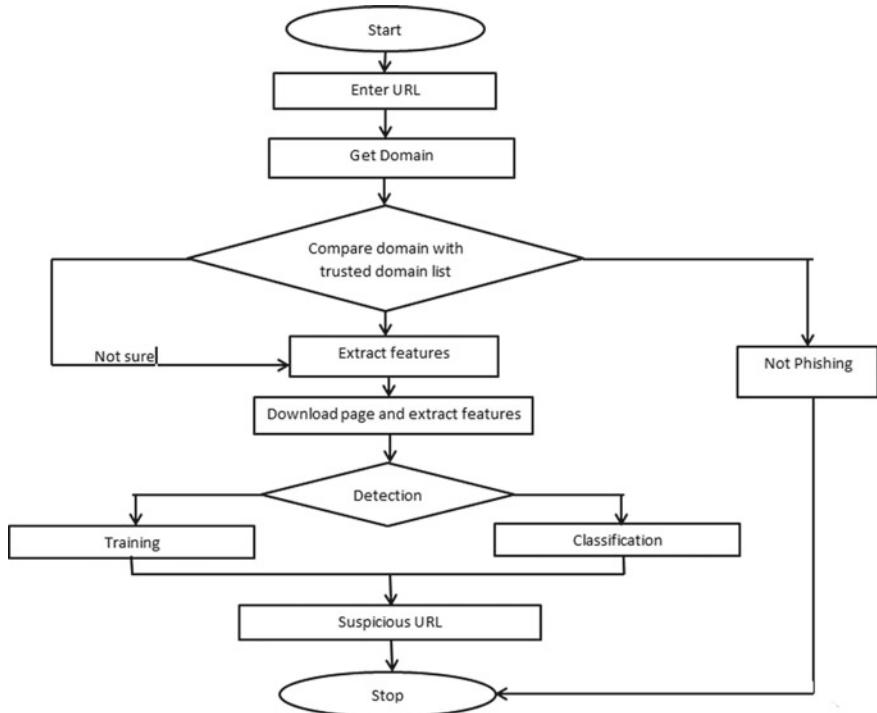
$$\begin{aligned} p(c|x) &= \frac{p(x|c)p(c)}{p(x)} \\ p(c|x) &= p(x_1|c) * p(x_2|c) * \dots * p(x_n|c) * p(c) \end{aligned} \quad (2)$$

where each of the parameters is defined as follows:

$P(\text{clx})$  can be defined as posterior probability of class ( $c$ , target) given predictor ( $x$ , attributes).

## 4 Analysis of Results

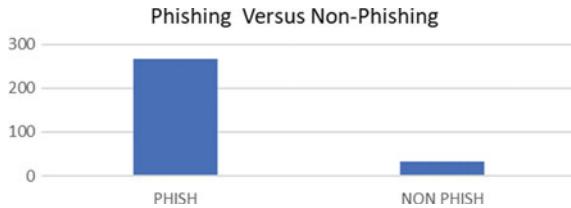
The system designed is web-based. To evaluate the system, efforts were made to download a dataset from Phish Tank. Meanwhile, One Hundred and Ten (120) URLs checked and confirmed to be phishing were tested and system was able to identify and

**Fig. 1** Flowchart of the system

detect eighty eight (88) as phishing sites while twenty two (22) was finally detected and identified as non-phishing URLs. With the results obtained, the accuracy level of the system is at 80% (Table 1).

**Table 1** Outcomes of the phishing detector

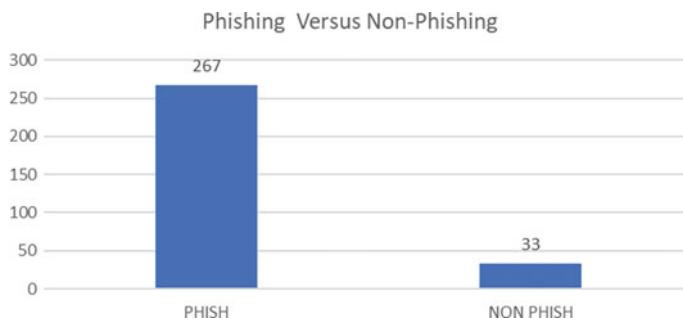
Stage	URLs	Phish	Non phish
A	12	10	2
B	22	11	3
C	32	8	5
D	42	12	0
E	52	8	5
F	62	11	2
G	72	8	4
H	82	12	0
I	92	10	3
J	120	18	5



**Fig. 2** Visual interpretation of the phish detector for 50 URLs

Figure 2, provides the graphical representation of phishing and non-phishing values after the first experimentation.

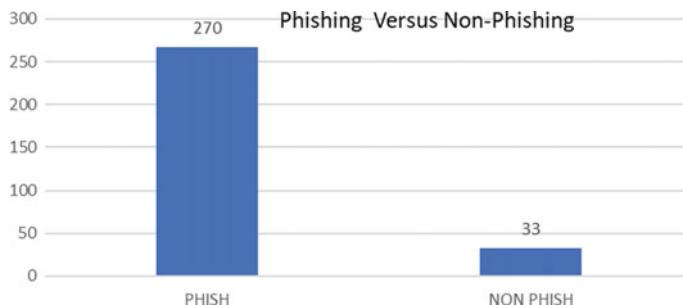
Figure 3 shows a total number of 267 phishing URLs to 33 non-phishing URLs.



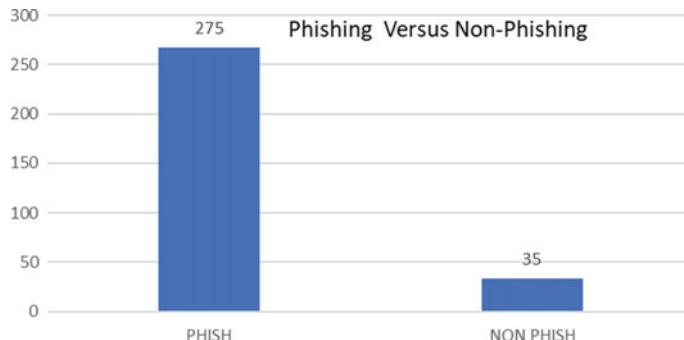
**Fig. 3** Visual interpretation of the phish detector for 60 urls

Figure 4 shows a total number of 270 phishing URLs to 33 non-phishing URLs.

Figure 5 shows a total number of 270 phishing URLs to 35 non-phishing URLs. The detector designed for phishing was able to prove its accuracy and efficiency at 89% based on the features considered for its implementation.



**Fig. 4** Visual interpretation of the phish detector for 100 URLs



**Fig. 5** Visual interpretation of the phish detector for 300 URLs

The experimental results prove that how our system performs better. As the results shows that our framework can detect phishing with 80% accuracy, which is an outstanding result because the phishing sites are always flexible, and in such variations our results are very good. With such good results, we can easily conclude that the other methods are not so effective and robust for phishing sites.

## 5 Conclusion

Having realized the effects and monumental loss realized through various handiwork of cybercriminals, the authors considered it a point of duty to proffer solution different from those already in place. It is as a result of the importance and urgency in the study that we considered address bar-based features and web content. At the end of the experimentation, this system could be able to detect phishing with 80% accuracy. The main reason why the system could not be able to determine above 80% accuracy is because of the flexible nature any phishing site. It changes overtime. It is the belief of the authors that the full-scale implementation of the system will go a long way at curbing phishing and protecting cyber users from falling into the hands of cybercriminals.

## References

1. Kirda, E. & Kruegel, C. (2005). Protecting users against phishing attacks with antiphish. In *Computer Software and Applications Conference, COMPSAC*.
2. Saberi, A., Vahidi, M., & Bidgoli, B. M. (2007). Learn to detect phishing scams using learning and ensemble methods. In *Proceedings of the 2007 IEEE/WIC/ACM*.
3. Azeez, N. A. & Venter, I. M. (2013). Towards ensuring scalability, interoperability and efficient access control in a multi-domain grid-based environment. *South Africa Research*, 104(2), 54–68.

4. Suganya, V., (2016, April). A review on phishing attacks and various anti phishing techniques. *International Journal of Computer Applications*, 139(1), 0975–8887
5. Azeez, N. A. & Ademolu. O. (2016). CyberProtector: Identifying compromised URLs in electronic mails with bayesian classification. *International Conference Computational Science and Computational Intelligence*, pp. 959–965 (2016).
6. Zhang, H. (2004). *The optimality of naive bayes*. New Brunswick, Canada: University of New Brunswick.
7. Pan, X. & Ding, X. (2006). Anomaly based web phishing page detection. In *Proceedings of the 22nd Annual Computer Security Applications Conference (AC SAC'06)*.
8. Azeez, N. A., Olayinka, A. F., Fasina, E. P., & Venter, I. M. (2015). Evaluation of a flexible column-based access control security model for medical-based information. *Journal of Computer Science and Its Application*, 22(1), 14–25.
9. Denning, E. D. (1987). An Intrusion-detection model. *IEEE Transaction on software Engineering*, 222–232.
10. Azeez, N. A. & Babatope, A. B. (2016). AANtID: an alternative approach to network intrusion detection. *Journal of Computer Science and Its Application*, 23(1).
11. Ayofe, A. N., Adebayo, S. B., Ajetola, A. R., & Abdulwahab, A. F. (2010). A framework for computer aided investigation of ATM fraud in Nigeria. *International Journal of Soft Computing*, 5(3), 78–82.
12. Madhusudhanan, C., Ramkumar, C., & Upadhyaya, S. (2006). Phoney: Mimicking user response to detect phishing attacks. In *WOWMOM '06 Proceedings of the 2006 International Symposium on on World of Wireless, Mobile and Multimedia Networks* (pp. 668–672), IEEE Computer Society Washington.
13. Adel, N. T. & Mohsen, K. (2007). A new approach to intrusion detection based on an evolutionary. *Computer Communications*, 2201–2212.
14. Nureni, A. A. & Irwin, B. (2010). Cyber security: Challenges and the way forward. *GESJ: Computer Science and Telecommunications*, 1512–1232.
15. Fazelpour, A. (2016) Ensemble learning algorithms for the analysis of bioinformatics data. A Dissertation Submitted to the Faculty of The College of Engineering and Computer Science in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy, pp. 1–177.
16. Kan, M.-Y., & Thi, H.O.N. (2005). Fast webpage classification using URL features In *CIKM '05 Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 325–326), Bremen, Germany, October 31–November 05, 2005.
17. Ali, M. & Rajamani, L. (2012) *APD: ARM deceptive phishing detector system phishing detection in instant messengers using data mining approach*. Berlin Heidelberg: Springer.
18. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009) Beyond blacklists: Learning to detect malicious Web sites from suspicious URLs. 1245–1254. *KDD'09*, June 28–July 1, 2009, Paris, France.
19. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2011). Learning to detect malicious URLs. In *ACM Transactions on Intelligent Systems and Technology*, (Vol. 2, No. 3, Article 30), Publication date: April 2011.
20. Surana, S. & Warade (2015). Detection and prevention of phishing attacks in web. *International Journal of Scientific Engineering and Technology Research*, 04(08), 1595–1598. ISSN 2319-8885
21. Chen, Y.-S., Yu, Y.-H., Liu, H.-S., & Wang, P.-C. (2014) Detect phishing by checking content consistency. In *2014 IEEE 15th International Conference on Information Reuse and Integration (IRI)*, pp 109–119, IEEE.
22. Sumaiya, T. I. & Aswani, K. C. (2016). Intrusion detection model using fusion of chi-square and multi class SVM. *Journal of King Saud University*.
23. Nidhi, S., Krishna, R., & Rama, K. C. (2013). Novel intrusion detection system integrating layered framework with neural network. In *IEEE 3rd International Advance Computing Conference (IACC)*, Ghaziabad.
24. Anti-phishing working group, (2004). What is Phishing? HYPERLINK “<http://www.antiphishing.org/>” <http://www.antiphishing.org/>.

25. Anti-Phishing Working Group. “Origins of the Word Phishing.”, 2004. URL: HYPERLINK “[http://www.antiphishing.org/word\\_phish.htm](http://www.antiphishing.org/word_phish.htm)” [http://www.antiphishing.org/word\\_phish.htm](http://www.antiphishing.org/word_phish.htm).
26. Smaha, R. E. (1998). Haystack.: An intrusion detection system. In *Proceeding of the IEEE Fourth Aerospace*. Orlando, FL.
27. Chandrasekhar, A. M. & Raghuvee, K. (2013). Intrusion detection technique by using k-means, fuzzy neural network and SVM classifiers. In *2013 International Conference Computer Communication and Informatics (ICCCI)*. Coimbatore, India.

# A Unified Framework for Outfit Design and Advice



**Adewole Adewumi, Adebola Taiwo, Sanjay Misra, Rytis Maskeliunas,  
Robertas Damasevicius, Ravin Ahuja and Foluso Ayeni**

**Abstract** The application of technology in the apparel industry has received significant attention from the research community in recent times. Technology is being leveraged to support the various processes in the supply chain of the industry. On the consumer side, choosing the right outfit for occasions can be quite challenging. It is for this reason that researchers have proposed a number of fashion recommender systems in the literature. Although the proposals in literature cover a number of areas, they are yet to touch on recommendation based on weather. It is also important to harmonise all of the proposals into a unified framework that will help guide developers. The aim of this study therefore is to propose a unified framework for outfit design and advice. The framework is developed using Unified Modelling Language (UML) diagrams and notations, which are globally recognised. In addition, a prototype of an aspect of the framework has also been implemented as proof of concept. We believe that this framework can be leveraged by online fashion stores to better serve their

---

A. Adewumi · A. Taiwo · S. Misra (✉)  
Covenant University, Ota, Nigeria  
e-mail: [sanjay.misra@covenantuniversity.edu.ng](mailto:sanjay.misra@covenantuniversity.edu.ng)

A. Adewumi  
e-mail: [wole.adewumi@covenantuniversity.edu.ng](mailto:wole.adewumi@covenantuniversity.edu.ng)

A. Taiwo  
e-mail: [adebola.taiwo@covenantuniversity.edu.ng](mailto:adebola.taiwo@covenantuniversity.edu.ng)

R. Maskeliunas · R. Damasevicius  
Kaunas University of Technology, Kaunas, Lithuania  
e-mail: [Rytis.Maskeliunas@ktu.lt](mailto:Rytis.Maskeliunas@ktu.lt)

R. Damasevicius  
e-mail: [robertas.damasevicius@ktu.lt](mailto:robertas.damasevicius@ktu.lt)

R. Ahuja  
University of Delhi, Delhi, India  
e-mail: [ravinahujadce@gmail.com](mailto:ravinahujadce@gmail.com)

F. Ayeni  
Southern University, Baton Rouge, USA  
e-mail: [foluso.ayeni@ictd.com](mailto:foluso.ayeni@ictd.com)

customers and can also be implemented as a mobile app to give suitable advice to its end users.

**Keywords** Apparel industry · Fashion recommender · Open source software · Outfit advice · Unified framework

## 1 Introduction

Fashion in contemporary times is in a state of flux and covers a number of areas including—clothing, footwear, accessories and beauty care products [1]. The field of fashion designing is thus an innovative business that involves delivering cutting-edge designs and clothing to customers of all ages, gender [2] and taste. The processes involved in the business of fashion designing have mostly been carried out in an un-automated way until recently. These processes include: the keeping of records about a customer and his/her measurements allowing for updates as and when necessary; the ability of customers to make choices about material for sowing and cloth style from a catalogue as well as ability of customers to pay and get their finished cloths without having to visit the tailor's shop.

In recent times, technology is being leveraged to support the various activities that occur in the supply chain of the fashion industry and in particular, help the consumer in being able to choose the right outfit for occasions, which constitutes a major challenge for most consumers [3]. It is for this reason that researchers have proposed fashion recommender systems in the literature. Recommender systems are different from each other based on “the approach used to analyse the collected data sources to establish notions of affinity among users and items, which can be applied to determine well-matched pairs”. “Collaborative Filtering systems are based on historical interactions analysis alone, while Content-based Filtering systems analyse based on concrete attributes; and Hybrid systems are a combination of the two approaches. The recommender system architecture and their evaluation on real-world problems is an active area of research” [4].

Recommendation systems that have been proposed in the fashion domain use searching techniques that analyse the suggestion, which is a desired or similar clothing feature from online databases. Some recent ones employ techniques such as long short-term memory (LSTM) [5], deep learning [6, 7], statistics, user modelling, rule construction [8, 9], ontology [10] to aid recommendation. Existing websites are based on “either shop statistics (collaborative filtering) or on simple key word matching. These systems do not consider important attributes such as style, fashion, age and importance of the features for users. In this work, the implemented recommendation system is based on concrete attributes of clothing descriptions, including weather, gender, occasion, and complexion”.

Although the proposals cover a number of areas, they mostly do not touch on recommendation based on weather. In addition, given the numerous propositions made in literature, it is important to harmonise them into a singular unit hence

making it easier to implement by developers. The aim of this study therefore is to propose a unified framework for outfit design and advice.

The rest of this paper is structured as follows: Sect. 2 reviews relevant and related literature while Sect. 3 goes on to propose the unified framework. In Sect. 4, the framework is demonstrated by developing a proof of concept. Section 5 discusses the findings and makes recommendation based on this. Section 6 concludes the paper.

## 2 Related Works

Fashion recommender systems have been proposed for two categories of users namely: the fashion designers as well as the consumers.

Tu and Dong in [11] proposed an intelligent personalised fashion recommendation system. The framework comprises of three distinct models namely: “interaction and recommender model, evolutionary hierarchical fashion multimedia mining model and colour tone analysis model which work together to give recommendations”. Style, favourite colour and skin colour were considered as the personalised index when recommending clothing matching.

Vogiatzis et al. [12] proposed a personalised clothing recommendation system for users that combines, “knowledge derived from fashion experts with the preferences of users towards garments”. The knowledge gathered from the experts is encoded as ontology.

Zeng et al. [13] proposed a “perception-based fashion recommender system for supporting fashion designers in selecting the best-personalised fashion design scheme”. It comprises of two distinct models, which work together to give recommendations. The two models characterise the relation between human body measurements and human perceptions on human body shapes.

Liu et al. [14] introduces a fashion recommendation system comprising of two subsystems. The first is the magic closet, which is an occasion-oriented clothing recommendation system while the second is Beauty E-expert for facial hairstyle and makeup recommendations. The system employs a latent Support Vector Machine-based recommendation model.

Ajmani et al. [15] presented a “method for content-based recommendation of media-rich commodities using probabilistic multimedia ontology. The ontology encodes subjective knowledge of experts that enables interpretation of media based and semantic product features in context of domain concepts. As a result, the recommendation is based on the semantic compatibility between the products and user profile in context of use. The architecture is extensible to other domains of media-rich products, where selection is primarily guided by the aesthetics of the media contents”.

Wakita et al. [16] proposed “a fashion-brand recommendation method that is based on both fashion features and fashion association rules”. The study was aimed at improving the accuracy of recommendation in Web services that sell fashion clothes.

The study found that combining the two attributes gave a better recommendation as against using the attributes individually.

Wang et al. [17] proposed a “fuzzy logic-based fashion recommender system to select the most relevant garment design scheme for a specific consumer in order to deliver new personalised garment products”. The system finds relevance in recommending fashion-related products and human-centred products in e-shopping.

Piazza et al. [18] designed and implemented an intuitive user interface for browsing a fashion product range based on visual preferences. The goal is to provide an assistive consumer technology to consumers who are hitherto unaware of their actual preferences and are overwhelmed with choice-overload.

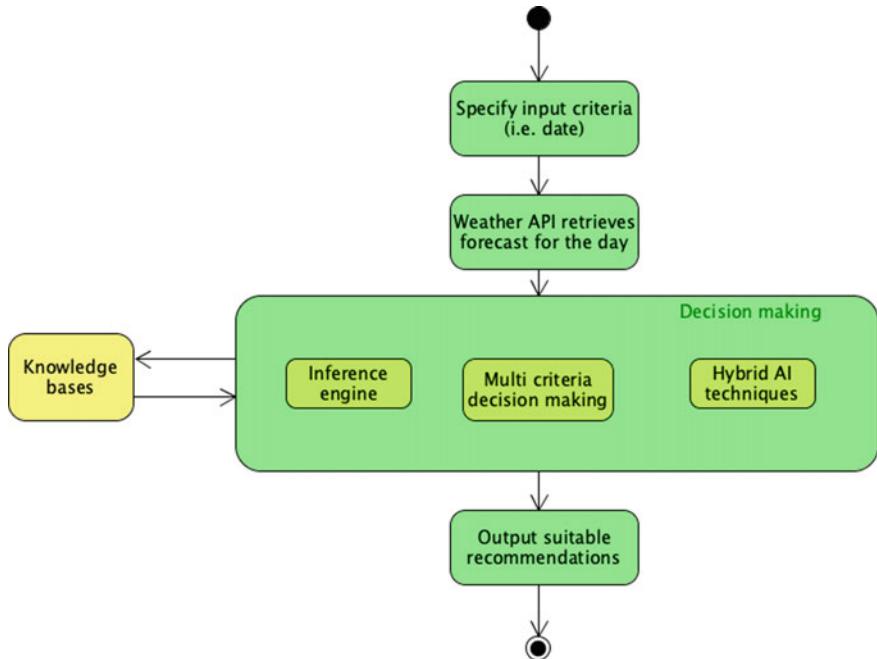
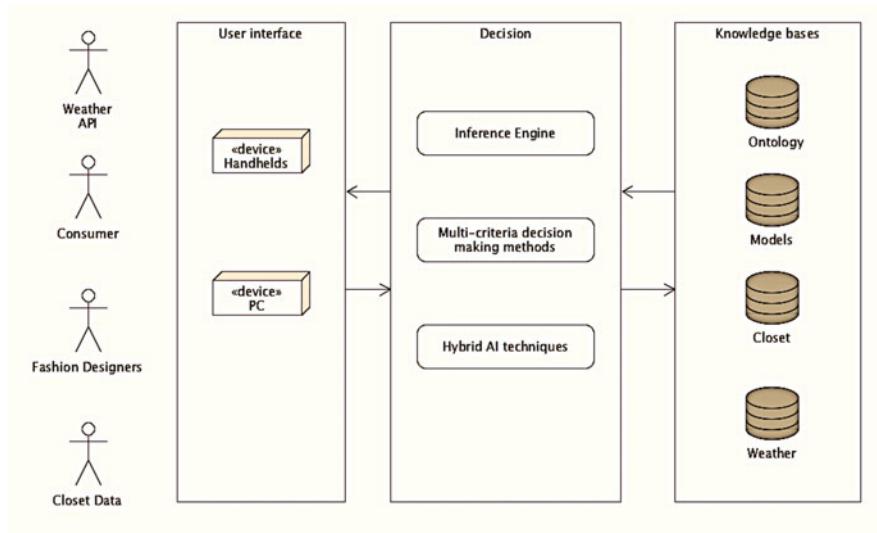
Wang et al. [19] proposed a “method to recommend handbags to e-shoppers based on the images the shopper clicks”. It was observed in the study that online shoppers preferred handbags with similar attributes such as pattern, colour or style at different periods of time. For that reason, the authors formulated an experimental approach that enables the projection of high-dimensional feature space (various handbags) into lower dimensional attribute space (handbags grouped based on their pattern, colour or style). The approach proved to be promising given the positive outcome obtained. However, the focus was on handbags which is just one aspect of fashion recommendation.

Zhang et al. [20] attempted to support travellers’ in packing for their trips by addressing the challenge of clothing recommendation based on social photos from travel destinations. This was achieved by employing “a hybrid multi label convolutional neural network combined with the support vector machine approach” [20]. The essence of this combination was to comprehend the intrinsic and composite association between clothing attributes and location attributes. The results depicted that the approach outperformed a number of baseline and alternative methods. In addition, the viability of the approach was further validated through a case study.

### 3 Proposed Framework

The activity diagram of the proposed framework is presented in Fig. 1. The user specifies the input parameter which is the date. The weather API then uses the input to check the weather condition for that given date. The input is passed to the decision-making engine. This can either be an inference engine, multi-criteria decision-making methodology or hybrid AI technique. Based on the results generated by the decision-making module, the output recommendations are then displayed to guide the user in choosing the appropriate apparel for the given day.

A layered architecture is used to depict the framework as observed in [11, 16]. It comprises three main subsystems, which summarises the main processes in a recommendation system. The framework is depicted in Fig. 2.

**Fig. 1** Activity diagram of proposed framework**Fig. 2** Unified fashion recommender system

### 3.1 *The Actors*

This refers to every entity that interacts with the recommender system. They can be classified into two broad categories, which include: human agents and software agents. In Fig. 2, the human agents include the fashion designers and the consumers. The fashion designers interact with the framework so as to make informed decisions when designing clothes for their varied customers. The consumers interact with the framework so as to be able to select suitable clothing to suit various weather conditions. Closet data and weather application programming interface (API) are the non-human or software agents. Weather API is a service that collects the weather forecast of a given location. The data it collects is stored in the knowledge base of the proposed framework. Closet data on the other hand represents images of a user's closet, which is also stored in the knowledge base of the proposed framework and used for decision-making.

### 3.2 *User Interface*

This refers to the medium through which the actors described in the previous subsection interact with the recommender system (i.e. supplying input and getting feedback/suggestions as output). From the framework in Fig. 2, there are two main user interface media: handhelds and PC. Handhelds refer to all manner of portable devices such as tablets and smartphones. On the other hand PC refers to all kinds of desktop and notebook computers. These all perform the function of allowing users to supply input to the recommender system and get feedback.

### 3.3 *Decision*

This is the layer where recommendation is done based on the input from actors as well as the knowledge from the knowledge bases. Inference engines make decisions based on ontologies that have been crafted in the knowledge base of the recommender system [12]. These can be leveraged in reaching suitable decisions. Multi-criteria decision-making (MCDM) methods on the other hand are used to process hierarchical models. There are quite a number of such methods and they include: Analytical Hierarchy Process (AHP), Analytical Network Process, Decision Trees and Technique for order preference by similarity to ideal solution (TOPSIS) to mention a few.

Hybrid artificial intelligence (AI) techniques refer to an infinite combination of two or more AI techniques so as to analyse user preferences and give suitable recommendation. These AI techniques can include: “artificial neural networks, expert systems, genetic algorithms, and fuzzy-logic” to mention a few.

### 3.4 Knowledge Bases

This refers to the data store of the recommendation system, which comprise of closet data, weather data (as collected from the weather API) and either of ontology or models. Closet data is crucial as this forms the basis on which the consumer makes decisions. This data is represented as images, which can be analysed using convolutional neural network architectures [20]. The weather API as earlier mentioned gathers weather forecast of a given location and stores it in the weather data store. It also constantly updates the data.

Ontologies centred on the fashion domain can be constructed so as to capture domain expertise [12]. Similarly, models can be developed to optimise the analysis of fashion features and consumer or designer preferences [11]. Such models would then be evaluated using MCDM methods in the decision layer (see Fig. 2).

## 4 Implementation of the Proposed Framework

In order to evaluate the framework, we built a prototype fashion store by leveraging on an array of open source tools namely: PHP, Hypertext Markup Language (HTML) and Cascading Style Sheets (CSS). This section shows the various interfaces and discusses their functionality.

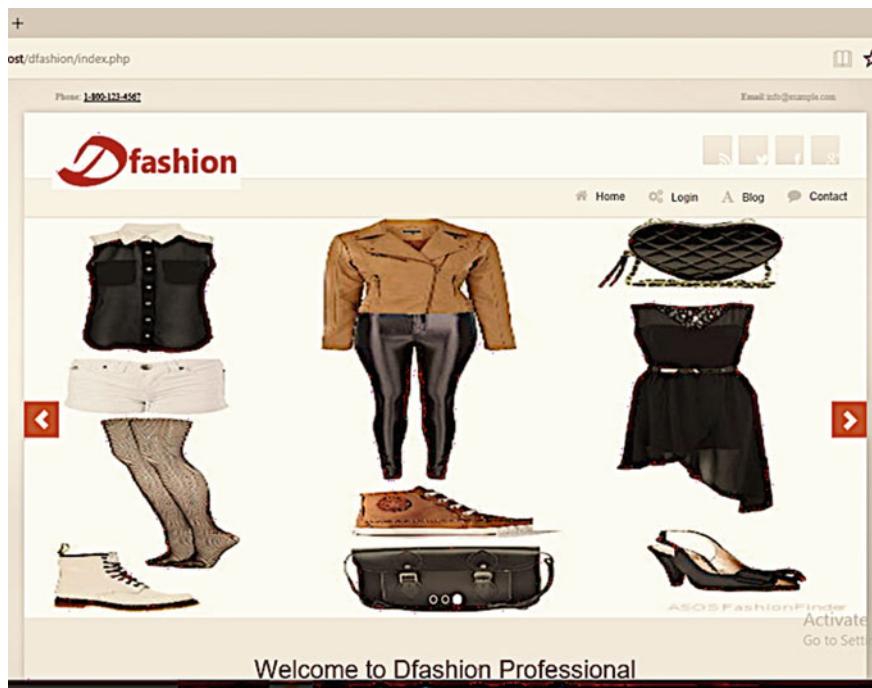
### 4.1 Home Page

This is the first page that a shopper comes across when s/he visits the site. Although it can be accessed using different interfaces as depicted in our proposed framework, Fig. 3 depicts the store's home page through a PC.

The home page being a window into the online store should allow shoppers to see first hand items sold in the store without needing to browse through every page of the site. This is demonstrated in Fig. 3. A shopper can simply swipe through the thumbnails to get a feel of what items the fashion store offers. When a specific item of choice is spotted, the user can proceed to get more details by clicking on that specific image. This helps to streamline decision making for shoppers [18].

### 4.2 Measurement Page

This page allows a user to specify custom designs by entering their style measurements into the provided fields. The system then stores it so that the designer can retrieve it. Figure 4 depicts the measurement page. Converting the various style inputs



**Fig. 3** Home page of the online store

**Fig. 4** Measurement page

into tags for the closet data store can also extend the concept of this page. By doing this, users will be able to receive recommendations based on style measurements.

### 4.3 Recommendation Page

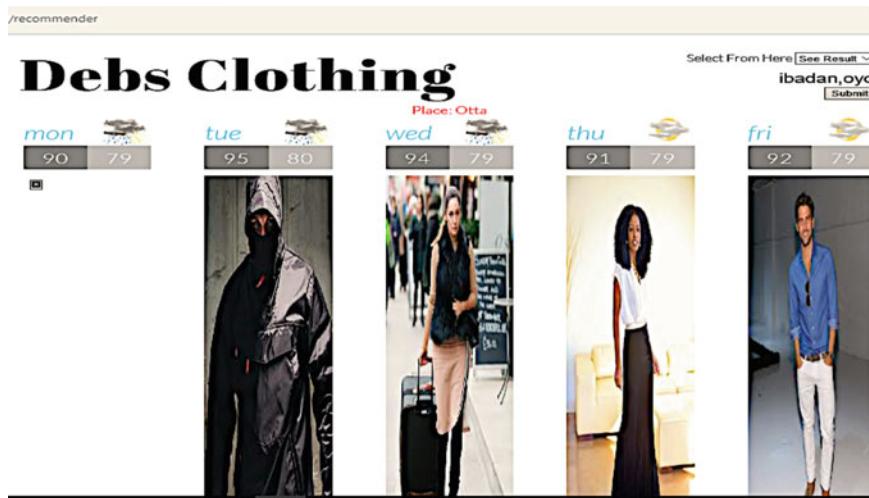
This page suggests to the user, suitable outfit to be worn for five working days of a week by leveraging on the data it obtains from the weather forecast API web service. This is depicted in Fig. 5.

From Fig. 5, we observe that the forecast of five days of the week (Monday to Friday) is shown. A picture below each day also shows the kind of clothing to be worn on such days. On Tuesday being a rainy day, for instance, the user is advised to wear thick clothing to protect from the downpour.

## 5 Discussion

The framework developed in this study has worked at unifying the various recommender system proposals in literature. We also provide the following recommendations to developers who would want to implement applications that leverage this framework:

1. In implementing the user interfaces for the recommender framework, developers can leverage cross-platform technologies such as Xamarin, Apache Cordova,



**Fig. 5** Recommendation page

- Corona, React Native. This will afford them the opportunity to have just one code base that can be adapted to different screen sizes.
2. Also, in implementing the framework, it is not compulsory that all the subsystems be implemented. For instance, a developer might wish to realise a recommender system by simply developing a hierarchy model (knowledge base) applying a multi-criteria decision-making method to analyse the model and developing a web application to both receive input and display recommendation on a PC [21, 22].
  3. A folksonomy can be generated from images that are captured into any application developed based on the framework. This can then be used to enhance retrieval from the system.

## 6 Conclusion

As a result of the growing interest of researchers in the fashion domain, there have been a number of proposals especially for fashion recommender systems. Given the challenge that this poses to developers who would want to implement them, this study has worked at unifying the proposals in a newly proposed framework. An instance of the framework has been developed to depict certain aspects of the framework. This framework instantiation has been done with the aid of open source tools for web development namely: HTML, CSS and PHP. As future work, we plan to extend the application further and also perform usability evaluation of the system.

**Acknowledgements** We acknowledge the support and sponsorship provided by Covenant University through the Centre for Research, Innovation and Discovery (CUCRID).

## References

1. Juhlin, O., Zhang, Y., Wang, J., & Andersson, A. (2016). Fashionable services for wearables: Inventing and investigating a new design path for smart watches. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (pp. 49–58). ACM.
2. Banerjee, D., Ganguly, N., Sural, S., & Rao, K. S. (2018). One for the road: Recommending male street attire. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (pp. 571–582) Springer.
3. Zeng, X., Zhu, Y., Koehl, L., Camargo, M., Fonteix, C., & Delmotte, F. (2010). A fuzzy multi-criteria evaluation method for designing fashion oriented industrial products. *Soft Computer*, 14, 1277–1285.
4. Konstan, J. A., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22, 101–123.
5. Jiang, Y., Xu, Q., Cao, X., & Huang, Q. (2018). Who to ask: An intelligent fashion consultant. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, (pp. 525–528).

6. Tangseng, P., Yamaguchi, K., & Okatani, T. (2017). Recommending outfits from personal closet. In *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)* (pp. 2275–2279).
7. Kalra, B., Srivastava, K., & Prateek, M. (2016). Computer vision based personalized clothing assistance system: A proposed model. In: *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)* (pp. 341–346) IEEE.
8. Iliukovich-Strakovskaia, A., Tsvetkova, V., Dral, E., & Dral, A. (2018). Non-personalized fashion outfit recommendations. In: *World Conference on Information Systems and Technologies* (pp. 41–52) Springer.
9. Han, X., Wu, Z., Jiang, Y. G., & Davis, L. S. (2017). Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 2017 ACM on Multimedia Conference* (pp. 1078–1086) ACM.
10. Goel, D., Chaudhury, S., & Ghosh, H. (2017). Multimedia ontology based complementary garment recommendation. In: *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (pp. 208–213).
11. Tu, Q. & Dong, L. (2010). An intelligent personalized fashion recommendation system. In: *2010 International Conference on Communications, Circuits and Systems (ICCCAS)* (pp. 479–485). IEEE.
12. Vogiatzis, D., Pierrakos, D., Palioras, G., Jenkyn-Jones, S., & Possen, B. J. H. H. A. (2012). Expert and community based style advice. *Expert Systems with Applications*, 39, 10647–10655.
13. Zeng, X., Koehl, L., Wang, L., Chen, Y. (2013). An intelligent recommender system for personalized fashion design. In: *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)* (pp. 760–765). IEEE.
14. Liu, S., Liu, L., & Yan, S. (2013). Magic mirror: An intelligent fashion recommendation system. In: *2013 2nd IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 11–15). IEEE.
15. Ajmani, S., Ghosh, H., Mallik, A., & Chaudhury, S. (2013). An ontology based personalized garment recommendation system. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (vol. 03, pp. 17–20). IEEE Computer Society.
16. Wakita, Y., Oku, K., Huang, H. H., & Kawagoe, K. (2015). A fashion-brand recommender system using brand association rules and features. In: *2015 IIAI 4th International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 719–720). IEEE.
17. Wang, L. C., Zeng, X. Y., Koehl, L., & Chen, Y. (2015). Intelligent fashion recommender system: Fuzzy logic in personalized garment design. *IEEE Transactions on Human-Machine Systems*, 45, 95–109.
18. Piazza, A., Zagel, C., Huber, S., Hille, M., & Bodendorf, F. (2015). Outfit browser—an image-data-driven user interface for self-service systems in fashion stores. *Procedia Manufacturing*, 3, 3521–3528.
19. Wang, Y., Li, S., & Kot, A. C. (2015). Joint learning for image-based handbag recommendation. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, (pp. 1–6). IEEE.
20. Zhang, X., Jia, J., Gao, K., Zhang, Y., Zhang, D., Li, J., et al. (2017). Trip Outfits advisor: Location-oriented clothing recommendation. *IEEE Transactions on Multimedia*, 19, 2533–2544.
21. Adewumi, A., Misra, S., Omoregbe, N., & Fernandez, L. (2013). Quantitative quality model for evaluating open source web applications: Case study of repository software. In *2013 IEEE 16th International Conference on Computational Science and Engineering (CSE)*.
22. Olokunde T., Misra S. & Adewumi A. (2017). Quality model for evaluating platform as a service in cloud computing. In R. Damasevičius & V. Mikäšyté (Eds.), *Information and Software Technologies. ICIST 2017. Communications in Computer and Information Science*, (vol. 756) Cham: Springer.

# Contextual Relevance of ICT-CALL Modules and Its Role in Promoting Employability Skills Among Engineering Students—A Study



**Yeddu Vijaya Babu**

**Abstract** This research paper deals with executing a task to examine contextual relevance of ICT-CALL modules in promoting employability skills among engineering students. This empirical study is based on ensuing perceptions of the learners in the context of imparting useful communication skills promoting employability skills through the ICT-CALL modules. This study probes mainly on how these skills enable individual interest in encouraging skills among the students based on the data collected from the students as a pilot study in one of the reputed national institutes of technologies in India. Theoretical groundwork for task has been demonstrated to the respondents. This task emphasizes mainly on the use of the learning modules and activities practiced in the language labs in promoting employability skills and to propose appropriate recommendations as well. This research asserts to be pragmatic, aiming to meet the objectives of the study, and this study is confined to the data collected from the 30 engineering students from one of the technology institutes in the state of Chhattisgarh in India.

**Keywords** ICT-CALL modules · Employability skills · Relevance · Suggestions

## 1 Introduction

The twenty-first-century science and technology envisages numerous changes in the pedagogy of English Language Teaching (ELT). It is a medium negotiator for enhancing effective employability skills among the learners. Information and Communication Technology (ICT) has prompted new possibilities into the classroom in innovative teaching–learning methodologies such as CALL (Computer-Assisted Language Learning) and MALL (Mobile-Assisted Language Learning). The National Knowledge Commission (NKC) report (2008), India emphasizes the need for effective communicative and entrepreneurial skills for upcoming professional students. Inno-

---

Y. V. Babu (✉)

Department of Humanities and Social Sciences, National Institute of Technology,  
Raipur 492010, India

e-mail: [yvbabu.eng@nitrr.ac.in](mailto:yvbabu.eng@nitrr.ac.in)

vation and start-ups are the instant calls for young professionals. These essential skills enable to create ample opportunities for the upcoming professional students to utilize resources, enhance competency, and accomplish in skilled job orientation. Effective communication skills are recognized as the single most important element which can “make” or “un-make” a person’s career.

## 2 EST (English for Science and Technology)

The students of science and technology need English language communication skills must be relevant to the technical field and also it should fulfill the communicative needs of the learners. It should be designed to cover the high-frequency technical terminology and should systematically impart the grammar used in the field.

## 3 Advantages of ICT-CALL Modules for Enhancing Employability Skill Among Engineering Students

The ICT-CALL module in the language laboratories is the effective medium which provides plentiful opportunities to utilize resources for the learners to obtain required skills and competencies for the learners by that they can land on skilled job platform. In fact, effective communication skills are recognized as the single most important element which can “make” or “un-make” an individual’s career.

Gardner and Lambert [1] highlighted the enormity of attitude in language learning. According to them, the institution of “family” plays a crucial role either in enhancing a positive or negative attitude toward other communities. The learners build up positive attitude with other communities if the family supports the community. Besides, they suggested that a learner’s orientation toward the other community develops with the family. This kind of orientation in turn helps the learner to learn the language successfully, i.e., the success in language depends on the “learners’ attitude” toward the target language.

ICT has become an integral part of Education Technology (ET), being implemented by the technical and professional institutions of learning to make use of digital technologies which are essential for facilitating useful communication skills and soft skills for learners. In contrast with the traditional ways of teaching, for example, teaching with chalk and talk method, educational technology in integrating with ICT modules becomes teaching–learning activity effectively.

According to Frey [2], “the project method originates from Pragmatism, the philosophical movement which appeared in the middle of the nineteenth century and promotes action and practical application of knowledge in everyday life.”

## 4 Goals and Objectives

This study has specific goals:

- To study and analyze the contextual relevance, strengths, and weaknesses of ICT-CALL course modules in empowering employability skills to be successful.
- To explore the reasons for learners' deficiencies in acquiring effective communication skills and to identify causes of CA (Communication Apprehension) during interview presentation.
- To facilitate alternative solutions to resolve the communication problems of the learners by developing a model CALL on employability skills course module (teaching–learning) to empower employability skills (especially those who come from L1 Medium of MTI as vernacular study background).

## 5 Methodology

Effective communication skills are essential employability skills which have been included at various levels in any engineering curriculum. First, the participants have explained the gradations of 5-point Likert scale assessment on contextual relevance of ICT-CALL modules in promoting employability skills based on learners' experiences: the groundwork, the procedure, etc. They were provided a mock practice session based on which they were prompted to review their weak points.

Here, in this study, the researcher has selected Likert scale for measuring the data. It is a convenient method as a Likert scale is a psychometric scale generally used in surveys and is extensively used in survey-based investigations. In the processes of responding questions in the Likert scale, respondents identify their point of concurrence to the given statement. Consisting of twenty items, the questionnaire was assembled in the form of a 5-point Likert scale (where 1 points out strongly disagree and 5 strongly agree)

**From 1–10:** Item 1 pertains to the question on the participants like their multimedia English language lab classes. Item 2 asked on whether the participants have computer-assisted language lab (Multimedia language lab) or similar facility in their college/institution and Item 3 on whether the participants attend language lab classes regularly; Item 4 is related to know whether CALL Lab (multimedia) lessons help them to develop LSRW, i.e., Listening, Speaking, Reading, and Writing skills; Item 5 explores whether the CALL Lab exposes to a variety of self-instructional learning. Item 6 elicits whether CALL modules enhance their pronunciation and conversational skills. Item 7 brings out whether multimedia lab develops learner friendly ways of language learning. Item 8 is to find out the need to improve students' CALL Lab modules develop their oral presentation–extempore skills. Item 9 is asked to find out the CALL modules that provide exercises on problem-solving, conflict management and leadership skills. Item 10 analyzes whether the use of multimedia language lab modules provides entrepreneurial and self-sustainability skills.

**From 11–15:** Item 11 relates to the question to know if CALL lab will help the learners to develop their interview (facing interview) skills; Item 12 asked on whether the CALL facilitates sufficient audio-visual exercises in enhancing their professional communication skills; Item 13 asks if their teachers in multimedia language lab are resourceful in conducting mock practice on debates, group discussions, and job interviews; Item 14 is related to know if there is a regular up gradation of modules in the multimedia language lab and Item 15 explores whether they need more CALL multimedia lab sessions to get trained in employability skills.

Computer-mediated language learning in the form of providing learning modules through effective digital medium has become essential methodology which included at various levels in teaching–learning of communication skills and soft skills engineering curriculum. First, the participants have explained the gradations of 5-point Likert scale assessment on writing skills based on learners' experiences: the groundwork, the procedure, etc. They were provided a mock practice session based on which they were prompted to assess their weak points.

This research study was based on the experiment conducted among thirty (30) respondents pursuing undergraduate engineering in one of the reputed National Institute of Technologies, India as part of the pilot research investigation. To assess the students' needs taking into account the specific purposes for which contextual relevance of ICT-CALL modules in promoting employability skills among engineering students, the scope of research provides the researcher to go further to make suggestions with this empirical study in order to develop effective employability skills among engineering students.

---

#### Research Tools

---

Researcher's observations

---

Questionnaire

---

Collection of feedback

---

The questionnaire consists of 15 questions which are based on the Likert-type scale. The questionnaire is constructed in a systematic manner. The process involves a number of interrelated steps. The steps used in constructing a questionnaire are as follows:

- Ranging from strongly disagree to disagree.
- Ranging from agree to strongly agree.
- The third rating is undecided where there is no compulsion on the respondent, and the compulsion is avoided on the respondent in giving the specific response. The responses are revealed in frankness by choosing the choices.

## 6 Analysis and Findings

The researcher has taken keen steps to collect the stratified samples from the participants. The participants' information related to data collection had been kept confidential according to the interest. The test questionnaire comprises 20 [3] questions.

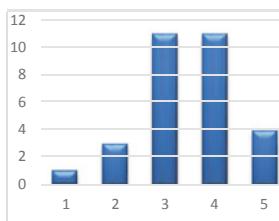
(The researcher has enclosed the questionnaire **annexure-I**).

## 7 Data Interpretation and Analysis

In the following, Likert scale graphical interpretation X-axis represents options (1—strongly disagree, 2—disagree, 3—not decided, 4—agree, 5—strongly agree), and Y-axis represents frequency and bars represent percentage.

- I like my multimedia English Language Lab classes.

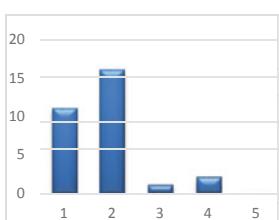
Most of the respondents **agreed** that they like multimedia English Language Lab classes.



Valid	Respon dent	Percent	Valid Percent	Cumulative Percent
1	1	3.3	3.3	3.3
2	3	10.0	10.0	13.3
3	11	36.7	36.7	50.0
4	11	36.7	36.7	86.7
5	4	13.3	13.3	100.0
Total	30	100.0	100	

- I have computer-assisted language lab (Multimedia language lab) facility in my college/institute.

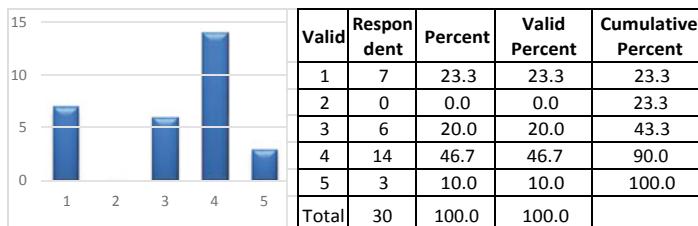
Most of the respondents **disagreed** that they have computer-assisted language lab (Multimedia language lab) facility in my college/institute.



Valid	Respon dent	Percent	Valid Percent	Cumulative Percent
1	11	36.7	36.7	36.7
2	16	53.3	53.3	90.0
3	1	3.3	3.3	93.3
4	2	6.7	6.7	100.0
5	0	0.0	0.0	100.0
Total	30	100.0	100.0	

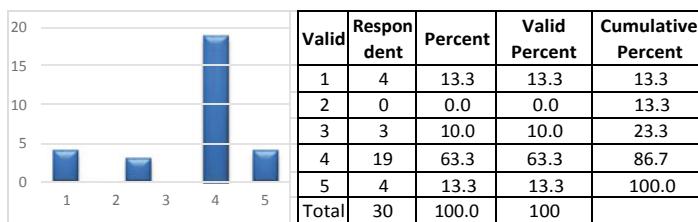
- I attend language lab classes regularly.

Most of the respondents **strongly agreed** that they attend language lab classes regularly.



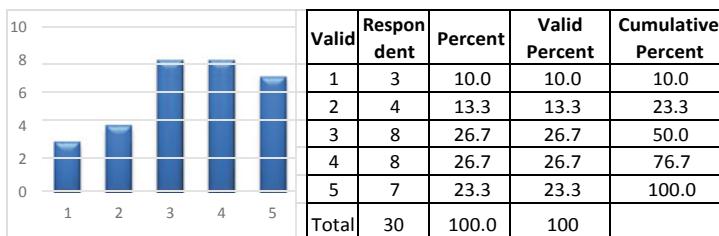
4. CALL Lab (multimedia) lessons help me to develop LSRW Listening, Speaking, Reading, and Writing skills.

Most of the respondents **agreed** that CALL Lab (multimedia) lessons help them to develop LSRW Listening, Speaking, Reading, and Writing skills.



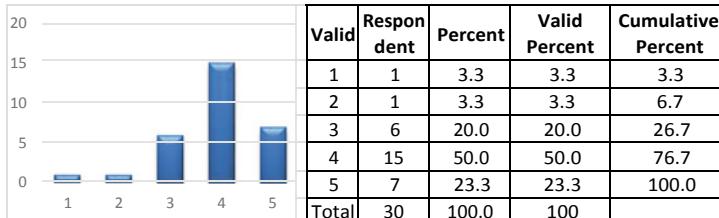
5. CALL Lab exposes me to a variety of self-instructional learning.

Most of the respondents **could not be able to give clear response as agreed with the same** that CALL Lab exposes them to a variety of self-instructional learning.



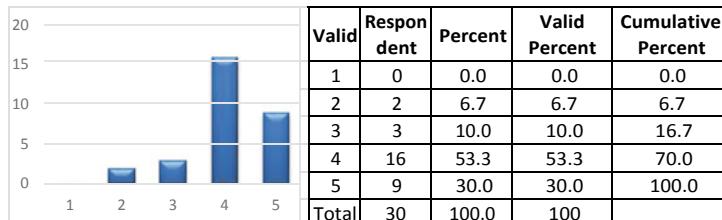
6. CALL modules enhance my pronunciation and conversational skills

Most of the respondents **agreed** that CALL modules enhance my pronunciation and conversational skills.



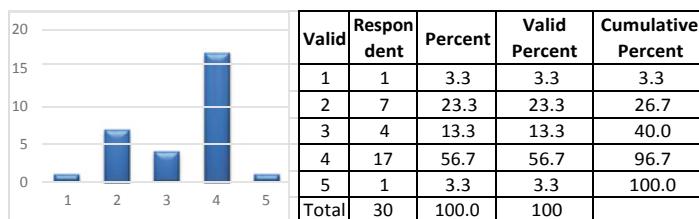
7. Multimedia lab develops learner-friendly ways of language learning.

Most of the respondents **agreed** that Multimedia lab develops learner-friendly ways of language learning.



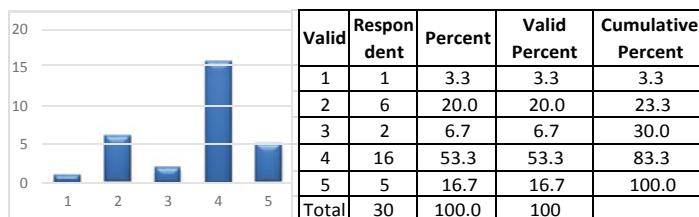
8. My CALL Lab modules develop my oral presentation—extempore skills.

Most of the respondents **agreed** that CALL Lab modules develop their oral presentation—extempore skills.



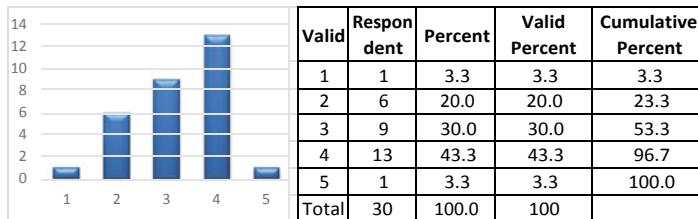
9. The CALL modules provide me exercises on problem-solving, conflict management, and leadership skills.

Most of the respondents **agreed** that the CALL modules provide me exercises on problem-solving, conflict management, and leadership skills.



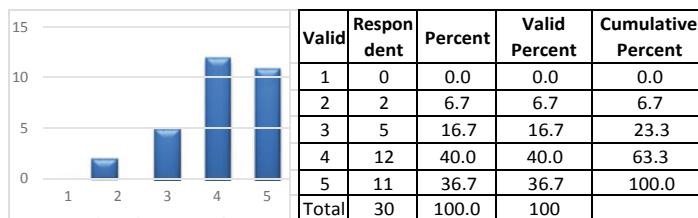
10. The use of multimedia language lab modules provides me entrepreneurial and self-sustainability skills.

Most of the respondents **agreed** that the use of multimedia language lab modules provides their entrepreneurial and self-sustainability skills.



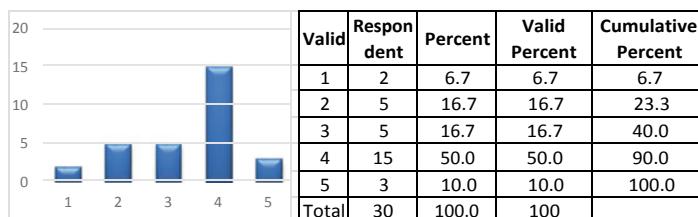
11. The CALL lab will help me to develop my interview (facing interview) skills.

Most of the respondents **agreed** the CALL lab will help them to develop their interview (facing interview) skills.



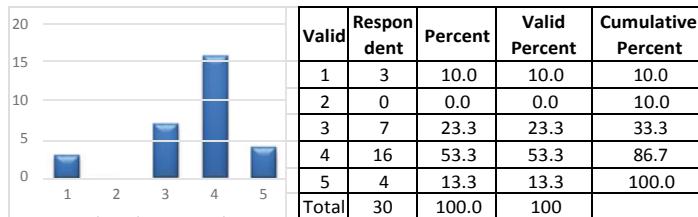
12. The CALL facilitates sufficient audio-visual exercises in enhancing my professional communication skills.

Most of the respondents **agreed** the CALL facilitates sufficient audio-visual exercises in enhancing my professional communication skills.



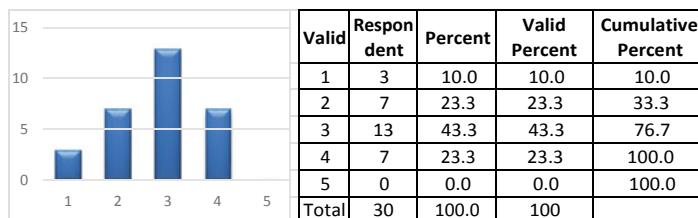
13. My teachers in multimedia language lab are resourceful in conducting mock practice on debates, group discussions, and job interviews.

Most of the respondents **agreed** that their teachers in multimedia language lab are resourceful in conducting mock practice on debates, group discussions, and job interviews.



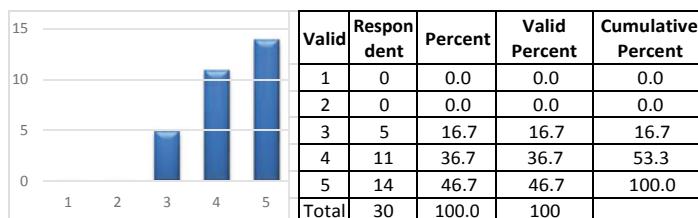
14. There is a regular up gradation of modules in the multimedia language lab.

Most of the respondents **agreed** that there is a regular up gradation of modules in the multimedia language lab.



15. I need more CALL multimedia lab sessions to get trained in employability skills.

Most of the respondents **strongly agreed** that they need more CALL multimedia lab sessions to get trained in employability skills.



## 8 Suggestions and Recommendations

- Overall, the learners expressed their need for computer-assisted language learning modules to learn and practice effective communication skills which would enhance employability skills.

- Institutes/colleges need to establish CALL labs for their engineering students.
- Institutes of engineering and technology may focus on the skill acquisition process for their students especially enhancing their students' employability skills through effective communication skills and soft skills.
- Teachers' objectives in imparting these skills are high end essential. They should be provided training and orientation on the ICT-related CALL lab modules.
- Time-to-time revisions and alternations are to be made to strengthen the modules to meet the requirements of the corporate world.
- Special sessions are required to get trained and acquainted with new technologies and new digital modules for learning.
- Special sessions are required for regional medium background students whose medium of learning is not English from the beginning (vernacular medium).
- Teachers and facilitators are suggested to arrange seminars and group discussion sessions on entrepreneurial and self-sustainability skills.
- There is a need to get a regular rapport with the industry officials to be familiar with the upcoming requirements and needs of effective documentation and corporate communication presentation which will be a significant academic and professional need for professional students.

This empirical research administered though pilot study discloses the fact that ICT-CALL modules for engineering students are essential for enhancing the employability skills and further the research can be extended to get the current status among the professional graduates pursuing engineering from science and technology institutes of learning in the state to find out whether the ICT-CALL modules are promoting employability skills among the students in the colleges of Chhattisgarh, in particular, Chhattisgarh state as a unit for research, as the empirical study may further be extended for the undergraduate engineering students in the districts of C.G. state which will further motivate the researcher to design a useful teaching–learning modules on the essential and useful topics for the students based on the existing needs.

## 9 Conclusion

The overall empirical study based on the perceptions of the respondents is evident that the effective ICT-CALL modules enable the learners to acquire employability skills. Further the study would focus on which has immense significance and implication of ICT-CALL module contents, materials, its relevance with the existing curriculum, level of the students' communication skills, language teachers' role in teaching and training their wards and their growth rate of success in campus placement drives, reasons for deficiencies in accomplishing effective communication abilities and alternative feasible solutions to the problem will be carried out in Chhattisgarh state. This current study witnesses the ICT-CALL modules in enhancing the employability skills among the engineering students in the state is needful and significant.

**Acknowledgements** The author expresses his heartfelt thanks to the honorable Director, Dean R & C and Head Department of HSS for their encouragement and support for writing this research paper as the part related to the SEED project. **Sponsoring Agency:** National Institute of Technology, Raipur, India, Project No: NITRR/Seed Grant/2016-17/007.

The questionnaire is framed and administered to the under graduate students studying engineering from the National Institute of Technology Raipur, Chhattisgarh, India. The author is a faculty member in the same institute. The questionnaire is framed as a part of above said project sanctioned from the same institute where there is no additional approval required for the project. The author is sole responsible for the statement.

## References

1. Gardner, R., & Lambert, W. (1972). *Attitudes and motivation in secondary language learning*. Rowley, M. A: Newbury House.
2. Frey, K. (1986). *The project method*. Thessaloniki, Kyriakidis. (In Greek).
3. Bloor, M. (1984). Identifying the components of a language syllabus: A problem for designers of courses in ESP or communication studies. In R. Williams, J. Swales, & J. Kirkman (Eds.), *Common ground—shared interests in ESP and communication studies* (ELT Documents 117, pp. 15–24). Oxford: Pergamon Press.

# A Clinically Applicable Automated Risk Classification Model for Pulmonary Nodules



**Triparna Poddar, Jhilam Mukherjee, Bhawati Ganguli,  
Madhuchanda Kar and Amlan Chakrabarti**

**Abstract** Lung cancer has the highest prevalence in cancer-related deaths due to its rapid progression and it is detected at advanced stages. The paper proposes a novel method for predicting the risk of being malignant of Pulmonary Nodule (PN), presence of which can be an indication of lung cancer, with the motive to reduce the number of unnecessary biopsies and prevent anxiety among the patients. The study has considered different morphological features along with the clinical history of the patient having the particular nodule as described in medical literature. Depending upon these features, we have classified the risk of being malignant of pulmonary nodule into two classes, namely, low-risked or benign and high-risked or malignant. The entire dataset required to design the model is collected from a retrospective dataset, containing 476 (401 Malignant or high-risked and 75 low-risked or benign) PNs. The classification is performed by Recursive Partitioning Algorithm (RPA). RPA not only improves the accuracy but also helps to interpret how the morphological features are classifying the true risk of being malignant of the nodules.

**Keywords** Pulmonary nodule · Morphological features · Decision tree · Recursive partitioning · Imbalance class problem · ROC curve · Low-risked · High-risked

---

T. Poddar · B. Ganguli

Department of Statistics, University of Calcutta, Kolkata, India

e-mail: [tri.poddar19@gmail.com](mailto:tri.poddar19@gmail.com)

B. Ganguli

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

J. Mukherjee (✉) · A. Chakrabarti

A.K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India

e-mail: [jhilam.mukherjee20@gmail.com](mailto:jhilam.mukherjee20@gmail.com)

A. Chakrabarti

e-mail: [acakes@caluniv.ac.in](mailto:acakes@caluniv.ac.in)

M. Kar

Peerless Hospitex Hospital, Kolkata, India

e-mail: [madhuchandakar@yahoo.com](mailto:madhuchandakar@yahoo.com)

## 1 Introduction

Lung cancer is considered as a life-threatening disease toward mankind which is likely to become an epidemic within 2020 [1]. One of the important factors behind the high mortality rate of lung cancer patients is due to the detection of the disease at its advanced stages [2]. It has been observed that the initiation of PN on lung parenchyma may be an indication of lung cancer. The majority of PN represents a variety of low-risked nodule or benign condition, requiring little intervention, but for some cases, they are indications of being high-risked or malignant. Hence, designing a methodology to predict the risk of being malignant of pulmonary nodule using different morphological features of the lesions and clinical history of the corresponding patients may be an useful option for the doctors.

The aim of this study is to propose an accurate prediction model predicting the risk of being malignant of PN. The state-of-the-art prediction models consider different shape and texture feature descriptors of computer vision as a feature vector for classification. In this methodology, we have used different morphological features along with the clinical history of the patient having the particular nodule used by doctors to confirm whether that pulmonary nodule should be sent for performing the FNAC or biopsy test. Hence, this can be considered as a key contribution of this proposed work.

Rest of the paper is organized as follows: Sect. 2 represents present state-of-the-art of the proposed work. Sect. 3 illustrates the algorithms adopted to implement the prediction model. In Sect. 4, the output result and a brief discussion of the proposed methodology have been represented. The conclusions are drawn in Sect. 5.

## 2 Related Work

Setio et al. [3] proposed a novel methodology to reduce the false positive detection rate using multi-view convolutional neural network. In their methodology, they have extracted 2D patches from nine symmetrical planes of a cube. The ROI is at the center of the patch with size of  $64 \times 64$  pixels. The authors Froz et al. in the research citation [4] proposed a novel method to detect pulmonary nodules from low-dose computed tomography images. The authors have introduced a 3D texture descriptors using artificial crawlers and rose diagram to extract the required features for classification. Classification using SVM [5] helps to achieve mean system accuracy to 94.30%. The methodology described in [6] proposed a novel pulmonary nodule detection system from Multidetector Computed Tomography (MDCT) images. The research work focuses on the temporal subtraction technique, in order to remove the normal structure and visualizations of new lesions generated in the thorax. The method also employed watershed algorithm, gradient vector flow snake, to segment lung nodules from surrounding structures. With the intention of procure accurate nodule detection system,

the authors have used the following classifier Artificial Neural Network (ANN), Mahalanobis Distance, FLD, CLAFIC, and a combined classifier for this purpose.

The CAD methodology described in [7] has been focused to propose a novel method to detect sub-solid nodules from thoracic CT images. The authors have combined very simple methods of computer vision, namely, double-threshold density mask, connected component analysis, and mathematical morphology [8] to segment nodules from the surrounding structures. In order to obtain a highly accurate system, they have executed a wide number of classification algorithms, namely, k-NN [9], random forest [10], Gentle Boost [11], nearest mean classifier [12], support vector machine with RBF kernel [13], and LDA using 245 extracted features (intensity, contextual, shape, texture, moments).

The study of Kuruvilla et al. [14] has considered to design a novel methodology to classify the pulmonary nodule into malignant and benign using feedforward neural network and backpropagation neural network. Prior to perform the classification into benign and malignant candidates, the nodules are segmented using Otsu's global thresholding [15] and mathematical morphology [8].

In another novel method of classification of PN [16] into malignant and benign, the authors have considered 45 mm × 45 mm patch of LIDC [17] database images as training/testing cases instead of extracting shape and texture feature descriptors from the segmented nodule. The deep residual neural network is then used for classification purpose.

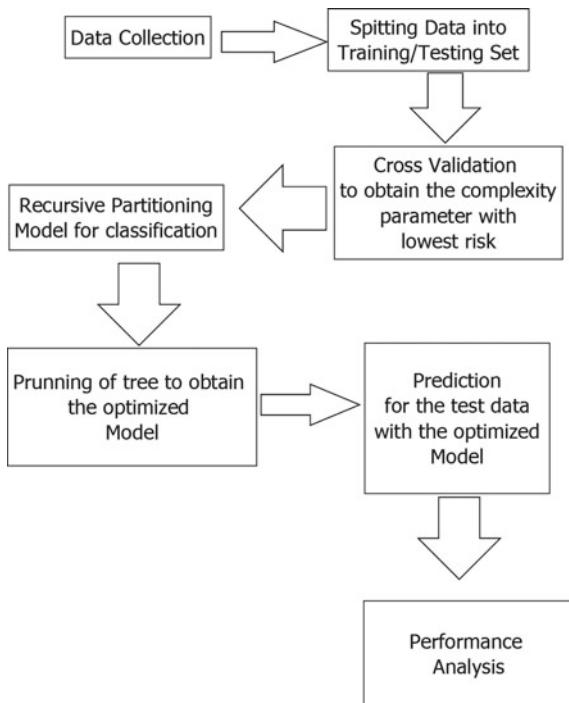
### 3 Methodology

This section illustrated a detailed description of the adopted methodology to design the proposed model. Figure 1 represents the workflow of our proposed model.

#### 3.1 Data Description

In order to design the CAD model, the method utilized *476(75 Low-risked and 401 High-risked)* PN of 256 different human subjects from an independent dataset named “Swash Image Dataset” available at (<http://www.cu-coe.in/datalist.php>). As this study is a retrospective study, hence, as per the Helsinki guidelines [18] of human research, the ethical committee has waived out the issue of informed consent form. All the images are collected retrospectively from Peerless hospital, Kolkata. All the patients either attended OPD or were admitted in the IPD of the mentioned institute with the presence of pulmonary nodule(s) in chest X-ray images from February 2015 to May 2018. Both the High-Resolution Computed Tomography (HRCT) and Contrast-Enhanced Computed Tomography (CECT) scans of each patient were acquired through Siemens 64 slice sonogram machine. The exposure setting is 140 kVp tube voltage in a single breath hold and 0.6 mm collimated width

**Fig. 1** Workflow of the prediction model



during suspended full inspiration. Image matrix size is  $512 \times 512$  in DICOM (Digital Image Communication in Medicine) format. Slice thickness of the dataset is either 1 mm or 5 mm. Images of varying slice thickness and window settings were reconstructed in multiple planes. Iodinated I.V. is used as a contrast-enhancing agent, while CECT is performed.

The morphological features of PN are manually annotated by two specialist radiologists and an oncologist. The state of cancer information is confirmed through biopsy or Fine-Needle Aspiration Cytology (FNAC) report. All these information are considered as the ground truth of the dataset in this study.

The study of Khan et al. [19] has illustrated the categories of different morphological features toward the prediction of chances of malignancy of PN. We have considered clinical and demographic features like age, gender, and previous history of chances of malignancy and morphological features like size, shape, margin of the pulmonary nodule, presence of calcification, necrotic pattern, types of nodule (based on density), number of nodule, location of the nodules, and contrast enhancement pattern. A detailed description of these features is given below.

- Age: Patient's age (in years).
- Gender: Gender of the patient (Male/Female).
- Number of Nodules Present: Number of nodules present in the entire thorax. Number of nodules generated on thorax can be multiple, usually, multiple lung

nodules occur when malignant cells spread to the lungs from other organs of the body. However, there are many benign (noncancerous) cases of multiple nodules as well.

- Previous history of malignancy: Whether the patient has been diagnosed with cancer in other organs (Yes/No).
- Size: Diameter of the PN in mm.
- Shape: The study considered round-, oval-, and irregular-shaped pulmonary nodule.
- Margin: Depending on the LUNG-Reporting and Data System, it is grouped among four classes, namely, smooth, lobulation, speculation, and irregular.
- Presence of necrosis: In some PN, cavity is present at the center. This cavity is filled with either blood or water. A cavitary pulmonary nodule containing blood is considered to have necrosis.
- Presence of calcification: It is an automatic procedure of deposition of calcium in the human cell. PN is calcified or not (Yes/No).
- Enhancement pattern of PN: The intensity value of CT images is expressed by Hounsfield Unit(HU). When a contrast-enhancing agent is injected in the body, the HU value increases. As a result, the intensity value in CECT scan is greater than the intensity value of HRCT scan of THE same patient. This phenomenon is known as contrast enhancement. Based on the difference between the HU values, the enhancement pattern is grouped into four categories, namely, Unenhanced, mildly enhanced, homogeneously enhanced, and heterogeneously enhanced. In case of unenhanced image, the difference of HU value is 0; for mildly enhanced pattern, the difference is not as much as expected. On the other hand, in homogeneous pattern, the difference in HU value is the same for each pixel and in heterogeneous pattern, the difference varies heterogeneously.
- Positions of the PN: The PN can occur at any of five lobes, namely, right upper, right middle, right lower, left lower, and left upper.
- Types of nodule: The PNs are categorized into solid, sub-solid and Ground Glass Opacity (GGO), based on density of tissues. The solid PN contains abnormal human cell, GGO contains air or water, whereas sub-solid nodule contains both of these.
- Risk of cancer: Type of Risk (High-Risked/Low-Risked).

### ***3.2 Recursive Partitioning***

Decision tree [20, 21] is a supervised learning algorithm which is used for nonparametric classification and regression. When the dependent variable is categorical, it is called classification tree and it is continuous, it is referred as regression tree. It is advantageous to choose decision tree algorithm over other classification techniques for several reasons.

- It can efficiently deal with large, complicated datasets. It is capable of manipulating highly skewed continuous variables into ranges and collapsing the large number of a categorical variable.
- Out of a large number of predictor variables, it is able to select the most important variables to construct the decision tree and calculates the relative importance of all the variables based on the model accuracy.
- Decision trees are able to deal with multicollinearity in the dataset.
- The method is robust even in the presence of missing data as it handles missing data by identifying surrogate splits in the modeling process. The model of decision tree is built in which the variables with lots of missing value are taken as dependent variables for prediction and the predicted values are used in place of the missing one.

Decision tree constructs an inverted tree with a root node, internal nodes, and leaf nodes. Recursive partitioning method creates the decision tree by binary splitting of the nodes. It is very flexible, allowing a user to find not only splits that are optimal in a global sense but also node-specific splits that satisfy various criteria.

The root node or the decision node is at the top of the tree which branches into two daughter nodes. Internal nodes, also called chance nodes, represent one of the possible choices available at that point in the tree structure which connects the parent node and the child nodes or the leaf nodes. The leaf nodes or the terminal nodes refer to the class of the target variables predicted from where further branching of nodes is not possible.

The tree is built by taking a single variable at first which best splits the data into two groups, and then this process is applied recursively to each subgroup until the subgroups either reach a minimum size or until no improvement can be made. These subgroups tend to be homogeneous. For each split, the predictor variable used is denoted as the splitting variable, and the set of values for the predictor variable, which are split between the left child node and the right child node, is called the split point. Here, the tree is split depending on the Gini's index of the classification problem.

The variable importance of the predictor variables is computed considering the impurity measure on exclusion of them. It is defined by the sum of the goodness of split measures for each split when it is the primary variable and goodness of fit for all splits in which it is a surrogate.

Let there exist  $n$  observations in the dataset and the target variable has  $C$  classes.

Let  $\pi_i, i = 1, \dots, C = \text{prior probability of each class}$ . If not specified by user, by default it is proportional to the data counts,

$L(i, j) = \text{loss function matrix}$ , where losses by default are taken to be 1.

$E = \text{Node at any point of the tree}$ .  $\tau(x) = \text{true class of an observation where } x \text{ is the vector of predictor variables}$ .

$\tau(E) = \text{predicted class}$

$n_i = \text{number of observations in the sample that are class } i$

$n_E = \text{number of observations in node } E$ .

$$P(E) = \sum_{i=1}^C \pi_i \times (Px \in E | \tau(x)) = \sum_{i=1}^C \pi_i \frac{n_{iE}}{n_i} \quad (1)$$

$$p(i|E) = \frac{\pi_i P(x \in E | \tau(x) = i)}{P(x \in E)} = \frac{\pi_i \frac{n_{iE}}{n_i}}{\sum_{i=1}^C \pi_i \frac{n_{iE}}{n_i}} \quad (2)$$

$$R(E) = \sum_{i=1}^C p(i|E) L(i, \tau(A)) = \sum_{i=1}^C \pi_i L_i \tau(E) \frac{n_{iE}}{n_i} \frac{n}{n_E} \quad (3)$$

where  $R(E)$  = is the risk of A and  $\tau(E)$  is chosen to minimize the risk,

$I(E) = \text{Impurity of node}$ ,  $E = \sum_{i=1}^C f(P_{iE})$ , where  $P_{iE}$  = proportion of obser-

vations in E that belong to class i for future samples and f is the impurity function.

In general, impurity function is defined by Ginis index as  $p(1 - p)$

In case the loss function is to be specified by the user, then it is defined as

$$L(i, j) = \begin{cases} L_i & i \neq j \\ 0 & i = j \end{cases}$$

This holds for  $C = 2$  and for  $C > 2$ ,  $L_i = \sum_j L(i, j)$ . Then the altered prior is defined as

$$\tilde{\pi}_i = \frac{\pi_i L_i}{\sum_j L(i, j)} \quad (4)$$

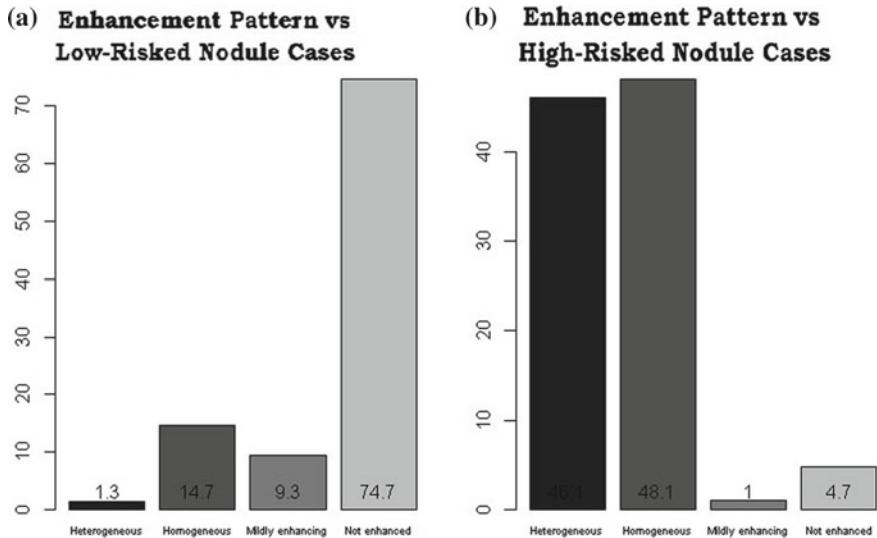
At node A, the best split is defined by the maximum decrease of impurity at node E.

$$\Delta I = p(E)I(E) - p(E_L)I(E_L) - p(E_R)I(E_R) \quad (5)$$

where  $E_L$  and  $E_R$  are the left and right sons split at node E.

A full-grown tree leads to overfitting and increases the misclassification for the future prediction of the class of the target variable. To solve this problem, stopping rule is introduced in the model to restrict splitting of the tree. Stopping rule can be defined in multiple ways like restricting the overall depth of the tree, limiting the minimum number of samples to present at each node to allow splitting or specifying the complexity parameter. The more complex the tree model is, the more are the chances of misclassification. To measure the accuracy of the model, tenfold cross-validation is performed. Thus, for different complexity parameters, different number of total splits of the tree are calculated, along with cross-validated error rate and error standard deviation, obtained over the cross-validation sets. The tree with the least cross-validated error rate is chosen to best fit the model, and the corresponding complexity parameter is chosen.

After choosing the complexity parameter, pruning is done to reduce overfitting, i.e., the tree formed originally is trimmed back.



**Fig. 2** Enhancement pattern distribution (in%) for two types of risk of cancer

## 4 Results and Discussion

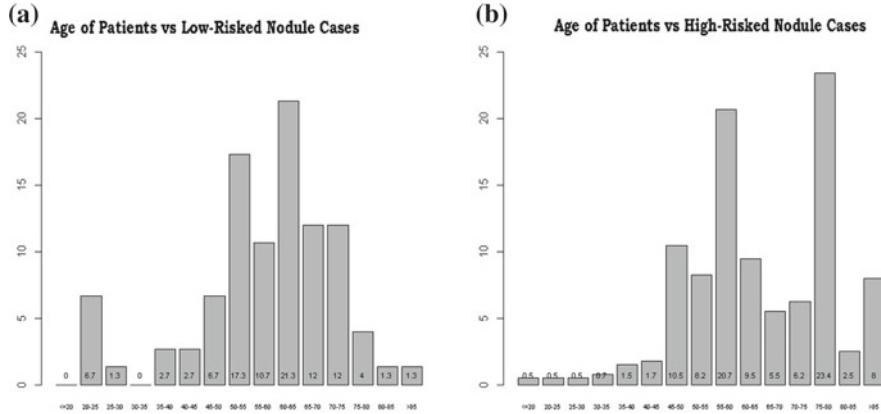
This section exhibits the results generated by the proposed model. The entire methodology has been designed on a computer having third-generation i5 processor with 12 GB DDR3 RAM, and R3.5.1 is used as a design platform.

### 4.1 Data Interpretation

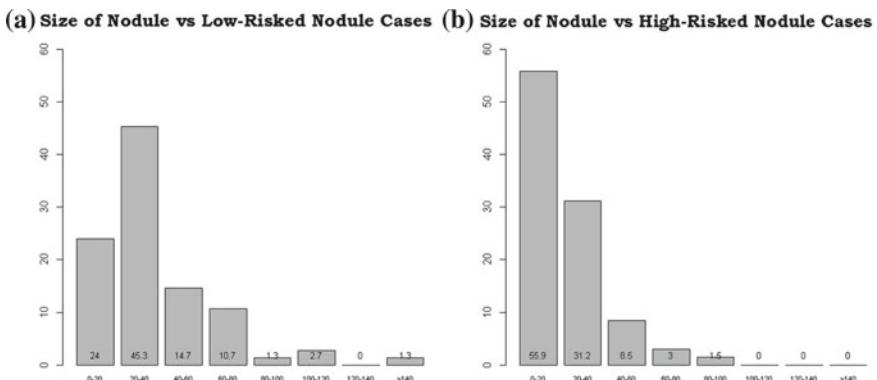
According to Fig. 2, in case of contrast enhancement pattern, for most of the cases, homogeneous and heterogeneous patterns usually indicate high-risked pulmonary nodule and unenhanced pattern along with mildly enhancement pattern tends toward benignity of the same.

Figure 3 illustrates the age distribution of patients (in %) between two types of risk of cancer. In case of high-risked PNs, the majority of the patients belong to the age groups of 75–80 and 55–60. However, looking at the age distribution alone, it is not efficient to determine the risk of cancer. Both the graphs also show that there are rare cases of this disease belonging to the lower age group.

Figure 4 illustrated the size of nodules with the two types of risk of cancer. Here, in the second graph, it is seen that over 55% of the cases, i.e., the majority of them have nodules of size less than 21 mm. While for the benign state of the nodules over 45%, the majority of them have nodule size of 20–40 mm. However, looking at the



**Fig. 3** Age distribution (in%) of two types of risk of cancer



**Fig. 4** Size distribution (in%) of two types of risk of cancer

entire graph and the unpruned tree, it is hard to conclude if size can alone significantly determine the risk of cancer of the nodule.

Table 1 represents the predictor variables, which do not contribute significantly to form the tree, and their distribution among the two classes of risk of cancer. It is evident that the different categories of these predictor variables cannot distinctly classify the nodule into low-risk or high-risk state.

## 4.2 Results of the Prediction Model

In order to implement an efficient prediction model, we have divided the entire dataset into 80/20, 70/30, and 60/40 training/ testing proportions. We have evaluated some standard metric, namely, sensitivity, specificity, and accuracy as described

**Table 1** Distribution of different features for two types of risk of cancer

Feature	Categories	Participation in low-risked class (in%)	Participation in high-risked class (in%)
Previous history of malignancy	Yes	48.13	51.87
Previous history of malignancy	No	100.00	0.00
Margin	Smooth	38.67	22.44
Margin	Lobulation	50.67	36.16
Margin	Speculation	10.67	41.40
Shape	Round	42.67	44.39
Shape	Oval	52.00	34.66
Shape	Irregular	5.33	20.95
Calcification	Yes	88.00	96.76
Calcification	No	12.00	3.24
Type of nodule	Solid	96	98.5
Type of nodule	GGO	4.0	1.5
Necrosis pattern	No Cavity	78.67	92.52
Necrosis pattern	Not necrosed	17.33	3.99
Necrosis pattern	Necrosed	43.99	3.49
Position	Right upper	22.67	18.45
Position	Right middle	25.33	16.21
Position	Right lower	14.67	19.20
Position	Left upper	18.67	23.94
Position	Left lower	18.67	22.19

in Eqs. 6, 7, and 8. Accuracy measures the overall performance of the proposed model, whereas specificity and sensitivity measure true detection of negative class and positive class, respectively.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (7)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (8)$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative outcomes of the prediction model.

**Fig. 5** Classification tree without pruning

In the model, benign state has been considered as the positive class. Table 2 concludes that the proposed methodology can classify two types of nodule very accurately. However, according to Table 3, due to the imbalancedness of the collected dataset, the Misclassification Rate (MCR) of benign class is higher than the MCR of high-risked class. This increases the overall MCR of the proposed model.

Figures 5 and 7 represent classification tree without pruning and after pruning. Hence, to obtain an accurate classification tree, we have adjusted the Complexity Parameter (cp) within a range for each training/testing proportion as described in Fig. 6. The best cp value is then chosen to obtain optimized pruned tree for the proposed model.

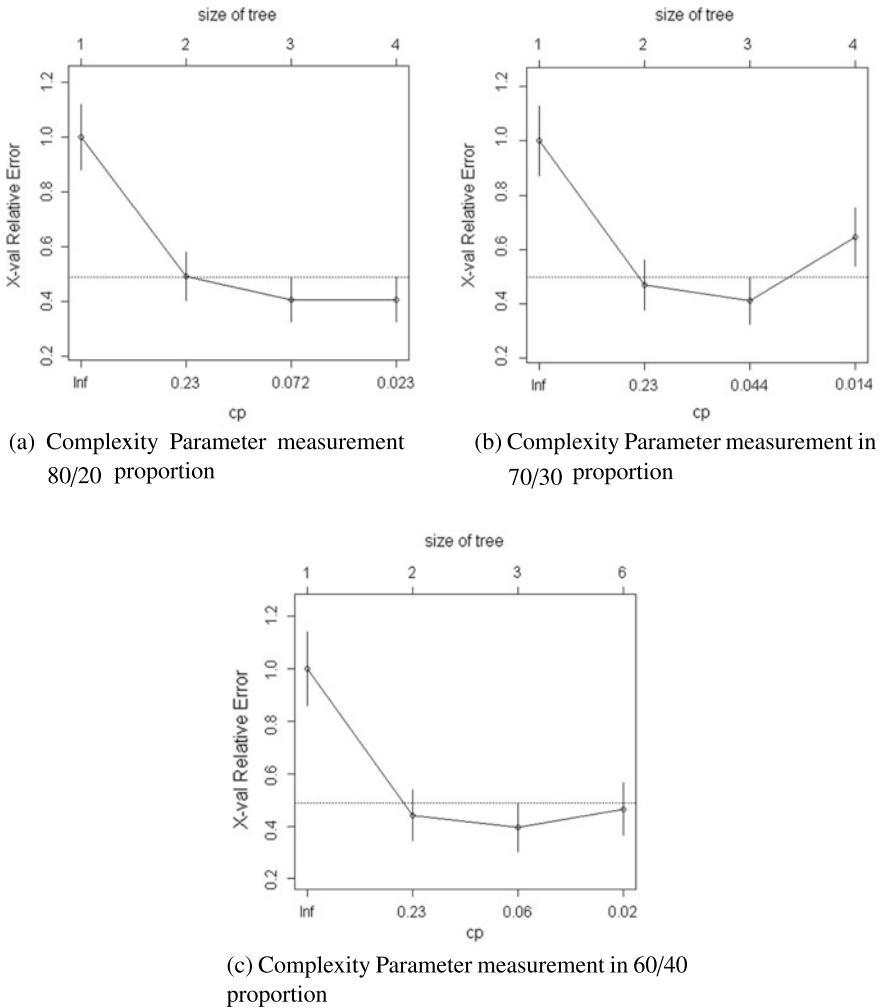
Table 4 represents the variable importance of the predictor variables, with importance factor greater than 0. ROC curves of our model have been shown in Fig. 8.

**Table 2** Performance analysis of the proposed model

Metric	80/20	70/30	60/40
Sensitivity	96.25	92.50	92.50
Specificity	97.33	86.36	86.67
Accuracy	95.79	91.55	91.58

**Table 3** Class-wise misclassification rate with 80/20 training and testing proportion

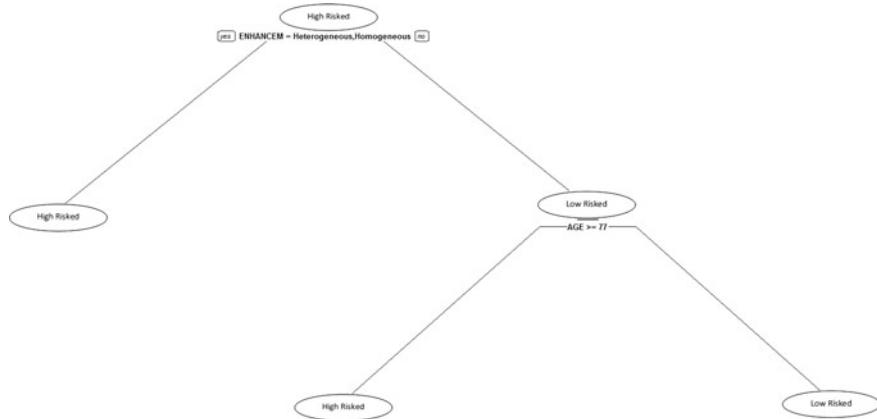
Class	MCR(%)
Low-risked	6.6
High-risked	2.5
Overall	3.15



**Fig. 6** Complexity parameter measurement with different training/testing proportions

Figure 7 tree reveals that taking different proportions of training data, the model generate same tree. In each of the cases, the size of the tree is 3 and we are getting least cross-validated error rate; hence, we have chosen the corresponding error rate to prune our tree.

The study of Khan et al. [19] provides an elaborate discussion about the characteristics of the different morphological features as well as its importance for manual prediction of the disease. The proposed methodology is also capable of finding different important variables involved in prediction. On comparing our results of variable importance with medical perspective, more emphasis has been given on contrast enhancement pattern, age, size, margin, and position of the nodule. Apart from these



**Fig. 7** Classification tree formed after pruning

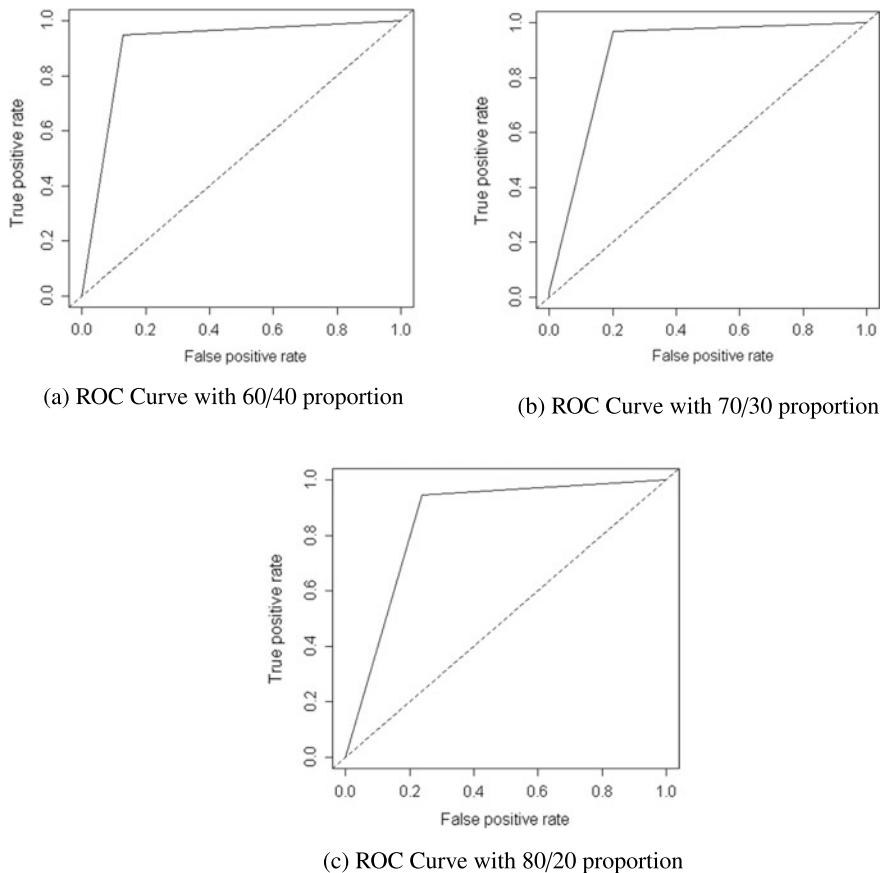
**Table 4** Variable importance

Variable	Importance
Enhancement pattern	61
Age	13
Size	12
Position	3
Margin	3
Necrosis	2
Gender	2
Previous History	1
Number of Nodule	1

features, medical science gives emphasis on types of nodule as another important features but our proposed methodology is failed to select it as an important variable as our collected dataset does not have much variation in its type of occurrence.

#### 4.3 Comparison with Existing Algorithm

In order to establish the accuracy of our prediction model, we compared our proposed methodology with some existing research work. However, these methodologies have been implemented using different shape and texture feature descriptors of computer vision; hence, we executed these methodologies using our collected image dataset, and the comparison of the results is shown in Table 5.



**Fig. 8** ROC curve in different training/testing proportions

**Table 5** Comparison with other algorithm

Method	Sp(%)	Se(%)	MCR
[14]	90.53	92.45	4.92
[22]	89.89	90.56	5.08
[16]	89.45	90.39	4.54
Proposed	96.25	93.33	3.15

## 5 Conclusion

The proposed study aimed to design an efficient prediction model that will automatically predict the chances of malignancy of PN using different clinical features and morphological features. Recursive partitioning decision tree has been used to design the model. However, due to the presence of imbalancedness in the collected database, the misclassification of low-risked class is slightly higher than the misclassification rate of high-risked class. Moreover, the entire dataset is collected retrospectively from Hospital. Hence, there are some limitations to incorporate more clinical features in the model. In future, we will try to execute the model with more number of benign nodule, so that the misclassification rate of benign class can reduce. Furthermore, we will perform the study prospectively to design more efficient model.

**Acknowledgements** We are thankful to Centre of Excellence in Systems Biology and Biomedical Engineering (TEQIP II and III), UGC UPE-II projects of University of Calcutta for providing the financial support of this research, and Peerless Hospital for providing their valuable dataset.

**Compliance with Ethical Standard.** The collection of patient images and pathological report for research purpose was approved by the Ethical Committee of Peerless Hospital and B. K. Roy Research Centre Ltd.

## References

1. Formdan, D., Bray, F., Brewster, D. H., Mbalawa, C. G., Kohler, B., Pieros, M., et al. (2013). *Cancer incidence in five continents*, vol. X (electronic version). Lyon: IARC (2013).
2. Bach, P. B., Mirkin, J. N., Oliver, T. K., Azzoli, C. G., Berry, D. A., Brawley, O. W., et al. (2012). Benefits and harms of CT screening for lung cancer: A systematic review. *Jama*, 307(22), 2418–2429.
3. Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., et al. (2016). Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5), 1160–1169.
4. Froz, B. R., de Carvalho Filho, A. O., Silva, A. C., de Paiva, A. C., Nunes, R. A., & Gattass, M. (2017). Lung nodule classification using artificial crawlers, directional texture and support vector machine. *Expert Systems with Applications*, 69, 176–188.
5. Smola, A. J., & Schlkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222.
6. Yoshino, Y., Miyajima, T., Huimin, L., Tan, J., Kim, H., Murakami, S., et al. (2017). Automatic classification of lung nodules on MDCT images with the temporal subtraction technique. *International Journal of Computer Assisted Radiology and Surgery*, 12(10), 1789–1798.
7. Jacobs, C., van Rikxoort, E. M., Twellmann, T., Th Scholten, E., de Jong, P. A., Kuhnigk, J.-M., et al. (2014). Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical Image Analysis*, 18(2), 374–384.
8. Jones, R., & Svalbe, I. D. (1994). Basis algorithms in mathematical morphology. In *Advances in electronics and electron physics* (vol. 89, pp. 325–390). Academic Press.
9. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.
10. Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8) (1998).

11. Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2), 337–407.
12. Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Elsevier.
13. Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5), 988–999.
14. Kuruvilla, J., & Gunavathi, K. (2014). Lung cancer classification using neural networks for CT images. *Computer Methods and Programs in Biomedicine*, 113(1), 202–209.
15. Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.
16. Nibali, A., He, Z., & Wollersheim, D. (2017). Pulmonary nodule classification with deep residual networks. *International Journal of Computer Assisted Radiology and Surgery*, 12(10), 1799–1808.
17. Armato, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., et al. (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2), 915–931.
18. World Medical Association. (2001). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, 79(4), 373.
19. Khan, A. N., Al-Jahdali, H. H., Irion, K. L., Arabi, M., & Koteyar, S. S. (2011). Solitary pulmonary nodule: A diagnostic algorithm in the light of current imaging technique. *Avicenna Journal of Medicine*, 1(2), 39.
20. Terry Therneau and Beth Atkinson. (2018). Rpart: Recursive partitioning and regression trees. R package version 4.1–13. <https://CRAN.R-project.org/package=rpart>.
21. Therneau, T. M., & Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines.
22. Gong, J., Gao, T., Bu, R.-R., Wang, X.-F., Nie, S.-D. (2014). An automatic pulmonary nodules detection method using 3d adaptive template matching. In *International Conference on Life System Modeling and Simulation and International Conference on Intelligent Computing for Sustainable Energy and Environment* (pp. 39–49). Springer, Berlin, Heidelberg.

# A Generalized Ensemble Machine Learning Approach for Landslide Susceptibility Modeling



**Akila Bandara, Yashodha Hettiarachchi, Kusal Hettiarachchi,  
Sidath Munasinghe, Ishara Wijesinghe and Uthayasanker Thayavasivam**

**Abstract** This paper presents a novel machine learning approach backed by ensembling machine learning algorithms to build landslide susceptibility maps. The results reveal that this approach outperforms prior machine learning-based approaches in terms of precision, recall, and F-score for landslide susceptibility modeling. In this research, three ensemble machine learning algorithms were tested for their applicability in landslide prediction domain, namely, random forest, rotation forest, and XGBoost. A comparison between these ensemble models and the machine learning algorithms used in previous researches was also performed. In order to evaluate the model's ability to generalize results, two different study areas were used in this study, which are Ratnapura district in Sri Lanka and Glenmalure in Ireland. Several landslide conditioning features including land use, landform, vegetation index, elevation, overburden, aspect, curvature, catchment area, drainage density, distance to water streams, soil, bedrock condition, lithology and rainfall prepared by surveying, remote sensing, and deriving from Digital Elevation Model (DEM) were utilized in building the spatial database. Importantly, this study introduces new landslide conditioning factors like overburden and water catchment areas which have good importance values. Further, research applies dynamic factors like rainfall and vegetation

---

A. Bandara · Y. Hettiarachchi · K. Hettiarachchi (✉) · S. Munasinghe · I. Wijesinghe · U. Thayavasivam

Department of Computer Science and Engineering, University of Moratuwa, Moratuwa, Sri Lanka

e-mail: [kusal.14@cse.mrt.ac.lk](mailto:kusal.14@cse.mrt.ac.lk)

A. Bandara

e-mail: [akilanbandara.14@cse.mrt.ac.lk](mailto:akilanbandara.14@cse.mrt.ac.lk)

Y. Hettiarachchi

e-mail: [yashodha.14@cse.mrt.ac.lk](mailto:yashodha.14@cse.mrt.ac.lk)

S. Munasinghe

e-mail: [sidath.14@cse.mrt.ac.lk](mailto:sidath.14@cse.mrt.ac.lk)

I. Wijesinghe

e-mail: [ishara.14@cse.mrt.ac.lk](mailto:ishara.14@cse.mrt.ac.lk)

U. Thayavasivam

e-mail: [rtuthaya@cse.mrt.ac.lk](mailto:rtuthaya@cse.mrt.ac.lk)

index for susceptibility map building, by making use of remote sensing data which is updated periodically. The study emphasizes the capability of ensemble approaches in generalizing results well for both study areas which inherit completely different environmental properties, and its ability to provide a scalable map building mechanism. Also, useful insights and guidelines are also provided for fellow researchers who are interested in building susceptibility maps using machine learning approaches.

**Keywords** Ensemble · GIS · Landslide susceptibility map · Machine learning · Random forest · Rotation forest · XGBoost

## 1 Introduction

Landslides are a major geological hazard, causing billions worth property damage and hundreds of casualties annually [1, 2]. Although landslides are observed in many regions around the globe, the probability of a landslide occurrence is not uniformly distributed geographically. Certain natural, geological factors, as well as human-caused factors, can cause a particular area to be more vulnerable to landslides than others [3]. Thus, taking such factors and the number of recorded landslide events into consideration, a given area can be categorized as a low-, medium-, or high-frequency event area. Evidently, the areas which have undergone unplanned constructions are more susceptible to landslides. Earthquakes or heavy rainfall can trigger a landslide in such areas. Due to the abundance of such areas, the risk of a landslide occurrence has increased over the past few decades. To mitigate the extensive losses to property and life, landslide susceptibility should be predicted accurately in advance.

However, predicting a landslide occurrence is an extremely difficult task due to the unpredictability of the underlying process which associates with multiple time-variant factors. Without proper analysis, it is hard to determine which factors have a higher impact on landslides than others, as the nature of the underlying process changes with locality. The most challenging task in building a landslide susceptibility map is discovering the prominent factors and training a model which can establish a correlation between them. The divergence of the landslide event frequency in the training dataset is another challenge that has to be overcome when designing a solution. Ideally, the proposed solution should have accurate predictions despite the nature of the training dataset, balanced or skewed. Otherwise, the scalability of the model will become an issue.

Data-driven machine learning approaches can detect the hidden relationships among these factors efficiently with the use of historical data. Several machine learning approaches were taken to model landslides, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Naïve Bayes, functional trees, etc. [3–7]. But none of these approaches has legitimately proven itself to be the optimal solution for this problem yet. A sizable amount of research has been done in the area of ANN-based solutions, where such models have performed reasonably good. These landslide models also require constant evaluation on account for the landslide-related

factors that tend to change over time. But, with limited available information on landslides and the factors which cause them, prediction power and robustness are the two main aspects to consider when selecting an algorithm to train a model for better results [8]. Although the aforementioned models predict with high accuracy on a balanced landslide dataset, their predictions deviate from the true values when provided with a skewed dataset such as a landslide dataset with a low-frequency event rate. Such scaling issues raise a concern about the robustness of the said models. This study highlights the outstanding performance of the proposed ensemble models, in both cases.

Geographic Information Systems (GIS), which can be employed to capture, store, manipulate, and represent spatially distributed data, have proven themselves to be a powerful tool in decision-making by providing insights of spatial analysis. A Landslide Susceptibility Map (LSM) visualizes the likelihood of a landslide occurrence in an area from low to high, based mostly on terrain details. Landslide Research and Risk Management division of National Building Research Organization (NBRO) [9] is the central body in Sri Lanka for monitoring the landslides and conducting field surveys on landslides. Spatial-digitalized GIS maps (Ex. landslides, slope, landform, land use, hydrology) based on the field survey conducted by NBRO were utilized in this study to create the spatial database for the study area in Sri Lanka. Data provided by Ireland Geological Survey (IGS) [10] were used to create the spatial database for the study area in Ireland. Additionally, a Digital Elevation Model (DEM) was used to derive new features like aspect, curvature, and stream network in the study area. Remote sensing Landsat images were used to calculate the vegetation index in the areas as well. In this research, ArcGIS 10.4 was used to handle the feature datasets, extract data, and construct LSMs according to the results generated by the models.

In this study, ensemble machine learning algorithms were used since they combine several classifiers to build a much powerful classifier. According to recent research, ensemble methods have outperformed individual models in several prediction tasks, proving their significant improvement over the years [6, 8].

Synthetic Minority Over-sampling Technique (SMOTE) [11] was used in this study as a solution for handling imbalanced datasets. This technique enhances a classifier's sensitivity to the minority class by over-sampling the minority class. This study contributes to the domain of landslide susceptibility modeling in two significant ways: First, by introducing highly accurate ensemble machine learning models to the domain of susceptibility map building; and second, by introducing potential landslide factors such as overburden, water catchment area which may help in predicting a landslide occurrence.

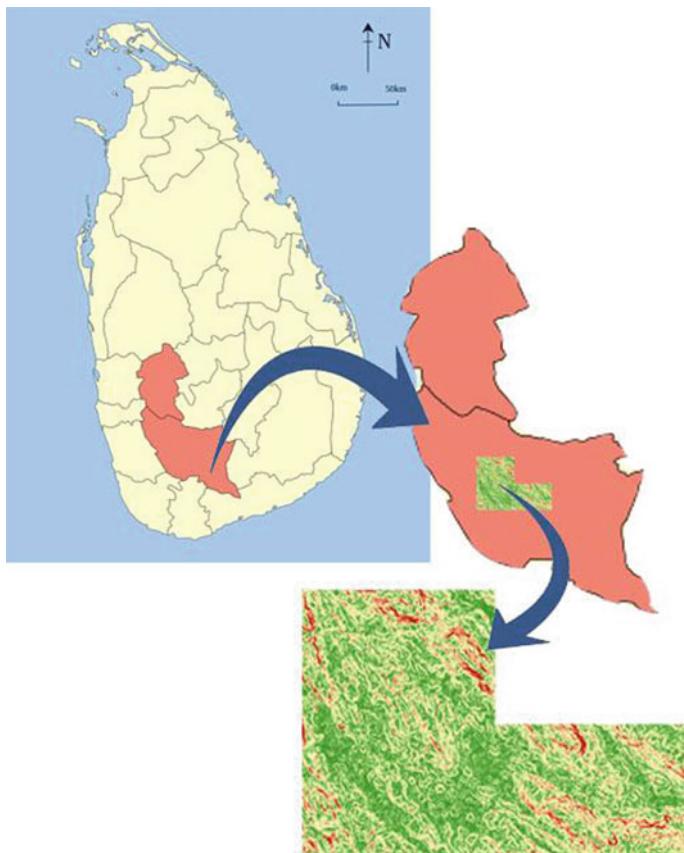
## 2 Study Area

The study area, Ratnapura, is one of the areas in Sri Lanka suffering from regular landslides. It is located amidst the mountain ranges with steep slopes, between the latitudes of  $6.53^{\circ}$ – $6.55^{\circ}$  and longitudes of  $80.33^{\circ}$ – $80.48^{\circ}$ . The daily mean tem-

perature of the area is 27 °C, and annual rainfall is roughly 3679 mm. During the monsoon season (April to July and September to November), a heavy rainfall occurs and causes floods and triggers landslides.

Figure 1 depicts the Ratnapura study area which covers 120 km<sup>2</sup> containing a total of 5333 landslide cells in the 10 m × 10 m landslide raster map. Most of the landslides are located near the water streams at high elevation levels.

Figure 2 depicts Glenmalure in Ireland which is also an area highly prone to landslides. The study area is located between the latitudes of 52.403°–53.056° and longitudes of –6.631° to –6.31°. The annual rainfall of the area is between 750 and 1000 mm, and the daily mean temperature is around 15 °C. The selected area covers 248.24 km<sup>2</sup> and contains 18,685 landslide cells.



**Fig. 1** Ratnapura study area in Sri Lanka covering 120 km<sup>2</sup> and 5333 landslide cells



**Fig. 2** Glenmalure study area in Ireland covering 248.24 km<sup>2</sup> containing 18,685 landslide cells

### 3 Literature Review

A landslide is one of the most destructive natural hazards due to its unpredictability. The casualties and the property damage can be minimized if there is a mechanism to accurately predict the occurrence of a landslide. But prediction of a landslide with both spatial and temporal accuracies is not an easy task as a landslide depends on a number of factors which we do not understand the correlation of [3]. The problem of landslide prediction can be addressed in two different ways. A landslide susceptibility map, which can map the probability of a landslide occurrence to a particular location, is one static solution. Another approach would be to develop a dynamic landslide prediction system which can generate temporal alerts based on dynamically updated data.

In the earlier days, prediction mechanisms were solely based on probability and mathematical models which can generate a landslide susceptibility map [12, 13]. Nevertheless, recent attempts to tackle this problem are grounded in various technologies. Hong et al. [14] introduce a methodology for early temporal and spatial

detection of landslides using a combination of a surface landslide susceptibility map based on geospatial data and real-time space-based rainfall analysis system. After establishing the landslide susceptibility map, they focused on developing the empirical relationship between rainfall intensity and the landslide occurrences. Chang et al. [15] also have taken a different approach in predicting landslides by combining multi-temporal satellite images and a cloud platform to aid in landslide vulnerability study. This method introduces a three-step process which includes a time series vegetation index change analysis followed by interpretation of landslide-prone areas and then updating the dedicated digital platform in the cloud.

However, the emergence of machine learning has proved that this problem can be efficiently addressed with machine learning-based solutions. Artificial Neural Networks (ANNs) are well known for its adaptive learning ability and their remarkable ability to derive meaning from complicated or imprecise data. Many researchers have attempted to develop a landslide prediction system based on ANN [3, 4, 8, 16]. Feedforward neural network architecture is ideal for modeling relationships between a set of predictors and one or more responses as Subhashini & Premaratne stated, “a Neural Network is a better solution as it handles uncertainty to a very high degree to the prediction of Landslide Disasters as a dynamic prediction model” in her research [3]. The multilayer feedforward neural networks also called Multilayer Perceptron (MLP) is the most widely studied and used artificial neural network model in practice. In their research [17], Pratik et al. attempted to predict the slope deformation of a potential landslide site located near Tangni village, using a Multilayer Perceptron (MLP)-based Backpropagation Neural Network (BPNN). Feedback or Recurrent Neural Networks (RNNs) can have signals traveling in both directions by introducing loops in the network and currently are being used for time series predictions. However, RNNs have not been used as much as other technologies in this domain according to Chen et al. [7]. On the other hand, a drawback in this approach is the growth of the relative error with respect to time, resulting in the system to be counterproductive in the long run. In addition to neural network-based solutions, Hyun-Joo et al. [8] trained a boosted tree model which ultimately outperformed an ANN model which was trained with the same dataset. In another study, Binh Thai et al. [4] compared a Naïve Bayes model and a functional tree model against an ANN model only to observe the ANN model marginally outperforming the rest. Wu et al. present a novel framework based on functional networks for landslide prediction [16]. They use an approach involving data mining paradigm based on functional networks to accurately predict the deformations of the slopes in Baishuihe, China.

Random forest is a powerful machine learning algorithm which stays unexcelled in accuracy among current machine learning algorithms, with the capability of handling a large amount of data efficiently. Generated forests can be saved for future use on other data which is very applicable in landslide prediction. Most importantly, it gives estimates of the relative importance of the variables during the classification, which can be useful to identify the most prominent factors causing landslides. Further, XGBoost (Extreme Gradient Boosting) algorithm is also considered in this research since gradient boosting trees have shown good results in recent research [18, 19]. XGBoost is based on the original boosted tree concept and the consid-

erations in system optimization and principles in machine learning. However, this model has not been applied in landslide prediction yet. Rotation forest is also an ensemble classification model introduced recently. The idea of the rotation approach is to encourage individual accuracy and diversity within the ensemble simultaneously. Diversity is promoted through the feature extraction for each base classifier. Rodriguez, Kuncheva & Alonso (2006) revealed that rotation forest ensembles individual classifiers which are more accurate than AdaBoost and random forest, and more diverse than Bagging, sometimes more accurate as well [20]. Despite their prediction power, these algorithms have not been used in the domain of landslide prediction until now.

Predicting landslide susceptibility involves identifying patterns among factors, which can increase the potential of a landslide occurrence at any given location. Nevertheless, identifying the potential landslide factors itself is a challenging task without the help of experts in the field. In the study [3], several factors were identified with the assistance of domain experts by Subhashini and Premaratne. The identified list of factors consists of 12 static factors such as geology, soil material, and slope, and dynamic factors like rainfall and land use. In another research, Chen and Hsiung [21] had identified a total of 14 factors which affected the landslides occurred in Hubei Baishuihe, Yangtze River. Further, in a study, Pratik Chaturvedi, Shikha Srivastava and Neetu Tyagi [17] successfully validated the theory that the landslides occurred near Tangni village, India, are highly dependent on recurring rainfalls by developing an accurate landslide prediction model based on daily and antecedent rainfall data. According to some researchers [22], features with zero impact on landslide occurrences need to be removed in order to achieve a higher accuracy. However, restricting the feature set to a lesser number of features is not encouraged due to the unpredictability of the relationship between a landslide and the factors. Although the impact of a certain feature may vary with the location [23], features such as geology, stream network, rainfall, and drainage density are recognized as major landslide factors by several independent researchers [3, 7].

The frequency of the event rate in the dataset is another important aspect to consider when designing and developing a prediction system. Chen et al. propose a method [21] for dynamically switching between pre-trained models in case of the existence of a minor class with fewer data samples. They discuss the use of Adaptive Synthetic Sampling (ADASYN) for improving the class balance by synthetically over-sampling new data points for the minor class via linear interpolation. Then, a BPNN-based event-class predictor is used to switch between these two models to achieve the most accurate prediction. Concerning the rainfall-induced landslides, Devi et al. proposed a method [24] to use Backpropagation Neural Network (BPNN) to predict the rainfall a day in advance, so that the predicted values can be used in a trained model to predict landslides. In case of a low-frequency event rate dataset, a suitable over-sampling or under-sampling mechanism must be implemented, or the chances of a misclassification are high. SMOTE [11], a more flexible over-sampling algorithm, will be used in this study to bring balance to the training dataset. All the aforesaid studies are restricted to a single study area, and their results have a high probability of being biased. In this study, several machine learning approaches

are tested for their applicability in two different geological locations to prove their performance and scalability.

## 4 Methodology

Figure 3 discloses an outline of the approach which was used for landslide susceptibility mapping in both study areas. The procedure which is demonstrated in the flowchart can be put under four main phases:

- (a) Data preparation and integration,
- (b) Landslide conditioning factor selection,
- (c) Landslide prediction modeling, and
- (d) Model validation.

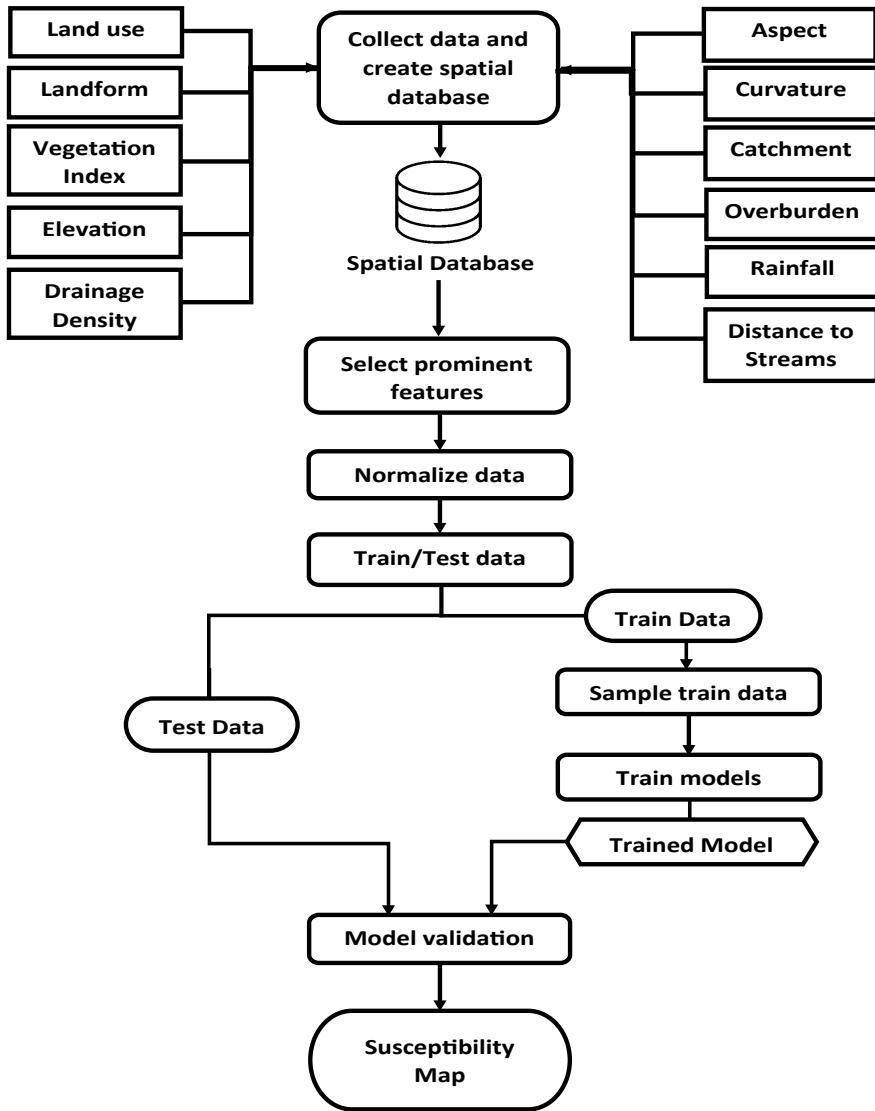
### 4.1 Data Preparation and Integration

The terrain condition in the study area was modeled using several different features (Fig. 4 and Table 1) and a landslide spatial database was constructed.

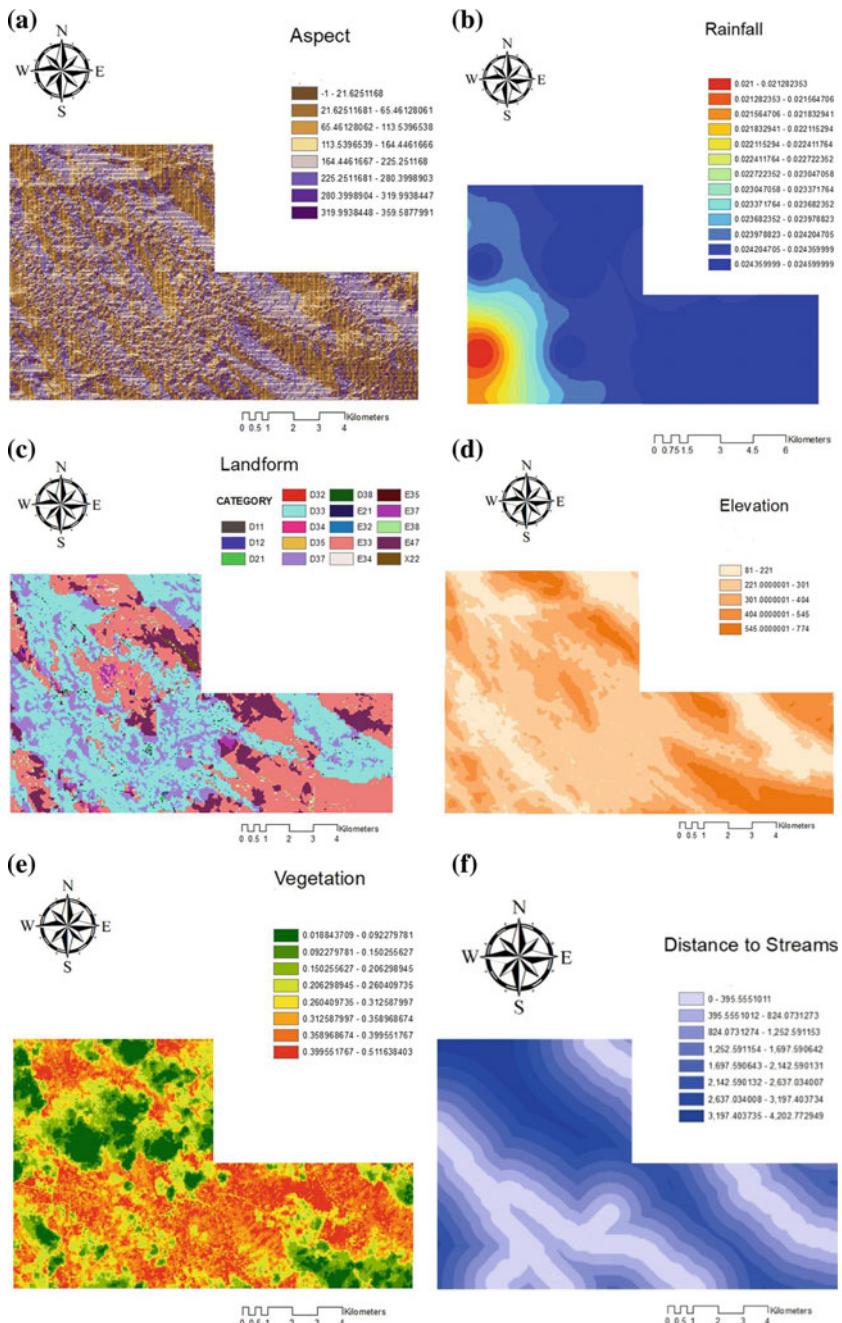
For Ratnapura study area, data were collected by surveying, by deriving maps from DEM and by remote sensing. All the surveys were conducted by NBRO and data were represented in GIS polygon maps for slope, land use, landform, overburden, hydrology, and historical landslide locations. A Digital Elevation Map (DEM) provided by U.S. Geological Survey [25] with one arc-second resolution was used to derive the features: elevation, aspect, curvature, and distance to water streams. Similarly, Ireland dataset was built using GIS maps provided by Ireland Geological Survey (IGS) [10], using a DEM and by remote sensing.

The vegetation index map was prepared by conducting the Normalized Differential Vegetation Index (NDVI) [26] calculation on Landsat satellite images provided by U.S. Geological Survey [25]. NDVI is a standardized vegetation index which indicates the relative biomass of a particular area based on the chlorophyll absorption in the red band and the relatively high reflectance of Near-Infrared band (NIR). The U.S. Geological Survey provides Landsat 8 satellite images periodically with separate bands for red, green, blue, infrared, etc. NDVI calculation was performed on those bands using the formula  $NDVI = (NIR - RED) / (NIR + RED)$  [26] to create a separate map containing vegetation index.

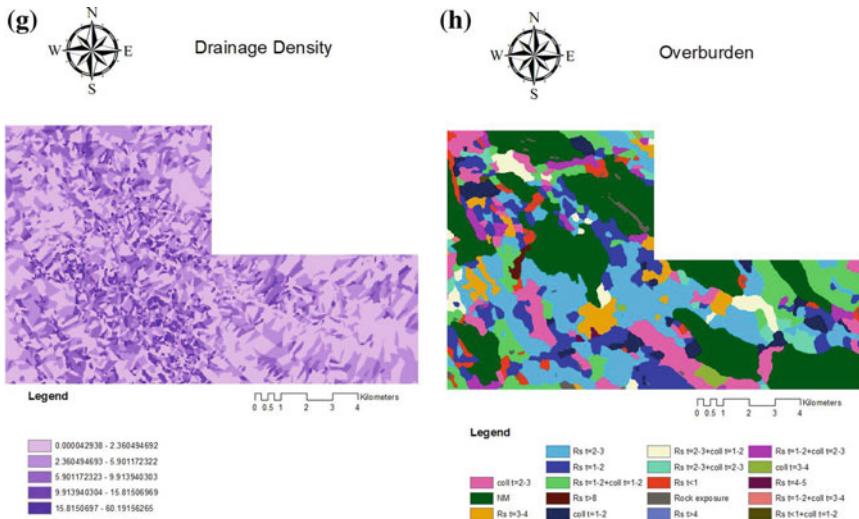
The stream network of the area was constructed using the DEM by analyzing the variation of slope in the area. Then, the distance from a cell to its nearest stream was calculated. This calculation was applied to every cell in the raster map and a separate layer was created representing the distance to streams. The annual rainfall data map for Ratnapura area was constructed by summing up the precipitation as provided by the Dark Sky API [27], in daily basis throughout the year. The study area was divided



**Fig. 3** Methodological flow conducted in this study from data collection to susceptibility map building in top-down manner



**Fig. 4** Landslide conditioning factors of Ratnapura **a** Aspect map, **b** Rainfall map, **c** Landform map, **d** Elevation map, **e** Vegetation index map, **f** Distance to streams map, **g** Drainage density map, **h** Overburden map

**Fig. 4** (continued)**Table 1** Collected feature data and their corresponding sources for each study area

Ratnapura–Sri Lanka		Glenmalure–Ireland	
Factor	Source	Factor	Source
Distance to water streams	Derived map from DEM	Distance to water streams	Derived map from DEM
Annual rainfall	Dark Sky API	Annual rainfall	Dark Sky API
Vegetation index	Derived from Landsat images	Vegetation Index	Derived from Landsat images
Elevation	U.S. Geological Survey	Elevation	U.S. Geological Survey
Overburden	NBRO Sri Lanka	Bedrock	Ireland Geological Survey
Aspect	Derived map from DEM	Aspect	Derived map from DEM
Curvature	Derived map from DEM	Curvature	Derived map from DEM
Water catchment area	NBRO Sri Lanka	Land cover	Ireland Geological Survey
Drainage density	NBRO Sri Lanka	Lithology	Ireland Geological Survey
Land use	NBRO Sri Lanka	Soil type	Ireland Geological Survey
Landform	NBRO Sri Lanka	Slope	Derived from DEM

into blocks and total precipitation was calculated. Since a rainfall distribution for the near proximity points is alike than those that are further apart, Inverse Distance Weighted (IDW) interpolation method [28] was used to create a continuous map. To predict the value of an unmetered location, the values of metered locations around it were used. Metered locations which are closer were given higher weight than the far locations to get a fair approximation of the rainfall distribution.

The spatial database was created from the prepared GIS maps for each factor. Then all maps were converted into ArcGIS raster maps with  $10 \times 10$  m cell size. Each layer represented the spatial distribution of a single factor. Value of each cell per layer was used to build a matrix of values where each column represents a factor, to model the study area. This matrix was used to train the machine learning models. Inherently, the predictions which were issued also were cell wise.

## 4.2 *Landslide Conditioning Factor Selection*

When evaluating landslide susceptibility using machine learning models, results are highly dependent on input features and quality of the input features [23]. Selected features may or may not have a significant impact on a landslide occurrence as well as where features with noise, which may also be called low-quality features, may even reduce the predictive capability of the trained model. In order to achieve a higher accuracy, features which have a high impact on a landslide occurrence needed to be identified and features with no or very low impact needed to be removed from the training set of features. In this research, relative feature importance based on information gain was used to quantify the predictive capability of conditioning factors. These methods could identify the most important factors and improve the overall classification accuracy of the model.

Feature importance provides insights into the impact of a feature on the trained model (Figs. 6, 7). Random forest and gradient boosting algorithms calculate the information gain of each feature, which is used as the basis of calculating the relative importance scores. Using these scores, a cut-off score was determined to filter the less impactful features.

## 4.3 *Landslide Susceptibility Modeling*

In this research, the following three ensemble models were tested on their ability in predicting the landslide susceptibility. Ensemble learning improves the result of a machine learning problem by combining the predictions of multiple models and calculating the final result based on a voting scheme of either bagging or boosting. This approach allows for achieving a better prediction accuracy compared to a single model. Further to get optimal results, grid search mechanism was utilized to tune the hyperparameters of each model.

(a) *Random Forest*

A random forest is an algorithmic structure consisting of multiple decision trees which are trained by the bootstrap aggregation algorithm. The final result of classification is achieved by the bagging the vote of each tree. Hence, it contains additional benefits compared to a single model such as the virtual immunity to overfitting. The random forest algorithm introduces extra randomness into the model during the training phase, causing extra diversity which ultimately leads to better predictions.

The bagging algorithm repeatedly selects a random sample with replacement from the training set and tries to fit trees to the sample. If there are  $n$  number of trees, each  $i$ th tree is trained with a random set of training data ( $X_i$ ) and targets ( $Y_i$ ) and will be denoted as the  $i$ th classifier ( $f_i$ ). When training is completed, predictions on unseen data are issued by taking the majority vote by individual trees ( $f_i$ ). This can lead to a higher accuracy by reducing the variance of the classifier. Compared to single model, this is more robust to noise since not every individual tree in the forest is correlated.

In random forest algorithm, feature bagging is also introduced to make sure that each tree uses only a subset of the features. If any particular feature or a set of features is identified to have a relatively bigger influence on the target, presumably they will be selected in most of the trees. Other than being used as a prediction tool, random forest model possess another distinctive ability to calculate the relative importance of the features by monitoring how effectively each feature reduces the sample impurity across the training set, which is known as information gain. This is very much convenient to extract the most important features from a large feature set.

(b) *Extreme Gradient Boosting (XGBoost)*

XGBoost model started as a research project conducted by Tianqi Chen and Carlos Guestrin, University of Washington [19]. XGBoost became a well-known machine learning model since its success in several machine learning competitions. XGBoost is a gradient boosting framework which provides parallel tree boosting. Specifically, XGBoost uses a more regularized model formalization than gradient boosting to control overfitting, which allows it to perform better.

Gradient boosting algorithm can be viewed as a pipeline of three main steps.

1. Identify an adequate loss function for the given problem.
2. Create a weak learner for predictions. A decision tree is selected as a weak learner in gradient boosting.
3. Create an additive model to add the predictions of the weak learners and reduce the loss function.

By following the principle of ensemble, XGBoost ultimately constructs a very strong classifier based on a set of weak learners which are not generalized optimally to predict the actual class. Unlike other ensemble algorithms, XGBoost constructs weak learners in a special manner such that each model fixes the errors by previous models. This is achieved by adding models in multiple iterations.

Throughout several iterations, models are added sequentially such that each descendent model predicts with a less rate of error. This process is executed until the

model is generalized on the training dataset. Identifying the adequate loss function to evaluate every single model has a big impact on the overall accuracy of the final model. Though the training process is sequential, XGBoost takes advantage of the multithreaded execution for faster performance.

XGBoost algorithm was implemented in such a way that it utilizes resources and computational time efficiently. The ultimate goal was to optimally use available resources during the model training. Hence, the XGBoost algorithm was well engineered with some key features, to achieve some of these design goals. For example, when the dataset is sparse, XGBoost internally handles missing data values. Also, the block structure of the algorithm supports the parallelizing of the construction of a single tree. Additionally, continuous training mechanism makes sure of the model's ability to boost further for new data as well [18].

### (c) *Rotation Forest*

Rotation forest is a machine learning algorithm for generating an ensembled classifier based on a feature extraction methodology known as Principal Component Analysis (PCI). In this technique, the training data set is divided randomly into a preconfigured,  $k$  number of subsets and then PCI is applied to each subset to form a rotation sparse matrix. The idea of the rotation forest can be explained as below.

Let us define  $X$  as the training set,  $Y$  as the class labels and  $F$  as the set of features. Suppose there are  $N$  number of training examples with each having  $n$  features which form  $N * n$  matrix. Let  $Y$  be the relevant class labels for the data which take values from the set of class labels  $\{w_1, w_2 \dots\}$ . Denoted by  $D_1$  to  $D_n$  are the classifiers in the ensembling method which are also decision trees. Steps mentioned below can be used to construct a training set for the  $D_i$ .

1. Split  $F$  randomly into  $k$  subsets. Subsets may or may not be disjoint. For simplicity, we choose  $k$  as a factor of  $n$  which leads to having  $M = n/k$  features in each subset. Let  $F_{i,j}$  be the  $j$ th subset of features for a given classifier  $D_i$ .
2. Randomly select a nonempty subset of classes for each subset of  $F$  and draw a bootstrap sample of objects which contains a 75% of the data count. Let  $X_{i,j}$  be the data set for the  $F_{i,j}$  and  $X'_{i,j}$  be the new set.
3. Store the coefficients of the principal components  $a_{i,j}^{(1)}, \dots, a_{i,j}^{(M_j)}$ , each of size  $M * 1$  by running PCA on  $X'_{i,j}$ . Denote the coefficient matrix by  $C_{i,j}$ .
4. For  $j = 1 \dots k$  arrange  $C_{i,j}$  in a rotation matrix denoted by  $R_i$

$$R_i = \begin{bmatrix} a_{i,1}^{(1)}, a_{i,1}^{(2)}, \dots, a_{i,1}^{(M_1)} & [0] & \dots & [0] \\ [0] & a_{i,2}^{(1)}, a_{i,2}^{(2)}, \dots, a_{i,2}^{(M_2)} & \dots & [0] \\ \vdots & \vdots & \ddots & \vdots \\ [0] & [0] & \dots & a_{i,K}^{(1)}, a_{i,K}^{(2)}, \dots, a_{i,K}^{(M_K)} \end{bmatrix} \quad (1)$$

5. Construct  $R_i$  by rearranging the columns of  $R_i$  to have the same order of features in the feature set.

The success of the rotation forest depends on the application of this rotation matrix constructed by linearly transformed subsets. Thus, the final classification accuracy will be greatly improved by an effective and efficient transformation [20].

#### 4.4 Model Validation

In machine learning-based modeling, one of the most critical phases is the validation phase of the prediction model. Validation help in two ways, it quantifies the ability of the model to work well with unseen examples and it also quantifies how accurately the model can perform for both seen and unseen examples. In this research, all the models were tested thoroughly to verify that the model is properly fitted to the training dataset without overfitting or underfitting.

Models can be assessed by referring to the known historical landslides data and comparing with the model predictions. A common approach is to split the dataset into two subsets labeled as a training and testing dataset with 80–20 split. 80% portion will be used as training set, while the other unseen portion is used for testing. Since a landslide dataset is imbalanced, SMOTE sampling method was applied before the model training process in this study. This technique increases the number of samples for the minor class by interpolation.

However, the problem of overfitting in machine learning can still reside in the trained model which leads to being less precise on unseen data. The tenfold cross-validation was conducted to make sure that the model is not overfitted. Data is divided into ten subsets such that each time, one of the subsets is used as a test set while other nine subsets are put together to form the training set.

In this study, landslide mapping was treated as a binary classification which produces two outputs as either as a landslide occurrence or a non-landslide occurrence. Four possible prediction types are shown in the confusion matrix in Fig. 5.

TP (True Positive) and TN (True Negative) are the numbers of landslide cells that are correctly classified, and FP (False Positive) and FN (False Negative) are the numbers of landslide cells incorrectly classified.

**Fig. 5** Confusion matrix which describes the possible prediction cases of a binary classifier

		Predicted	
		Positive	Negative
Observed	Positive	Number of True Positives	Number of False Negatives
	Negative	Number of False Positives	Number of True Negatives

Accuracy defined in Eq. 2 is the proportion of landslide and non-landslide pixels that the model has correctly classified. This is a good measure for assessing most of the binary classification models for equally distributed datasets. However, for landslide model assessment, it can be misleading since the dataset is skewed. Due to the skewness of the ratio between landslide positive cells and negative cells, the model scores a high accuracy by only predicting negative cells, which composes the majority of the dataset.

$$\text{Accuracy} = \frac{\sum \text{TP} + \sum \text{TN}}{\sum \text{TP} + \sum \text{TN} + \sum \text{FP} + \sum \text{FN}} \quad (2)$$

This kind of binary classifiers performance can be measured by precision and recall which are defined in Eqs. 3 and 4 where precision measures the results relevancy,

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FP}} \quad (3)$$

while recall measure of how many truly relevant results are returned.

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FN}} \quad (4)$$

In uneven class distribution scenario, *F*-score is a useful measure of test accuracy which considers both precision and recall which varies between 0 and 1.

$$F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

## 5 Results

### 5.1 *Landslide Conditioning Factor Selection Results*

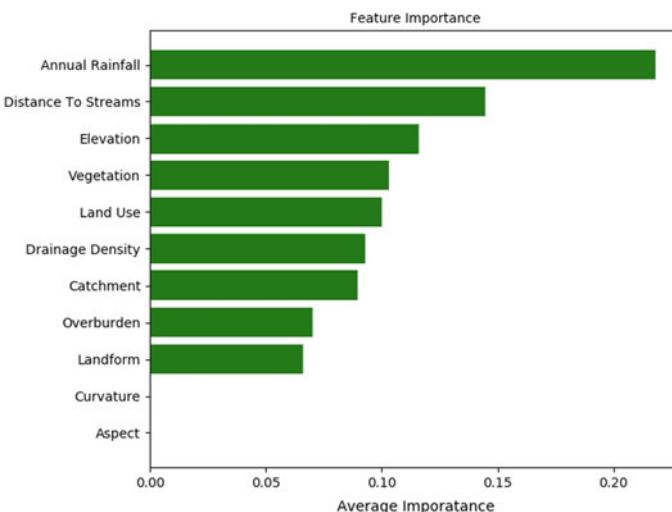
The prediction capability of collected conditioning factors was evaluated using a tree-based feature importance method. As per the results shown in Figs. 6 and 7, rainfall has the highest prediction capability among all conditioning factors in both study areas. This result validates a previous research in the literature about rainfall being a primary factor which causes landslides [15] compared to other conditioning factors which have less impact compared to rainfall. In comparison to Ireland, aspect and curvature score in Ratnapura have very small average importance values.

If any conditioning factor scores a negligible value for Average Importance (AI), that factor needs to be removed [22] from the feature set. Since the curvature and

aspect have shown a very low importance in Ratnapura area, they were not utilized in building the susceptibility map for Ratnapura. Also, the rest of the factors were selected to train the models since they all got significant AI score. This observation proves the hypothesis from previous research that the impact on these conditioning factors can vary based on the geological location [23]. Hence, these average feature importance values are most likely to change for a new landslide zone with different geological properties. Thus, it is highly recommended that the susceptibility map building researchers use the proposed feature importance calculating technique to quantify the importance of factors for each zone afresh.

## 5.2 Model Performance Evaluation and Comparison

In landslide modeling, it is essential to evaluate and assess the quality and productivity of the trained models.  $F$ -score, precision, and recall measurements scored by the trained models for both training dataset and test dataset are included in Tables 2 and 3 for Ratnapura and Ireland, respectively. All three models exhibit reasonably good predictive capability. For the test set, the highest  $F$ -score and precision values were scored by the random forest classifier. XGBoost produced higher recall value compared to the random forest and rotation forest classifiers. These observations can be seen in both study areas. Cross-validation results justify the fact that random forest and XGBoost models are not overfitted. However, rotation forest model can be considered as overfitted since it scores low performance on the test set relative to its train set.



**Fig. 6** Feature importance of landslide conditioning factors in Ratnapura study area having zero importance for aspect and curvature

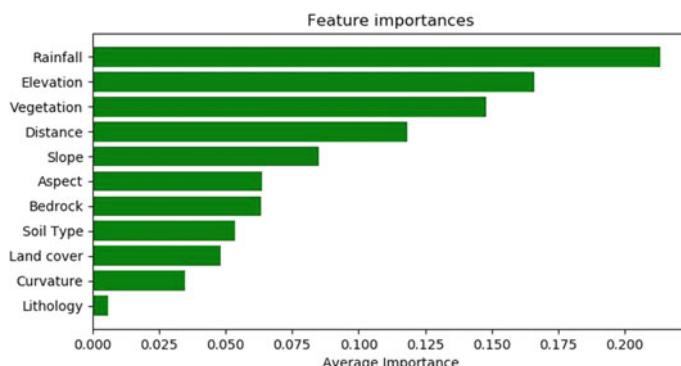
### 5.3 Comparison with Other Machine Learning Models

Machine learning algorithms which were effective according to the previous researches were also applied to the Ratnapura dataset in this study in order to compare their performance. Artificial Neural Network (ANN), SVM models, and Naïve Bayes (NB) were the considered models.

As per results in Table 4 and comparison chart in Fig. 8, ANN shows a significantly better performance over SVM and NB. Even though NB scored a higher recall value, due to the huge over-prediction, it ended up with a very low score for precision. Altogether, ensemble algorithms outperform ANN, SVM, and NB models as per the above results.

### 5.4 Landslide Susceptibility Map

A landslide susceptibility map visualizes the likelihood of a landslide occurrence, with each cell of the raster map classified into different risk levels. Originally, the problem was treated as a binary classification problem, where each cell was catego-



**Fig. 7** Feature importance of landslide conditioning factors in Ireland study area having the highest importance for rainfall and least for lithology

**Table 2** Train and test data results for Ratnapura study area for precision, recall, and *F*-score measures

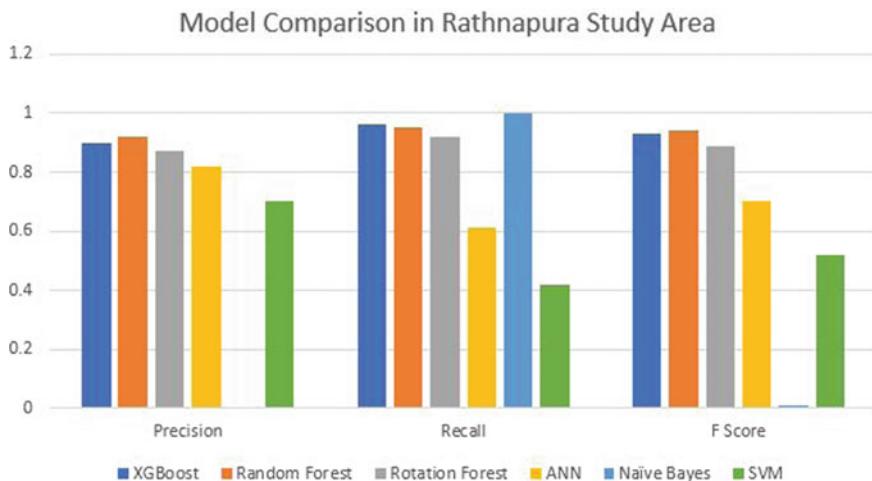
Parameters	XGBoost		Random forest classifier		Rotation forest	
	Train	Test	Train	Test	Train	Test
<i>F</i> -score	0.98	0.93	1.00	0.94	0.99	0.89
Precision	0.97	0.90	1.00	0.92	0.99	0.87
Recall	1.00	0.96	1.00	0.95	0.99	0.92

**Table 3** Train and test data results for Ireland study area for precision, recall, and *F*-score measures

Parameter	XGBoost		Random forest classifier		Rotation forest	
	Train	Test	Train	Test	Train	Test
<i>F</i> -score	0.96	0.86	0.99	0.89	0.96	0.77
Precision	0.93	0.83	0.99	0.92	0.93	0.72
Recall	0.99	0.89	1.00	0.88	0.99	0.82

**Table 4** Results of models which used in prior researches prominently (applied in Ratnapura study area)

Model	Parameter		
	Precision	Recall	<i>F</i> -score
ANN	0.82	0.61	0.70
SVM	0.70	0.42	0.525
NB	0.004	1.00	0.008

**Fig. 8** Model performance comparison graph between the tested ensembling models and prominent models in prior researches measured in terms of precision, recall, and *F*-score

rized into either the “landslide” or “non-landslide” class. Later, in the susceptibility map building phase, risk levels were classified into six categories as extremely low, low, moderate, high, very high, and extremely high, and each cell was assigned one of these six values.

Landslide susceptibility map was constructed using the predictions by the random forest classifier which was the best performing model. The study area was split into 10 subareas such that, in each iteration, one of the subareas was used for prediction, while other nine subareas were utilized to train the model. The prediction matrix by random forest classifier contains ones and zeroes indicating whether a given cell

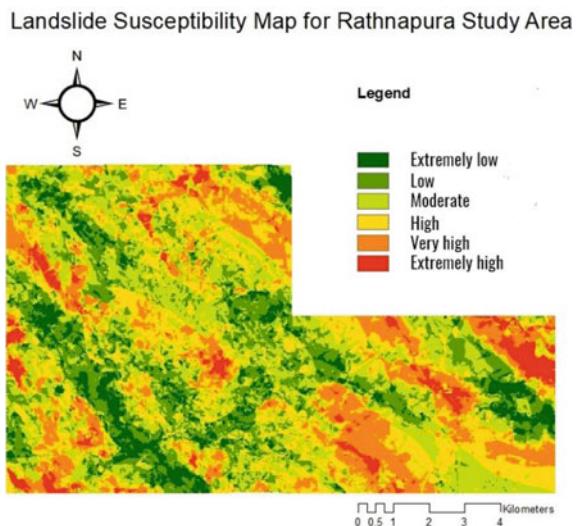
is prone to a landslide occurrence or not. This resulting matrix was converted into a map using ArcGIS as shown in the predicted landslide map in Fig. 9. The actual landslides' map in the bottom section of Fig. 9 contains the originally recorded landslides in the study area of Ratnapura. This comparison between the two maps itself is a proof of the performance of the model.

This prediction map (Fig. 9) was utilized to generate the landslide susceptibility map using ArcGIS. The susceptibility map identifies each cell into one of the six aforementioned different risk levels, and the resultant landslide susceptibility map is shown in Fig. 10. An identical process was followed to construct the landslide susceptibility map for the study area in Ireland which is shown in Fig. 11.

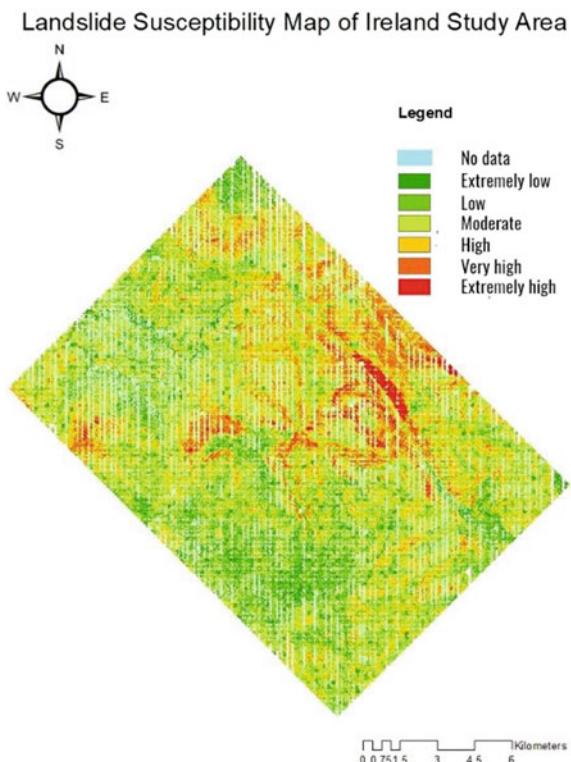


**Fig. 9** Comparison between actual landslides map and the binary predicted landslides map from random forest classifier for Ratnapura study area

**Fig. 10** Landslide susceptibility map for Rathnapura study area based on the prediction of random forest classifier visualizing the risk levels from low to high



**Fig. 11** Landslide susceptibility map for Ireland study area based on the results of random forest classifier visualizing the risk levels from low to high



## 6 Conclusion

Landslide susceptibility prediction is one of the growing areas of study in machine learning in the recent past. This study was carried out to test the applicability of three ensemble machine learning models in two different geological locations. The proposed ensemble models were able to outperform machine learning models which were predominantly known to be the best performers in landslide prediction domain. In addition to that, they were able to perform well in both study areas as well. Thus, it is safe to conclude that ensemble models have the ability to generalize the problem well and predict with high accuracy in landslide susceptibility modeling.

Further, this study validates the claim that the impact of landslide factors varies between zones by rendering some factors virtually useless in a low-frequency zone. Thus, selecting landslide factors which actually contribute to a landslide is of vital importance. In this study, relative feature importance was calculated based on the information gain of each feature.

Handling class imbalance in landslide modeling is a key step to develop an accurate model. In this study, several methods were tested, and SMOTE method provided the best performance.

This research introduced new landslide conditioning factors such as overburden, and water catchment areas which had good feature importance values. Also, this research brought forward dynamic landslide factors which can be used to extend this system to implement a real-time landslide early warning system. On the whole, two landslide susceptibility maps were produced utilizing the predictions from the most accurate predictive model developed during the research.

**Acknowledgements** Authors are thankful to Director, National Building Research Organization—Sri Lanka (NBRO) for their support for providing required spatial maps for the study area.

## References

1. Karunaratne, M. (2017). *Sri Lanka floods and landslides* (p. 2017). Colombo: IOM Sri Lanka.
2. U. W. L. Chandradasa, Mallawatantri, A., & Wijethunga, R. (2009). Sri Lanka national report on disaster risk, poverty and human development relationship. Ministry of Disaster Management and Human Rights, Sri Lanka.
3. Subhashini, L. D. C. S., & Premaratne, H. L. (2013). Landslide prediction using artificial neural networks. In *ICSBE-2012: International Conference on Sustainable Built Environment*, Kandy, Sri Lanka.
4. Pham, B. T., Bui, D. T., Pourghasemi, H. R., Indra, P., & Dholakia, M. B. (2015). Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes.
5. Pradhan, B., & Lee, S. (2009). Landslide risk analysis using artificial neural network model focusing on different training sites. *International Journal Physics Sciences*, 3(11), 1–15.
6. Hong, H., Liu, J., Bui, D. T., Pradhan, B., Acharya, T. D., Pham, B. T., et al. (2017). Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China).

7. Chen, H., Tang, H. &, Zeng, Z. (2013). Landslide deformation prediction based on recurrent neural network. In *International Conference on Neural Information Processing*, China.
8. Oh, H.-J. & Lee, S. (2017). Shallow Landslide susceptibility modeling using the data mining models artificial neural network and boosted tree.
9. NBRO. (2018). Sri Lankan government research and development institute web resource <http://nbro.gov.lk/index.php?lang=en>.
10. Dept. of Geology | SIU. (2018). Geology.Siu.Edu. [https://geology.siu.edu/?gclid=EA1alQobChMI7sG84wIV1YRwCh2zqQXCEAYASAAEgIIGPD\\_BwE](https://geology.siu.edu/?gclid=EA1alQobChMI7sG84wIV1YRwCh2zqQXCEAYASAAEgIIGPD_BwE).
11. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *JAIR*, 16, 321–357.
12. Chung, C.-J. F. & Fabbri, A. G. (1999). Probabilistic prediction models for landslide hazard mapping.
13. Ward, T. J., Li, R.-M., & Simons, D. B. Mathematical modeling approach for delineating landslide hazards in watersheds.
14. Hong, Y., Adler, R. F., Huffman, G. (2007). An experimental global prediction system for rainfall-triggered landslides using satellite remote sensing and geospatial datasets. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6), 1671–1680.
15. Chang, K., Liu, J., Kuo, C., Wang, H., & Chang, Y. (2017). Combining multi-temporal satellite images and a cloud platform to develop new evaluating procedures for landslide vulnerability study. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Fort Worth, TX, pp. 1912–1915.
16. Wu, A., Zeng, Z., & Fu, C. (2014). Data mining paradigm based on functional networks with applications in landslide prediction. In *2014 International Joint Conference on Neural Networks (IJCNN)*, Beijing, pp. 2826–2830.
17. Chaturvedi, P., Srivastava, S., & Tyagi, N. (2015) Prediction of landslide deformation using back-propagation neural network. In *IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions*, India.
18. Sa, R., Uzirb, N., Rb, S., & Banerjeeb, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets 9(40).
19. Chen T., & Guestrin C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, California, USA, pp. 785–794.
20. Rodriguez, J. J. & Kuncheva, L. I. (2006). Rotation forest: A new classifier ensemble method.
21. Chen, S. F. & Hsiung, P. A (2017). Landslide prediction with model switching. In *The 2018 IEEE Conference on Dependable and Secure Computing*, Taiwan.
22. Bui, D. T., Tuan, T. A., Klempe, H., Pradhan, B., & Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* 9, 11, 27 January 2015.
23. Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., et al. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *CATENA*, 151, 147–160.
24. Devi, S. R., Venkatesh, C., Agarwal, P. & Arulmozhivarman, P. (2014). Daily rainfall forecasting using artificial neural networks for early warning of landslides. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2218–2224). IEEE, September, 2014.
25. United States Geological Survey. (2018). A scientific agency of the United States government, web resource <https://earthexplorer.usgs.gov/>.
26. Matsushita, B., Yang, W., Chen, J., Onda, Y. & Qiu, G. (2007). Sensitivity of the enhanced vegetation index (EVI) and normalized difference vegetation index (NDVI) to topographic effects: A case study in high-density cypress forest.
27. Dark Sky. (2018). web resource <https://darksky.net/dev>.
28. Bhunia, G. S., Shit, P. K., & Maiti R. (2016). Comparison of GIS-based interpolation methods for spatial distribution of soil organic carbon (SOC). *Journal of the Saudi Society of Agricultural Sciences* (in press). <http://dx.doi.org/10.1016/j.jssas.2016.02.001>

# Comparative Evaluation of AVIRIS-NG and Hyperion Hyperspectral Image for Talc Mineral Identification



Himanshu Govil, Mahesh Kumar Tripathi, Prabhat Diwan and Monika

**Abstract** The advancement and progressive development in hyperspectral remote sensing technology enhance the capability to measure the minor variations in spectral feature on the small and big scale of measurement. The remote sensing images provide information of earth's features according to the capability of regional coverage and larger synoptic view, but hyperspectral image provides detailed, subtle variation in the spectral resolution. In this study, Hyperion and AVIRIS-NG data were used for identification of talc mineral in and around the Jahajpur city, Chabbadiya village, India. The Hyperion hyperspectral data show absorption features of talc minerals but AVIRIS-NG minerals show the multiple absorption features with iron-bearing talc at the same location. In comparison, observed result shows that AVIRIS-NG has much accurate capability for identification of different minerals in the visible region along with the short-wave infrared portion of the electromagnetic spectrum.

**Keywords** AVIRIS-NG · Hyperion · Talc · Jahajpur · Spectroscopy

## 1 Introduction

On November 21, 2000, NASA launched push broom and sun-synchronous Hyperion hyperspectral sensor under the Earth observation (EO-1) mission. Hyperion hyperspectral remote sensing sensor became first hyperspectral spaceborne sensor [13]. Since its launch, Hyperion is only continuous source for hyperspectral remote sensing data in full coverage from 400 to 2500 nm of electromagnetic spectrum with 242 bands [15]. According to Mitchell et al. [13], the Hyperion has a problematic issue of cross track calibration and low signal-to-noise ratio.

In the success series of airborne and spaceborne hyperspectral remote sensing sensors such as AVIRIS and Hyperion, JPL NASA started a new mission of AVIRIS -NG [6] in the year 2014 in U.S.A. On achievement of positive result of AVIRIS-NG

---

H. Govil · M. K. Tripathi (✉) · P. Diwan · Monika

Department of Applied Geology, National Institute of Technology Raipur (C.G.), Raipur 492010, India

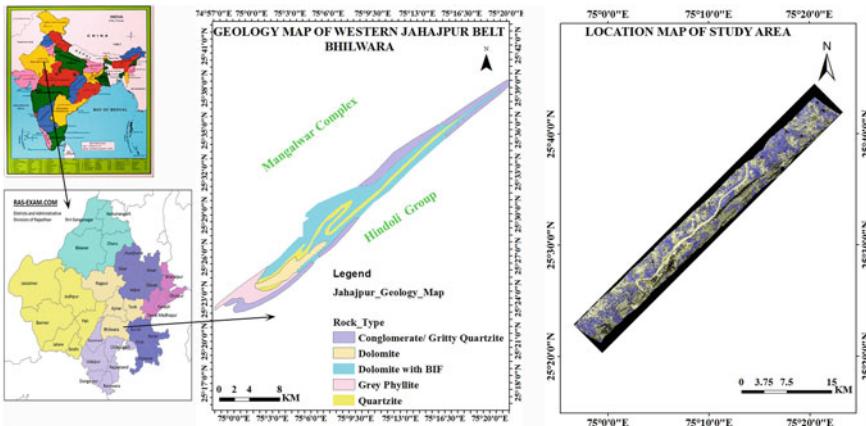
e-mail: [tripathi.mahesh1@gmail.com](mailto:tripathi.mahesh1@gmail.com)

hyperspectral sensor in the USA, NASA-ISRO started a joint mission with some flight mission in India in the year 2016. On February 04, 2016, Jahajpur area of Bhilwara, Rajasthan, India was acquired by the airborne AVIRIS-NG hyperspectral sensor. ISRO-NASA has made a series of flights in that area for mineral identification [9, 16]. The image is captured by AVIRIS-NG sensor in 425 contiguous channels in spectral range of 0.38–2.51  $\mu\text{m}$  including spectral resolution 5 nm and capability of high signal-to-noise ratio [2, 20].

According to various researchers [3, 8], the iron oxides/hydroxides and hydroxyl minerals which contain  $\text{Fe}^{3+}$ ,  $\text{Mg-OH}$ ,  $\text{Al-OH}$ ,  $\text{SO}_4$ , and  $\text{CO}_3$  can be identified and mapped with the hyperspectral remote sensing data. On the basis of spectral characteristics, the EMR region is divided into two intervals such as VNIR and SWIR. VNIR interval ranges from 0.4 to 1.1  $\mu\text{m}$  and SWIR interval ranges from 2.0 to 2.5  $\mu\text{m}$ .  $\text{Fe}^{2+}$  and  $\text{Fe}^{3+}$  show electronic charge transfer absorption and crystal field absorptions in VNIR region and  $\text{Al-OH}$ ,  $\text{Mg-OH}$ ,  $\text{CO}_3$ , and  $\text{SO}_4$  show vibrational OH and HOH overtones and stretches in SWIR region [3]. Kiran [10] described that the iron oxides/hydroxides such as goethite, limonite, and hematite show characteristic absorptions between 0.845 and 0.95  $\mu\text{m}$  but hematite mineral is differentiated to other iron oxides on the basis of different absorptions features such as at 0.465, 0.650, and 0.850–0.950  $\mu\text{m}$ . According to Meer ([12], the carbonate shows strong absorption at 2.3 and 2.35  $\mu\text{m}$  (talc, calcite) and weaker absorption 2.12–2.16, 1.85–1.97, and 1.97–2.0  $\mu\text{m}$  [12]. Meer [12, 21] suggested that talc shows the absorption at 2.3  $\mu\text{m}$  due to  $\text{Mg-OH}$ . Meer et al. [21] have given description that in the SWIR region “vibration gives rise to narrow and more pronounced features roughly at 1.400  $\mu\text{m}$  combined with 1.900  $\mu\text{m}$  due to molecular water, 1.400  $\mu\text{m}$  due to OH, 2.200  $\mu\text{m}$  due to  $\text{Al-OH}$ , 2.300  $\mu\text{m}$  due to  $\text{Mg-OH}$ , and 2.320–2.350  $\mu\text{m}$  due to  $\text{CaCO}_3$ .” Between the range of absorption, features of dolomite to calcite, epidote, talc, and pyrophyllite show absorption at 2.32, 2.33, 2.30, and 2.32  $\mu\text{m}$ , respectively.

## 2 Geology of the Study Area

The geology of the Jahajpur area is classified in three zones named as Mangalwar complex situated in west from Archean basement of Jahajpur group, Hindoli group intruded the Archean basement rocks of Jahajpur groups in east, and two parallel dolomitic limestones and quartzite ride striking in NE–SW direction of Archean basement rocks along the river Banas. The main rocks of the Jahajpur group are quartzite, phyllite, banded iron formation (BIF), and quaternary sediments [7, 17–19, 23, 22] (Fig. 1).



**Fig. 1** Location and geology map of the study area [4, 9]

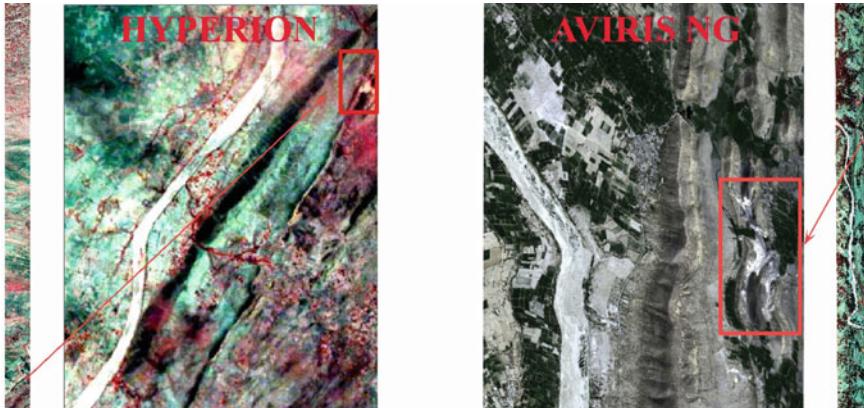
### 3 Materials and Methodology

The used scene of Hyperion hyperspectral remote sensing image have contiguous 242 narrow bands in the range of 0.4–2.5  $\mu\text{m}$  with capability of spatial and spectral resolutions such as 30 m and 10 nm, respectively, with narrow swath coverage of  $7.5 \times 100$  km including high radiometric accuracy of 12 bit quantization [1, 5, 11, 1214].

The AVIRIS-NG hyperspectral remote sensing data have 425 narrow contiguous bands in spectral region of VNIR and SWIR ranging from 0.38 to 2.51  $\mu\text{m}$  of electromagnetic spectrum. The spatial and spectral resolutions are 8.1 m and 5 nm, respectively [2, 6, 9, 16]. The adopted methodology for mineral identification is divided into three parts such as pre-image processing, extraction of mineral spectra, and validation. The stage of prefield-image processing contains de-striping, bad band removal, and atmospheric correction. The second stage related to extraction of mineral spectra through image. Third stage involves in matching of image-extracted spectra and measured spectra of minerals through USGS mineral library spectra and with field samples spectra. There are 155 bands and 390 bands were found suitable for mineral identification and extraction of spectra by Hyperion and AVIRIS-NG hyperspectral image, respectively (Fig. 2).

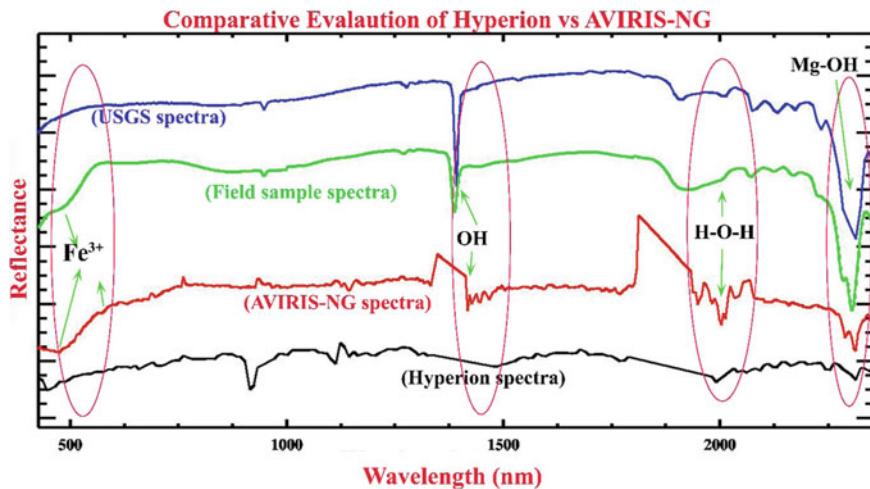
### 4 Result and Discussion

The comparative analysis of extracted spectra of talc mineral by Hyperion hyperspectral remote sensing image of Chabhadiya area of Jahajpur, Bhilwara, Rajasthan shows diagnostic absorption features at 2.31  $\mu\text{m}$  due to Mg-OH [3, 5, 12, 21]. Absorptions



**Fig. 2** Hyperion and AVIRIS-NG imagery

present at 1.4 and 1.9  $\mu\text{m}$  due to OH and HOH overtones. There is some minor absorption also present between 2.0 and 2.2  $\mu\text{m}$ . In spectral range of 0.4–2.5  $\mu\text{m}$ , some major absorption such as at 0.45  $\mu\text{m}$  due to  $\text{Fe}^{3+}$  and at 0.9  $\mu\text{m}$ , 1.1  $\mu\text{m}$  due to  $\text{O}_2$  and  $\text{H}_2\text{O}$  also identified in VNIR region of EMR. The Hyperion image spectra showing deep absorption at 0.85–0.95  $\mu\text{m}$  which is indication of the presence of hematite [24] and showing normal absorption at near 0.456  $\mu\text{m}$  [10]. The spectral analysis of AVIRIS-NG extracted talc mineral spectra shows diagnostic absorption at 2.31  $\mu\text{m}$  with doublet 2.29  $\mu\text{m}$  due to Mg-OH, and 1.4 and 1.9 due to overtones of OH and HOH. At near 1.4–1.5  $\mu\text{m}$  and 2.0  $\mu\text{m}$  shows maximum variation in absorption due to  $\text{O}_2$  and  $\text{CO}_2$  which is not present in the Hyperion image. Absorption at 0.47 and 0.6  $\mu\text{m}$  indicating the presence of  $\text{Fe}^{3+}$ , (Hematite) and lesser deep absorption at 0.9  $\mu\text{m}$  which is unavailable in Hyperion image. Various minor absorptions are also present minutely compared to minor absorption features of Hyperion image. The shape and reflectance value are also differentiating the characteristics and properties of both spectra. The subtle changes and subtle variation in shape, reflectance, and absorption can be observed very minutely in AVIRIS-NG than the Hyperion image. The spectroscopic spectra of talc field samples also validating the potential and capability of identification of talc and hematite with absorption at 2.31  $\mu\text{m}$  with doublet at 2.28  $\mu\text{m}$  with overtones of OH and HOH at 1.4 and 1.9  $\mu\text{m}$  and  $\text{Fe}^{3+}$  presence identified at absorptions at near 0.56 and 0.9  $\mu\text{m}$ , respectively. The shape, absorption, reflectance, and depth of image spectra of Hyperion and AVIRIS-NG in correlation of spectroscopic result of field sample spectra including comparison with USGS talc minerals spectra shows the most accurate and acute capability and performance of AVIRIS-NG hyperspectral images. The USGS talc mineral library spectra show complete matching of identified and measured spectra of talc through Hyperion, AVIRIS-NG, and spectroscopic field sample spectra (Fig. 3).



**Fig. 3** Comparative evaluation of spectral profile of Chabbadiya talc mineral

## 5 Conclusion

The observed spectra of USGS talc mineral identified spectra of talc minerals through Hyperion and AVIRIS-NG and measured spectra of talc field sample by spectroscopy show the variation in absorption and reflectance. The AVIRIS-NG and field sample spectra show the presence of hematite in talc mineral spectra, but Hyperion talc mineral spectra have only capability to identify the presence of talc minerals. Some minor absorptions and range of absorptions also vary with subtle shift (changes) in absorptions. USGS talc mineral spectra validating the capability of identification of talc minerals through Hyperion and AVIRIS-NG but shows maximum similarity with AVIRIS-NG compared to Hyperion image. These comparative analytical results show that coarse spatial, spectral resolution, and low signal-to-noise ratio of Hyperion is unable to identify the minor absorption features of other minerals such as hematite in talc minerals, but AVIRIS- NG fine spatial and spectral resolutions and higher signal-to-noise ratio show capability and potential of the presence of minor minerals concentration with major minerals and their abundance. The observed result measured field sample spectra also verifying the potential and capability of AVIRIS-NG compared to Hyperion hyperspectral remote sensing data.

**Acknowledgements** This work is supported by Space Application Centre, Indian Space Research organization India grant EPSA/4.2/2017.

## References

1. Agar, B., & Coulter, D. (2007). Remote sensing for mineral exploration—A decade perspective 1997–2007 plenary session : The leading edge. In B. Milkereit (Ed.), *Proceedings of Exploration 07: Fifth Decennial International Conference on Mineral Exploration* (pp. 109–136).
2. Bhattacharya, B. K. (2016). *AVIRIS programme and science plan*. No. February, 18–22.
3. Clark, R. N. (1999). *Spectroscopy of rocks and minerals, and principles of spectroscopy. Remote sensing for the earth sciences: Manual of remote sensing* (Vol. 3). <https://doi.org/10.1111/j.1945-5100.2004.tb00079.x>.
4. Dey, B., Das, K., Dasgupta, N., Bose, S., & Ghatak, H. (2016). Zircon U-Pb SHRIMP dating of the Jahazpur granite and its implications on the stratigraphic status of the Hindoli-Jahazpur group. In *Seminar Abstract Volume: Developments in Geosciences in the Past Decade*.
5. Ducart, D. F., Silva, A. M., Toledo, C. L. B., de Assis, L. M., Ducart, D. F., Silva, A. M., Toledo, C. L. B., et al. (2016). Mapping iron oxides with Landsat-8/OLI and EO-1/Hyperion imagery from the Serra Norte iron deposits in the Carajás mineral province, Brazil. *Brazilian Journal of Geology*, 46(3), 331–49. (Sociedade Brasileira de Geologia). <https://doi.org/10.1590/2317-4889201620160023>.
6. Hamlin, L., Green, R. O., Mouroulis, P., Eastwood, M., Wilson, D., Dudik, M., & Paine, C. (2011). Imaging spectrometer science measurements for terrestrial ecology: AVIRIS and new developments. In *IEEE Aerospace Conference Proceedings*, No. August. <https://doi.org/10.1109/AERO.2011.5747395>.
7. Heron, A. M. (1935). Synopsis of the Pre-Vindhyan Geology of Rajputana. *Trans. Nat. Instt. Sci. India. I*, 17–33.
8. Jing, C., Bokun, Y., Runsheng, W., Feng, T., Yingjun, Z., & Dechang, L. (2014). Regional-scale mineral mapping using ASTER VNIR/SWIR data and validation of reflectance and mineral map products using airborne hyperspectral CASI/SASI Data. *International Journal of Applied Earth Observations and Geoinformation*, 33, 127–141 (Elsevier B.V.). <https://doi.org/10.1016/j.jag.2014.04.014>.
9. JPL NASA. (2015). *ISRO - NASA AVIRIS – NG Airborne Flights over India sciene plan document for hyperspectral remote sensing*.
10. Kiran Raj, S., Ahmed, S. A., Srivatsav, S. K., & Gupta, P. K. (2015). Iron Oxides mapping from E0-1 hyperion data. *Journal of the Geological Society of India*, 86(6), 717–725. <https://doi.org/10.1007/s12594-015-0364-7>.
11. Kruse, F. A., Boardman, J. W., & Huntington, J. F. (2003). Comparison of airborne hyperspectral data and EO-1 hyperion for mineral mapping. *IEEE Transactions on Geoscience and Remote Sensing*, 41(6 PART I), 1388–1400. <https://doi.org/10.1109/TGRS.2003.812908>.
12. Meer, F. V. (2004). Analysis of spectral absorption features in hyperspectral imagery. *International Journal of Applied Earth Observation and Geoinformation*, 5(1), 55–68.
13. Mitchell, J. J., Shrestha, R., Spaete, L. P., & Glenn, N. F. (2015). Combining airborne hyperspectral and LiDAR data across local sites for upscaling shrubland structural information: Lessons for HyspIRI. *Remote Sensing of Environment*, 167, 98–110 (Elsevier Inc.). <https://doi.org/10.1016/j.rse.2015.04.015>.
14. Pargal, S. (2011). *Hyperspectral subspace identification and endmember extraction by integration of spatial-spectral information*. University of Twente.
15. Pour, A. B., & Hashim, M. (2014). Exploration of gold mineralization in a tropical region using Earth Observing-1 (EO1) and JERS-1 SAR Data : A case study from Bau Gold Field, Sarawak, Malaysia, 1, 2393–2406. <https://doi.org/10.1007/s12517-013-0969-3>.
16. SAC, ISRO. (2016). Space application Center.Pdf. SAC COURIER, 41(3). <http://www.sac.gov.in/SACSITE/SACCourier/July2016.pdf>.
17. Saxena, A. S. H. A., & Pandit, M. K. (2012). Geochemistry of Hindoli group metasediments, SE Aravalli Craton, NW India : Implications for palaeoweathering and provenance. *Journal Geological Society of India*, 79, 267–278. [http://mocl.gov.in/Reports/EXE\\_SUMM\\_BANERA.pdf](http://mocl.gov.in/Reports/EXE_SUMM_BANERA.pdf).

18. Shekhawat, L.S., & Sharma, V. (2001). *Basemetal exploration in Pachanpura-Chhabriya Block, Umedpura-Manoharpura Block, Gelaji (East) and Amargarh Blocks, Jahajpur Belt, Bhilwara District, Rajasthan* (Final Report For The Field Seasons 1999–2000 & 2000–2001).
19. Sinha Roy, S., & Malhotra, G. (1988). Structural relations of proterozoic cover and its basement: An example from the Jahazpur Belt, Rajasthan. *Jour. Geol. Sec. India* (in Press).
20. Thorpe, A. K., Frankenberg, C., Aubrey, A. D., Roberts, D. A., Nottrott, A. A., Rahn, T. A., et al. (2016). Mapping methane concentrations from a controlled release experiment using the next Generation Airborne Visible/Infrared Imaging Spectrometer (AVIRIS-NG). *Remote Sensing of Environment*, 179, 104–115. <https://doi.org/10.1016/j.rse.2016.03.032>. (Elsevier Inc.).
21. van der Meer, F. D., Harald, M. A., van der Werff, van Ruitenbeek, F. J. A., Hecker, C. A., Bakker, W. H., et al. (2012). Multi- and hyperspectral geologic remote sensing: A review. *International Journal of Applied Earth Observation and Geoinformation*, 14(1), 112–128. <https://doi.org/10.1016/J.JAG.2011.08.002>.
22. Yadav, S. K., & Rao, R. S. (2009). Assessment of iron ore potentiality of Jahazpur Belt, Central Rajasthan—A case study. *Geological Survey of India*, 128(3239), 885–886. <https://doi.org/10.1038/128885a0>.
23. Yadav, O. P., Babu, T.B., Shrivastava, P.K., Pande, A. K., & Gupta, K. R. (2001). Short communications and its significance in West Jahajpur Basin, Bhilwara District, Rajasthan. *Journal Geological Society of India*, 58, 0–3.
24. Zhang, T., Yi, G., Li, H., Wang, Z., Tang, J., Zhong, K., Li, Y., et al. (2016). Integrating data of ASTER and Landsat-8 OLI (AO) for hydrothermal alteration mineral mapping in Duolong Porphyry Cu-Au deposit, Tibetan Plateau, China. *Remote Sensing*, 8(11). <https://doi.org/10.3390/rs8110890>.

# Real-Time Scheduling Approach for IoT-Based Home Automation System



Rishab Bhattacharyya, Aditya Das, Atanu Majumdar and Pramit Ghosh

**Abstract** Internet of Things (IoT) is one of the most disruptive technologies nowadays which can efficiently connect, control, and manage intelligent objects that are connected to the Internet. IoT-based applications like smart education, smart agriculture, smart health care, smart homes, etc., which can deliver services without manual intervention and in a more effective manner. In this work, we have proposed an IoT-based smart home automation system using a microcontroller-based Arduino board and mobile-based Short Message Service (SMS) application working functionality. Wi-Fi connectivity has been used to establish communication between the Arduino module and automated home appliances. We have proposed a real-time scheduling strategy that offers a novel communication protocol to control the home environment with the switching functionality. Our simulation results show that the proposed strategy is quite capable to achieve high performance with different simulation scenarios.

**Keywords** Task scheduling · Real-time · Deadline · Signal · Down counter clock

## 1 Introduction

Internet of Things (IoT) paradigm is a composition of intelligent and self-configuring devices. It represents the intertwining of physical objects—devices, vehicles, build-

---

R. Bhattacharyya (✉)

School of Electronics, Kalinga Institute of Industrial Technology, Bhubaneswar, India  
e-mail: [rishabbhattacharyya2009@gmail.com](mailto:rishabbhattacharyya2009@gmail.com)

A. Das (✉) · P. Ghosh

Department of Computer Science and Engineering, RCC Institute of Information Technology, Kolkata, India  
e-mail: [findaditya1639@gmail.com](mailto:findaditya1639@gmail.com)

P. Ghosh

e-mail: [pramitghosh2002@yahoo.in](mailto:pramitghosh2002@yahoo.in)

A. Majumdar

A.K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India  
e-mail: [a.majumdar007@gmail.com](mailto:a.majumdar007@gmail.com)

ings and other items embedded with electronics, software, sensors, and network connectivity that enables these objects to collect and exchange data [1]. The IoT object allows to send and control remotely through the existing network infrastructure. Combination of the current network technology and IoT platforms provides a large amount of space and innovative service based on wireless communication. Smart systems basically incorporate functions such as sensing, actuation, and control in order to analyze a situation and also to make decisions in a highly predictive or adaptive manner. [2] In a similar way, smart housing system or home automation will control lighting, climate, entertainment systems, and other appliances. When such a system is combined with the Internet, it will thereby become an important constituent of the Internet of things [3]. Home automation systems that integrate an astronomic time clock and perform events in a timed fashion implement the concept of real-time scheduling [4–6]. So, we need new technologies for handling such devices remotely and to establish communication, we require GSM, mobile technology, Short Message Service (SMS), and certain hardware resources. Since cellular phone is a recently invented technology, a home automation system based on cellular phone is extremely important for the human generation [7–9].

IoT smart objects' configuration and their management as well as their integration with real-time applications are complex challenges which require suitable scheduling models and techniques [10]. Scheduling is a process by which we can fulfill the users' requirement by allocating their request to different processes on the basis of the interest [11–13]. In this paper, we have proposed a priority-based real-time scheduling approach Shortest Deadline First—Real-Time Task Scheduling (SDF-RTTS) for controlling of smart home automation systems. Arduino-based platform has been used to perform proposed work which lies on the inherent irrelevancy of distance or remoteness, in controlling sensor-based technologies. Short Message Service (SMS) that triggers the appliances can be sent from anywhere regardless of the distance between the source and the appliance. We have not considered the security issues and network-related issues in this work. Our focus is only on the scheduling part mainly to optimize the overall execution time.

The organization of the paper is as follows: Related work has been discussed in Sect. 2. System model and assumptions are described in Sect. 3. A real-time priority-based scheduling strategy has been discussed in Sect. 4. Working strategy is explained with an example in Sect. 5. Results and experimental setup are illustrated in Sect. 6. Finally, Sect. 7 concludes the paper.

## 2 Related Work

The authors in [14] described a home automation system which is composed of smartphones and microcontrollers. Home appliances are controlled and monitored by the smartphone applications through various types of communication techniques. Different types of communication techniques such as ZigBee, Wi-Fi, Bluetooth, EnOcean, and GSM have been explored and compared. The advantage of this system is to

provide security so that the system is accessible only to the authorized users. In [15], the authors made a survey to understand the topography of devices used in the home automation system. They also compared the proposed programming languages for different systems which are based on ECA structures but presenting subtle differences. The authors in [16] established a connection between a temperature sensor and a microcontroller, where the temperature sensor was used to measure the temperature of the room and the speed of the fan varied according to the temperature using pulse width modulation technique. Although the advantages of the system were that it was simple, cost-effective, and provided automatic control, it has very specific limitations.

Scheduling algorithm is an important aspect of real-time systems. According to system environment, real-time system can be classified into uniprocessor scheduling, centralized multiprocessor scheduling, and distributed scheduling as shown in [17]. Real-time scheduling algorithms such as RMS, EDF, and LLF are discussed for uniprocessor systems. In multiprocessor systems, scheduling thought and strategies are investigated. GRMS and DSR are discussed for the distributed real-time scheduling algorithms. The authors in [18] said that an ideal scheduling algorithm minimizes the response time, maximizes the throughput, minimizes the overhead (in terms of CPU utilization, disk, and memory), and maximizes fairness.

However, as established by the same paper, there are no such algorithms in existence which meet all the criteria of an ideal algorithm. The authors in [19] further stated that the four model tenets are static table-driven scheduling, static priority preemptive scheduling, dynamic planning-based scheduling, and dynamic best effort scheduling when we talk about the best efforts that are to be used in scheduling. The authors in [20] established a home automation system that operated on the basis of sending Short Message Service (SMS); however, the specific limitation of this work was profound because there was no established scheduling algorithm to regulate the entry of multiple SMSs.

The authors in [21] clearly underscored the importance of the duality of wireless security and home automation system. The usage of text messages as “alerts” can be a very important security initiative in domain of home automation. This dual was well addressed by the authors. Moreover, IoT-based home automation systems serve to mitigate the limitations of existing Bluetooth-controlled systems as the system can be accessed from anywhere (even from places which do not have Wi-Fi or Internet connectivity, as only the board is required to have Internet connectivity). Lastly, the authors cited the importance of the flexibility of the system as it does not use any of the traditional user interfaces of smartphones, but a few digits of the keypad of the phone.

### 3 System Model and Assumptions

The system model gives a schematic description of the idea presented in this paper. It incorporates multiple users in the system who can send SMSs (tasks). These SMSs

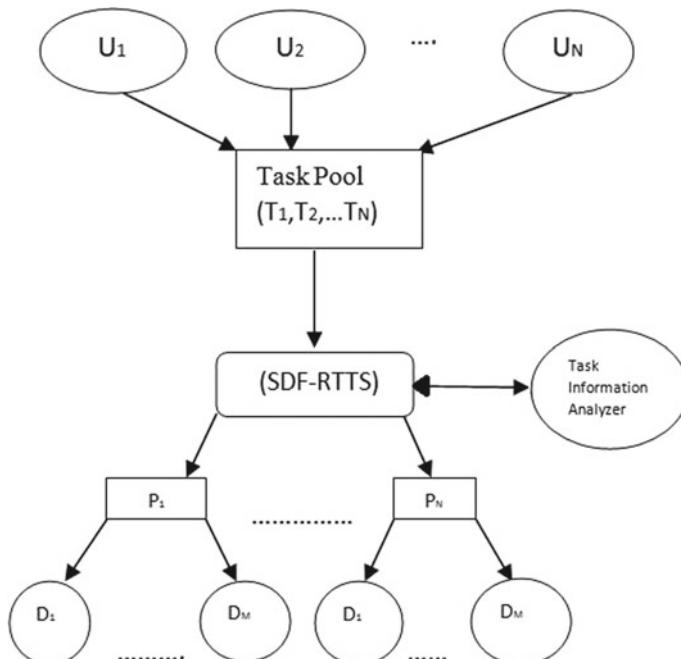
are stored in the task pool. We use a priority scheduling algorithm SDF-RTTS based on Shortest Deadline First (SDF) to sort the SMSs. The working of our priority scheduling algorithm SDF-RTTS is later explained in the working section. Without loss of generality, we are assuming that (Fig. 1).

- Each user can send one signal which will be treated as a task at a time.
- Device actuation time for each home appliance is known to the processors at design time.
- Task to processor mapping is fixed. Each task will be processed by a unique processor which will remain the same if that task has multiple occurrences.

The scheduler allocates task to the processing elements based on their priority. Each processor has control over the appliances connected to it.

#### 4 SDF-RTTS: Working Strategy

Let us assume that at time instance  $t$ ,  $N$  number real-time tasks arrive for possible allotment on the processors. Our proposed SDF-RTTS will attempt to meet the timing requirements of all tasks by completing them within their soft deadline. Each task consists of two parameters:



**Fig. 1** SDF-RTTS working module

- { 1) Signal with ON/OFF value (S)
- { 2) Deadline of the Task (T)

ON/OFF value is a binary number. ON signal indicates the value “1” and OFF is “0”. Deadline of a task can be represented by multiple bits. Assume if deadline is 5 unit, then signal can be represented as “101”.

We also assume that  $M$  number of home automation appliances are connected to each processor. A down counter clock is maintained to track the deadline. Each appliance or device actuation time is known to its connected processor. First, processor checks the deadline with the devices’ actuation time. Devices are partitioned into two categories based on the deadline. One partition, let say SET1, consists of devices whose actuation time is less than the deadline and second SET2 consists of devices whose actuation time is greater than the deadline. Appliances or devices of SET1 ensure that it will be activated with the desired condition with the stipulated deadline. On the other side, SET2 devices will be activated but without reaching the desired conditions.

Our proposed algorithm will check if any task has arrived at SET1 or not. If a user request or task arrived, then it will start the down counter clock (DL). Each device connected with the processor has different actuations, and it is known in prior. User request or task mapping to the processor is fixed. Down counter clock (DL) starting value will be the deadline and continues up to zero (0). In between that, devices those actuation time matches with the down counter clock (DL) will receive an activation signal. SET2 devices will get activate signal randomly when the processor is free or within the time gap between two consecutive devices actuation time or in between down counter clock (DL) start time and first device activation.

The pseudocode for the SDF-RTTS scheduling strategy is shown in Algorithm 1.

## 5 SDF-RTTS: An Example

Let us assume that four users A, B, C, D want to start their home automation appliances and make the home environment comfortable before they reach home. Hence, users A, B, C, and D send a text message which is known as a task to switch on the devices, say  $D_1$ ,  $D_2$ ,  $D_3$ , and  $D_4$  in their respective homes. In text message, users have mentioned the deadline or the time within he/she will reach home. Let us say the deadlines for reaching the destination are 5, 10, 8, and 12 min. Devices’ actuation time is also known to the scheduler SDF-RTTS. Let us say 1, 2, 5, and 10 min are the actuation time for the devices  $D_1$ ,  $D_2$ ,  $D_3$ , and  $D_4$

### **Algorithm 1: SDF-RTTS Scheduling Strategy**

Input:  $N$  number of signals  $[S, T]$ , Number of devices  $M$

Output: Priority-based device actuation

1. Initialize DL = deadline; Detect signals  $[S, T]$  from multiple users;
2. Set priority of the tasks based on the deadline ( $T$ ) given by the user (Smallest deadline will have the highest priority);
3. Assign the tasks to the predefined processors based on the priority;
4. for(each devices in N)do
5. if ( $deadline \geq device actuation time$ ) then{
6. move device to SET 1;
7. else{
8. if ( $deadline \leq device actuation time$ ) then
9. Move device to SET 2;}}
10. if (SET 1 != NULL) then{
11. Set a down counter clock(DL);
12. if ( $T == device actuation time$ ) then{
13. Send the signal to activate the device;
14. Decrement the clock till  $DL = 0$ ;}}
15. if (SET 2 != NULL) then{
16. if ( $T \neq device actuation time$ ) then
17. Send the signal randomly to activate the devices;}
18. End

respectively. Hence, according to the algorithm, user A has the highest priority as it has shortest deadline, then C, B, and D. Now for the first task, overall deadline is 5 min but device actuation time for  $D_4$  is 10. Hence, classify the device set into two subsets. One set, SET1, has devices whose actuation time is less or equal to the deadline, and the other set named as SET2 has devices whose actuation time is greater than the deadline. Algorithm will check whether SET1 is null or not, if it has devices in the set it will activate the down counter clock (DL). According to the actuation, time processor will send the activation signal to the devices. At down counter clock (DL), five processors will activate  $D_3$ , at two  $D_2$ , and at one  $D_1$ , respectively. SET2 devices will be activated randomly by the processor when it is free or there is no request for SET1 device activation. This similar procedure will be followed by the other processors also.

## 6 Evaluation and Results

We have presented simulation-based results from the solution of our proposed SDF-RTTS strategy for real-time scheduling of home automation appliances. We have analyzed the execution time and performance for the proposed solutions. Before presenting the detailed results, we now discuss our experimental setup.

## 6.1 Experimental Setup

**Arduino:** It is a microcontroller board that happens to be based on the AT-mega328. There are a total number of fourteen digital input/output pins where six of them can be used as PWM outputs. It also consists of six analog inputs, 16 MHz ceramic resonators, a USB connection, a power jack, an ICSP header, and a reset button [17]. Arduino Uno is the main processing element in this experiment.

**GSM SIM 300:** Sim300 is well known and is often used in quite a lot of projects, and hence various types of development boards with regard to this have been developed. These developmental boards encompass various features that make it easily communicable with SIM 300 module [18]. The GSM module used in this experiment comprises two parts: a TTL interface and RS232 interface. The TTL interface is used for interfacing and communicating with the microcontroller [22]. The RS 232 interface uses a MAX232 IC to allow communication with the PC. It also consists of SIM slot. This module in this current application is used as a data circuit-terminating equipment and PC as a data terminal equipment [23].

**Relay Drive:** Relay can be used as an electromagnetic device that is very commonly used to separate two circuits electrically and connect them magnetically [24–27]. They are very popular and equally useful devices which enable one circuit to switch to another one while they are completely isolated [28]. These devices are mostly preferred to be used while interfacing an electronic circuit (working at a low voltage) to an electrical circuit which works at very high voltage [29]. Henceforth, a small sensor circuit can drive, say, a refrigerator or an electric bulb [30–32].

**BC 547:** BC547 is a bipolar junction transistor (NPN). A transistor represents a transference of resistance, which is commonly used to increase current many folds [33].

**Wi-Fi module:** Wi-Fi connectivity has been used to establish the communication between the Arduino and home-automated appliances.

**Number of Tasks ( $T$ ):** Totally, 4–20 number of tasks or user requests are considered with different appliances used for simulation.

**Number of Processors ( $P$ ):** As task and processor mapping is fixed,, each task has a dedicated processor to perform the execution. Hence, a number of processors are also varied from 4–20.

**Number of Devices (D):** Totally, 4–16 number of different types of devices have been used for simulation.

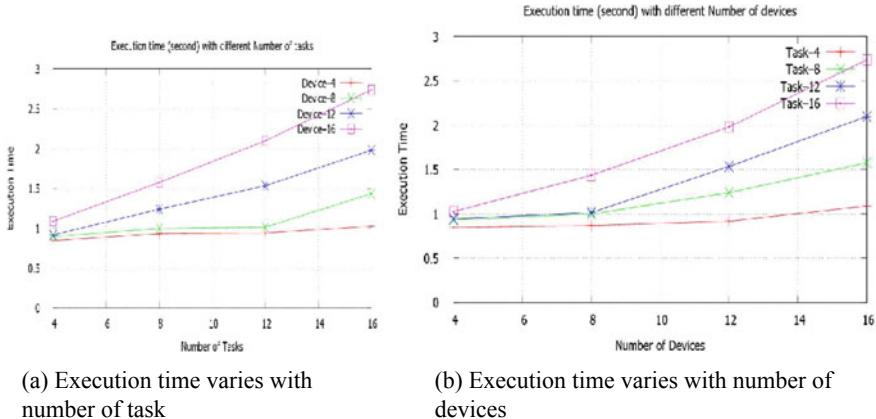
Each result is generated by executing 100 different instances of each data set type and then taking the average over these 100 runs.

Table 1 shows the performance of the SDF-RTTS strategy over different number of processors, tasks, and devices. Run time of SDF-RTTS strategy is measured in terms of millisecond.

As the number of tasks increases, execution time of our proposed strategy SDF-RTTS also increases. For each task, few verifications are done before it gets scheduled to the processor. As task number goes high, it takes more time to verify and schedule it to the appropriate processor. Figure 2a clearly depicts that when the number of

**Table 1** Task execution time (second)

<i>D</i>	<i>T</i>	<i>P</i>	Runtime									
4			0.847			0.932			0.943			1.026
8	4	4	0.896	8	8	0.998	12	12	1.014	16	16	1.435
12			0.917			1.239			1.534			1.980
16			1.089			1.576			2.098			2.737



**Fig. 2** Performances of SDF-RTTS strategy

tasks increases (without increasing the device number), overall execution time also increases.

Figure 2b shows that if we increase the number of devices for the same number of tasks, execution time also increases. As the number of devices increases, device count in SET1 and SET2 also increases. Hence, it has to send more number of signals to activate all the devices.

## 7 Conclusion

In this paper, we presented a strategy for allocating real-time tasks on Arduino-based platform using mobile SMS application such that we can minimize the overall execution time. Our proposed scheduling strategy SDF-RTTS is based on the “Shortest Deadline First” approach, which ensures that scheduling will fulfill all the real-time constraints and will optimize the overall execution time. Arduino-based test bed which provides the exibility of controlling the IoT devices through Wi-Fi connectivity. We designed, implemented, and evaluated the algorithms using simulation-based experiments, and results are promising. Though there are some drawbacks like security issues, network issues which we have not covered are presented in this work. But in future work, we will focus on these issues.

## References

1. Byun, J., Kim, S., Sa, J., Kim, S., Shin, Y. T., & Kim, J. B. (2016). Smart city implementation models based on IoT technology. *Advanced Science and Technology Letters*, 129(41), 209–212.
2. Zhou, H. (2012). The Internet of Things in the cloud: A middleware perspective, 1st edn. Boca Raton, FL: CRC Press. ISBN: 1439892997, 9781439892992.
3. Bin, S., Yuan, L., & Xiaoyi, W. (2010). Research on data mining models for the Internet of Things. In *International Conference on Image Analysis and Signal Processing* (pp. 127–132).
4. Dickerson, R., Gorlin, E., & Stankovic, J. (2011). Empath: A continuous remote emotional health monitoring system for depressive illness. *Wireless Health*.
5. Hishama, A. A. B., Ishaka, M. H. I., Teika, C. K., Mohameda, Z., & Idrisb, N. H. (2014). Bluetooth-based home automation system using an android phone. *Jurnal Teknologi (Sciences & Engineering)*, 70(3), 57–61.
6. Pavana, H., Radhika, G., & Ramesan, R. (2014). PLC based monitoring and controlling system using WiFi device. *IOSR Journal of Electronics and Communication Engineering*, 9(4), 29–34.
7. Tang, S., Kalavally, V., Ng, K. Y., & Parkkinen, J. (2017). Development of a prototype smart home intelligent lighting control architecture using sensors onboard a mobile computing system. *Energy and Buildings*, 138, 368–376.
8. Panth, S., Jivani, M., et al. *Home Automation System (HAS) using android for mobile phone*. IJECSE (Vol. 3, Issue 1). ISSN 2277-1956/V3N1-01-11.
9. He, W., Yan, G., & Xu, L. Developing vehicular data cloud services in the IoT environment. In: *IEEE Transactions on Industrial Informatics* (pp. 1–1). <https://doi.org/10.1109/ii.2014.2299233>.
10. Matlak, S., Bogdan, R. (2016). *Reducing energy consumption in home automation based on STM32F407 microcontroller*. IEEE.
11. Deng, L. (2010). Research of intelligent home control system. In *International Conference on Electrical and Control Engineering*.
12. Delgado, A. R., Picking, R., & Grout, V. *Remote-controlled home automation systems with different network technologies*. Centre for Applied Internet Research (CAIR), University of Wales, NEWI, Wrexham, UK. <http://www.glyndwr.ac.uk/groutv/papers/p5.pdf>.
13. Piyare, R., Tazil, M. (2011). Bluetooth based home automation system using cell phone. In *2011 IEEE 15th International Symposium on Consumer Electronics*.
14. Asadullah, M., & Raza, A. (2016) An overview of home automation systems. In: *2016 2nd International Conference on Robotics and Artificial Intelligence (ICRAI)* (pp. 27–31), IEEE.
15. Demeure, A., Caaú, S., Elias, E., & Roux, C. (2015). Building and using home automation systems: A field study. In: *International Symposium on End User Development*. Heidelberg: Springer (pp. 125–140).
16. Das, A. (2018). Fan speed controlled system by temperature using pulse width modulation (pwm). *International Journal of Current Research*, 10(4), 68021–68024.
17. Jie, L., Ruifeng, G., Zhixiang, S. (2010). The research of scheduling algorithms in real-time system. In: *2010 International Conference on Computer and Communication Technologies in Agriculture Engineering (CCTAE)* (Vol. 1, pp. 333–336), IEEE.
18. Adekunle, Y., Ogunwobi, Z., Jerry, A. S., Efuwape, B., Ebiesuwa, S., & Ainam, J. P. (2014). A comparative study of scheduling algorithms for multiprogramming in real-time systems. *International Journal of Innovation and Scientific Research*, 12(1), 180–185.
19. Ramamritham, K., & Stankovic, J. A. (1994). Scheduling algorithms and operating systems support for real-time systems. *Proceedings of the IEEE*, 82(1), 55–67.
20. Teymourzadeh, R., et al. (2013). Smart GSM based home automation system. In *2013 IEEE Conference on Systems, Process & Control (ICSPC)*, IEEE.
21. Kodali, R., & Jain, V., Bose, S., & Boppana, L. (2016). *IoT based smart security and home automation system* (pp. 1286–1289). <https://doi.org/10.1109/ccaa.2016.7813916>.
22. Martinez, K., Hart, J. K., & Ong, R. (2004). Environmental sensor networks. *Computer*, 37(8), 50–56.

23. Ma, Y., Richards, M., Ghanem, M., Guo, Y., & Hassard, J. (2008). Air pollution monitoring and mining based on sensor grid in london. *Sensors*, 8(6), 3601–3623.
24. Ghosh, P., Bhattacharjee, D., & Nasipuri, M. (2017). Automatic system for plasmodium species identification from microscopic images of blood-smear samples. *Journal of Healthcare Informatics Research*, 1(2), 231–259.
25. Datta, S., Bhattacharjee, D., & Ghosh, P. (2009). Path detection of a moving object. *International Journal of Recent Trends in Engineering*, 2(3), 37.
26. Ghosh, P., Bhattacharjee, D., & Nasipuri, M. (2015). *An automatic non-invasive system for diagnosis of tuberculosis*. *Applied computation and security systems* (pp. 59–70). New Delhi: Springer.
27. Bhide, V. H., & Wagh, S. (2015). i-learning IoT: An intelligent self-learning system for home automation using IoT. In *2015 International Conference on Communications and Signal Processing (ICCP)*. IEEE.
28. Stankovic, J. (2014). Research directions for the internet of things. *IEEE Internet of Things Journal*, 1(1), 3–9.
29. Gubbi, J., et al. (2014). Understanding the Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7), 1645–1660.
30. Chattoraj, S. (2015). Smart Home Automation based on different sensors and Arduino as the master controller. *International Journal of Scientific and Research Publications* (Vol. 5, Issue 10).
31. Kushalnagar, N., Montenegro, G., & Schumacher, C. (2007). *IPv6 over low-power wireless personal area networks (LoWPANs): Overview, assumptions, problem statement, and goals*. RFC 4919.
32. Sweatha, K. N., Poornima, M., Vinutha, M. H. (2013). Advance home automation using FPGA controller. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(7).
33. Deore, R. K., Sonawane, V. R., Satpute, P. H. (2015). Internet of Thing based home appliances control. In *International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 898–902).
34. Nonaka, T., Shimano, M., Uesugi, Y., & Tomohiro. (2010). Embedded server and client system for home appliances on real-time operating systems. IEEE.

# Floating Car Data Map-Matching Utilizing the Dijkstra's Algorithm



Vít Ptosek, Lukáš Rapant and Jan Martinovič

**Abstract** Floating car data (FCD) are one of the most important sources of traffic data. However, their processing requires several steps that may seem trivial but have far-reaching consequences. One such step is map-matching, i.e. assignment of the FCD measurement to the correct road segment. While it can be done very simply by assigning the point of measurement to the closest road, this approach may produce a highly undesirable level of error. The second challenge connected with processing of FCD measurements is missing measurements. They are usually caused by the shortcomings of GPS technology (e.g. the satellites can be obscured by buildings or bridges) and may deny us many measurements during longer downtimes. The last problem we will solve is the assignment of measurements to very short segments. FCD measurements are taken in periodic steps for several seconds long. However, some road segments are very short and can be passed by a car in the shorter interval. Such segments are therefore very difficult to monitor. We plan to solve all these problems through a combination of geometric map-matching with the Dijkstra's shortest path algorithm.

**Keywords** Floating car data · Map-matching · Traffic routing · Dijkstra algorithm

## 1 Introduction

In the world of Intelligent Transportation System (ITS) [1], one of the biggest challenges is to be able to express actual traffic status correctly. This information does matter in certain traffic-modelling-related areas, and its application significantly helps to analyse, for example, travel time prediction, which leads to routing time management improvement and traffic congestion avoidance and prevention, etc.

---

We certify that this manuscript is our original contribution and it was not submitted or accepted anywhere else.

---

V. Ptosek (✉) · L. Rapant · J. Martinovič

IT4Innovations, VŠB - Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava,  
Czech Republic

e-mail: [vit.ptosek@vsb.cz](mailto:vit.ptosek@vsb.cz)

Generally, data sources utilized for this monitoring can be seen as two main classifications—stationary and floating ones. Stationary data contain representations of sources as profile sensors and toll gates. They are very accurate and monitor the traffic flow in its entirety. Nevertheless, in the case of countries such as the Czech Republic, where the sensor network is too sparse, potential risks exist which need to be taken into account. Because every measuring point, as ASIM sensor [2], comes with unnecessary expenses, there are only up to 120 toll gates equipped with them, and they can only be found on motorways. Additionally, they lack any flexibility. If there is some modification of the traffic network, they cannot be moved to more important or interesting places. Also, toll gates are often placed in such a manner that only allows the splitting of roads into various length fragments, sometimes reaching many kilometres between each other. The result is that data obtained from ASIM sensors only describes the traffic state reliably in the area around tollgates and does not give any idea otherwise.

The opposite of stationary data is so-called Floating Car Data (FCD) [3]. The floating car data technique is based on floating cars informing a central data system about their travel on a road network. The floating cars repeatedly update their recent accumulated position data and sometimes send travel intentions. The central data system then tracks such data and uses information gained along the travelled routes.

Speed, timestamp, and average travel time are essential information for most FCD techniques. ITS usually provides only speeds and average travel times lining road links (for example, see Khan et al. [4]). FCD can manage to predict (short-term) travel conditions as critical situations and incident detection (for example, see Wu et al. [5]) or determine patterns based on origin–destination (for example, see Ma et al. [6]).

Specific FCD systems aimed to present short-term traffic forecasting solutions based on FCD history where the software was encouraged heavily by car flotillas data in return to offer information from large cities in real time. The information such as speed showed urban arterials or particular traffic segments, for example, the Italian motorway [7], Berlin [8], Beijing [9], Vienna [10], and many others.

This approach, as already mentioned, is based on the measurements of certain vehicle metrics such as geolocation, speed, bearing, and timestamp information within the traffic. This information is received from the GPS module placed inside the car. The transmitter is usually a cell phone or a radio unit nowadays. These broadcasts are collected and temporally and spatially aggregated by an FCD engine, which then computes average speed for each part of the road in a certain time interval. The traffic speed along with other metrics is calculated preferably every minute as the mean of the speed of all floating cars belonging to the very same segment in the last period of time.

The described approach may not always be sufficient, but it certainly has advantages. Currently, a built-in GPS unit is quite common and standard at least for company flotillas. Additionally, the number of personal cars equipped with a GPS module has doubled in the last 5 years. Also, the availability of smartphone technology increases the amount of source data and improves FCD contributions, and changes

are that the trend will continue. In contrast with stationary data, GPS receivers are not fixed to predefined places, which improves coverage.

On the other hand, some disadvantages exist as well and are to be found especially in the two following sources: collecting the data and the GNSS [11] technology itself. The very same principle that enables FCD to cover most of the traffic network can be its greatest fault. When a segment contains too few floating cars or none at all, the quality of the output severely diminishes. This problem can easily happen even on the motorway during off-peak hours and is quite usual on less used roads. It drastically complicates usage of the FCD data from certain times and all the less used roads. Also, it may be very difficult to cover short road segments with any kind of measurement, because there is a low probability that a car will report its speed when driving along them. The other disadvantage comes from the GPS devices as they can fail to produce accurate outputs. The output accuracy relies upon many circumstances. Some circumstances which have a strong influence are the quality of a given device, the location of the device, and the weather conditions. When these combine, it can lead to positioning flaws of several metres. Satellite-based technology such as GNSS needs a GPS receiver to be able to connect with several satellites, which transpires to be difficult or even impossible in some cases such as urban areas with tall buildings and wide walls creating obstacles between satellites and receivers. When a GPS receiver is unable to resolve its position then, of course, it cannot report its coordinates.

All these faults, however, can be mitigated to a certain extent by smart map-matching and aggregation of FCD measurements. Map-matching the FCD measurement to a certain road segment can be done very easily; it is simply finding the geometrically nearest road segment. This approach sadly introduces much error into the measurements because, due to the GPS properties, the point can be snapped to a vastly different road. We propose to combine this snapping approach with the Dijkstra's routing algorithm [12]. The task of the Dijkstra's routing algorithm is to check the viability of neighbouring segments and therefore filter those which cannot be reached by a said car within a given time from the last measurement. It also maps the route of the car between measurements, so it can create dummy measurements for very short segments and fill in the missing measurements.

This article has the following structure. Section 2 contains a short description of the current state of the art. Section 3 provides an explanation of our map-matching algorithm. Section 4 presents our experimental results and their setting. The article is concluded with an evaluation of our proposed method and some remarks regarding future work.

## 2 State of the Art

We have chosen to utilize classification of methods found in the article [13], i.e. we will speak separately about the algorithms working only with geometry, algorithms

working with some additional information and utilizing weights, and advanced algorithms utilizing more complex solutions.

## 2.1 *Geometry Based Approaches*

These approaches were the first developed solutions to the problem. In general, they are not attractive because they do not offer any advantage over more complicated approaches except for their speed of calculation (which can be important in some cases). These approaches usually have a very simple concept and therefore are far too inaccurate to be used in parts of the traffic network with a high density of roads (i.e. cities) because there are problems with a high margin of error where there are too many similar candidate solutions. These were simple algorithms utilized prior to the advent of more computationally intensive approaches. However, they are still used in some cases as parts of more complex solutions, so we will mention some selected approaches.

The first presented article is written by White et al. [14]. It sets the baseline for the map-matching problem. It describes its basic principles and presents four possible geometrical and topological map-matching solutions to this problem. The first solution is very simple. It translates the point of measurement onto the closest edge by its geographical distance. The second solution can be described as being similar to the first one. In addition, it utilizes a direction difference between the measurement and the candidate segment to identify impossible segments. The third and fourth solutions are more topological and improve the previous one with the utilization of information about the topology of the traffic network to remove unconnected edges or use of curve-to-curve matching, respectively. However, these proposed solutions are, in some sense, considered outdated because current solutions present different and more beneficial approaches to solve many problems such as the first session point or measurement points close to or in the intersection.

Another example of these approaches is work written by Taylor et al. [15]. In this article, they present a curve-to-curve line-snapping approach, which is based solely on geometric properties of the road and the measurements. The main advance presented in this paper is the introduction of a least-squares optimization algorithm to model the vehicle position in the Euclidean space (i.e. utilizing 3D space). This estimation enabled the use of height information to improve the accuracy of the method.

In another paper, the last one presented in this part, Srinivasan et al. [16] present an approach based on point-to-curve line-snapping approach. This task is done by a combination of a Kalman filter and both the GPS location of the measurement and the vehicle's last known position.

## 2.2 Weight-Based Approaches

Most line snapping algorithms presented in the following paragraphs have multiple variants of their progress for map-matching the raw measurement to the correct segment in various situations. There are different approaches for matching the first segment, subsequent segments, or the segments in or adjacent to the intersections. Correct application of these rules significantly increases the percentage of correct segment identification. It also does not make said algorithm overly elaborate or resource demanding. As for the representatives of weight-based approaches, we have chosen the following articles.

In their article, Quddus et al. [17] proposed a method that, for choosing the best segment, utilizes several simple weighted criteria such as angle difference between a direction of the measurements and direction of candidate segments or geographical distance of GPS measurement to each candidate segment. In the case of the intersections, their approach uses adjacency in the traffic network graph to identify the candidate edges.

Another example of these approaches was proposed in the paper written by Velaga et al. [18]. It proposes a topological map-matching algorithm for intersections based on weights of four criteria. The best candidates' segment at intersections is chosen by its distance from the measurement, the direction variance calculated as the cosine of the angle between the direction of the last measurement and the candidate segment, a bonus score which is determined by the possibility of the legal turnings from the last known measurement (in cases where the turn is illegal, the segment suffers a penalty), and another bonus score determined by linear connectivity to the last known segment (again, unconnected segments suffer a penalty).

The last paper in this section was written by Li et al. [19]. They also present a map-matching approach based on weights of inputs. In the first step of their approach, they proposed the utilization of an extended Kalman filter that is used to assimilate a digital elevation model, the location of measurement and its last known position. In the second step, the correct candidate segment is identified by a correct weighting of its distance from the measurement and difference of its direction to the direction of measurements.

## 2.3 Advanced Approaches

These approaches usually integrate approaches from the previous categories with some advanced data processing and machine learning methods. Most of the currently developed approaches fall into this category because they tend to be both very accurate and computationally intensive. However, due to the growing availability of computational power, the later fact slowly ceases to be much of a problem. We have chosen the following articles as representatives of these approaches.

The first representative of this approach is the article written by Kim and Kim [20]. It proposes an approach based on a curve-to-curve mapping. Its main benefit is a calculation of the C-measure index for each possible segment. It is determined by a combination of the geographical distance from the measurement, the distance of the last measurements from the segments (i.e. their trajectory) and its adjacency to the previously known segment. The C-measure is then utilized to identify the best candidate segment. This process was implemented as an adaptive fuzzy inference system.

Another example of advanced approaches is the work of Quddus et al. [21]. They implemented a fuzzy inference system utilizing a very broad spectrum of information, which is intended to cope with signal outages. They organized their fuzzy rules into three sets, where each set is intended for a different map-matching situation. The first set of rules is intended to solve the problem of mapping the initial measurement, the second set is used for mapping the subsequent measurements, and the last set is used in case of proximity of intersections. The main disadvantage of this work is the computational complexity of this system. It has 23 fuzzy rules for each segment that are used to determine its probability. Due to this fact, this algorithm is considered quite difficult to apply in real-world map-matching, where fast calculation is one of the more important considerations. However, it is also considered to be one of the most accurate algorithms.

Another possible approach was presented by Yuan et al. [22]. The proposed algorithm is based on a calculation of the error region base of the GPS measurement error. This calculation is done from the information about both the FCD and the traffic network and is used to determine the potential candidate segments. Then, they utilize parallel computing to solve the problem of finding a matching path graph with spatial and temporal analyses.

One of the other possible methods among the advanced approaches is the Hidden Markov Model (HMM). There are many articles concerning this approach. For example, both Newson et al. [23] and Ren and Karimi [24] proposed the use of HMM in a map-matching problem. Their works, however, prove that the use of HMM is complicated in the case of map-matching. HMM works with the most recent measurements (usually at least four of them) to determine the sequence of the most probable segments. It requires the storage of a lot of data and slows down the response of the entire algorithm (even HMM itself is not considered to be the fastest algorithm). This problem seems to be solved by Che et al. [25] who propose a different structure for the HMM, which they called enhanced hidden Markov map-matching. They seem to have overcome the problems of previous HMM-based works with improved performance and slightly improved accuracy.

### 3 Algorithm

As it was already mentioned, our algorithm has two purposes:

1. Provide reliable map-matching, even in urban conditions (i.e. we must be able to reliably solve areas around the intersection, mapping the point onto the segments with the right direction, etc.).
2. Create dummy measurements for very short segments and fill in the missing measurements.

Our map-matching algorithm is based on traffic routing, which is the task of finding the shortest path in the traffic network. This firmly places it amongst the advanced approaches to map-matching. There are many algorithms that can be used to perform this task; however, due to the demands of this routing task (i.e. finding many very short routes) we have chosen to utilize the basic Dijkstra's algorithm [12] because properties of algorithms such as A\* are beneficial for substantially larger tasks. For the sake of completeness, we present a short description of the Dijkstra's algorithm.

It is possible to formalize the general course of our algorithm into the four following tasks:

1. Finding the possible candidate segments
2. Perform the routing to these segments
3. Find the most probable segment and eventually correct the previous segment
4. Insert the artificial FCD measurements along the route to the chosen segments

Let us go through these steps in more detail. The first step is a classic point-to-curve search where we try to find all segments that satisfy the condition

$$\text{dist}(p_m, \text{seg}_n) \leq v_{\max} \cdot \Delta t,$$

where  $\text{dist}(p_m, \text{seg}_n)$  is Euclidean distance between measured point  $p_m$  and segment  $\text{seg}_n$ ,  $v_{\max}$  is maximal allowed speed in the network in reasonable distance from  $p_m$  and  $\Delta t$  is time interval between the current and the previous measurement. Please note that if  $\Delta t < 60$  s, we terminate the current session and start a new one taking this point as the first one. This decision was made because if  $\Delta t$  is too large, there can be a lot of variability in the possible routes and resulting mapping error undesirably increases.

Potential segments (and exact points on them) within this bound are identified by simple geometrical line snapping implemented within the Spatialite library. In the case of the first point of the session, this step is simplified to finding the closest segment and rectifying the potential error in the third step of processing of the next point. Also, the next steps are skipped in this case.

In the second step of the algorithm, we perform the one to many routings from the last snapped point to all potential points identified in the first step. We try to find the fastest route, i.e. not the shortest one from the spatial perspective, but from the

temporal perspective. For the calculation of the travel time, we utilize free flow speed obtained from Open Street maps [26]. This may produce some inaccuracy due to the other traffic involved, but this problem can be rectified by the addition of traffic speed profiles. After the routing step, we attain the list of travel times  $\text{TT}_n$  to all candidate segments.

In the third step, we compare these travel times to  $\Delta t$ . For the final step, we choose only segments satisfying the condition

$$\frac{\text{TT}_n}{\Delta t} \in \langle 1 - \alpha_1, 1 + \alpha_2 \rangle, \quad (1)$$

where  $\alpha_1$  is the coefficient of travel time calculation error which accounts for the possible delay caused by traffic conditions and  $\alpha_2$  is the coefficient of travel time calculation error which accounts for the possible speedup caused by driver's misbehaviour.  $\alpha_1$  is usually set to be higher than  $\alpha_2$ . Applying this condition from Eq. 1 produces a much smaller candidate set, where the best candidate is chosen again based on the distance from  $p_m$ . This step is, however, open for future improvement by application of more complex decision-making processes.

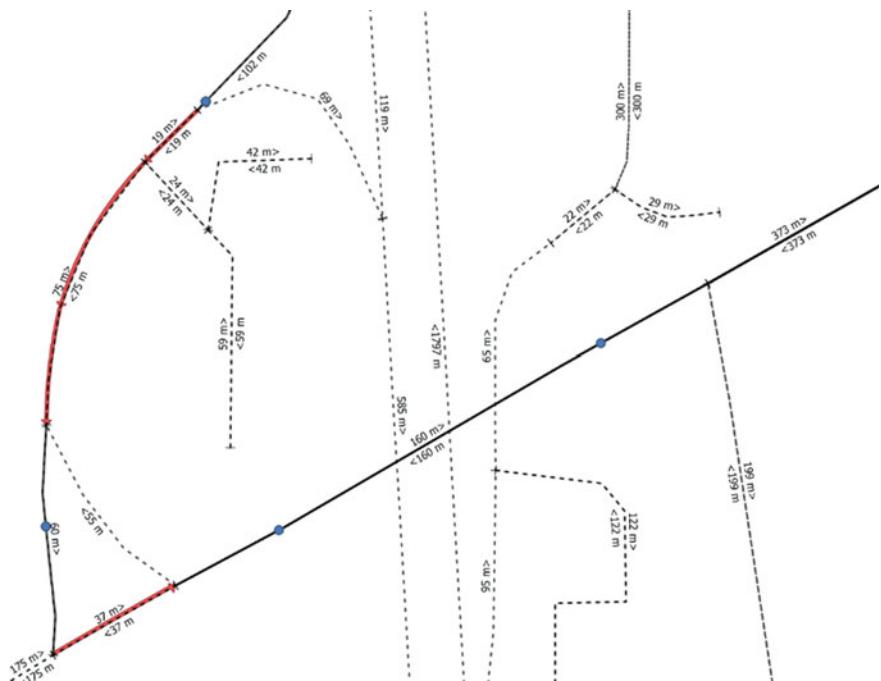
In case that no candidate segment satisfies the condition from Eq. 1, the most probable explanation is that there was an incorrect match in the previous matching step. In that case, the algorithm backtracks to the previous matched point and chooses the second closest one from the previous step. Then it again performs the routing part and again applies the condition from Eq. 1. If there is a fitting candidate, the algorithm replaces the previous matching with the updated one, and again chooses the closest segment from amongst the possible solutions for the current matching. If there is still no segment satisfying the condition from Eq. 1, we try the same procedure with the third closest in the previous matching etc. If there is no acceptable solution, we terminate the current session and start a new one.

In the last step of the algorithm, we create dummy FCD measurements along all segments that were in the chosen route and do not contain a start or an end. In mathematical terms, it means that

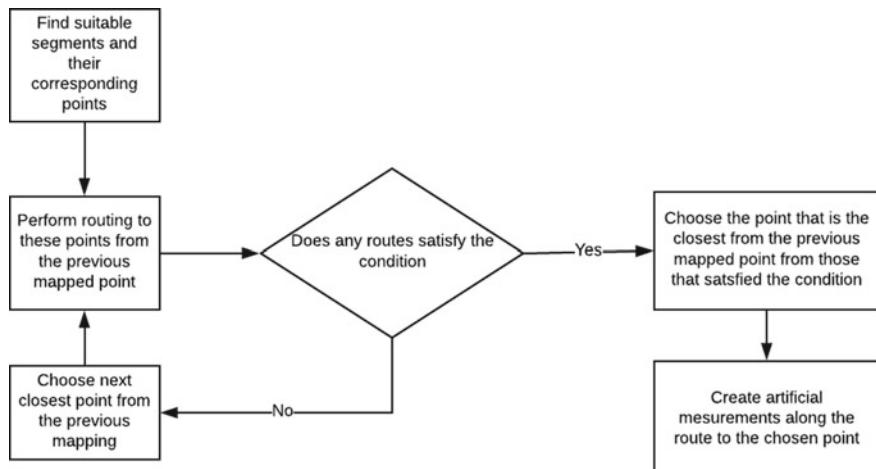
$$\forall \text{seg}_i \in S \wedge \text{seg}_i \neq \text{seg}_s \vee \text{seg}_t : v_{\text{seg}_i} = \frac{\text{length}(S)}{\Delta t},$$

where  $S$  is the entire route,  $s$  is its start point,  $t$  is its end point,  $\text{length}(S)$  is its length in metres,  $\Delta t$  is period between measurements (i.e. how long does the route take) and  $v_{\text{seg}_i}$  is approximated speed on the segment. Example of how this part works can be seen in Fig. 1 (red lines represent interpolated segments, blue dots represent known points and their position).

The progress of our entire algorithm can be summarized by the schema in Fig. 2.



**Fig. 1** Example of segment interpolation



**Fig. 2** Diagram of the proposed algorithm

## 4 Experiments

Experiments for verification of our map-matching algorithm were all performed on results from our recently developed traffic simulator (described in [27]) in the city of Brno (use case 1) and intercity simulation of traffic between Brno, Ostrava, Olomouc and Zlín (use case 2). The explained datasets containing FCD of both use cases are to be found publicly under DOI 10.5281/zenodo.2250119<sup>1</sup> and available for a download and further usage.

Hundreds of origin and destination points were placed in the city and a number of cars were generated in the simulator. These cars then reported their position with a predefined time gap just as would real floating cars do. For our experiment, we used 9000 cars and gaps of 5 s. These values then served as our baseline as they were one hundred percent correct. Resulting traffic flows can be seen in Figs. 3 and 4.

The thickness of the line represents the traffic load of the given road along with colours that represent a bottleneck levels (red is worst case).

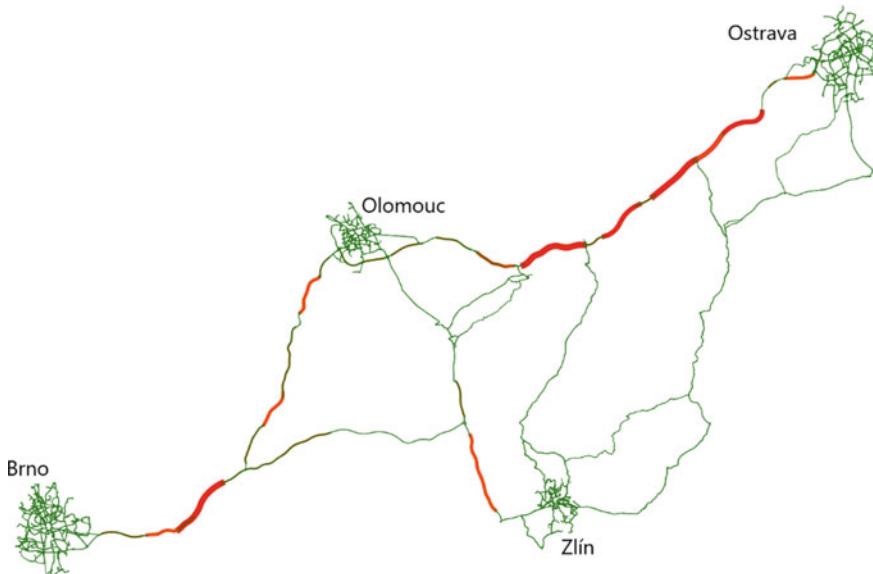
We post-processed raw measurements from the experiment in two ways. The first way was based on original data from our simulator. Because there is always a possibility that a simulated car drives through a (especially short) segment between chosen gaps of 5 s, we decided to fill missing segments with help of interpolation to reach high-quality model. We did this post-processing with both caching enabled and disabled. By caching is meant that we always store the last known segment of a car in a memory which allows us to only have one first history point per session.



**Fig. 3** Map of Brno city with traffic load from the simulator after map-matching

---

<sup>1</sup><https://zenodo.org/record/2250119>.



**Fig. 4** Map of inter-city with traffic load from the simulator after map-matching

When disabled though, we only memorize the previous point of a car for a period of given dataset time window (5 min in our case) and then we lose a segment history of a car and start over with first history point every 5 min. This has a significant impact even in case of high-quality dataset because map-matching the very first point with no history is the most challenging part. We decided to proceed to second post-processing with caching enabled only for better results.

Our first thought of the second post-processing way was to simply round the latitude and longitude of each point to the fourth decimal place. It introduced a random error from approximately 0 to 10 m and was not found sophisticated enough for our needs and thus we came up with more realistic one where each measured point was moved in a random direction a number of metres given by drawing a number from a Gaussian distribution. We utilized two Gaussian distributions, one for the roads outside the city ( $N(0, 10)$ ) and one for the roads inside the city ( $N(0, 15)$ ). Then some predefined number of randomly chosen points were removed (3% in our case). This approach should roughly represent real conditions encountered by FCD data as described by El Abbous and Samanta [28].

Two experiments were performed, all with 9000 cars. The first one utilized a obfuscation mentioned in paragraph above and the second one was similar to the first but with doubled standard deviation of both distributions. Results of these experiments are shown in Tables 1 and 3 for Use Case 1 (city) and Tables 2 and 4 for Use Case 2 (intercity). The difference between levels of obfuscation can be seen in Figs. 5 and 6.

**Table 1** Use case 1—Missing and Excessive map-matching compared to simulator

Dataset	FCD count	Segments missing		Segments interpolated ratio true positive/false negative	
		Unique	Total	Unique	Total
Simulator	4,968,119	0	0	0	0
High accuracy—no caching	5,205,980	6	39	1.15	0.57
High accuracy	5,214,369	7	463	3.53	1.67
Obfuscated	5,078,927	33	91	0.39	0.29
Obfuscated high	5,044,492	39	74	0.38	0.28

**Table 2** Use case 2—Missing and Excessive map-matching compared to simulator

Dataset	FCD count	Segments missing		Segments interpolated ratio true positive/false negative	
		Unique	Total	Unique	Total
Simulator	1,045,167	0	0	0	0
High accuracy—no caching	1,160,752	0	0	0.82	0.55
High accuracy	1,167,135	0	0	2.48	0.98
Obfuscated	994,232	4	8	0.36	0.02
Obfuscated high	1,058,563	7	14	0.36	0.05

**Table 3** Use case 1—Map-matching success rate compared to simulator

Dataset	Map-matching comparison (%)				
	Overall		Per segment occurrences		
	No interpolation	Interpolation	Top 100	Top 250	Top 500
Simulator	100	100	100	100	100
High accuracy—no caching	98.90**	98.79	98	97.6	96.6
High accuracy	99.64**	99.39	99	98.8	96.6
Obfuscated	66.37	85.12	87	88*	88.2
Obfuscated high	56.04	73.56	80	80	80.2

\* Indicates that the segment occurring in Top 100 has been found on 101st up 250th place

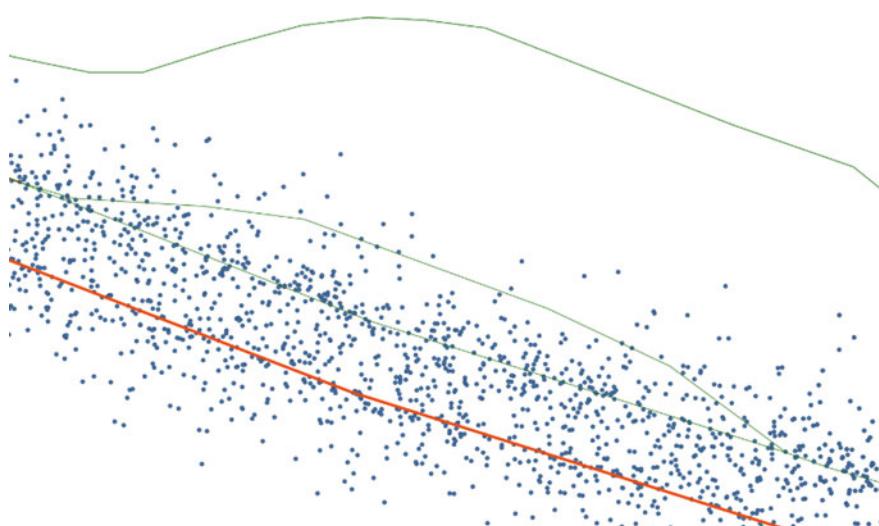
\*\* Dataset quality was too high for interpolation causing true positives/self-correction being not so helpful compared to more obfuscated datasets

**Table 4** Use case 2—Map-matching success rate compared to Simulator

Dataset	Map-matching comparison (%)				
	Overall		Per segment occurrences		
	No interpolation	Interpolation	Top 100	Top 250	Top 500
Simulator	100	100	100	100	100
High accuracy—no caching	99.38 <sup>b</sup>	99.39	99	100 <sup>a</sup>	98.4
High accuracy	99.87 <sup>b</sup>	99.80	100	100	99
Obfuscated	88.01	92.67	97	96.8	96.2
Obfuscated high	87.11	79.80	95	94	92.2

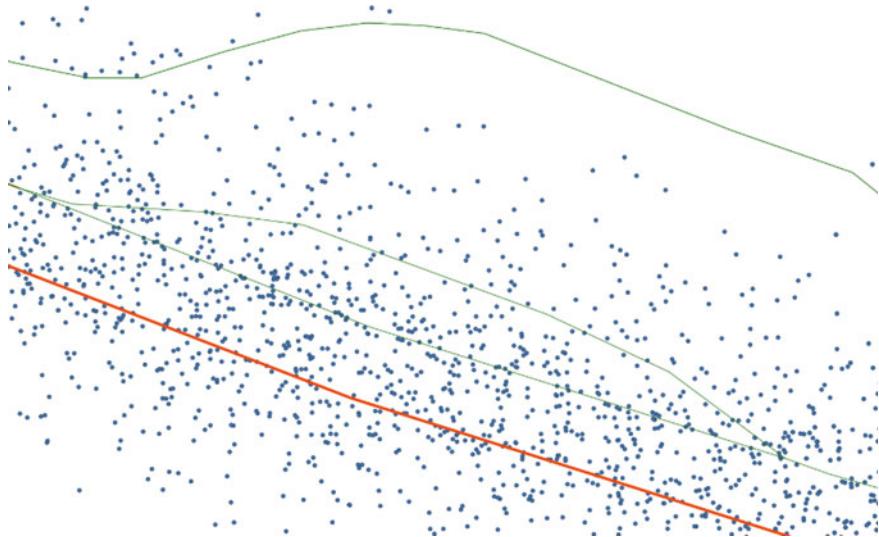
<sup>a</sup>Indicates that the segment occurring in Top 100 has been found on 101st up 250th place

<sup>b</sup>Dataset quality was too high for interpolation causing true positives/self-correction being not so helpful compared to more obfuscated datasets

**Fig. 5** Obfuscated quality FCD

In case of missing segments, we are interested in how many distinct (unique) segments were map-mismatched (point assigned to wrong segment or ruled out when found not reliable) and how many occurrences of a given missing segment we observe (total). In case of excessive segment (skipping segment due to time gasps mentioned above) caused either by correct or wrong interpolation we track ratio of those two (again distinct and in total).

The successful mapping percentage represents the proportion of correctly mapped segments. We measured overall mapping where we compared matching segments of every single session in time and observed mapping of top segment occurrences—bottleneck detection.



**Fig. 6** Obfuscated quality FCD—high

Even if interpolation was correct, it may have seemed to be mismatch because of a concrete segment missing in the original dataset. However, in this case we are able to fix a previous segment match if necessary, which improves success rate.

From the results, it is evident that our algorithm is performing well in case of individual map-matching and very well in case of global map-matching discovering traffic bottlenecks where occasional false positives can be neglected as they do not tend to be too much repetitive. False positives are i.e. incorrectly routed cars mostly in the case of the highly obfuscated experiment or the very first point of a session with no history to look back to. Under these circumstances, points can be assigned to wrong segment as seen in Figs. 5 and 6, where four different segments seemed to be reasonable, although in high accuracy model, it was only lower two of them because points were not so close to a gas station while having low speed. This area may require some improvement.

## 5 Conclusion

From the results of the experiments, it is evident, that our proposed map-matching algorithm is performing well. It has some shortcomings in terms of incorrect interpolations and first session point segment classification decisions, but these are not excessively high and have little-to-no effect on route network bottleneck detection. From the performance perspective, the current implementation of the algorithm can map several thousand cars in real time. While not exactly a low number, from the

applicational perspective, it must be able to process tens of thousands of cars. Therefore, the implementation also has some potential for improvement.

From the perspective of future work, we see two areas of potential additional research. The first one is thorough parallelization utilizing HPC infrastructure. This parallelization should handle mostly the routing part of the algorithm, and with it, we should be able to map an almost unlimited number of cars. The second area is the utilization of additional information from the floating cars such as acceleration, heading and other computed metrics available, to improve decision-making regarding choosing the most probable road. This could lead to a much smaller number of incorrectly interpolated roads.

**Acknowledgements** This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project “IT4Innovations excellence in science - LQ1602”. This work has been partially funded by ANTAREX, a project supported by the EU H2020 FET-HPC programme under grant 671623.

## References

1. Fujise, M., Kato, A., Sato, K., & Harada, H. (2002). Intelligent transport systems. Wireless communication technologies: New multimedia systems. In *The International Series in Engineering and Computer Science* (Vol. 564).
2. <https://xtralis.com/p.cfm?s=22&p=381>. 30.3.2018.
3. Pfoser, D. (2008). *Floating car data, Encyclopedia of GIS*. US: Springer.
4. Khan, R., Landfeldt, B., & Dhamdher, A. (2012). *Predicting travel times in dense and highly varying road traffic networks using starima models*. Technical report, School of Information Technologies, The University of Sydney and National ICT Australia.
5. Wu, Y., Chen, F., Lu, C., & Smith, B. (2011) Traffic flow prediction for urban network using spatiotemporal random effects model. In *91st Annual Meeting of the Transportation Research Board*.
6. Ma, Y., van Zuylen, H. J., & van Dalen, J. (2012). Freight origin-destination matrix estimation based on multiple data sources: Methodological study. In *TRB 2012 Annual Meeting*.
7. de Fabritiis, C., Ragona, R., & Valenti, G. (2008). Traffic estimation and prediction based on real time floating car data. In *Proceedings of 11th International IEEE Conference on Intelligent Transportation Systems*. ITSC 2008.
8. Kuhns, G., Ebendt, R., Wagner, P., Sohr, A., & Brockfeld, E. (2011). Self-evaluation of floating car data based on travel times from actual vehicle trajectories. In *IEEE Forum on Integrated and Sustainable Transportation Systems*.
9. Li, M., Zhang, Y., & Wang, W. (2009). Analysis of congestion points based on probe car data. In *Proceedings of International IEEE Conference on Intelligent Transportation Systems*, ITSC '09.
10. Graser, A., Dragaschnig, M., Ponweiser, W., Koller, H., Marcinek, M., & Widhalm, P. (2012) FCD in the real world—System capabilities and applications. In *Proceedings of 19th ITS World Congress* (p. 7), Vienna, Austria.
11. Hofmann-Wellenhof, B., Lichtenegger, H., & Wasle, E. (2008). *GNSS—Global navigation satellite systems*. Wien: Springer.
12. Knuth, D. E. (1977). A generalization of Dijkstra's algorithm. *Information Processing Letters*, 6(1).

13. Hashemi, M., & Karimi, H. A. (2014). A critical review of real-time map-matching algorithms: Current issues and future directions. *Computers, Environment and Urban Systems*, 48.
14. White, C. E., & Bernstein, D., Kornhauser, A. L. (2000). Some map matching algorithms for personal navigation assistants. *Transportation Research Part C: Emerging Technologies*, 8(1).
15. Taylor, G., Blewitt, G., Steup, D., & Corbett, S. (2001). Car road reduction filtering for GPS-GIS Navigation. *Transactions in GIS*, 5(3).
16. Srinivasan, D., Cheu, R. L., & Tan, C. W. (2003). Development of an improved ERP system using GPS and AI techniques. In *Proceedings of Intelligent Transportation Systems Conference* (Vol. 1).
17. Quddus, M. A., Ochieng, W. Y., Zhao, L., & Noland, R. B. (2003). A general map matching algorithm for transport telematics applications. *GPS Solutions*, 7(3).
18. Velaga, N. R., Quddus, M. A., & Bristow, A. L. (2009). Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems. *Transportation Research Part C: Emerging Technologies*, 17(6).
19. Li, L., Quddus, M., & Zhao, L. (2013). High accuracy tightly-coupled integrity monitoring algorithm for map-matching. *Transportation Research Part C*, 36.
20. Kim, S., & Kim, J.-H. (2001). Adaptive fuzzy-network-based C-measure map-matching algorithm for car navigation system. *IEEE Transactions on Industrial Electronics*, 48(2).
21. Quddus, M. A., Noland, R. B., & Ochieng, W. Y. (2006). A high accuracy fuzzy logic based map matching algorithm for road transport. *Journal of Intelligent Transportation Systems*, 10(3).
22. Yuan, L., Li, D., & Hu, S. (2018). A map-matching algorithm with low-frequency floating car data based on matching path. *EURASIP Journal on Wireless Communications and Networking*, 146(1).
23. Newson, P., & Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.
24. Ren, M., & Karimi, H. A. (2009). A hidden Markov model-based map-matching algorithm for wheelchair navigation. *Journal of Navigation*, 62(3).
25. Che, M., Wang, Y., Zhang, C., & Cao, X. (2018). An enhanced hidden Markov map matching model for floating car data. *Sensors*, 18(6).
26. Haklay, M., & Weber, P. (2008). OpenStreetMap: User-generated street maps. In *IEEE Pervasive Computing*, 7(4).
27. Ptošek, V., Ševčík, J., Martinovič, J., Sláničová, K., Rapant, L., & Čmar, R. (2018). Real time traffic simulator for self-adaptive navigation system validation. In *Proceedings of EMSS-HMS: Modeling & Simulation in Logistics, Traffic & Transportation*.
28. El Abbous, A., & Samanta, N. (2017). A modeling of GPS error distributions. In *Proceedings of 2017 European Navigation Conference (ENC)*.

# Calculating AQI Using Secondary Pollutants for Smart Air Management System



Gautam Jyoti, Malsa Nitima, Singhal Vikas and Malsa Komal

**Abstract** With the onslaught of the industrial revolution, the environment is suffering from severe pollution leading to major imbalances. Air Quality Dispersion Modelling can be done through one of the most efficient model “Eulerian Grid based model”. Various existing methods of prediction work on the basis of models resulting in satisfactory outcomes but with some certain loopholes. This project involves methods of predicting pollutants’ concentration and air quality using machine learning. The data of different sites are collected and the pollutants contributing maximum to the pollution is elucidated using machine learning based methods. Also in this project, a user-friendly, smart application system is developed which can be used to monitor the pollution produced at an individual level. The analysis of the feature stimulating the pollution level (to reach at a dangerous level) can be done with the help of machine learning tools. This paper involves calculating the amount of harmful pollutants released by any individual during their journey. Further solutions can be identified at government level to reduce these pollutants raising the pollution level.

**Keywords** Air pollutants · AQI · Eulerian grid based model · Machine learning · Smart air pollution system

---

G. Jyoti · M. Nitima (✉) · S. Vikas · M. Komal  
JSS Academy of Technical Education, Noida, India  
e-mail: [nitima.malsa@gmail.com](mailto:nitima.malsa@gmail.com)

G. Jyoti  
e-mail: [jyotig@jssaten.ac.in](mailto:jyotig@jssaten.ac.in)

S. Vikas  
e-mail: [vikassinghal75@gmail.com](mailto:vikassinghal75@gmail.com)

M. Komal  
e-mail: [erkomalmalsa@gmail.com](mailto:erkomalmalsa@gmail.com)

## 1 Introduction

Air Pollution is growing at a fast rate due to globally increasing industrial development. Industries have come out to be one of the major contributors for increasing the pollution levels and bringing about a drastic change in the pollution pattern. Transportation system even though after being efficient is continuing to cause pollution, which is accelerating day by day. The proliferating number of vehicles and population is leading the world to a more pollution intense zone, which is causing many types of health hazards as well as decreasing natural resources. Air pollution is one of the most dangerous forms of pollution, which needs to be handled immediately. Secondary pollutants due to their reactive nature are the leading patrons for all problems and need to be dealt with.

The paper outlines the major differences based on certain parameters between the three popular models: Gaussian model, Lagrangian model, and Eulerian model. The paper involves the study of the dataset of the concentrations of air pollutants at certain sites of Delhi, Gujarat, and Rajasthan, which is used to predict the maximum pollution causing pollutants and to determine the most dangerous factor stimulating the pollutants to cause more pollution.

Now to further analyze the most harmful pollutants, Air Quality Index is calculated. The air quality index (AQI) is an index used by government firms to aware the public the pollutant concentration in the air currently or to predict the future concentrations. As the AQI increases, a huge percentage of the population is likely to experience its adverse effects. We have demonstrated a procedure to determine the air quality in the methodology section.

Further an android application will be developed in which the users can register themselves by using any of their identity cards. The registered users can measure the amount of pollution released by their vehicle during any of their journey. It is also helpful in maintaining a record as to what amount of pollution has been released by the user on a daily or monthly basis. This application can be used at a higher level in future to control the pollution levels. The purpose of this paper is to perform comparative analysis between Gaussian, Eulerian, and Lagrangian model on the basis of air pollutants, to predict the factors stimulating the pollutants to cause more pollution and also to develop an application to monitor pollution at individual level.

## 2 Related Work

Our study mainly focuses on outdoor sensing. Many wearable sensors have been used to monitor air quality, for instance in CitiSense [1, pp. 23–25] and Common Sense [2, pp. 4–6]. Both these solutions rely on small, battery-powered sensor nodes that measure the concentrations of pollutants and send air quality data to user's smart phones through Bluetooth. Conversely, in [3, pp. 268–285] the authors propose a solution to derive high-resolution air pollution maps for urban areas, using nodes

provided with several sensors, such as UFP (ultrafine particles), CO, O<sub>3</sub>, and NO<sub>2</sub> sensors.

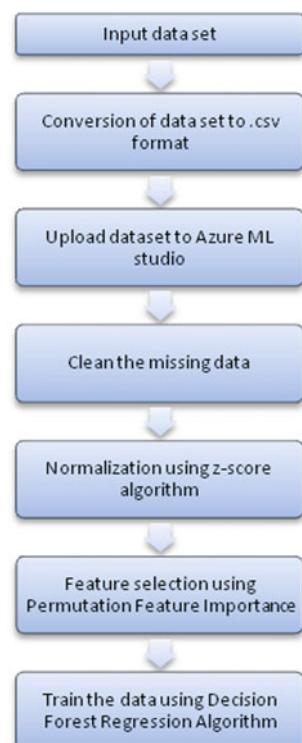
Different learning techniques can be applied to check the accuracy of prediction of pollutants. An application was proposed in Spain to forecast the air quality considering the traffic intensity and meteorological parameters including wind speed. A globally accepted model named THOR [4, pp. 117–122] includes models which can operate on different applications and scales. This model can be used to predict the weather and pollution levels up to three days. It helps in emission reduction and pollution management.

### 3 Methodology

To achieve our objectives we need to follow certain procedures as explained in the flowchart (Fig. 1):

- A dataset is prepared consisting of the concentration of pollutants in Azure Machine Learning understandable format.

**Fig. 1** Process chart for prediction of most stimulating feature



- Data Preprocessing techniques to make the dataset consistent need to be applied in order to make raw, noisy and unreliable data clean and normalize the results thus obtained.
- Now the data is normalized using Z-score algorithm with the following formula:

$$\text{Normalized}(xi) = \frac{x(i) - \bar{x}}{s(x)} \quad (1)$$

where  $x(i)$  is the mean and  $s(x)$  is the standard deviation.

- Further Permutation Feature Importance Algorithm is applied to determine the feature stimulating the pollutants to cause more pollution.
- The model is trained using the Decision Forest Regression algorithm. Decision trees are created which work as nonparametric models and perform a sequence of tests for every instance of the variable until the complete tree is traversed and a goal or decision is found (Fig. 2).

Further for the calculation of AQI, following methodology is adopted:

1. Building the Database: The concentration of pollutants from CPCB is collected.
2. Calculation of Air Quality or AQI: The procedure to calculate the air quality index varies from region to region and also from country to country. We have demonstrated a procedure to determine the air quality in the methodology section.
3. Determination of safety levels: Based on the AQI value calculated the range of safety levels for different human groups is determined.

The formula used for calculation of AQI is as follows:

$$I = \frac{I_{\text{high}} - I_{\text{low}}}{C_{\text{high}} - C_{\text{low}}} * (C - C_{\text{low}}) + I_{\text{low}} \quad (2)$$

$I$   
 $C$   
 $C_{\text{low}}$

the Air Quality Index  
the pollutant concentration  
the concentration breakpoint that is  $\leq C$

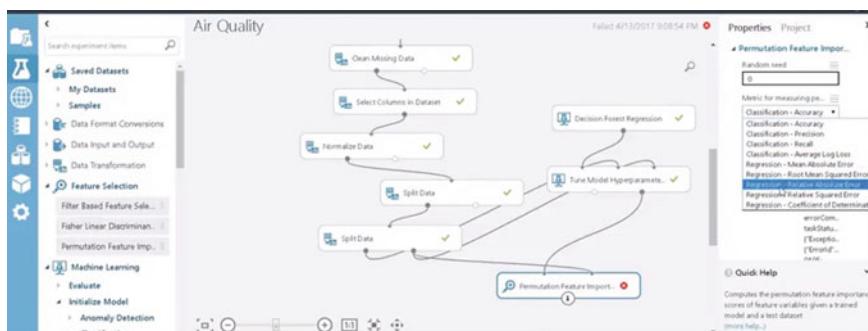


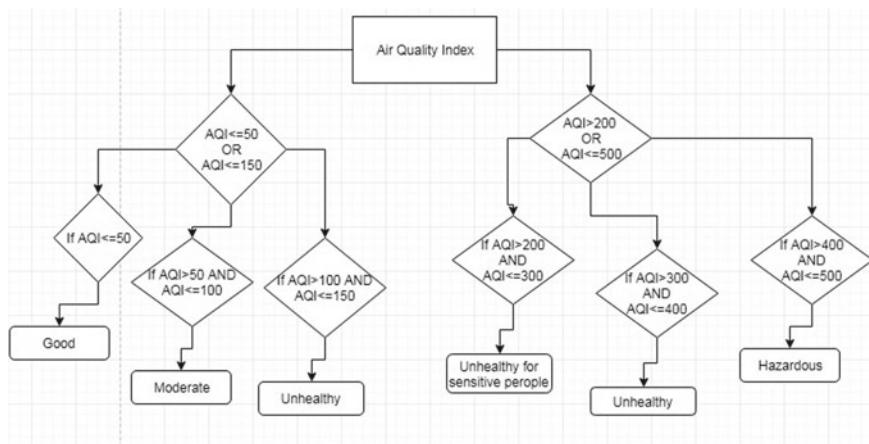
Fig. 2 Azure ML studio screenshot

$C_{high}$  the concentration breakpoint that is  $\geq C$   
 $I_{low}$  the index breakpoint corresponding to  $C_{low}$   
 $I_{high}$  the index breakpoint corresponding to  $C_{high}$   
 $C_{low}, C_{high}, I_{low}, I_{high}$  are from the US EPA Pollutant Breakpoint

The AQI determined is displayed to the users along with the various consequences of that AQI value. The flowchart elucidates the range of AQI values along with their effects (Fig. 3).

An android application is under development which will aware its users about the level of pollution in their area and accordingly control measures can be performed by the users. The main steps involved in the working of the application are:

1. Users access the application by registering himself.
2. There will be a start/stop button on the home screen. By clicking on the start button, the tracker will automatically start tracking the distance traveled by the vehicle as well as the sensors will start recording the concentration of air pollutants.
3. When the destination arrives the user will press stop button. As soon as the user presses the stop button based on the distance traveled, the respective average percentage of the pollutants emitted during the journey will be displayed on the screen.
4. Monthly, daily, or yearly receipt will be generated. The receipts will contain data regarding the pollutants released during the particular time span. Accordingly, the alert notifications will be displayed if the total pollutant released is higher than its critical limit.
5. If the user crosses the critical limit, heavy taxes and fines may be imposed by government. This will encourage the user to use public transport. Also the government can keep a check on every user.



**Fig. 3** Flow chart for AQI safety levels

Interface for distance calculator application (Figs. 4 and 5).

## 4 Results

The results of the project are as shown below:

1. Graphical depiction of concentrations of pollutants over the time span along with final predicted feature score (Figs. 6, 7, 8).
2. AQI Calculation (Figs. 9, 10, 11; Table 1).
3. Graphical Depiction of pollutants' contribution to AQI (Fig. 12):

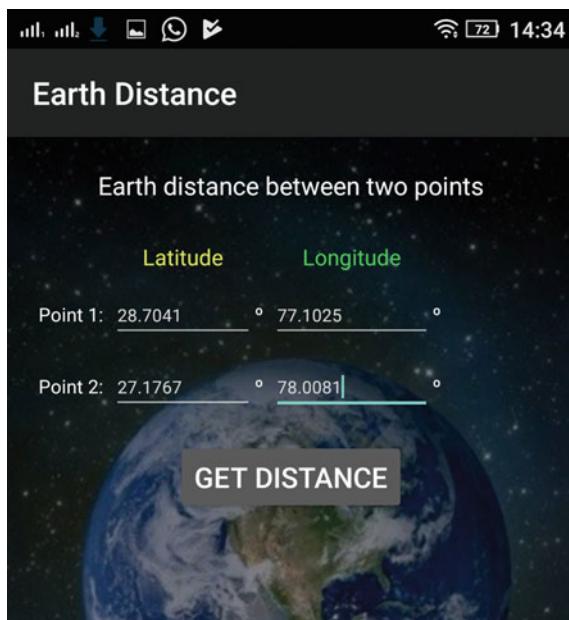
```
Location locationA = new Location("point A");
locationA.setLatitude(latA);
locationA.setLongitude(lngA);

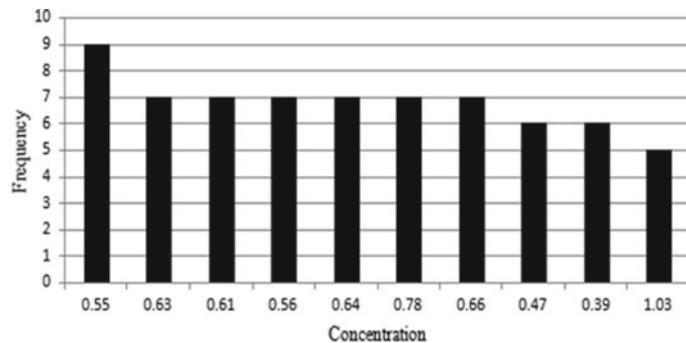
Location locationB = new Location("point B");
locationB.setLatitude(latB);
locationB.setLongitude(lngB);

float distance = locationA.distanceTo(locationB);
```

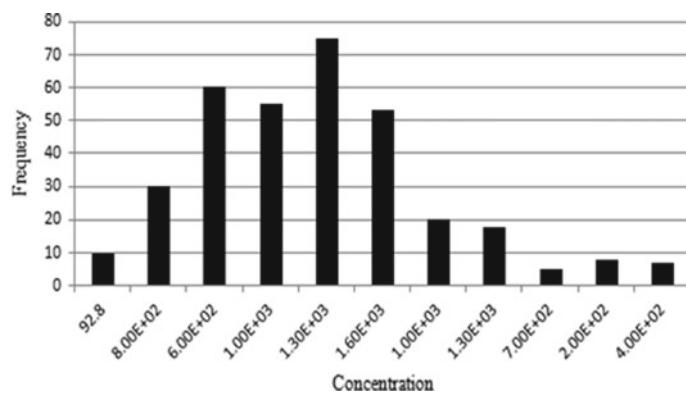
**Fig. 4** Method used to calculate distance between two points

**Fig. 5** Distance calculator application screenshot

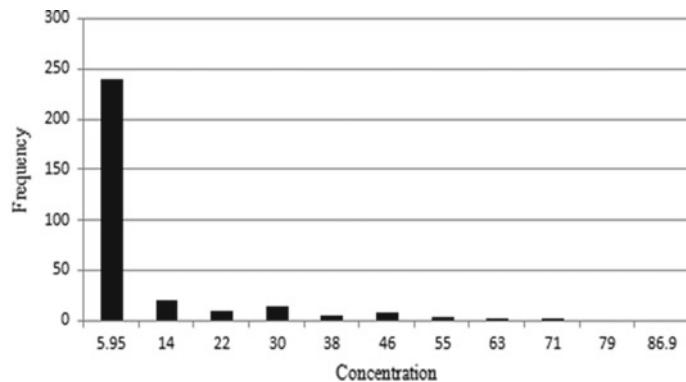




**Fig. 6** CO



**Fig. 7** PM2.5



**Fig. 8** O<sub>3</sub>

**PM 2.5 AQI Calculation**

Enter Concentration from 12 AM to 2 AM	1
Enter Concentration from 2 AM to 4 AM	3
Enter Concentration from 4 AM to 6 AM	11
Enter Concentration from 6 AM to 8 AM	14
Enter Concentration from 8 AM to 10 AM	9
Enter Concentration from 10 AM to 12 PM	21
Enter Concentration from 12 PM to 14 PM	12
Enter Concentration from 14 PM to 16 PM	5
Enter Concentration from 16 PM to 18 PM	7
Enter Concentration from 18 PM to 20 PM	5
Enter Concentration from 20 PM to 22 PM	3
Enter Concentration from 22 PM to 24 PM	23

AQI	AQI Category
4.2	18
Good	

Sensitive Groups	Health Effects Statements	Cautionary Statements
People with respiratory or heart disease, the elderly and children are the groups most at risk.	None	None

**Fig. 9** Concentration range for good AQI

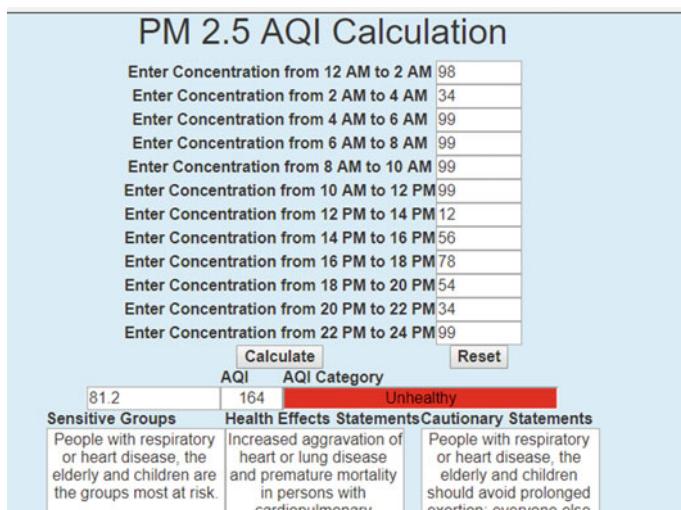
**PM 2.5 AQI Calculation**

Enter Concentration from 12 AM to 2 AM	23
Enter Concentration from 2 AM to 4 AM	34
Enter Concentration from 4 AM to 6 AM	45
Enter Concentration from 6 AM to 8 AM	43
Enter Concentration from 8 AM to 10 AM	43
Enter Concentration from 10 AM to 12 PM	23
Enter Concentration from 12 PM to 14 PM	12
Enter Concentration from 14 PM to 16 PM	56
Enter Concentration from 16 PM to 18 PM	76
Enter Concentration from 18 PM to 20 PM	54
Enter Concentration from 20 PM to 22 PM	34
Enter Concentration from 22 PM to 24 PM	23

AQI	AQI Category
30.5	90
Moderate	

Sensitive Groups	Health Effects Statements	Cautionary Statements
People with respiratory or heart disease, the elderly and children are the groups most at risk.	None	None

**Fig. 10** Concentration range for moderate AQI

**Fig. 11** Concentration range for dangerous AQI**Table 1** Concentration range according to AQI level

Pollutant	Concentration( $\mu\text{mg}^{-3}$ )	AQI level
PM2.5	1–25	Good
	25–77	Moderate
	78–100	Unhealthy
$\text{O}_3$	1–30	Good
	30–75	Moderate
	76–100	Unhealthy
CO	1–40	Good
	40–80	Moderate
	80–100	Unhealthy

Initial parts of our paper have already been achieved and are in the process of publishing under the ICETEAS-2018, JECRC. We have further implemented our work with more datasets. Earlier we had taken two year (2016–2018) data for prediction, now we have taken ten year (2008–2018) data. The results achieved with modified datasets differ slightly thus proving that relative humidity is the main factor to stimulate the production of pollutants.

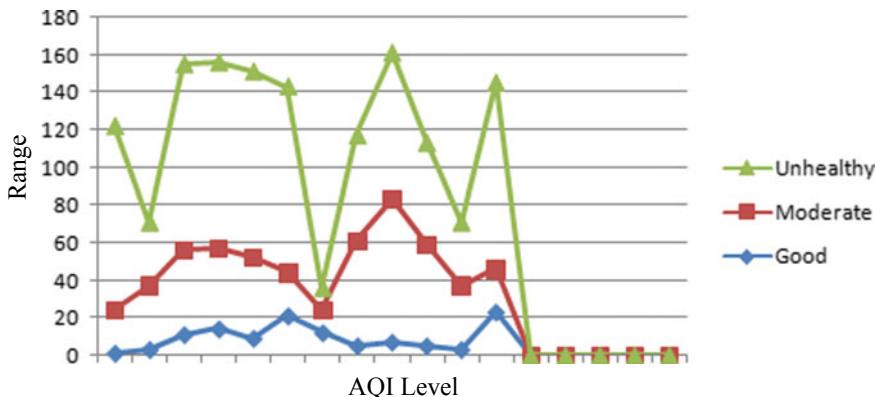


Fig. 12 Depiction of pollutants contribution to AQI

## References

1. Nikzad, N., Verma, N., Ziftci, C., Bales, E., Quick, N., Zappi, P., et al. (2012). CitiSense: Improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system. In *Proceedings of the ACM Conference on Wireless Health*, San Diego, CA, USA, October 23–25, 2012.
2. Dutta, P., Aoki, P. M., Kumar, N., Mainwaring, A., Myers, C., Willett, W., & Woodruff, A. (2009). Common sense: Participatory urban sensing using a network of handheld air quality monitors. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, Berkeley, CA, USA, November 4–6, 2009.
3. Hasenfratz, D., Saukh, O., Walser, C., Hueglin, C., Fierz, M., Arn, T., et al. (2015). Deriving high-resolution urban air pollution maps using mobile sensor nodes. *Pervasive Mobile Computer*, 16, 268–285.
4. Brandt, J., Christensen, J. H., Frohn, L. M., & Zlatev, Z. (2002). *Operational air pollution forecast modelling using the THOR system*. National Environmental Research Institute, Department of Atmospheric Environment.

# Enhanced Prediction of Heart Disease Using Particle Swarm Optimization and Rough Sets with Transductive Support Vector Machines Classifier



M. Thiagaraj and G. Suseendran

**Abstract** Over the last decade heart disease has significantly increased and it has emerged to be the primary reason behind the mortality in people living in many nations across the world. The computer-assisted systems act as a tool for the doctors in the prediction and diagnosis of heart disease. In the medical domain, Data Mining yields a variety of techniques that are extensively employed in the medical and clinical decision support systems that has to be quite useful in diagnosing and predicting the heart diseases with less time and good accuracy to improve their health. The previous system designed a radial basis function with support vector machine for heart disease prediction. However it does not provide a satisfactory classification result. To solve this problem the proposed system designed a Particle Swarm Optimization and Rough Sets with Transductive Support Vector Machines (PSO and RS with TSVM) based prediction is performed. In this proposed work, the dataset of the heart disease is collected from UCI repository. In order to reduce data redundancy and improve data integrity, the data normalization is performed by using Zero-Score (Z-Score). Then Particle Swarm Optimization (PSO) algorithm and Rough Sets (RS) based attribute reduction technique is used for selecting the optimal subset of attributes that, in turn, minimizes the computational hurdles and improves the performance of the prediction system. Finally, the Radial Basis Function-Transductive Support Vector Machines (RBF-TSVM) classifier is used for heart disease prediction. The results obtained from the experiments indicate that the system proposed accomplishes a superior performance in comparison.

**Keywords** Particle swarm optimization (PSO) · Rough sets (RS) · Radial basis Function-Transductive support vector machines (RBF-TSVM)

---

M. Thiagaraj (✉)

Department of Information and Technology, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India  
e-mail: [mthyagarajphd@gmail.com](mailto:mthyagarajphd@gmail.com)

G. Suseendran

Department of Information and Technology, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India  
e-mail: [suseendar\\_1234@yahoo.co.in](mailto:suseendar_1234@yahoo.co.in)

## 1 Introduction

“Data Mining refers to the sensitive extraction of inherent, earlier unknown and potentially helpful knowledge about data” [1]. Briefly said, it is defined as the process of evaluating the data from various points of view and collecting. Data mining offers a number of approaches for the discovery of hidden patterns out of the data. One important problem encountered in the Healthcare domain is associated with quality of service. Quality of service indicates the correct diagnosis of the disease and yielding treatments that are effective to patients. Incorrect diagnosis can result in dangerous results that cannot be accepted [2].

As per the survey of WHO, an overall of 17 million deaths globally are because of heart attacks and strokes. The deaths occurring because of heart disease in several nations are the result of work overload, emotional stress, and several other issues. Overall, it has been observed to be the major cause behind the death in adult people [3]. The diagnosis is a complex and significant task, which has to be executed with accuracy and efficiency. The diagnosis is frequently done, on the basis of the experience and knowledge of the doctor. This results in incorrect outcomes and expensive medical treatments given to patients [4].

Cardiovascular disease is a type of critical health-endangering and often occurring disease. The world health organization has made an estimation that 12 million deaths happen across the entire world, each year the reason being the cardiovascular disease. Progress made in the medical field over the last few decades facilitated the recognition of risk factors, which may have contributed to the cardiovascular diseases [5]. The most typical reason of heart disease includes the narrowing or blockage of the coronary arteries, the blood vessels, which are responsible for supplying blood to the heart itself. This is known as coronary artery disease and it occurs gradually as time passes. It’s the primary reason that people suffer from heart attacks. A blockage, which does not get treatment within a couple of hours makes the affected heart muscle to face death. Nearly 30% of all those who suffered from heart attacks experienced no symptoms. But, apparent symptoms of the attack stay in the blood-stream for many days. Medical diagnosis is a vital but sophisticated task, which has to be performed with accuracy and efficiency and its automation would prove to be very helpful. Unfortunately, all the doctors do not have equal skills in all the subspecialties and in several regions. With these many factors used for the analysis of the heart attacks’ diagnosis, physicians usually diagnose by analyzing the current test results of the patient [6]. The physicians also investigate the earlier diagnoses done on other patients with the same kind of results. These sophisticated procedures are nontrivial. Hence, a physician has to be experienced and hugely expertized for diagnosing the patients’ heart attacks. Subsequently, the undertakings made to utilize the learning and experience of different experts and the clinical screening information of patients aggregated in databases for enabling the analysis procedure is respected to be a useful framework, which is the mix of clinical choice help and PC based patient records and could limit the therapeutic mistakes, enhance the safety of patient, reduce unnecessary rehearse contrasts, and improve the patient outcomes.

The rest of the paper is organized as given below: Section 2 discusses the various methods that have been used for heart diseases prediction. Section 3 focuses on our proposed methodology of classification method, and PSO algorithm that is used for attribute reduction was performed. Section 4 provides the experimental results that were conducted, and finally Sect. 5 concludes the work.

## 2 Literature Review

Dangare et al. [7] assessed the forecast frameworks utilized for Heart ailment making use of a broad number of data attributes. The structure applies remedial terms like sex, circulatory strain, cholesterol-like 13 characteristics for foreseeing likelihood being influenced by illness. Till now, 13 properties are used for forecast purposes. This exploration work has included another two characteristics, which are corpulence and smoking. The information mining arrangement strategies.

Anooj et al. [8] portrayed a weighted fluffy control based clinical choice emotionally supportive network (CDSS) for diagnosing the coronary illness, via naturally procuring the learning from the clinical information of the patient. The recently presented clinical choice emotionally supportive network utilized for hazard expectation of heart patients involves two phases, (1) robotized plot for creating the weighted fluffy tenets, and (2) plan of a fluffy govern based choice emotionally supportive network. In the main stage, the mining approach, quality choice and trait weight age system are used forgetting the weighted fluffy principles. After this, the fluffy framework is planned by the weighted fluffy guidelines and chose characteristics. Finally, the investigation is performed on the recently presented framework utilizing framework as far as exactness, affectability, and specificity.

Tahseen et al. [9] utilized information mining approaches. It was assumed that the reliant variable was taken to be the conclusion—having dichotomous qualities showing the diseases existence or nonattendance. Twofold relapse has been connected to the components of the reliant variable. The informational collection has been procured from two diverse cardiovascular doctor's facilities in Karachi, Pakistan. An aggregate of sixteen factors out of which one was thought to be reliant and the rest of the 15 free factors are utilized. Information Reduction methodologies, for example, rule segment examination was utilized for having a superior execution of the relapse demonstrate in the forecast of Acute Coronary Syndrome. Only 14 out of sixteen variables have been viewed as in light of the results of information decrease.

Making use of neural network, Shantakumar et al. [10] presented a brilliant and effective heart assault expectation framework. For extricating the critical examples from the coronary illness databases for doing heart assault forecast, an efficient technique has been presented. At point when the preprocessing is done, the coronary illness vault was bunched with the assistance of the K-means grouping calculation that will play out the extraction of the information suitable for heart assault from the store. Hence, with the assistance of the MAFIA calculation, the incessant examples that are a match to coronary illness are mined from the information extricated. Heart

assault expectation proficiently, the neural system was prepared with the picked noteworthy examples. Utilizing the Back-spread as the preparation calculation, the Multi-layer Perceptron Neural Network has been utilized. Outcomes accordingly accomplished have demonstrated that the new expectation framework was proficient in the forecast of the heart assault with effectiveness.

Parthiban et al. [2] drawn closer a coactive neuro-fluffy surmising framework (CANFIS) for coronary illness forecast. The recently presented CANFIS demonstrate coordinated the neural system versatile potential and the fluffy rationale subjective strategy that, thus, is joined with hereditary calculation for diagnosing the presence of the ailment. The exhibitions of the CANFIS display were evaluated as far as the preparation exhibitions and order correctnesses and the outcomes uncovered that the CANFIS show proposed has amazing capability with respect to the coronary illness forecast.

Anbarasi et al. [11] portrayed about the forecast of the nearness of heart disease with more exactness using limited number of properties. As a matter of fact, thirteen qualities were utilized in the forecast of the coronary illness. This exploration work, Genetic calculation is used for deciding the traits that help more in the diagnosing the heart sicknesses that in a roundabout way restricts the quantity of tests required. Ascribes constrained characteristics with the assistance of hereditary hunt. Next, three classifiers, for example, Naive Bayes, Classification by bunching and Decision Tree are used for anticipating the finding of patients as precisely as it is accomplished before the minimization of the quantity of properties. What's more, the deductions demonstrate that the execution of Decision Tree information mining procedure is vastly improved than other two information mining approaches once the highlight subset choice is consolidated with significantly high model development time. The execution of Naïve Bayes is reliable earlier and after the decrease in the number of traits with a similar time taken for display development. The execution of Classification through bunching is poor in examination with other two methods.

Soni et al. [12] developed a GUI based Interface for entering the patient record and for anticipating if the patient is experiencing Heart contamination or not utilizing Weighted Association poisonous Classifier. The longing is done through the mining of certain information or information storeroom of the patient. It has just been built up that the execution of the Associative Classifiers is great in correlation with conventional classifiers methods including choice tree and manage acceptance.

### 3 Proposed Methodology

In PC helped coronary illness finding systems, the information is obtained from a couple of different sources and is analyzed using PC based applications. PCs have for the most part been used for building learning based clinical choice emotionally supportive networks that made utilization of the data acquired from therapeutic expert, and the exchange of this information into PC calculations was performed with manual intercession. This procedure devours much time and is basically dependent on

the supposition of the restorative master that might be abstract. In order to deal with this issue, machine learning approaches have been designed to obtained knowledge by automatic means from examples or unprocessed data.

The proposed system designed a Particle Swarm optimization algorithm with Rough set and Radial Basis Function based Transductive Support Vector Machines (TSVM) for heart diseases prediction. The elaborate design of heart disease diagnosis system comprises of three important stages: Normalization, attribute reduction, feature extraction, and classification (Fig. 1).

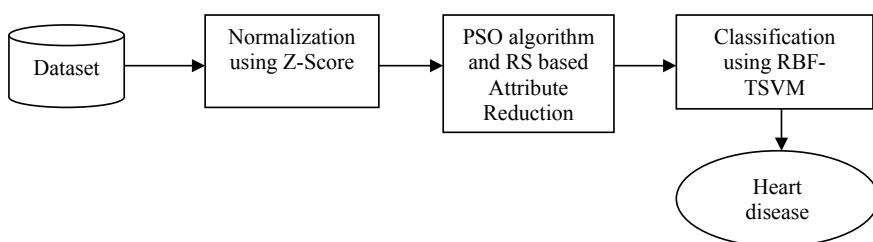
### 3.1 Normalization Using Z-Score Normalization

All the input and output data underwent normalization prior training and the testing processes so as to guarantee that there is no overwhelming of data in terms of distance metric.

In Z-Score standardization, ordinarily it is helpful. For the standardization of an arrangement of scores making utilization of the standard deviation, each score is isolated by the standard deviation of this arrangement of scores. Appearance of this, the vast majority of the occasions, the mean of the scores is subtracted from each score before isolating by the standard deviation. This standardization is called as Z-scores [13]. Mathematically, each set of  $N$  scores is represented by  $Y_n$  and whose mean is equivalent to  $M$  and whose standard deviation is equivalent to  $\hat{S}$  gets transformed in Z-scores to be

$$Z_n = \frac{Y_n - M}{\hat{S}} \quad (1)$$

Using basic algebraic expressions, it can be proven that a set of Z-score has a mean equivalent to zero and a standard deviation of one. Hence, Z-scores have a unit free measure that can be utilized for comparing the observations that are measured with various units. Once the normalization is done, the transformed datasets are applied for attribute reduction technique.



**Fig. 1** Block diagram of the proposed methodology

### 3.2 Particle Swarm Optimization (PSO)

Fundamentally a streamlining calculation in the swarm insight field. Accept in D measurements seek space, n particles are introduced in irregular [14, 15]. The PSO considers the present data of the populace to be its examination arrangement to be exploration field, a few times of data cycles and exchange, it will get the current worldwide ideal arrangement. Each molecule at that point refreshes its speed and position amid each cycle in view of Eqs. (2) and (3)

$$v_{id}^{t+1} = v_{id}^t + c_1 R_1 * (p_{id}^t - X_{id}^t) + c_2 R_2 * (p_d^t - X_{id}^t) \quad (2)$$

$$X_{id}^{t+1} = X_{id}^t + v_{id}^{t+1} \quad (3)$$

In above conditions, the nearby and worldwide optima. R1 and R2 stand for an irregular number somewhere in the range of 0 and 1. C1 and C2 specify a positive consistent that is known as quickened factor utilized for altering the progression length in the nearby and worldwide ideal heading and generally its esteem is somewhere in the range of 0 and 4.

$$v_{ij}(t+1) = \omega * v_{ij}(t) + c_1 R_1 * (p_{best}(t) - X_{ij}(t)) + c_2 R_2 * (p_{best}(t) - X_{ij}(t)) \quad (4)$$

$$X_{ij}(t) = \begin{cases} 1, & \rho < s(v_{ij}(t)), \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

The calculation will perform better when  $\omega$  that is known as weight factor esteems goes somewhere in the range of 0.4 and 0.9 [16].  $\rho$  refers to an irregular number somewhere in the range of 0 and 1.  $s(v_{ij}(t))$  refers to a fluffy capacity by and large utilized in neural system, and its capacity recipe is as communicated as underneath:

$$s(v_{ij}(t)) = \frac{1}{1 + e^{-v_{ij}(t)}} \quad (6)$$

$p_{best}$  and  $g_{best}$  refers to the individual extremum and global optimal solution correspondingly that are represented by Eqs. (7) and (8):

$$p_{best} = \max(p_{best}, \text{fitness}(i)) \quad (7)$$

$$g_{best} = \max(p_{best}, g_{best}) \quad (8)$$

Well-being goes about as the sole marker for reflecting and managing the swarm molecule to keep moving toward the ideal arrangement. Since the decrease calculation is subject to positive area decrease calculation [17], it modifies the wellness work for accomplishing a base decrease with various outcomes much

advantageously Eq. (9).

$$\text{fitness}(i) = \begin{cases} |i|, & \text{if } (\text{pos}'_{|i|}(D)) = U'_{\text{POS}} \\ |c|, & \text{if } (\text{pos}'_{|i|}(D)) \neq U'_{\text{POS}} \end{cases} \quad (9)$$

### 3.2.1 Rough Set Algorithm for Attribute Reduction Through Particle Swarm Optimization

Twofold division encoding mapping technique is utilized to unite the obnoxious set and the molecule swarm figuring in this examination work. Qualities estimation of the molecules are mapped specifically and the measurement is set “0” or “1” where “0” shows its particular quality ought to be decreased in a definitive outcome when “1” won’t. From there on, it considers the BPSO calculation to be the heuristic data and tosses it into the positive area quality decrease calculation. In this way, the calculation gives arrangement to the issue all the more effortlessly and productively, and the outcomes incorporate different learning decrease, and the calculation will step if the most extreme number of cycle is reached [18].

## 3.3 Classification Using Radial Basis Function-Transductive Support Vector Machines (RBF-TSVM)

Classification of RBF based TSVM support to use for heart diseases prediction. In previous work briefly discussed about the RBF method.

As TSVM algorithm makes utilization of the idea of transductive learning with effectiveness, it can combine the showed dispersion data having a place with unlabeled examples and preparing tests much better. Subsequently, in correlation with the traditional help vector machine calculation, TSVM algorithm offers more noteworthy arrangement precision. Be that as it may, there are a couple of disadvantages with TSVM calculation, similar to the number  $N$  of positive name tests present in the unlabeled examples for TSVM algorithm must be physically indicated, however  $N$  esteem is for the most part hard to make an important gauge [19].

TSVM algorithm makes utilization of a basic technique for assessing the estimation of  $N$  that shows the estimation of the proportion of tests with positive names to all the unlabeled examples in view of the proportion of tests with positive marks and all the named tests. Be that as it may, if the quantity of tests with marks is less, it turns out to be difficult for the method to get a more precise estimation of  $N$  evaluated. At the point when the pre-characterized estimation of  $N$  is very dissimilar from the first number of tests having positive names, the execution of the TSVM calculation will have a tendency to end up exceptionally temperamental, and also, the order exactness of the calculation can’t be guaranteed to be successful [20].

Before long, the major immense work of transductive end in the zone of assistance vector purchasing consolidates Transductive Support Vector Machine (TSVM) that will be in no time clarified in whatever is left of this division. Given a course of action of self-ruling, correspondingly scattered named points of reference.

$$(x_1, y_1), \dots, (x_n, y_n), x_i \in R^m, y_i \in \{-1, +1\} \quad (10)$$

and another set of unlabeled examples from the identical sharing,

$$x_1^*, x_2^*, x_3^*, \dots, x_k^*$$

:

$$(y_1^*, \dots, y_k^*, w, b, \xi_1, \dots, \xi_n, \xi_1^*, \dots, \xi_k^*)$$

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i + C^* \sum_{j=1}^n \xi_j^* \quad (11)$$

Subject to:

$$\begin{aligned} \forall_{i=1}^n : y_i t[w \cdot x_i + b] &\geq 1 - \xi_i \\ \forall_{j=1}^k : y_j^* [w \cdot x_j^* + b] &\geq 1 - \xi_j^* \\ \forall_{i=1}^n : \xi_i &\geq 0 \\ \forall_{j=1}^k : \xi_j^* &\geq 0 \end{aligned}$$

The Radial premise work (RBF) based piece otherwise called the Gaussian part does the mapping of the lower dimensional component space onto an unending dimensional space. At the point when directly indistinguishable highlights are mapped onto higher dimensional space they oftentimes turn out to be straight differentiable.

The RBF based TSVM serves the purpose of classification is used to find the heart diseases prediction.

## 4 Experimental Results

Efficient Heart Disease Prediction is the most significant method to detect heart diseases. Here proposed PSO and RBF-TSVM approach and existing system IT2FLS [21] and modified FA and RBF-SVM are compared in terms of Sensitivity, Specificity, and Accuracy.

### False Positive Rate (FPR)

Defined as the percentage of cases in which an image was segmented to tumor portion, but actually did not.

$$FPR = \frac{FP}{FP + TN} \quad (12)$$

### False Negative Rate (FNR)

Defined as the percentage of cases in which an image was segmented to non-tumor portion, but actually it did.

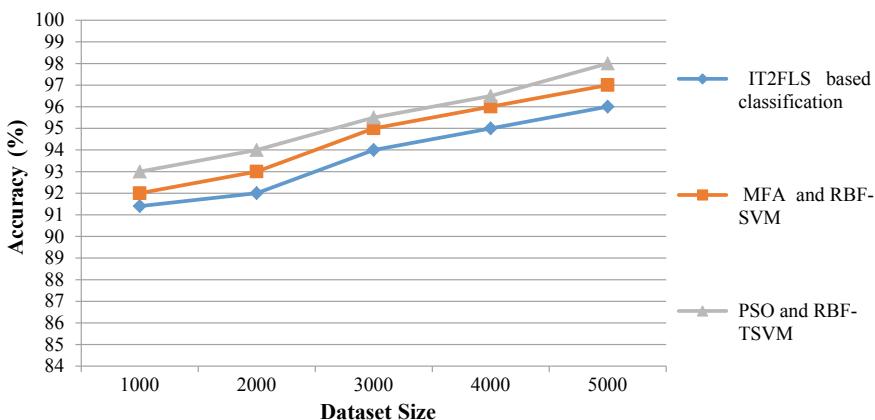
#### 1. Accuracy

The weighted percentage of tumor parts in images that is segmented correctly is measured by the metric accuracy. It is expressed as,

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (13)$$

Figure 2 illustrates that the proposed PSO and RBF-TSVM based classification approach is compared with the existing IT2FLS and MFA and RBF-SVM based classification approach in terms of accuracy [22]. The size of dataset is taken as *X*-axis and accuracy is taken in *Y*-axis. In order to achieve high accuracy the proposed system used PSO is used for attribute reduction. It concludes that the PSO and based RBF-TSVM classification approach has show the high accuracy results for all size of dataset compared with the existing method.

#### 2. Sensitivity



**Fig. 2** Accuracy comparison

The sensitivity measure is defined as the ratio of actual positives that are properly identified. It is associated with the capability of test to identify positive results.

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \times 100 \quad (14)$$

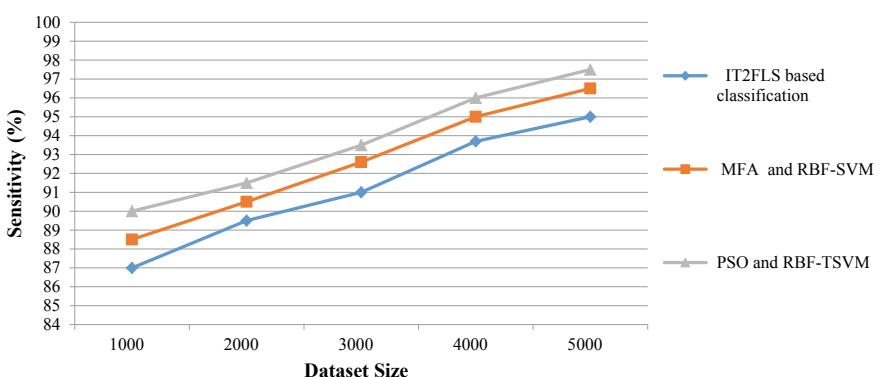
Figure 3 illustrates that the proposed PSO and RBF-TSVM based classification approach is compared with the existing IT2FLS and MFA and RBF-SVM based classification approach in terms of accuracy. The size of dataset is taken as X-axis and sensitivity is taken in Y-axis. Z-Score algorithm is used for normalization purpose. And also an efficient classification is done by using RBF-TSVM. It improves the true positive rate. For all size of dataset the proposed PSO and based RBF-TSVM classification approach has shown the high sensitivity results compared with the existing system.

### 3. Specificity

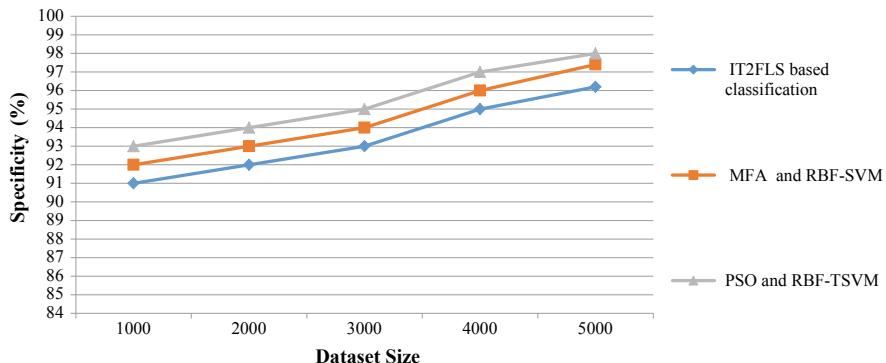
The specificity measure is defined as the ratio of negatives that are properly identified. It is associated with the capability of test to identify negative results.

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{number of false positives}} \times 100 \quad (15)$$

Figure 4 illustrates that the proposed PSO and RBF-TSVM based classification approach is compared with the existing IT2FLS and MFA and RBF-SVM based classification approach in terms of accuracy [23]. The size of dataset is taken as X-axis and specificity is taken in Y-axis. For all size of dataset the proposed PSO and based RBF-TSVM classification approach has shown the higher specificity results compared with the existing system.



**Fig. 3** Sensitivity comparison



**Fig. 4** Specificity comparison

## 5 Conclusion

In this research work, the proposed system designed a PSO algorithm and RBF based TSVM approach for predicting the heart disease with intelligence and efficiency, and to get over manual labor. After the normalization process, in turn, minimizes the redundancy and improves the performance of classifier and attribute features are also extracted through. Finally the classification is performed by using RBF-TSVM, it predicts the heart diseases. The truth is that humans cannot be replaced by computers and through the comparison of the computer-assisted detection results obtained from the pathological observations, doctors can be informed about the best means of evaluating the areas, which is focused by computer-assisted detection. The results of experiments indicate that the newly introduced system accomplishes a superior execution as far as precision, affectability, and specificity.

## References

1. Senthil, D., Suseendran, G. (2017). Data mining techniques using time series analysis. In *Proceedings of the 11th INDIACom; INDIACom-2017; IEEE Conference ID: 40353 2017 4th International Conference on "Computing for Sustainable Global Development* (pp. 2864–2872), March 01st–03rd, 2017 BharatiVidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA) ISSN 0973-7529; ISBN 978-93-80544-24-3.
2. Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. *International Journal of Biological, Biomedical and Medical Sciences*, 3(3).
3. Thiagaraj, M., Suseendran, G. (2017). Review of chronic kidney disease based on data mining. In *Proceedings of the 11th INDIACom; INDIACom-2017; IEEE Conference ID: 40353 2017 4th International Conference on Computing for Sustainable Global Development* (pp. 2873–2878), March 01st–03rd, 2017 BharatiVidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA) ISSN 0973-7529; ISBN 978-93-80544-24-3.

4. Thiagaraj, M., Suseendran, G. (2017). Survey on heart disease prediction system based on data mining techniques. *Indian Journal of Innovations and Developments*, 6(1), 1–9.
5. Senthil, D., Suseendran, G. (2018). Efficient time series data classification using sliding window technique based improved association rule mining with enhanced support vector machine. *International Journal of Engineering & Technology*, 7(2.33), 218–223. <https://doi.org/10.14419/ijet.v7i2.33.13890>.
6. Rohini, K., Suseendran, G. (2016). Aggregated K means clustering and decision tree algorithm for spirometry data. *Indian Journal of Science and Technology*, 9(44).
7. Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, 47(10), 44–48.
8. Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using decision tree fuzzy rules. *Int J Res Rev ComputSci*, 3(3), 1659–1667.
9. Jilani, T. A., Yasin, H., Yasin, M., Arدل, C. (2009). Acute coronary syndrome prediction using data mining techniques—An application. *World Academy of Science, Engineering and Technology*, 59.
10. Patil, S. B., Kumaraswamy, Y. S. (2009). Intelligent and effective heart attack prediction system using data mining and artificial neural network. *European Journal of Scientific Research*, 31(4), 642–656.
11. Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, 2(10), 5370–5376.
12. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Intelligent and effective heart disease prediction system using weighted associative classifiers. *International Journal on Computer Science and Engineering*, 3(6), 2385–2392.
13. Al Shalabi, L., Shaaban, Z., & Kasabeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, 2(9), 735–739.
14. Kennedy, & Eberhart, R. (1995). Particle swarm optimization. In *IEEE International Conference on Neural Networks* (pp. 129–132).
15. Abraham, A., Guo, H., & Liu, H. (2006). Swarm intelligence: Foundations, perspectives and applications. In *Swarm intelligent systems* (pp. 3–25). Heidelberg: Springer.
16. Liu, H., Abraham, A., & Clerc, M. (2007). Chaotic dynamic characteristics in swarm intelligence. *Applied Soft Computing*, 7(3), 1019–1026.
17. Guan, W., & Bell, D. A. (1998). Rough computational methods for information. *Artificial Intelligence*, 105(98), 77–103.
18. Thyagaraj, M., & Suseendran, G. (2018). An efficient heart disease prediction system using modified firefly algorithm based radial basis function with support vector machine. *International Journal of Engineering & Technology*, 7(2.33), 1040–1045.
19. Bruzzone, L., Chi, M., & Marconcini, M. (2006). A novel transductive SVM for semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11), 3363–3373.
20. <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>.
21. Long, N. C., Meesad, P., & Unger, H. (2015). A highly accurate firefly based algorithm for heart disease prediction. *Expert Systems with Applications*, 42(21), 8221–8231.
22. Zheng, X., Zeng, B., & Liu, S. (2008, December). Rough set based attribute reduction and extension data mining. In *Second International Symposium on Intelligent Information Technology Application* (Vol. 2, pp. 109–112), IITA'08. IEEE.
23. Kennedy, J., & Eberhart, R. C. (1997). A discrete binary version of the particle swarm algorithm. In *1997 IEEE International Conference on Systems, Man, and Cybernetics* (Vol. 5, pp. 4104–4108). Computational Cybernetics and Simulation.

# DataCan: Robust Approach for Genome Cancer Data Analysis



Varun Goel, Vishal Jangir and Venkatesh Gauri Shankar

**Abstract** While we glance in the past twenty years, it can be evidently noticed that biological sciences have brought about an active analytical research in high-dimensional data. Recently, many new approaches in Data Science and Machine Learning fields have emerged to handle the ultrahigh-dimensional genome data. Several cancer data types together with the availability of pertinent studies going on similar types of cancers adds to the complexity of the data. It is of commentator biological and clinical interest to understand what subtypes a cancer has, how a patient's genomic profiles and survival rates vary among subtypes, whether a survival of a patient can be predicted from his or her genomic profile, and the correlation between different genomic profiles. It is of utmost importance to identify types of cancer mutations as they play a very significant role in divulging useful observations into disease pathogenesis and advancing therapy varying from person to person. In this paper we focus on finding the cancer-causing genes and their specific mutations and classifying the genes on the 9 classes of cancer. This will help in predicting which genetic mutation causes which type of cancer. We have used Sci-kit Learn and NLTK for this project to analyze what each class means by classifying all genetic mutations into 17 major mutation types (according to dataset). Dataset is in two formats: CSV and Text, where csv containing the genes and their mutations and text file containing the description of these mutations. Our approach merged the two datasets and used Random Forest, with GridSearchCv and ten-fold Cross-Validation, to perform a supervised classification analysis and has provided with an accuracy score of 68.36%. This is not much accurate as the genes & their variations don't follow the HGVS Nomenclature of genes because of which conversion of text to numerical format resulted in loss of some important features. Our findings suggest that classes 1, 4 and 7 contribute the most for causing cancer.

**Keywords** Cancer · Genes · Random-Forest · SVM · Variations · Data analytics

---

V. Goel (✉) · V. Jangir · V. G. Shankar

Department of Information Technology, Manipal University Jaipur, Jaipur, India  
e-mail: [varungoel122@gmail.com](mailto:varungoel122@gmail.com)

## 1 Introduction

Over the last few decades, cancer-related research has evolved continuously. Screening in early stage was one of the most applied methods which Scientists used, such as, in order to find types of cancer before symptoms occur. Earlier cancer prediction has always been clinical based and morphological [1]. However, accurately predicting the presence of disease is one of the most interesting and challenging tasks for any physician. Data Extraction is an essential step in the process of finding knowledge in databases on which intelligent methods can be applied to extract patterns [2].

The gene expression data has several differential factors as compared to any other data as: (1) Gene expression data is very high dimensional containing thousands of genes. (2) The publicly available data contains a lot of noisy data. (3) Most of the genes in human body are irrelevant to cancer distinction [1]. A cancer tumor can have thousands of genetic mutations after sequencing. But the main problem lies in distinguishing the mutations contributing to growth of tumor (called drives) from the neutral mutations (called passengers). It has been impossible to determine meaningful cancer subtypes without manipulating the innate statistical and numerical features within these distinct input modalities [3]. On the other hand, “cross-platform genomic data” for the same tumor sample is improbable to be independent [3]. Bhola [1], has proposed an extensive overview of several cancer classification methods and has also evaluated these proposed methods based on their classification model accuracy, computational time and capability to reveal cancer gene information. In this paper we have proposed Random Forest with 10-fold Cross-validation and SVM with Feature Extraction and Count Vectorizer for our dataset [4–6].

The rest of the paper is structured as follows. In Sect. 2, we describe the related work proposed by other authors working or doing research in this field. Section 3.1 & 3.2, gives an overview of the data we acquired, it's preprocessing and the algorithms we propose for our problem. Section 4 describes the tools and various libraries used to successfully implement our work. Section 5, describes the dataset. Our dataset contains 9 classes of cancer, present in numerical form. Section 6, describes all the classes and the analysis done to derive their meanings. In Sect. 7, we describe our approach to the problem and the proposed algorithms and comparative analysis of our approach and other existing methods for integrative cancer data analysis.

## 2 Related Work

Researchers have devoted immense time in studying Classification problems in the area of databases, statistics, and machine learning. For the past few years, researchers have been intensively working on cancer classification using gene expression which suggests relation of the gene expression changes to different types of cancers. Diagnostics on molecular level with “microarray gene expression” [1] profiles are capable of suggesting the objective, efficient and accurate cancer classification methodology.

**Table 1** Relevant works on cancer classification of gene expression data for single type of cancer

Author(s)	Dataset	Methods	
		Feature	Classifier
Dev et al. [10]	Breast	Signature composition	BPN FLANN PSO-FLANN
Castano et al. [11]	Breast	BARS	EGRBF LR
	Gum		
	Colon		
	Leukemia		
	Lung		
	CNS		
Sharma and Paliwal [12]	Leukemia	Proposed algorithm	Bayesian classification
	Lung		
	Breast		

Some of the best suggested methods for cancer classification are Naive Bayes, k-Nearest Neighbour, Support Vector Machine, Random Forest, Bagging, AdaBoost and Linear Discriminant Analysis. The problem with Naïve Bayes [1, 2] is that it considers each and every attribute of data as independent and strong of each other. Mishra [7] has used k-NN [1] and LDA to assess the biased relevance of DE proteins with pan-cancer DE proteins to determine if analysis of pan-cancer yielded any relevant proteins for a specific cancer type. The LDA models generally performed better than the KNN models, which may be a result of sensitivity nature of the KNN algorithm to noise. SVM [1, 2, 8] in our study is used for feature selection [8] and ranking of text-based data and also for classification [1, 2]. SVM was chosen because without reducing space dimensionality and using regularization, it avoids overfitting of the data to some extent. Still, it didn't produce good results as a classifier. Random Forest Classifier [1] used in our study, provides the best accuracy as it uses random split selection which is faster than bagging [1] or boosting [9]. Its accuracy is better than Adaboost [1] and is relatively robust to outliers and noise in the data.

Table 1 shows relevant works on cancer gene expression classification for a single type of cancer. Dev et al. [10] have focused on three contrasting classification techniques: FLANN, PSOFLANN, and BPN and found that the integrated approach of Functional Link Artificial Neural Network (FLANN) and Particle Swarm Optimization (PSO) could be used to predict the disease as compared to other method.

Sharma [12] has proposed an algorithm for analysis of gene expression data. Initially the algorithm divides genes into subsets of comparatively smaller size, then selects informative smaller subsets and combines them. Process was repeated until all subsets were merged to create one informative subset.

Castano [11] has proposed classification technique for microarray gene expression, produced of genomic material by the analysis of light reflection. This proposed

algorithm was in two subsequent stages. In the first stage, salient expression genes are identified using two filter algorithms from thousands of genes. During the second stage, the proposed methodology was performed by the new input variables are selected from gene subsets. The methodology was composed of a union of (EGRBF) Evolutionary Generalized Radial Basis Function and (LR) Logistic Regression neural networks. The modeling of high-dimensional patterns has shown to be highly accurate in earlier research.

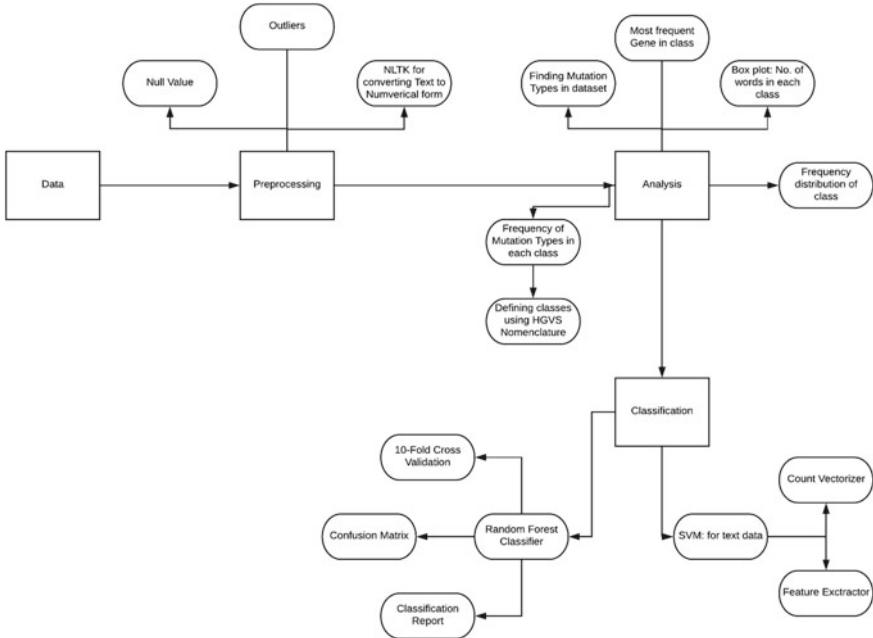
Some previous researches also suggest clustering as a solution. Rajput [13] proposed an algorithm in which most relevant dimensions from the high-dimensional data is selected using the minimum value of Median Absolute Deviation (MAD) and then select the efficient initial cluster centers using mode value of sorted and reduced data set. Finally, K-means algorithm is used to perform the clustering. Jiang et al. [14] proposed (1) clustering based on gene, where genes are data object and samples are features; (2) clustering based on sample, where samples are data objects to be clustered and genes are features; (3) clustering based on subspace, where both, genes and samples can be viewed as objects or features. The unsupervised learning algorithms used were K-Means, Hierarchical Clustering, and Model-Based Clustering.

Liang et al. [3] integrated multiple cross-platform genomic and clinical data, with distinct statistical properties, for analysis and disease subtype identification. Multimodal Deep Belief Network was used to achieve their goal, which is a network of stacked Restricted Boltzmann Machines (RBMs), in which bottom level RBMs take multimodal data as input, and the top-level RBMs contain hidden variables representing the common features from multiple cross-platform data.

### 3 Methodology

#### 3.1 Data

Dataset on cancer genes and their mutations is provided by Memorial Sloan Kettering Cancer Center (MSKCC) and ONCOKB, cancer research organizations based in US. The dataset is of two types: Comma Separated Values (CSV) Format and Text-Based data. After some preprocessing we discovered that there is a total of 8989 IDs, 264 unique genes, 2996 unique variations, and 9 classes and also found out the maximal and minimal occurring genes. First, countplot of the class frequency was plotted and found Class 7 to be the most frequent. Then plotted the maximal occurring genes against each class to find out which gene is dominating in which class. After studying HGVS Nomenclature and referring to OncoKB, learned about the major types of cancer gene mutations. Using regular expression, classified all the mutations in respective types and created a separate feature for it. Then, bar graph of classes against the frequency of mutation types in each class was plotted, which helped to determine the meaning of each class. After this, countplot of mutation types of train variants data was plotted, to know the mutation type with maximum occurrence and



**Fig. 1** Overall working system for performing genome cancer analysis

its contribution to the occurrence of cancer. Point mutations were found to have the maximum frequency.

### 3.2 Feature Extraction and Classification Modeling

Figure 1 shows the overall working system of our study. NLTK was used for extracting important genetic words from the text data by removing the stopwords and special characters and then tokenizing the words and convert them into numerical values to feed into our classification models. Gene-like words from the list of tokenized words were extracted for each genetic mutation and stored in separate table. LabelEncoder and OneHotEncoder was used to create a Gene table with unique genes as columns and IDs as rows and a Mutation table with mutation types as columns and IDs as rows, to indicate the presence of unique genes and mutation types for each ID. Merged all tables on IDs to create a Features table on which classification models were developed using Random Forest Classifier (RFC) and Support Vector Classifier (SVC). Random Forest Classifier achieved better accuracy and was further improved using GridSearchCV for parameter tuning and 10-Fold Cross-Validation for checking model's accuracies for different splitting of test and train data and for avoiding overfitting. SVC was implemented using pipeline library of Sklearn [15, 16].

Used BaseEstimator and TransformerMixin, for defining classes FeatureExtractor (text data) and CategoricalExtractor (gene & mutation data). Then created three pipelines for text, gene, and mutation data. In text pipeline, first extracted features as tokens from the text and then countvectorized it by removing the stopwords and special characters and finally converted into a matrix. In gene pipeline, first genes were extracted from the dataset, one-hot encoded them and then stored in another matrix. In mutation pipeline, variations were extracted and stored in a matrix after encoding them. Then, all three pipelines were joined. Finally, a model pipeline was created of the joined pipeline and SVC [17–19].

## 4 Tools/Supporting Libraries for Methodology

- A. Numpy is used for mathematical computation of matrices and array manipulations.
- B. Pandas is used for reading dataset from local directory and storing it in dataframes. It is also used for dataframe manipulations and data cleaning.
- C. Matplotlib & Seaborn is used for graphical representation of data and for gaining insights, e.g., plotting Distribution of genetic mutation classes, Gene frequency in each class, etc.
- D. Regular Expressions is used for finding major mutation types in the dataset and for data cleaning as the mutation types present in the dataset didn't follow the correct HGVS nomenclature.
- E. NLTK is used for extracting important genetic words from the text data by removing the stopwords and special characters and then tokenizing the words and convert them into numerical values to feed into our classification models.
- F. Pickle is used to save and load our models.
- G. ScikitLearn is the most popular library for machine learning as it contains a great collection of various sub libraries which help in preprocessing (feature\_extraction, preprocessing, etc.) and applying machine learning algorithms (ensemble, svm, etc.). ScikitLearn also contains various libraries for model evaluation (model\_selection, metrics, etc.).
- H. Jupyter Notebook is used as it has the best open-source web application for machine learning projects as it has the ability to rerun individual snippets several times and can even edit those snippets before rerun.

## 5 Dataset for DataCan

We have tested our proposed data analysis methods on two types of genomic datasets, including a csv format containing variants dataset and text-based dataset. Text-based dataset contains the description of the genetic mutations. Fields are ID (row number to link the mutation to the clinical evidence), Gene (the gene where the genetic

mutation is located), Variation (the amino acid change for this mutations), Class (9 classes in which the genetic mutations has been classified into). Text dataset contains the clinical evidence (text) used to classify genetic mutations. Fields are ID (row number to link the clinical evidence to the genetic mutation), Text (the clinical evidence used to classify the genetic mutation). These datasets are provided by Memorial Sloan Kettering Cancer Center, USA.

## 6 Classes for DataCan

After the analysis of dataset and learning about cancer genes nomenclature and their features, we concluded that the classes appear to be as follows:

1. Likely Loss-of-function
2. Likely Gain-of-function
3. Neutral
4. Loss-of-function
5. Likely Neutral
6. Inconclusive
7. Gain-of-function
8. Likely Switch-of-function
9. Switch-of-function

Likely Loss of Function—Contains Truncation and Deletion as well as Duplication Mutations.

- Truncation Mutation, is a type of stop codon and can cut short the protein. Deletion Mutation causes a high loss of genes within a chromosomal region. A Duplication Mutation abnormally copies a piece of DNA one or more times in the same region.

Likely Gain of function—Contains Fusion, Deletion and Amplification Mutations.

- A gene formed from joining of two previously separate genes is a hybrid gene resulting from Fusion Mutation. It can be occurred by several processes: translocation of gene, interstitial deletion, or chromosomal inversion within the gene. Deletion Mutation causes a high loss of genes within a chromosomal region. Amplification Mutation (also called gene duplications), similar to duplication mutation leads to multiple copying of all chromosomal regions, which results in rising quantity of the genes located within them.

Neutral—Contains Wildtype Mutations.

- A Wildtype Mutation (also called homozygous non-mutated organism), is a mutation identical to both the parent alleles. In this neither allele is mutated. Geneticists call the individuals with the common or “normal” version of a characteristic the wild type, and the individuals with a unique or contrasting characteristic of the mutants.

**Loss-of-function—Contains Deletion and Frame Shift Mutations.**

- Deletion Mutation causes a high loss of genes within a chromosomal region. Frameshift Mutation (also called Framing Error or Reading Frame Shift) is caused by the insertion or deletion of several nucleotides in a DNA sequence that is not evenly divisible by three.

**Likely Neutral—Contains Promoter Mutations.**

- Point Mutations occurring within the protein chromosomal coding region of a gene is classified into three kinds depending upon what the error causing codon codes for: Silent mutations, resulting in same amino acids. Missense Mutations, resulting in different amino acids. Truncation Mutation, is a type of stop codon and can cut short the protein.

**Inconclusive—Contains Others Mutations.**

- When there is no proper conclusion as to which mutation type it is. Some individuals receive a genetic test resulting in “variant of uncertain significance” or “VUS.” This means that, during the testing, the laboratory can’t conclude whether the change in gene is a “deleterious change,” increasing the risk for cancer or a ‘benign’ variant, not increasing cancer risk.

**Gain of function—Contains Fusion, Amplification and Insertion Mutations.**

- A gene formed from joining of two previously separate genes is a hybrid gene resulting from Fusion Mutation. Amplification Mutation(also called gene duplications), similar to duplication mutation leads to multiple copying of all chromosomal regions, which results in rising quantity of the genes located within them. Insertion Mutation adds extra nucleotides to the DNA which are caused by element transposition, or errors occurred due to replication of elements.

**Likely Switch-of-function—Contains Fusion, Gene Subtype, and Amplification Mutations.**

- A gene formed from joining of two previously separate genes is a hybrid gene resulting from Fusion Mutation, occurring as a consequence of: translocation in gene, interstitial deletion, or chromosomal inversion of gene. Amplification Mutation leads to multiple copying of all chromosomal regions, which results in rising quantity of the genes located within them. Gene Subtype means various subtypes of genes which might be a contributing factor for SOF type of cancer.

**Switch-of-function—Contains only Fusion Mutation.**

- A gene formed from joining of two previously separate genes is a hybrid gene resulting from Fusion Mutation. It can be occurred by several processes: translocation of gene, interstitial deletion, or chromosomal inversion within the gene.

## 7 Results and Discussion

### 7.1 Datacan: Random Forest Classifier

GridSearchCV was used to find the perfect parameters for the classification model. It is an extensive searching method which over stated parameter values for any estimator. These parameters are then optimized by “cross-validated grid-search” over a grid of parameters. The selected parameters maximize the score of the left-out data. The best value found for “max\_depth” & “n\_estimators” after using GridSearchCV:

```
{'max_depth': 50, 'n_estimators': 200}
```

10-Fold Cross-Validation was used, in which the given data is randomly partitioned into 10 subsamples of equal sizes, from which a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. This process is then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. Table 2 shows the accuracy score of the model for 10-Fold Cross-Validation. Random Forest Classifier got the best accuracy score of 68.36%, which is not much accurate as the genes and their variations don’t follow the HGVS Nomenclature of genes because of which mutations couldn’t be classified correctly and conversion of text to numerical format resulted in loss of some important features.

**Table 2** Accuracy Score of Random Forest for 10-Fold Cross-Validation

Dataset split	Accuracy score (%)
1	64.686
2	68.211
3	68.106
4	67.000
5	64.333
6	64.214
7	65.771
8	65.540
9	65.762
10	68.367
	Average: 66.199

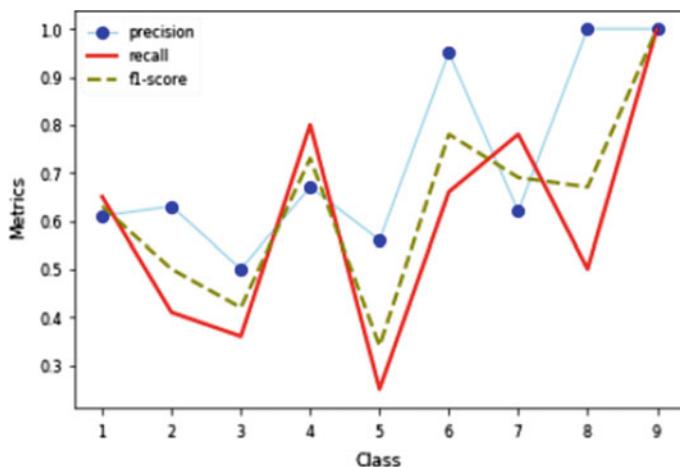
## 7.2 Random Forest Classifier: Classification Report

Precision indicates the proportion of positive identifications that are actually correct and Recall indicates the proportion of actual positives that are identified correctly. We measured our model's efficiency on the basis of "10-Fold Cross-Validation" using precision, recall, and f-measure. Class 6 has high precision and low recall, which means most of its predicted labels are correct when compared to the training labels. Calculated metrics on the basis of "10-Fold Cross-Validation" is shown in Table 3. Figure 2 presents the accuracy and ability of our proposed frame work for each class.

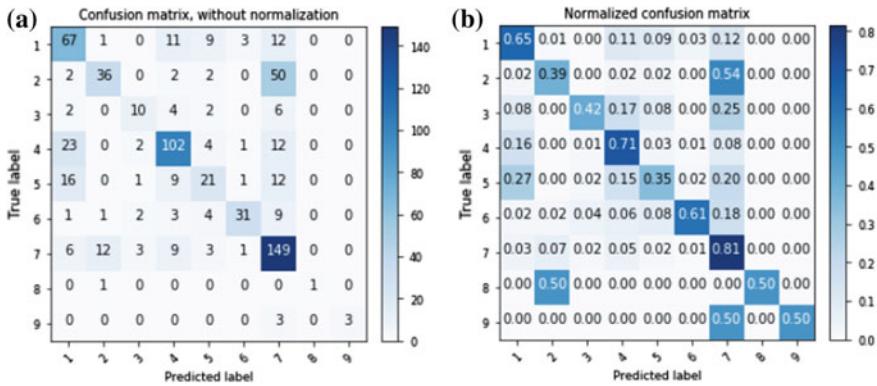
**Random Forest Classifier: Confusion Matrix.** A confusion matrix represents a summary of predicted results in a classification problem. The total number of correct and incorrect predictions are summarized with count values and grouped by each

**Table 3** Classification report for random forest classifier

Classes	Precision	Recall	F1-Score	Support
1	0.61	0.65	0.63	51
2	0.63	0.41	0.50	46
3	0.50	0.36	0.42	11
4	0.67	0.80	0.73	75
5	0.56	0.25	0.34	20
6	0.95	0.66	0.78	32
7	0.62	0.78	0.69	94
8	1.00	0.50	0.67	2
9	1.00	1.00	1.00	2
Average	0.66	0.65	0.64	333



**Fig. 2** Precision, recall and F1-Score



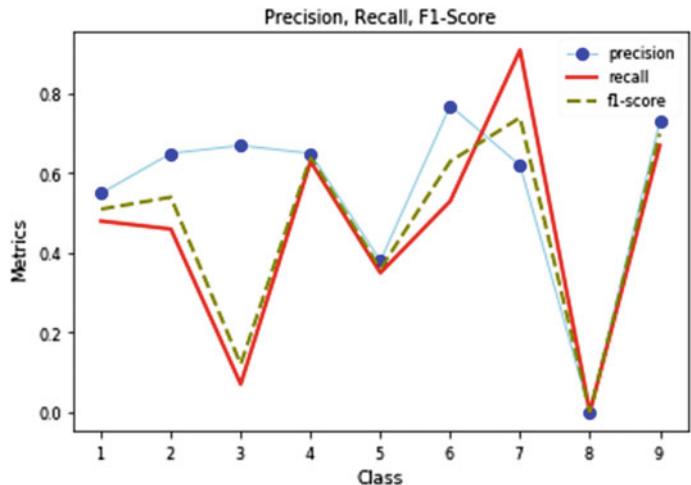
**Fig. 3** The confusion matrix: **a** Without normalization, **b** normalized

**Table 4** Classification report for support vector classifier

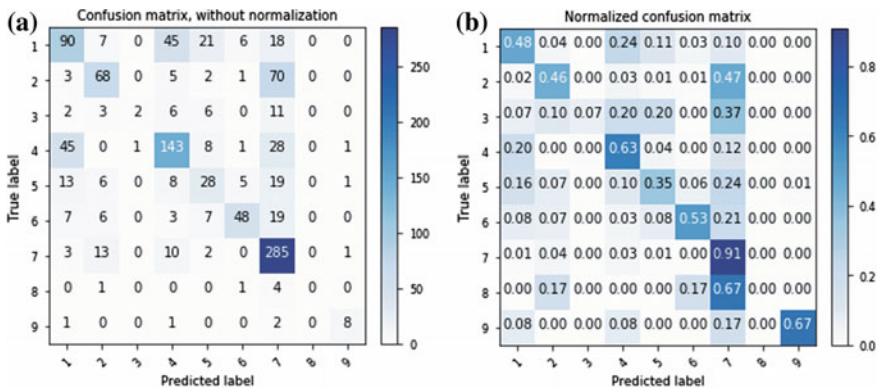
Classes	Precision	Recall	F1-Score	Support
1	0.55	0.48	0.51	187
2	0.65	0.46	0.54	149
3	0.67	0.07	0.12	30
4	0.65	0.63	0.64	227
5	0.38	0.35	0.36	80
6	0.77	0.53	0.63	90
7	0.62	0.91	0.74	314
8	0.00	0.00	0.00	6
9	0.73	0.67	0.70	12
Average	0.61	0.61	0.59	1095

class. We applied confusion matrix on our predicted values and true values, which we gathered by applying Random Forest. Figure 3 shows the Confusion Matrix: Normalized, Without Normalization.

*DataCan: Support Vector Classifier.* Preprocessing for support vector classifier was similar to what we did for random forest classifier, but here we used pipeline which helps in smooth transformation of raw dataset before applying final estimator. We can call a single fit and predict method on the data to fit a whole sequence of estimators. Pipelines also help avoiding leaking statistics from your test data into trained model in cross-validation, by ensuring that same samples are used to train the transformers and predictors. Still, RFC performed better than SVC as SVC is not the best choice for multi-label and multiclass classification. Table 4 shows classification report of the svc. Figure 4 represents the accuracy and ability of the svc.



**Fig. 4** Precision, recall and F1-Score



**Fig. 5** The confusion matrix: **a** Without normalization, **b** normalized

*Support Vector Classifier: Confusion Matrix.* The diagonal of the normalized confusion matrix represents the recall values of the model for each class. Figure 5 shows the Confusion Matrix: Normalized, Without Normalization.

*Comparative Analysis.* Table 5 shows the comparative analysis of results achieved by our implementation and that of other authors working on the same problem.

**Table 5** Comparative analysis of results

Author(s)	Dataset	Methods		Accuracy (%)		
		Feature	Classifier			
Dev et al. [10]	Breast	Signature composition	BPN	56.12		
			FLANN	63.34		
			PSO-FLANN	92.36		
Castano et al. [11]	Breast	BARS	EGRBF LR	91.08		
	Gum					
	Colon					
	Leukemia	FCBF				
	Lung					
	CNS					
Sharma and Paliwal [12]	Leukemia	Proposed algorithm	Bayesian classification	96.3		
	Lung			100.0		
	Breast			100.0		
DataCan (Our-Approach)	Genome cancer dataset	Proposed algorithms	Random forest classifier	<b>68.36</b>		
			Support vector classifier	<b>61</b>		

## 8 Future Work

Our dataset contained Genes and their mutations from which we analyzed the mutation types. Using the mutation types, we were able to identify the meaning of each class. Now for future, we plan on using LSTM Recurrent Neural Network in Keras and Natural Language Processing for the preprocessing and classification of Text-Based data. Advanced NLP models like “Term Frequency-Inverse Document Frequency” will be used for feature extraction and information retrieval. Word2vec will be used for word embeddings and creating feature vectors for words that neural networks can understand. We believe that using Deep Learning algorithm such as RNN in our project will help in better text classification and increase our accuracy score. We also plan on working on the Pan-cancer dataset provided by TCGA and try to find its correlations with our dataset and do a combined research on it.

**Acknowledgements** Varun Goel is the corresponding author. It is our privilege to express our sincere thanks to Prof. Venkatesh Gauri Shankar (Assistant Professor) from Manipal University Jaipur for his helpful guidance and discussions on our data analysis methods. He provided with various resources to support us during the implementation of this work.

## References

1. Bhola, A., & Tiwari, A. K. (2015, December) Machine learning based approaches for cancer classification using gene expression data. *Machine Learning and Applications: An International Journal (MLAIJ)*, 2(3/4).
2. Kharya, S., (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, 2(2).
3. Liang, M., Li, Z., Chen, T., & Zeng, J. (2015, July/August) Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4).
4. Gregory, K. B., Momin, A. A., Coombes, K. R., & Baladandayuthapani, V. (2014, November/December) Latent feature decompositions for integrative analysis of multi-platform genomic data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(6).
5. Weitschek, E., Cumbo, F., Cappelli, E., & Felici, G. (2016). Genomic data integration: A case study on next generation sequencing of cancer. In *2016 27th International Workshop on Database and Expert Systems Applications*.
6. Huang, H.-Y., Ho, C.-M., Lin, C.-Y., Chang, Y.-S., Yang, C.-A., & Chang, J.-G. (2016). An integrative analysis for cancer studies. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering*.
7. Mishra, S., Kaddi, C. D., & Wang, M. D. (2015). Pan-cancer analysis for studying cancer stage using protein expression data. In *Conf Proc IEEE Eng. Med Biol Soc* (pp. 8189–8192).
8. Guyon, I., et al. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3), 389–422.
9. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
10. Dev, J., et al. (2012). A classification technique for microarray gene expression data using PSO-FLANN. *International Journal on Computer Science and Engineering*, 4(9), 1534.
11. Castaño, A., et al. (2011). Neuro-logistic models based on evolutionary generalized radial basis function for the microarray gene expression classification problem. *Neural Processing Letters*, 34(2), 117–131.
12. Sharma, A., Imoto, S., & Miyano, S. (2012). A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(3), 754–764.
13. Rajput, D. S., Singh, P., & Bhattacharya, M. (2011). Feature selection with efficient initialization of clusters centers for high dimensional data clustering. In *2011 International Conference on IEEE Communication Systems and Network Technologies (CSNT)* (pp. 293–297).
14. Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11), 1370–1386.
15. Devi, B., Kumar, S., Anuradha, Shankar, V.G. (2019). AnaData: A novel approach for data analytics using random forest tree and SVM. In B. Iyer, S. Nalbalwar, N. Pathak (Eds.), *Computing, communication and signal processing* (Vol. 810). Advances in Intelligent Systems and Computing. Singapore: Springer. [https://doi.org/10.1007/978-981-13-1513-8\\_53](https://doi.org/10.1007/978-981-13-1513-8_53).
16. Shankar, V. G., Jangid, M., Devi, B., Kabra, S. (2018). Mobile big data: Malware and its analysis. In *Proceedings of First International Conference on Smart System, Innovations and Computing* (Vol. 79, pp. 831–842). Smart Innovation, Systems and Technologies. Singapore: Springer. [https://doi.org/10.1007/978-981-10-5828-8\\_79](https://doi.org/10.1007/978-981-10-5828-8_79).
17. Priyanga, A., & Prakasam, S. (2013). Effectiveness of data mining—Based cancer prediction system (DMBCPS). *International Journal of Computer Applications*, 83(10), 0975–8887.
18. Azuaje, F. (1999). Interpretation of genome expression patterns: computational challenges and opportunities. In *IEEE Engineering in Medicine and Biology Magazine: The Quarterly Magazine of the Engineering in Medicine & Biology Society* (Vol. 19, Issue, 6, pp. 119–119).
19. Shankar, V. G., Devi, B., Srivastava, S. (2019). DataSpeak: Data extraction, aggregation, and classification using big data novel algorithm. In B. Iyer, S. Nalbalwar, N. Pathak (Eds.), *Computing, communication and signal processing* (Vol. 810). Advances in Intelligent Systems and Computing. Singapore: Springer. [https://doi.org/10.1007/978-981-13-1513-8\\_16](https://doi.org/10.1007/978-981-13-1513-8_16).

# DataAutism: An Early Detection Framework of Autism in Infants using Data Science



Venkatesh Gauri Shankar, Dilip Singh Sisodia and Preeti Chandrakar

**Abstract** Data Science with analytics and machine learning in the field of health care are the most prominent and emerging fields in today's scenario. Our research paper aims to the healthcare solution toward autism in infants. Autism is the neurodevelopment disorder categorized by diminished societal interaction, lingual and non-lingual communication, repetitive and antagonistic behavior. Autism neurodevelopment figures out in the infants nearly about one year of age. The overall process of autism detection is a very long and cost-oriented process that takes 6 months to 10 months in total. We are concentrating on two data set and developed a framework for early detection of autism in infants. Form the same above, we use the concept of data analytics with training of data model and inclusion of SVM classification. We have tested our model and novel algorithm "DataAutism" over large data set and figure out high precision, recall with accuracy approx. 89%.

**Keywords** Data analytics · Autism · Classification · Data science · Support vector machine

## 1 Introduction

To the advancement of technology and information in each conceivable domain, there has been a requirement with mechanizing the procedures in order to build up quick and productive. Medicinal services are robust ventures that assign extremely mind boggling regarding finding and procedures required with human medical problems [1]. This paper centers around identifying infants Autism, the data corpus are identified with the infants points of interest as content having the overall description of the side effects with which they (data set) are enduring. The data corpus contain descrip-

---

V. G. Shankar (✉)  
Manipal University Jaipur, Jaipur, Rajasthan, India  
e-mail: [venkateshgaurishankar@gmail.com](mailto:venkateshgaurishankar@gmail.com)

V. G. Shankar · D. S. Sisodia · P. Chandrakar  
National Institute of Technology, Raipur, Chhattisgarh, India

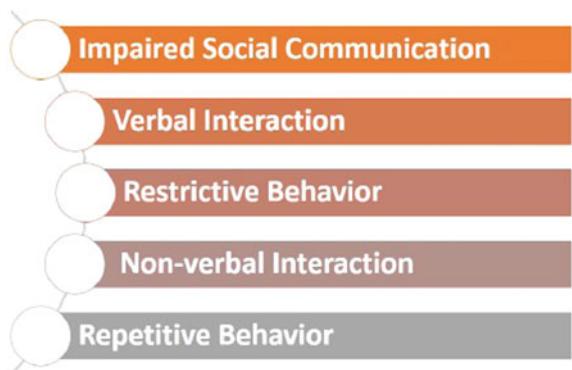
tion of both, infant's with Autism neurodisorder and without Autism neurodisorder, in order to prepare with the two sorts of datapoints [2].

The inspiration driving this paper is to have a quick discovery of danger of extreme autism. There are many methodologies which are utilized for detecting extreme ASD (Autism Spectrum Disorder), which have a higher legitimacy of detecting the infection, however the impediment implies exceptionally tedious and it creates a higher postponements in coming to a detection. The drawback of the current detection is their length or the many-sided approaches, because of which they take a considerable measure of time in long period. They likewise require medical treatment directed by the prepared experts and due to these components a long duration postponements in finding correct detection. Additionally, on account of such protracted and critical methodology, it is not worthy for those infants which need treatment. Henceforth this outcome is unequal and conflicting conveyance or inclusion. The clinical offices and preparedness experts have a tendency to be accessible more in significant urban areas. They are in general very not as much as the populace which needs treatment [2]. Because of absence of assets and time limitations, starting indicative screenings doesn't get led reliably. It very well may be severe to the point that families may need to hold up in long duration of 12 months from starting test to medical determination. An exhaustive survey estimated that 29% of the ASD problems stay undiscovered until the age of 2 years child. The normal time of a mental imbalance determination is over 5 years in many countries [3].

## 1.1 *Classification of Autism*

Some research papers proposed many latest approaches that are used for detecting autism with high precision of detecting the disorder, but the cons is that it is time taking and little bit complex, due to long duration in time it has high delays in approaching a decision. They also focused on a very small corpus data of children with and without ASD. On the basis of autism infants behavior, primarily they are classified into five types. Impaired Social Communication: it is also known as a social communication disorder in autism, which is also pragmatic communication disorder. It incorporates issues with social cooperation, social comprehension, and pragmatics. Verbal and Nonverbal Interaction: It incorporates unquestioning messages, regardless of whether purposeful or not, which are communicated through nonverbal practices. Nonverbal behavior incorporates facial expression, pitch and tone of the voice, gesture showed through body language or nonverbal communication (kinesics) and the physical separation between the communicators (proxemics). Restricted and Repetitive Behavior: Restricted, repetitive behaviors and interests are basic core symptoms of neural autism. They include repetitive gesture with objects, repeated body actions such as stunning and Self-stimulatory behaviors (stemming), ritualistic activities, receptive sensitivities, and confined interests (see Fig. 1). An infant with autism is also known as Autism Spectrum Disorder having following symptoms [1, 3]:

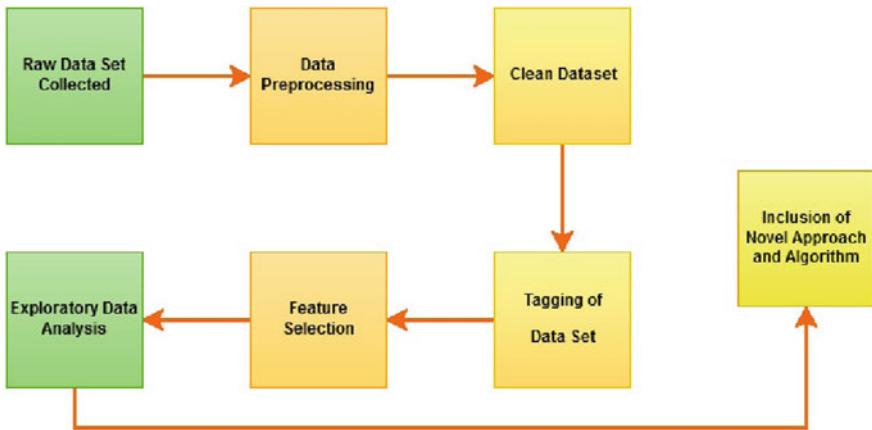
**Fig. 1** Classification of autism



- Not react to their name by a year
- Avoid eye to eye connection and want to be separated from everyone else
- Experience difficulty understanding the other individuals' sentiments or discussing their own emotions
- Not point at items to demonstrate curiosity (point at a plane flying over) by 16 months
- Not play “fictitious” games (fictitious to “feed” a doll) by 20 months
- Have infatuated interests
- Flap their hands, stun their body, or swirl in circles
- Have uncommon responses to the manner in which things sound, smell, taste, look, or feel
- Have postponed speech and linguistic skills
- Rehash words or expressions again and again
- Give random responses to questions
- Get annoyed with minor changes.

## 2 State-of-the-Art: Data Analytics and Training of Data Model

Data Analytics is the part of Data Science, in which raw or unstructured data are converted into structured data. After performing raw data collection, preprocessing start with the cleaning of data sets. Cleaning of data is required for removal of null values from the data field set [4, 5]. This cleaning of data is most important for best and effective throughput of the novel algorithm. On the other hand tagging of data set is also performed as preprocessing part of feature selection. The main purpose of feature selection and tagging of data is also important to calculate the real behavior or symptoms of infants. We have also used all above symptoms explained in Sect. 2 as Classification of Autism. We have done the exploratory data analytics using the

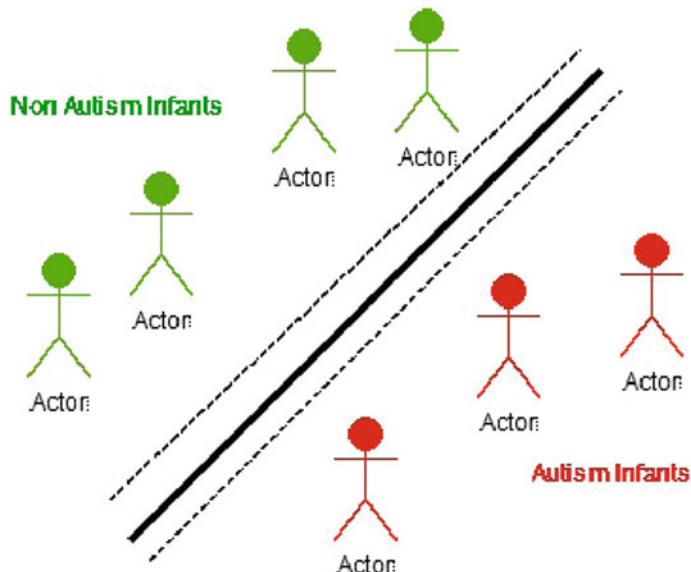


**Fig. 2** Data analytics and training of data model

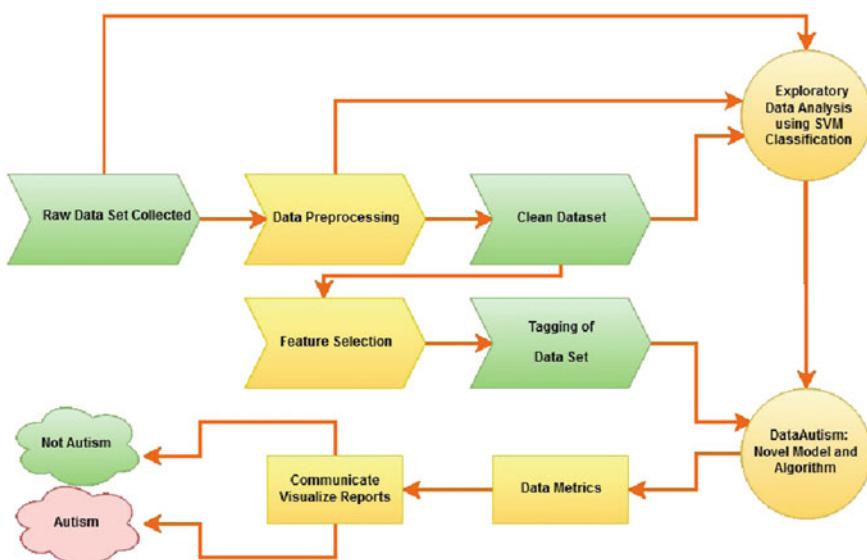
classification of data for that symptoms using SVM. The overall process of Data Analytics and training of the data model is given in Fig. 2.

### 3 State-of-the-Art: SVM: Support Vector Machine

We have done Exploratory Data Analysis with classification using SVM. We are using SVM on a labeled data set (tagged data set). As for the SVM, it is a classification Algorithm and also Machine learning algorithm. Support Vector Machines are supervised machine learning strategies used for classification and regression assignments that began from statistical learning hypothesis [5, 6]. As a classification strategy, SVM is a global classification framework that produces non-overlapping dividers and generally engagements all attributes. Visual techniques for exploratory information analysis can assist us with studying the outputs and supplement automated classification algorithm and model metrics calculations in information mining. We have used exploratory data analytics for better visualization of data model metrics and SVM for enhancing classification model with autism and not autism infants. The fundamental SVM classification is given in Fig. 3 and the inclusion of SVM in DataAutism is given in Fig. 4.



**Fig. 3** SVM: support vector machine



**Fig. 4** DataAutism: experimental setup

## 4 Related Work

Heinsfeld et al. [7] suggested an idea for autism spectrum disorder(ASD) using deep learning with ABIDE corpus. He has investigated pattern for functional connectivity that is used to detect autism spectrum disorder. They have improved result by achieving 70% accuracy. Abraham et al. [8] gives an idea for biomarker extraction from fMRI on ABIDE data set. He has used the concept of functional magnetic resonance imaging and to reveal functional biomarkers for neurodisorders. They have achieved accuracy as 68% by his fMRI framework. Duda et al. [9] presented an observation based classifier method for detection of autism risk. He has used machine learning-based classifier to detect approx. 72% of autism. Shankar et al. [10] suggested a way of data analytics with the use of k-nearest neighbor algorithm with spark. It uses the concept of data analytics, data aggregation, and classification based on data set collection. Wall et al. [11] proposed an idea for behavioral diagnosis of autism using Artificial Intelligence. He has evaluated autism genetic research exchange for 891 individuals. They have evaluated total of 86% of accuracy over the data set. Devi et al. [6] presented an enhanced platform for data analytics using classification techniques with the modified form of support vector machine and random forest tree. Jamal et al. [12] demonstrated an EEG system for discriminating typical autism spectrum disorder using brain connectivity with autism spectrum disorder states. His model achieved a total of 72% accuracy. Shankar et al. [13] demonstrated the concept of type of big data and how the feature extraction perform in the data set with data tagging. Martino et al. [14] proposed an idea for brain image data exchange using autism brain imaging data exchange and fMRI for 539 individual autism detection. They have calculated a total of 70% accuracy in his model.

In our framework as DataAutism , we have used the concept of data analytics with the classification. Classification is used for finding legitimate and non-legitimate unit of output. We have also performed exploratory data analytics using support vector machine (SVM) in the log file of tagged data set. We have calculated a total of 89% accuracy of our model, which is presented in the result and discussion section with the many metrics evaluations.

## 5 DataAutism: Experimental Setup and Algorithm

We are concentrating on two data set and developed a framework for early detection of autism in infants. For the same processing, we use the concept of data analytics with the training of the data model and the inclusion of SVM classification. To remove long processing of autism detection in normal treatment, we have suggested a novel data analytic approach in a way of fast detection process. In the processing of detection, our approach started with the feature extraction of data set with the tagging of the feature for analysis. We have also been using a linear classifier to classify the behavior of infants. In continuation with classifier, we have applied

SVM hybrid approach (linear and nonlinear both) to classify autism and not autism infants. After doing data analytics part we have applied SVM over the features of a data set and perform exploratory data analysis using SVM. We have stored the result of exploratory data analytics in a log file using Strace tool. We have applied our novel algorithm of DataAutism over this log file and applied some standard classification metrics over them. The framework architecture of DataAutism is given below in Fig. 4.

The framed novel algorithm of 'DataAutism' is given below in Algorithm 1.

---

**Algorithm 1** DataAutism: Exploratory Data Analysis and Classification
 

---

```

procedure DATA,LOG_FILE(Feature C1, r) Feature Extraction of the data set and log tagging
  for p = 1 → r do
    if C1 → ISP_VI_ResB == 1 then
      return to combined log file
    Execute the ISP,VI and ResB segment;
    Check for feature extraction;
    Start tagging to the feature;
    end if
  end for
  Find C1 in (1 → r,C1 → ISP_VI_ResB)
  for i = 1 → x do
    if Exploratory_query(C1, r) == TRUE then
      x ← C1p%
      for item x ∈ r do
        if x.features == C1 then Autism or Not
          include x as the second segment
        Execute the NVI,RepB and Mis segment;
        Check for feature extraction;
        Start tagging to the feature;
        end if
      end for
    end if
  end for
end procedure
  
```

---

## 6 DataAutism: Data set Selection

We have collected many data set for evaluating the trend and figure out the main features for autism. We have collected total of four data set, out of these four data set two are from UCI repository data set [15, 16], one is CDC data set [17] and other one is NDAR data set [18]. We have merged all the data set having approx. 1880 individuals. We have performed feature extraction and exploratory data analysis with classification over given data set. The overall data set structure is given below in Table 1.

**Table 1** DataAutism: data set collection

Data set collected	Category	No. of individuals selected
UCI Data Set	Autism_1	891
UCI Data Set	Autism_2	433
CDC Data Set	Autism_infant	355
NDAR	NIMH_Autism	201
<b>Total</b>		<b>1880</b>

## 7 DataAutism: Result and Discussion

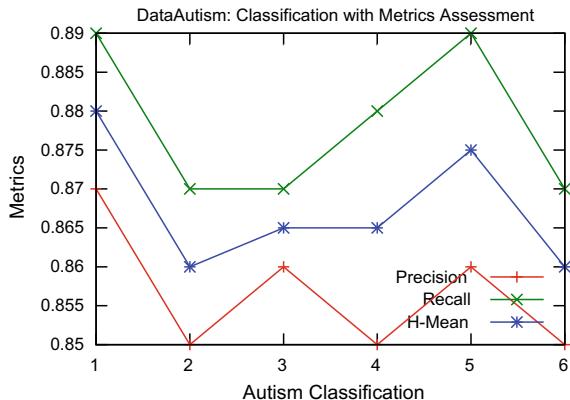
As for the accuracy and ability of our proposed framework DataAutism, we have used many performance metrics over the result and findings of data set. For the effectiveness of our model, we have tested our model over data assessment metrics such as precision, recall, specificity, sensitivity, and F-Measure. We have also performed all of the above metrics with cross validation for best evaluation and accuracy of DataAutism. As for precision, it is reflected as positive predictive value, whereas recall is the ability of model also known as true positive rat. On the other hand, specificity is the true negative rate and fault Measure(F-measure) is the harmonic mean of accuracy and ability. We have presented all metrics with classification findings with our novel algorithm DataAutism in Table 2 and in the Fig. 5.

- 1. Impaired Social Communication: ISP
- 2. Verbal Interaction: VI
- 3. Restrictive Behavior ResB
- 4. Non-verbal Interaction: NVI
- 5. Repetitive Behavior: RepB
- 6. Miscellaneous (Covering all other): Mis.

**Table 2** DataAutism: classification with metrics assessment

Assessment metrics	ISP	VI	ResB	NVI	RepB	Mis
Precision	0.87	0.85	0.86	0.85	0.86	0.85
Recall	0.89	0.87	0.87	0.88	0.89	0.87
Specificity	0.11	0.13	0.13	0.12	0.11	0.13
Sensitivity	0.89	0.87	0.87	0.88	0.89	0.88
Harmonic mean	0.88	0.86	0.865	0.865	0.875	0.86

**Fig. 5** DataAutism: classification with metrics assessment

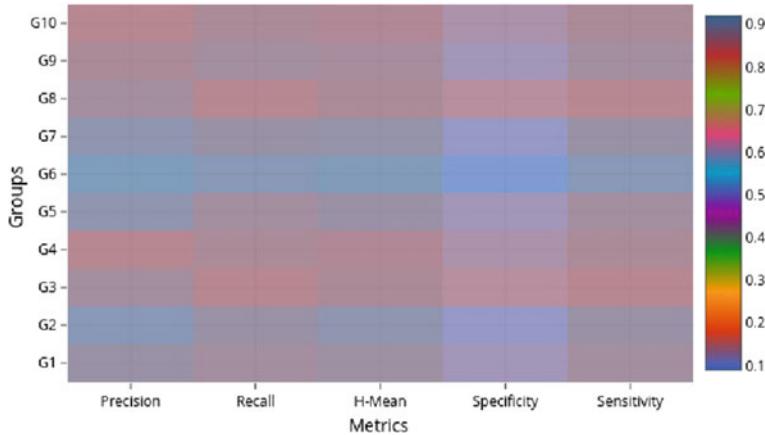


## 7.1 DataAutism: Cross Validation

We have also tested our result with cross validation for validating the random data from data set. In cross validation, we have considered one set as a test-set and other set as a training-set. In here, DataAutism, we have used 10-fold cross validation with the data set, where we have divided whole data set in ten equal part and at a time one data set as test-set and remaining nine as training-set. We have divided overall data set in ten equal groups and each having 188 individuals. We have also calculated all the metrics (precision, recall, specificity, sensitivity, and F-measure) with cross validation also. The overall evaluation and calculation of cross validation is given in Table 3 and Heat-map for the same is also given in Fig. 6.

**Table 3** DataAutism: cross validation

Groups (188)	Test Set	Training Set	Precision	Recall/sensitivity	H-mean	Specificity
Group1 (G1)	(G1)	(G2–G10)	0.89	0.88	0.885	0.12
Group2 (G2)	(G2)	(G1) (G3–G10)	0.91	0.89	0.90	0.11
Group3 (G3)	(G3)	(G1–G2) (G4–G10)	0.88	0.86	0.87	0.14
Group4 (G4)	(G4)	(G1–G3) (G5–G10)	0.86	0.87	0.865	0.13
Group5 (G5)	(G5)	(G1–G4) (G6–G10)	0.90	0.88	0.89	0.12
Group6 (G6)	(G6)	(G1–G5) (G7–G10)	0.92	0.91	0.915	0.09
Group7 (G7)	(G7)	(G1–G6) (G8–G10)	0.90	0.89	0.895	0.11
Group8 (G8)	(G8)	(G1–G7) (G9–G10)	0.88	0.86	0.87	0.14
Group9 (G9)	(G9)	(G1–G8) (G10)	0.87	0.88	0.875	0.12
Group10 (G10)	(G10)	(G1–G9)	0.86	0.87	0.865	0.13



**Fig. 6** DataAutism: cross validation visualization

## 7.2 DataAutism: Comparative Analysis

We have compared our Framework accuracy with other literature work. Our framework is based on data analytics with classification, which gives a total of 89% result. Other literature work and findings of research as Heinsfeld et al. [7] constituted total of 70% accuracy, whereas Duda et al. [9] proposed a network has also achieved the accuracy as 72%. On the other hand Wall et al. [11] also suggested an AI based model achieved 86% of accuracy. The overall comparative analysis is given below in Table 4.

**Table 4** DataAutism: comparative analysis

Autism research	Year	Model	Accuracy
Heinsfeld A.S. et al.	2018	Deep learning	70
Duda M. et al.	2014	Machine learning	72
Dennis P. Wall et al	2012	Behavioral analysis	86
Alexandre A. et al.	2017	fMRI Imaging	68
Our model (DataAutism)	–	Data analytics and classification	89

## 8 Conclusion and Future Work

Nowadays Data Science and analytics in the biomedical field is a robust and efficient methodology. We have used the concept of data analytics and exploratory data analytics with classification in DataAutism. We have also used the concept of SVM and produced a novel algorithm which give high precision and recall with the large data set. We have applied our framework over a large data set and performed the metric assessments using precision, recall, H-mean, specificity, sensitivity. We have also performed the cross validation metrics with the random data set with segments and evaluated all same metrics on the basis of classification. Our approach achieved the total accuracy of 89% over the 1880 individuals. We have also calculated individual classified autism category and it also has high accuracy and ability with high precision and recall. As for future direction we will work on the concept of Deep Learning, Biomedical Imaging, as well as neural network for enhancing result in respect to accuracy and ability of the model for Autism spectrum disorder. We have also explored the area of Autism in children as well as adult with the large data set.

**Acknowledgements** It is my privilege to express my sincere gratitude to National Institute of Technology, Raipur and Manipal University Jaipur for providing research platform and support to carry out research. We are thankful to National Institute of Mental Health and University of California, Irvine for making data available.

## References

1. *HealthLine autism classification*. <https://www.healthline.com/health/autism>. Accessed July 22, 2018.
2. *AutismSpeaks about autism*. <https://www.autismspeaks.org/what-autism>. Accessed August 02, 2018.
3. *Healthitanalytics autism\_types*. <https://healthitanalytics.com/news/ehr-data-analytics-reveal-subtypes-of-autism-in-children>. Accessed August 14, 2018.
4. Tang, J., Liu, J., Zhang, M., & Mei, Q. (2016). Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on WWW* (pp. 287–297).
5. Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning* (pp. 59–66). Washington, USA: ACM.
6. Devi, B., Kumar, S., & Anuradha, S. V. G. (2019). AnaData: A novel approach for data analytics using random forest tree and SVM. In B. Iyer, S. Nalbalwar, & N. Pathak (Eds.), *Computing, communication and signal processing. Advances in intelligent systems and computing* (Vol. 810). Singapore: Springer. [https://doi.org/10.1007/978-981-13-1513-8\\_53](https://doi.org/10.1007/978-981-13-1513-8_53).
7. Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17, 16–23. <https://doi.org/10.1016/j.nicl.2017.08.017>.
8. Abraham, A., Milham, M. P., Di Martino, A., Cameron Craddock, R., Samaras, D., Thirion, B., Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, 147, 736–745, ISSN 1053-8119. <https://doi.org/10.1016/j.neuroimage.2016.10.045>.

9. Duda, M., Kosmicki, J. A., & Wall, D. P. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Translational Psychiatry*, 4, e440. <https://doi.org/10.1038/tp.2014.65>.
10. Shankar, V. G., Devi, B., & Srivastava, S. DataSpeak: Data extraction, aggregation, and classification using big data novel algorithm. In B. Iyer, S. Nalbalwar, & N. Pathak (Eds.), *Computing, communication and signal processing. Advances in intelligent systems and computing* (Vol. 810). Singapore: Springer. [https://doi.org/10.1007/978-981-13-1513-8\\_16](https://doi.org/10.1007/978-981-13-1513-8_16).
11. Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., & DeLuca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS ONE*, 7(8), art. no. e43855.
12. Jamal, W., Das, S., Maharatna, K., Kuyucu, D., Sicca, F., Billeci, L., Apicella, F., & Muratori, F. (2013). Using brain connectivity measure of EEG synchronostates for discriminating typical and autism spectrum disorder. In *2013 6th International IEEE/EMBS Conference* (pp. 1402–1405), San Diego, CA. <https://doi.org/10.1109/NER.2013.6696205>.
13. Shankar, V. G., Jangid, M., Devi, B., Kabra, S. (2018). Mobile big data: Malware and its analysis. In *Proceedings of First International Conference on Smart System, Innovations and Computing. Smart Innovation, Systems and Technologies* (Vol. 79, pp. 831–842). Singapore: Springer. [https://doi.org/10.1007/978-981-10-5828-8\\_79](https://doi.org/10.1007/978-981-10-5828-8_79).
14. Di Martino, A., Yan, C.-G., & Milham, M. P. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19, 659–667. <https://doi.org/10.1038/mp.2013.78>.
15. UCI UCI Data Set 1. <http://archive.ics.uci.edu/ml/machine-learning-databases/00419/>. Accessed September 04, 2018.
16. UCI UCI Data Set 2. <http://archive.ics.uci.edu/ml/machine-learning-databases/00420/>. Accessed June 28, 2018.
17. CDC CDC Dataset. <https://www.cdc.gov/nccbddd/autism/data.html>. June 21, 2018.
18. NIMH NDAR Dataset. [https://ndar.nih.gov/edit\\_collection.html?QA=false&id=1880](https://ndar.nih.gov/edit_collection.html?QA=false&id=1880). Accessed July 13, 2018.

# AnaBus: A Proposed Sampling Retrieval Model for Business and Historical Data Analytics



Bali Devi, Venkatesh Gauri Shankar, Sumit Srivastava  
and Devesh K. Srivastava

**Abstract** Information Retrieval is a domain of transforming unstructured data to structured data by various strategies. These strategies are heuristic in nature. It is an implementation of data mining and warehousing. Probabilistic and Statistical Language Models already exist in the field of IR. This paper describes a new blend of retrieval model to the past concepts of information retrieval as the vector space information retrieval model, probabilistic retrieval model, and last one as statistical model. Our novel framework is centric around the Sampling Distribution of statistics. We are emphasizing on Statistical Model and Probabilistic Model that is basically Statistical cum Probabilistic Model. This work shows the presence of effective novel retrieval algorithms that use the common terms of statistical computation using expectation, mean, and variance. By using the concept of Sampling Distribution, we categorize our novel algorithm as ad hoc and filtering approaches. This paper deals with summarized statistical calculations and suggests a new model for relevancy and ranking of documents.

**Keywords** Information retrieval · Business analytics · Big data analytics · Historical data · Big data

## 1 Introduction and Existing Approach Overview

Information Retrieval has become an integral part of our lives as it is required for a wide range of applications such as World Wide Web, transactions, navigations, and search-based applications and so on. Information Retrieval is basically a study between data and information [1, 2]. In Salton's words, "This is a domain centric with structure retrieval and analysis, enterprise, storage, retrieval of data and relevancy of information" [3]. An IR (Information Retrieval) Model is a technique by which a

---

B. Devi (✉)  
SCIT, Manipal University Jaipur, Jaipur, Rajasthan, India  
e-mail: [baligupta03@gmail.com](mailto:baligupta03@gmail.com)

V. G. Shankar · S. Srivastava · D. K. Srivastava  
Department of Information Technology, Manipal University Jaipur, Jaipur, Rajasthan, India

**Table 1** Advantages and disadvantages of existing model

Dimensions	Boolean retrieval model	Vector space retrieval model	Probabilistic retrieval model	Statistical language retrieval model
Principle	Binary decision criteria	Cosine similarity	Maximize the overall probability of relevance	Probability distribution over a sequence of words
Matching	Exact	Best	Best	Best
Advantages	Clean usage, simplicity	Partial matching, improved retrieval performance	Documents ranked in decreasing order of probability	Overcomes the shortcomings of probabilistic model
Disadvantages	Exact matching, few retrieval of documents, Boolean expression required	Mutual independence between index terms	Guessing initial relevant and irrelevant, mutual Independence between index terms	favors documents that contain frequent (low content) words

relevance measure is obtained between queries and documents [1]. The major models used in IR are Boolean Retrieval and Relevancy Model, Vector Space Retrieval Model, Probabilistic Relevancy Model, and Statistical Language Retrieval Model; everyone has its own advantages and disadvantages as shown in Table 1.

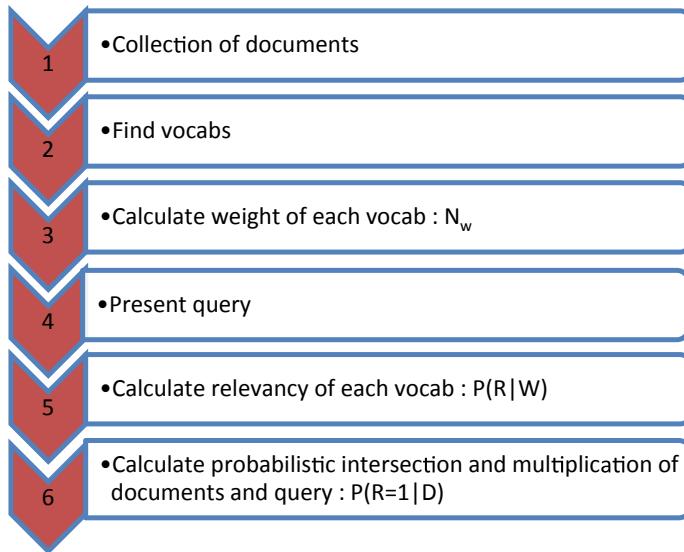
Depending on the model, two implementations can be used:

- (1) Ad hoc Approach: Relevancy is measured between documents and a query. On the basis of the relevancy with respect to the query, ranking of the documents is demonstrated (Fig. 1).
- (2) Filtering Approach: Relevancy of a new corpus is measured with respect to the existing corpus (Fig. 2).

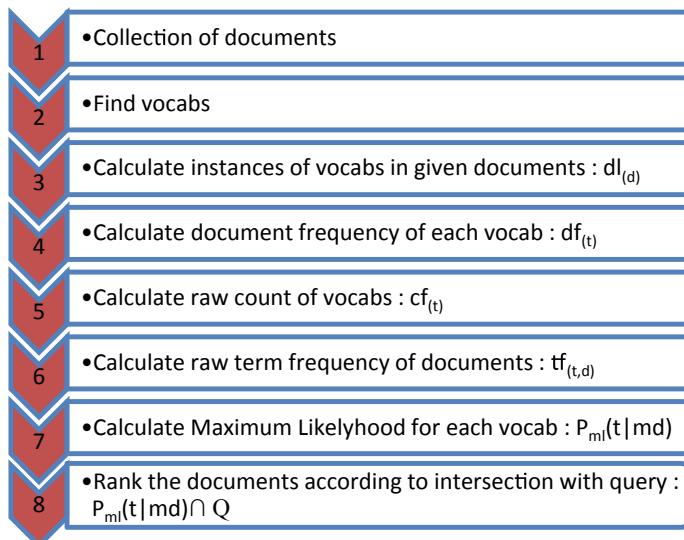
In this paper, we are using a new novel framework of Sampling Retrieval Distribution. The Sampling Retrieval Distribution is a delivery of a sample statistic of a given population that contains the probability retrieval delivery of sample statistics framed on randomly designated samples from a population.

In this paper, a population is referred to as a document and samples are the vocabs occurring in that document. The strategy involves calculation of expectation of sample mean and variance of sample mean for ad hoc and filtering approaches respectively.

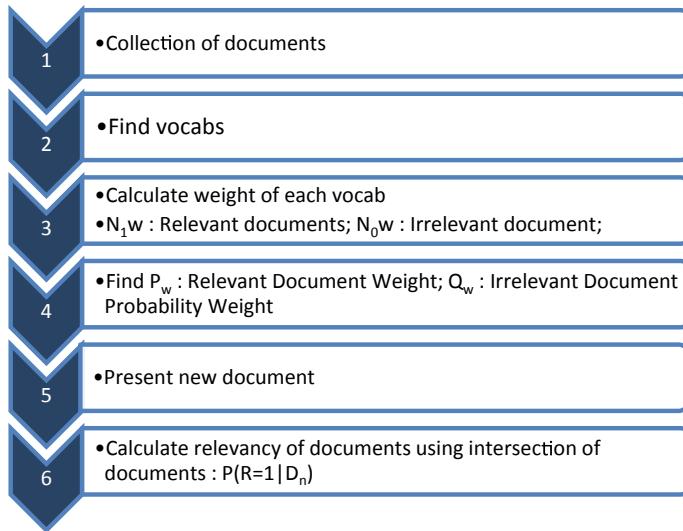
We are using following statistical formulae for our novel approach: (Figs. 3 and 4).



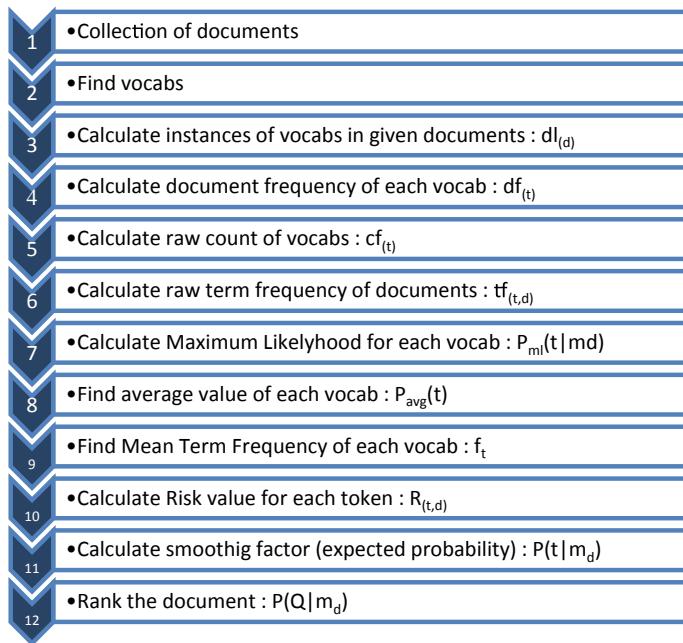
**Fig. 1** Illustrating probabilistic model: adhoc approach



**Fig. 2** Illustrating probabilistic model: filtering approach



**Fig. 3** Illustrating statistical language model: maximum likelihood method



**Fig. 4** Illustrating statistical model: mean term frequency method

**Formula 1**

$$E[p] = \sum P[p] \cdot p$$

**where p = Sample Mean**

**E[p] = Expectation of Sample Mean**

**P[p] = Probability of Sample Mean**

**Formula 2**

$$\text{Var}[p] = E[p^2] - E[p]^2$$

**Where**

**E[p<sup>2</sup>] = Expectation of square of sample mean**

**E[p]<sup>2</sup> = Square of Expectation of sample mean**

## 2 Related Work

The specific cons for corpus retrieval is rank of documents, generally, how to generate the rank based function (model) that is used to sort corpus based on relevancy order to the given new retrieval query [4]. There are many models buildup over the past years that have many different ways to rank new documents/query. Some of these include: vector space models for information retrieval [3], probabilistic retrieval models for document query [5], language models for document query and relevancy [6], divergence from randomness models for IR [7], learning models for retrieval [8, 9] and some other more eccentric models [10].

In document/query retrieval, the very prominent work is explained in [11], in which Chen and Karger discussed that the old probabilistic rank and relevancy principle (PRP) [11], that ranking artifacts on the basis of chances of relevance in the descending order, but for different user needs it is not always optimal. Among the top n documents, it was proposed to increase the chances of resulting relevant artifacts by the authors [12, 13].

We have set ourselves more ambitious goals, but our framework can be seen as novel work with the query retrieval and relevancy. We presented that relevancy of artifacts by inspecting their expectation of sample mean is insufficient. Therefore, we are taking the Variance of sample mean as the basis to do our relevancy check.

## 3 Experiment and Methodology

### 3.1 Algorithms

#### Ad hoc Approach

- (1) Input: = Query and collection of documents.
- (2) Output: = Ranking of documents w.r.t query.
- (3) Initialization: = words/tokens/vocabs/terms

- (4) Draft the incidence matrix against given set of documents and vocabs.
- (5) Calculate the total of each vocab in the given documents.

$$\text{Total}[Tvi] = vi(D_1 + D_2 + \dots + D_n)$$

- (6) Find the Sample Mean ( $x$ ) of each vocab.

$$x = Tvi/N(\text{total}).$$

- (7) For each document draw the frequency distribution table of Sample Mean.
- (8) Compute the Probability of sample mean for each entry made in the tables.

$$P[x] = fx/V(\text{total})$$

as,  $V = \text{All Vocab}$

- (9) Evaluate the Expectation of sample mean for each document.

$$E(D_i \cap Q)[x] = \sum P[x].x$$

where  $x = \text{Sample Mean}$

$E[x] = \text{Expectation of Sample Mean}$  and  $P[x] = \text{Probability of Sample Mean}.$

- (10) Rank the documents on the basis of calculated expectations.

### Filtering Approach

- (1) Input: = New document and collection of documents.
- (2) Output: = Relevancy of documents w.r.t New document.
- (3) Initialization: = words/tokens/vocabs/terms.
- (4) Draft the incidence matrix against given set of documents and vocabs.
- (5) Calculate the total of each vocab in the given documents.

$$\text{Total}[Tvi] = vi(D_1 + D_2 + \dots + D_n)$$

- (6) Find the Sample Mean ( $x$ ) of each vocab.  $x = Tvi/N$   
where  $N = \text{Total number of documents (excluding new document).}$
- (7) For each document (including new document), draw the frequency distribution table of Sample Mean.
- (8) Compute the Probability of sample mean for each entry made in the tables.

$$P[x] = fx/V(\text{total})$$

as,  $V = \text{Total Vocab}$

- (9) Evaluate the Expectation of sample mean for each document (including new document).

$$E[x] = \sum P[x] \cdot x$$

where  $x$  = Sample Mean

$E[x]$  = Expectation of Sample Mean  $P[x]$  = Probability of Sample Mean.

- (10) Determine Variance of sample mean for each document (including new document).  $x = p$  (for use)

$$\text{Var}[p] = E[p^2] - E[p]^2$$

Where  $E[p^2]$  = Expectation of square of sample mean  $E[p]^2$  = Square of Expectation of sample mean.

Compare the  $\text{Var}[p]$  of new document to the  $\text{Var}[x]$  of existing documents. Mark the document “most relevant” whose  $\text{Var}[x]$  is near to that of new document’s.

## 4 AnaBus: Result and Discussion

We have applied our ad hoc and filtering approach over the four sets of dataset, which are collected from different domains as Iris Data Set from UCI repository [14], Titanic Data Set from Kaggle dataset [15], Bigmart Sales DataSet from Analyticsvidhya [16] and Census Income DataSet from UCI repository [17]. All these datasets are tested with ad hoc approach as well as filtering approach in combined as well as individual form. We have also checked the relevancy of new data element with the model. The overall dataset collection is given in Table 2.

We have tested our experimental model and algorithm on the DataSet Sources with the evaluation in term of Precision, Recall,  $F$ -measure, Specificity, and Sensitivity. We have applied these metrics assessment for the accuracy and ability of our algorithm. Our proposed algorithm and model “AnaBus” gives higher accuracy and ability with the different sets of dataset. The overall finding of our algorithm with the all metrics assessment is given in Table 3.

**Table 2** AnaBus: Data Set Sources

Data Set	Type	Tagged Name	No. of individuals selected
UCI	Iris Data Set	$D_1$	1200
Kaggle	Titanic Data Set	$D_2$	890
Analyticsvidhya	Bigmart Sales Data Set	$D_3$	470
UCI	Census Income Data Set	$D_4$	390
Total			2950

**Table 3** AnaBus: metrics evaluation

Evaluation Metrics	$D_1$	$D_2$	$D_3$	$D_4$	Combined
Precision	0.77	0.76	0.75	0.78	0.765
Recall	0.75	0.77	0.73	0.75	0.76
Harmonic mean	0.76	0.765	0.74	0.765	0.7625
Specificity	0.25	0.23	0.27	0.25	0.24
Sensitivity	0.75	0.77	0.73	0.75	0.76

## 5 AnaBus: Conclusion and Future Scope

Information Retrieval is a robust area of document relevancy and selection. There are many models existing with their pros and cons such as statistical model, probabilistic model, Boolean model, etc. Information retrieval is also the part of intelligent data analytics and it uses many new models as Machine Learning and AI. Here in AnaBus, We have used the concept of Ad hoc and Filtering sampling theory with the large dataset. We have also collected a large business corpus for finding accuracy and ability of our approach “AnaBus”. For Accuracy and ability, we have applied the metrics of accuracy as precision and ability as recall and our model gives as approx. 75% of result accuracy with the calculation of this matrix. In future, we will use the concept of Machine Learning and Deep Learning with the huge dataset for business analytics.

## References

1. Imperial College London, IR. [https://www.doc.ic.ac.uk/~nd/surprise\\_97/journal/vol4/hks/inf\\_ret.html](https://www.doc.ic.ac.uk/~nd/surprise_97/journal/vol4/hks/inf_ret.html). Accessed 05 Aug 2018.
2. TechTarget, Business Analytics, <https://searchbusinessanalytics.techtarget.com/definition/business-analytics>—BA. Accessed July 13, 2018.
3. Salton, G., Wong, A., & C. S. Yang. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620. <http://dx.doi.org/10.1145/361219.361220>.
4. Devi B., Kumar S., Anuradha, & Shankar, V. G. (2019). AnaData: A novel approach for data analytics using random forest tree and SVM. In B. Iyer, S. Nalbalwar, & N. Pathak (Eds.), *Computing, communication and signal processing. advances in intelligent systems and computing*, (Vol. 810). Singapore: Springer. [https://doi.org/10.1007/978-981-13-1513-8\\_53](https://doi.org/10.1007/978-981-13-1513-8_53).
5. Robertson, S. (2005). On event spaces and probabilistic models in information retrieval. *Information Retrieval*, 8(2), 319–329. <http://dx.doi.org/10.1007/s10791-005-5665-9>.
6. Ponte, J. M., & Bruce Croft, W. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)* (pp. 275–281). New York, NY, USA: ACM. <https://doi.org/10.1145/290941.291008>.
7. Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4), 357–389. <http://dx.doi.org/10.1145/582415.582416>.

8. Radlinski, F., & Joachims, T. (2006). Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In A. Cohn (Ed.), *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06)* (Vol. 2, pp. 1406–1412). AAAI Press.
9. Fuhr, N., & Buckley, C. (1991). A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9(3), 223–248. <http://dx.doi.org/10.1145/125187.125189>.
10. Shi, S., Wen, J.-R., Yu, Q., Song, R., & Ma, W.-Y. (2005). Gravitation-based model for information retrieval. In *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 488–495). New York, NY, USA: ACM. <http://dx.doi.org/10.1145/1148170.1148245>.
11. Chen, H., Karger, D. R. (2006). Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information retrieval (SIGIR'06)* (pp. 429–436). New York, NY, USA: ACM. <http://dx.doi.org/10.1145/1148170.1148245>.
12. Shankar, V. G., Devi B., & Srivastava S. (2019). DataSpeak: Data extraction, aggregation, and classification using big data novel algorithm. In: B. Iyer, S. Nalbalwar, & N. Pathak (Eds.) *Computing, communication and signal processing. Advances in intelligent systems and computing*, (Vol 810). Singapore: Springer. [https://doi.org/10.1007/978-981-13-1513-8\\_16](https://doi.org/10.1007/978-981-13-1513-8_16).
13. Shankar, V. G., Jangid, M., Devi, B., & Kabra, S. (2018). Mobile big data: malware and its analysis. In *Proceedings of First International Conference on Smart System, Innovations and Computing. Smart Innovation, Systems and Technologies* (Vol. 79, pp. 831–842). Singapore: Springer. [https://doi.org/10.1007/978-981-10-5828-8\\_79](https://doi.org/10.1007/978-981-10-5828-8_79).
14. UCI, Iris Data Set. <http://archive.ics.uci.edu/ml/datasets/Iris>. Accessed July 03, 2018.
15. Kaggle, Titanic Data Set. <https://www.kaggle.com/c/titanic>. Accessed August 03, 2018.
16. Analyticsvidhya, Bigmart Sales Data Set. <https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/>, Online; Accessed July 27, 2018.
17. UCI, Census Income Data Set. <http://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/>. Accessed July 21, 2018.

# **Big Data Management**

# Abrupt Scene Change Detection Using Block Based Local Directional Pattern



T. Kar and P. Kanungo

**Abstract** In the present communication an ingenious and robust method of abrupt scene change detection in video sequences in presence of illumination variation based on local directional pattern is proposed. A similarity measure is developed by evaluating the difference between the new texture based feature descriptor which is compared to an automatically generated global threshold for evaluation of the scene change detection. The proposed framework is tested on few publicly available videos and TRECVID dataset. The encouraging results are in favor of the credibility of the proposed framework.

**Keywords** Abrupt scene change detection · LBP · LDP · Illumination variation

## 1 Introduction

The mammoth development in the fields of video and multimedia technology, has made available plentiful of cheap video acquisition tools to the common man. Thus video acquisition, uploading and downloading of video information to different hosting websites is no longer restricted to expert personals. This led to massive use of the video acquisition tools by the common man resulting in an impulsive growth of video database. However, the complex and unstructured characteristics of the video created the difficulty in retrieving the desired video data from the massive video data. Moreover, manual searching techniques are highly cumbersome in terms of processing time. To address this issue, several methods are devised for quick and systematic organization of the multimedia data. The first and foremost step toward such data organization is to identify the boundary between the shots which is popularly known

---

T. Kar (✉)

KIIT Deemed to be University, Bhubaneswar, Odisha, India

e-mail: [tkarfet@kiit.ac.in](mailto:tkarfet@kiit.ac.in)

P. Kanungo

C. V. Raman College of Engineering, Bhubaneswar, Odisha, India

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

as shot detection. All shot detection algorithms primarily focus on evaluating the discontinuity in the visual information between the frames. Splitting of the video data into convenient and primitive unit of video is known as a shot [1, 2]. A shot being the smallest and meaningful segment, that is captured through operation of a single camera in one continuous run. The frontiers between two consecutive shots can be either a cut transition or abrupt shot transition (AST) and gradual shot transition (GST). The efficiency of any shot detection technique primarily relies on three parameters namely (i) feature extraction, (ii) development of the inter frame similarity measure, and (iii) the criteria of threshold selection.

A remarkable number of works in the fields of shot boundary detection schemes can be found in [3–6]. The most clean way to find the discontinuity in visual content is to compare the nearby frames based on the intensity values [1]. But this shows high sensitivity toward large intensity variation and motion. Another popular, yet simple way to find the visual discontinuity is the difference of the histogram feature. In histogram based approach the histogram of the nearby frames is compared for AST detection. Since histogram evaluation depends only on frequency of occurrence of the pixels and ignores the spatial distribution of pixels, hence the possibility of two different frames having similar histogram cannot be avoided. Although it is a rare situation, but if present produces missed transitions. Many authors proposed edge and gradient feature [7–10] based approaches for SBD. Miller et al. [10] obtained inter frame similarity based on edge features using edge change ratio. Edge feature-based approaches have relatively higher insensitivity to slow light variation, but for videos involving high motion or large variation in illumination the edge features are lost, producing false transitions. Yoo et al. [7] used the distribution of edge pixel variance of frame sequence for exclusive detection of GSTs. They approximated the variance sequence to an ideal curve by extracting unique parabolic sequence of the local regions of the video frames. Motion-based approaches for SBD have also been exploited in literature [11, 12]. Moreover, many authors tried to combine the strengths of multiple features [13–15] for AST detection. Recently machine learning based approaches are gaining popularity due to their performances. Chasanis et al. [16] in 2009, proposed color histogram-based feature extraction followed by modified  $\chi^2$  square metric between frames at a frame separation of 1, 2 and 6 units. This is subsequently concatenated to form the final feature vector and fed to a SVM based classifier. Kar and Kanungo [17, 18] exploited spatial correlation among pixels which is proven to be good in handling videos with sudden light variation. Therefore, in this communication we proposed a new texture based similarity measure suitable for AST detection in presence of sudden illumination variation and proved to exhibit higher efficiency as compared to the LBP based approach.

The remaining paper is framed as follows. The basics of LDP feature, LDP feature descriptor generation are presented in Sect. 2. The strength of the proposed LDP feature to nonuniform illumination variation is presented in Sect. 3. The proposed algorithm is described in Sect. 4. The results and discussions followed by conclusions are presented in Sects. 5 and 6 respectively.

## 2 Local Directional Pattern (LDP)

The local binary pattern (LBP) is found to be a potential intensity invariant texture descriptor [19]. The LBP operator compares every individual pixel with the intensity of its neighboring pixels and subsequently generates a new binary number. If the intensity of the pixels in the neighborhood is same or higher than the center pixel value then a binary one is generated else a binary zero is generated. It favorably encodes the micro level information of local features such as edges, spots present in an image extracted from the intensity change information around different pixels. In contrast to LBP, Local directional pattern (LDP) [20, 21] is a comparatively improved local descriptor that uses Kirsch compass kernels to incorporate a directional component. It was proven to have comparatively lesser susceptibility to noise than the traditional LBP operator.

LDP code evaluates the textural feature of an image through edge response values obtained in different directions. It has successfully applied in the literature for face recognition and gender classification problem. LDP is a dense descriptor. It is a binary code of eight bits assigned to individual pixel of an image. It is evaluated through comparison of the relative edge response values of a pixel in 8 different directions. The directional edge response values of a particular pixel in 8 directions are obtained through masking operation using the eight *Kirsch directional masks* in 8 different orientations ( $M_1$ – $M_8$ ). The masks in various orientations are illustrated in Fig. 1. These responses are denoted by  $m_1, m_2, \dots, m_8$ . The response values do not possess equal significance in all the directions. Each response corresponds to an edge feature in respective direction. The edge or corner regions have comparatively higher response in specific directions. The 4 largest directional bit response values are encoded as 1 and the leftover bits are encoded as 0. The final LDP code is the decimal equivalent of the corresponding 8 bit binary code. Figure 2a shows the position of pixels in a  $3 \times 3$  neighborhood with center pixel position 0 and the position of eight neighboring pixels are  $p_1, p_2, \dots, p_8$ . Figure 2b shows the intensity of all the pixels in the  $3 \times 3$  neighborhood.

The response of the  $k$ th mask over a center pixel is given by  $m_k$

$$\begin{array}{cccc}
 \begin{matrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{matrix} &
 \begin{matrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{matrix} &
 \begin{matrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{matrix} &
 \begin{matrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{matrix} \\
 \text{East } M_1 & \text{North East } M_2 & \text{North } M_3 & \text{North West } M_4 \\
 \\ 
 \begin{matrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{matrix} &
 \begin{matrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{matrix} &
 \begin{matrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{matrix} &
 \begin{matrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{matrix} \\
 \text{West } M_5 & \text{South West } M_6 & \text{South } M_7 & \text{South East } M_8
 \end{array}$$

**Fig. 1** Kirsch masks in 8 orientations

$p_1$	$p_2$	$p_3$	82	38	25
$p_8$	$p_0$	$p_4$	57	40	20
$p_7$	$p_6$	$p_5$	50	35	46

(a)  $3 \times 3$  neighborhood positions

(b)  $3 \times 3$  neighborhood gray values

**Fig. 2**  $3 \times 3$  neighborhood positions and a sample image patch

$$m_k(x, y) = \sum_{n=1}^8 w_k(n) g_{(x,y)}(n), \quad (1)$$

where  $w_k(n)$  represents the weight of the  $n$ th position in the  $k$ th mask and  $g_{(x,y)}(n)$  represents the corresponding gray value of the pixel at positions in the neighborhood of  $(x, y)$ . The corresponding LDP value at  $(x, y)$  is evaluated as

$$\text{LDP}(x, y) = \sum_{n=0}^7 b_n(x, y) \times 2^n \quad (2)$$

$$b_n(x, y) = \begin{cases} 1 & \text{rank}(m_k(x, y)) \leq 4 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Where mask having the highest response value is ranked as 1. Similarly mask having lowest response is ranked as 8. The 8 bit binary code is generated by setting 4 largest ranks as one and the 4 smallest ranks as zero. The corresponding decimal value gives the LDP feature value of the pixel. In this process a new feature value is generated for every pixel value. The response of the 8 Kirsch masks, rank of the mask response values, corresponding assigned binary code and the final LDP code for a  $3 \times 3$  example neighborhood (Fig. 2b) is presented in Table 1.

**Table 1** Illustration of LDP code generation of a sample pixel value in a  $3 \times 3$  neighborhood for the center pixel of image in 2(b)

Mask index	$m_8$	$m_7$	$m_6$	$m_5$	$m_4$	$m_3$	$m_2$	$m_1$
Mask value	-251	-11	77	453	357	101	-395	-331
Rank of the position	6	5	4	1	2	3	8	7
Binary code	0	0	1	1	1	1	0	0
LDP code					60			

### 3 Materials and Methods

Since edge response values have comparatively better stability than intensity feature, hence LDP generates the same pattern even under noise and nonuniform illumination condition. Therefore LDP feature based scene representation can help to reduce the false scene change detection due to nonuniform illumination variation and noise between consecutive frames. The strength of the LDP feature over LBP feature is demonstrated in Fig. 3 using a sample  $3 \times 3$  image which is affected by sudden flash light in Fig. 3b and noise in Fig. 3c. In this example we considered an increment of intensity value by 10 units in each pixel value due to sudden flash light which is shown in Fig. 3b. After evaluation of the LBP and LDP values for 3(a), 3(b) and 3(c) images, it is observed that LDP value is not affected by sudden increment of intensity values of all the pixel values or by the effect of arbitrary noise. Therefore it motivated us to develop a similarity index based on LDP feature for AST detection to address the issues like flash light and noise. The proposed LDP based AST detection method has three steps, (i) LDP feature frame generation, (ii) Development of the similarity index and (iii) threshold based classification. Using the process defined in Sect. 2 each gray frame of the video is converted to LDP feature frame. The similarity measure can be evaluated as absolute sum of the difference of the successive feature frames. However, under high object motion it may produce high response resulting in false detection. Therefore instead of considering absolute difference of the feature frames we considered block based LDP feature descriptor for representing the visual content and obtained the absolute sum of the difference between the consecutive frames block-based LDP feature descriptor as a similarity index. The block based feature descriptor generation is described in next subsection.

<b>82</b>	<b>38</b>	<b>25</b>	<b>82+10</b>	<b>38+10</b>	<b>25+10</b>	<b>82-3</b>	<b>38+29</b>	<b>25-7</b>
<b>57</b>	<b>40</b>	<b>20</b>	<b>57+10</b>	<b>40+10</b>	<b>20+10</b>	<b>57+12</b>	<b>40+9</b>	<b>20+5</b>
<b>50</b>	<b>35</b>	<b>46</b>	<b>50+10</b>	<b>35+10</b>	<b>46+10</b>	<b>50+5</b>	<b>35+5</b>	<b>46-3</b>
<b>LBP code=10001011=139</b>			<b>LBP code=10001011=139</b>			<b>LBP code=11000011=195</b>		
<b>LDP code=00111100=60</b>			<b>LDP code=00111100=60</b>			<b>LDP code=00111100=60</b>		
(a) $3 \times 3$ neighborhood			(b) $3 \times 3$ neighborhood affected by flash light			(c) $3 \times 3$ neighborhood affected by arbitrary noise		

**Fig. 3** Illustration of the strength of the LDP feature

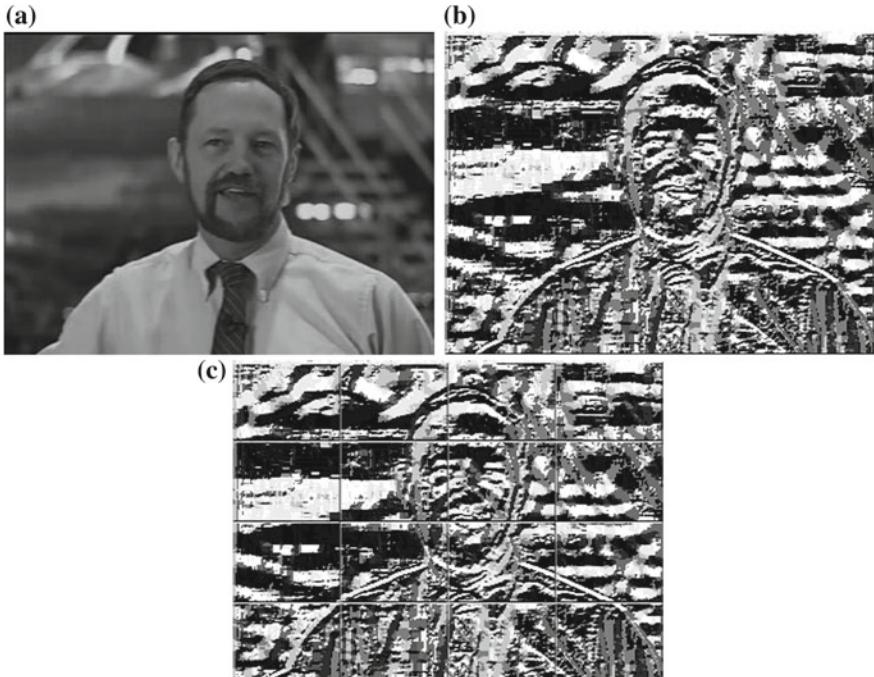
### 3.1 Block Based LDP (BBLDP) Descriptor

Each LDP feature frame is equally divided into 16 non overlapping blocks. Figure 4a shows the 11,000th frame of the video “NAD 57”. Figure 4b indicates its corresponding LDP feature image. Figure 4c shows the corresponding 16 non overlapping block LDP feature image. The histogram of every block is evaluated as follows.

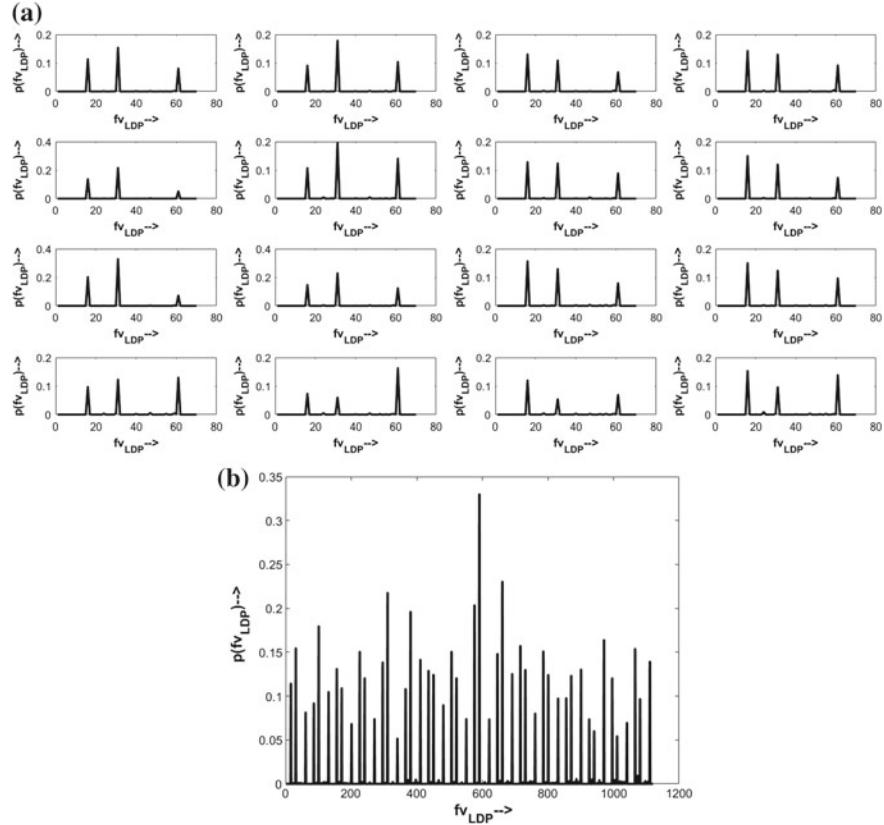
$$h_r(i) = \frac{n_i}{M_1 \times N_1}, \quad (4)$$

where  $r = 1, 2, 3, \dots, 16$ ;  $i = 0, 1, 2, \dots, n_i$  is the number of  $i$ th LDP value.  $M \times N$  represents size of the frame,  $M_1 \times N_1$  represents the size of each block such that  $M_1 = \frac{M}{16}$  and  $N_1 = \frac{N}{16}$ . The number of possible combinations of values in the frame representation by considering 1 at only four positions is 70. So a 70 bin histogram is generated for each block. The histogram of each block of Fig. 4c is shown in Fig. 5a. The LDP feature descriptor is generated by concatenating the histogram of each block as follows.

$$\mathbf{L}_{\text{DK}} = [\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \dots, \mathbf{h}_{16}] \quad (5)$$



**Fig. 4** **a** 11,000th frame of video “NAD 57” in Table 1, **b** LDP image of (a). Block wise LDP image of (a)



**Fig. 5** a Block wise LDP histogram of the sample frame considered in Fig. 4a, b. Concatenated LDP histogram of the blocks of the image in Fig. 4a

The length of the feature descriptor for each frame will be  $16 \times 70 = 1120$ . Figure 5b shows the feature descriptor  $L_{DK}$  for the image shown in Fig. 4a. The similarity index  $S(t)$  is developed based on the absolute sum of the LDP descriptor values as follows.

$$S(t) = \sum_{i=1}^{1120} |L_{DK(t)}(i) - L_{DK(t+1)}(i)| \quad (6)$$

If  $S(t) > Th$  then there exist an AST between  $t$ th and  $(t + 1)$ th frame, else both frames are within shot frames, where  $Th$  is the threshold.

## 4 Demonstration of the Proposed BBLDP Feature Based AST Detection Algorithm

In this section the LDP feature based AST detection algorithm is described.

### 4.1 BBLDP Based AST Detection Algorithm

The steps for the BBLDP based AST detection algorithm are as follows:

1. Initially all the frames are converted into gray frames.
2. For each gray frame, LDP feature frame is evaluated and divided into 16 equal sized non overlapping blocks (Fig. 4c).
3. For each block of the frames 70 bin LDP histogram (Fig. 5a) is obtained and they are concatenated row wise to build final frame feature descriptor known as block based LDP (BBLDP) feature descriptor (Fig. 5b).
4. Then the absolute value of the difference between consecutive frames LDP feature descriptor is evaluated which is then summed to obtain the similarity value for AST detection.
5. Threshold value is evaluated by taking mean and standard deviation of the similarity measure as given in (10).
6. The similarity value is compared to the threshold for identification of the final AST position.

## 5 Simulations and Discussions

For validation of the proposed method, we selected 10 videos collected from well known database TRECVID 2001 [22] and publicly available videos, comprising of a total of 78447 frames and 437 ASTs. The ground truth transition information of the test videos are obtained manually. The complete information about the test videos, video id, their ground truth transitions and their sources are tabulated in Table 1. The proposed BBLDP method is compared with intensity difference (ASID) method [1], the histogram difference (ASHD) method [1] and LBP based [17] method. The performances of the proposed BBLDP method is evaluated by Recall(Rec), Precision(Pr) and  $F_1$  parameters [2] defined by (7), (8) and (9) respectively.

$$\text{Rec} = \frac{\text{AT}}{\text{AT} + \text{MT}} \quad (7)$$

$$\text{Pr} = \frac{\text{AT}}{\text{AT} + \text{WT}} \quad (8)$$

$$F1 = \frac{2 \times \text{Rec} \times \text{Pr}}{(\text{Rec} + \text{Pr})}, \quad (9)$$

where the actual number of transitions detected by the algorithm is denoted as AT, number of transitions missed by the algorithm is denoted as MT and the number of transitions wrongly detected by the algorithm are denoted by WT. Rec, Pr and  $F1$  values can take a minimum of 0 to a maximum value of 1. When all transitions are correctly detected without any false or missed transition it takes a value of 1. The threshold value used for detecting ASTs for all the methods are as given by Zhang et al. [1].

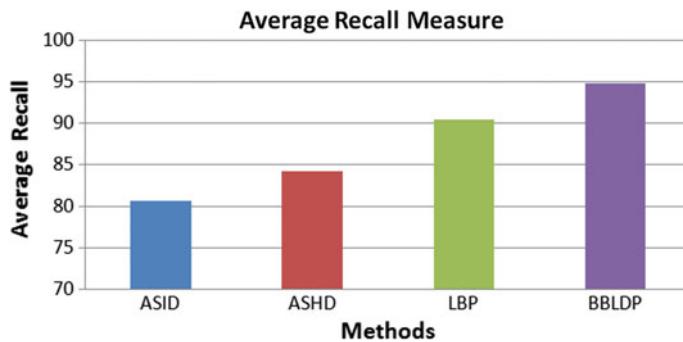
$$\text{Th} = m_s + \lambda \times s d_s$$

$$= \frac{1}{F_N - 1} \sum_{t=1}^{F_N-1} S(t) + \lambda \left[ \frac{1}{F_N - 1} \sum_{t=1}^{F_N-1} \left( S(t) - \frac{1}{F_N - 1} \sum_{k=1}^{F_N-1} S(k) \right)^2 \right]^{\frac{1}{2}} \quad (10)$$

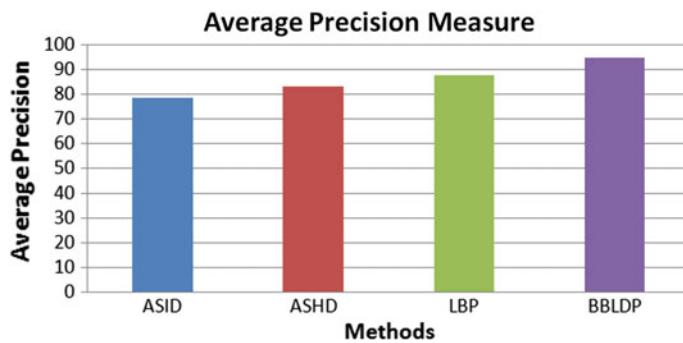
where Th is the threshold,  $S(t)$  is the feature similarity value for the  $t$ th frame,  $m_s$  and  $s d_s$  represent mean and standard deviation of the feature similarity index respectively,  $F_N$  is the total number of frames in a video and  $\lambda$  is a constant. Empirically it's value is set between 3 and 7. The performance comparison of the proposed BBLDP based method with ASID, ASHD and LBP based methods are presented in Table 2. It is observed from Table 2 that, the proposed BBLDP method has the highest performance in case of  $v_3$ ,  $v_4$ ,  $v_5$ ,  $v_7$ ,  $v_8$ ,  $v_9$  and  $v_{10}$  videos in comparison to other videos. The average performance of the proposed BBLDP method is found to be nearly 95%. The average performance measures are represented by bar charts in Figs. 6, 7, 8 respectively. It is observed from the Figs. 6, 7 and 8 that, in terms of average performance measure BBLDP has the highest Rec, Pr and  $F1$  values and

**Table 2** Information about test videos

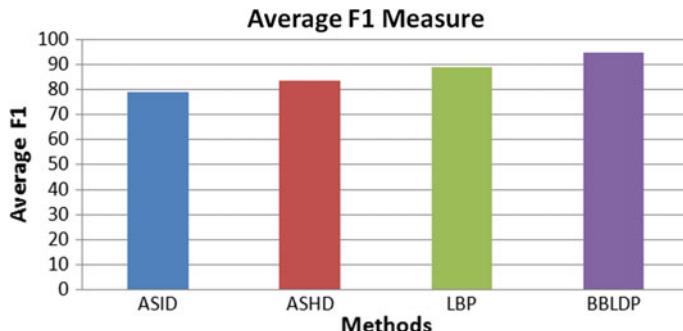
Video ID	Video name	Frames count	AST count	Sources
$V_1$	Little miss sunshine	4665	34	CD
$V_2$	2 Brothers	4798	38	CD
$V_3$	BS1	6146	25	CD
$V_4$	The big bang theory	14261	153	Sitcom
$V_5$	Soccer	4159	14	UEFA
$V_6$	Cartoon	3000	39	Youtube
$V_7$	anni 004	3895	13	TRECVID 2001
$V_8$	anni 005	11363	39	TRECVID 2001
$V_9$	NAD 57	12510	42	TRECVID 2001
$V_{10}$	NAD 58	13650	40	TRECVID 2001
	Total	78447	437	



**Fig. 6** Average recall measure



**Fig. 7** Average precision measure



**Fig. 8** Average F1 measure

LBP feature histogram based approach has the second highest Rec, Pr and *F*1 values among all the methods considered for comparison (Table 3).

**Table 3** Comparison of performance measures of ASID, ASHD, LBP and proposed LDP based methods

Video ID	ASID [1]			ASHD [1]			LBP [17]			Proposed BBLDP based method		
	Rec	Pr	F1	Rec	Pr	F1	Rec	Pr	F1	Rec	Pr	F1
$V_1$	70.59	72.73	71.64	88.24	90.91	89.55	<b>91.18</b>	<b>91.18</b>	88.24	88.24	88.24	88.24
$V_2$	78.95	46.15	58.25	92.11	89.74	<b>90.91</b>	<b>94.74</b>	81.82	87.80	89.47	<b>91.89</b>	90.67
$V_3$	96	92.31	94.12	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
$V_4$	98.69	94.37	96.49	90.85	95.21	92.98	94.77	92.95	93.85	<b>99.35</b>	<b>97.44</b>	<b>98.38</b>
$V_5$	78.57	68.75	73.33	92.86	92.86	92.86	92.86	92.86	92.86	<b>100</b>	<b>100</b>	<b>100</b>
$V_6$	56.41	81.48	66.67	<b>100</b>	90.70	<b>95.12</b>	94.87	<b>94.87</b>	94.87	94.87	88.10	91.36
$V_7$	92.31	75.00	82.76	84.62	84.62	84.62	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
$V_8$	79.49	83.78	81.58	43.59	42.50	43.04	92.31	85.71	88.89	<b>97.44</b>	<b>100</b>	<b>98.7</b>
$V_9$	76.19	84.21	80	71.43	71.43	80.95	75.56	78.16	90.48	<b>90.48</b>	<b>90.48</b>	<b>90.48</b>
$V_{10}$	80	86.49	83.12	77.50	72.09	74.70	62.5	62.5	87.50	<b>92.11</b>	<b>89.74</b>	<b>89.74</b>
<i>Avg</i>	80.72	78.52	78.8	84.12	83	83.6	90.41	87.74	89.01	<b>94.73</b>	<b>94.82</b>	<b>94.75</b>

## 6 Conclusions

In this communication, we presented a novel framework for abrupt scene change detection. Toward this goal the novelty lies in developing a new inter frame similarity measure based on the block LDP feature histogram for precise detection of boundaries between shots under sudden change in illumination and nonuniform noise such as nonuniform intensity variation condition. The proposed framework assumes high similarity between frames within a shot and discontinuity in inter frame similarity measure values for frames belonging to different shots. The proposed method is simulated on 10 test videos. The average performance of the proposed method is found to be closer to 95 percent. Promising results are in favor of the proposed algorithm. In the current scope of the work the parameter used for threshold evaluation is experimentally selected. As a future scope of the work, effort may be given to automate the threshold selection process with more test videos under challenging environment of high motion along with sudden illumination variation issues. Moreover, new models can be explored further to combine LDP features with other existing features for GST detection.

## References

1. Zhang, H. J., Kankanhalli, A., & Smoliar, S. W. (1993). Automatic partitioning of full motion video. *Multimedia Systems*, 1(1), 10–28. Jan.
2. Gargi, U., Kasturi, R., & Strayer, S. (2000). Performance characterisation of video shot change detection methods. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1), 1–13.
3. Cotsaces, C., Nikolaidis, N., Pitas, I. (2006). Video shot detection and condensed representation. a review. *IEEE Signal Processing Magazine*, 23(2), 28–37.
4. Koprinska, I., & Carrato, S. (2001). Temporal video segmentation: A survey. *Signal Processing: Image Communication*, 16(5), 477–500. January.
5. Hanjalic, A. Shot boundary detection: Unraveled and resolved. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2), 90–105.
6. Boreczky, J. S., & Rowe, L. A. (1996). Comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, 5(2), 122–128. April.
7. Yoo, H. W., Ryoo, H. J., & Jang, D. S. (2006). Gradual shot boundary detection using localised edge blocks. *Multimedia Tools and Applications*, 28(3), 283–300. March.
8. Ling, X., Chao, L., Huan, L., & Zhang, X. (2008) A general method for shot boundary detection. *Proceedings of International Conference of Multimedia and Ubiquitous Engineering* (pp. 394–397), 24–26 April 2008, Busan, Korea.
9. Adjeroh, D., Lee, M. C., Banda, N., & Kandaswamy, U. (2009). Adaptive edge-oriented shot boundary detection. *EURASIP Journal on Image and Video Processing*.
10. Zabih R., Miller, J., & Mai, K. (1995). A feature based algorithm for detecting and classifying scene breaks. In *Proceedings of the Third ACM International Conference on Multimedia* (pp. 189–200), San Francisco, California, USA.
11. Amel, A. M., Abdessalem, B. A., & Abdellatif, M. (2010). Video shot boundary detection using motion activity descriptor. *Journal of Telecommunication*, 2(1), 54–59. April.
12. Murai, Y., & Fujiyoshi, H. (2008). Shot boundary detection using co-occurrence of global motion in video stream. In *Proceedings of the 19th ICPR* (pp. 1–4). 23 January 2009.

13. Kawai, Y., Sumiyoshi, H., & Yagi, N. Shot boundary detection at TRECVID 2007. In *Proceedings of the TRECVID Workshop* (pp. 1–8).
14. Lian, Shiguo. (2011). Automatic video temporal segmentation based on multiple features. *Soft Computing*, 15(3), 469–482. March.
15. Lakshmi Priya, G. G., & Dommic, S. (2014). Walsh-Hadamard transform kernel-based feature vector for shot boundary detection. *IEEE Transactions on Image Processing*, 23(12), 5187–5197.
16. Chasanis, V., Likas, A., & Galatsanos, N. (2009). Simultaneous detection of abrupt cuts and dissolves in videos using support vector machines. *Pattern Recognition Letters*, 30(1), 55–65.
17. Kar, T., Kanungo, P. (2015). A texture based method for scene change detection. *2015 IEEE Power, Communication and Information Technology Conference (PCITC)* (pp. 72–77), 15–17 October, India.
18. Kar, T., & Kanungo, P. (2015). Cut detection using block based centre symmetric local binary pattern. In *2015 International Conference on Man and Machine Interfacing (MAMI)* (pp. 1–5). 17–19 December 2015.
19. Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern analysis and machine intelligence*, 24(7), 971–987.
20. Chakraborti, T., McCane, B., Mills, S., & Pal, U. (2018). Loop descriptor: Local optimal oriented pattern. *IEEE Signal Processing Letters*, 25(5), 635–639.
21. Jabid, T., Kabir, M. H., & Chae, O. (2010). Gender classification using local directional pattern (LDP). *2010 International Conference on Pattern Recognition, Istanbul, Turkey* (pp. 2162–2165), August 2010.
22. The open video project. <http://www.open-video.org>. Accessed March 2014.

# A Framework for Web Archiving and Guaranteed Retrieval



A. Devendran and K. Arunkumar

**Abstract** As of today, ‘web.archive.org’ has more than 338 billion web pages archived. How many of those pages are 100% retrieval. How many of the pages were left out or ignored just because the page doesn’t have some compatibility issue? How many of them were vernacular language and encoded in different formats (before UNICODE is standardized)? If we are talking about the content-type text. Consider other mime types which were encoded and decoded with different algorithms. The fundamental reason for this lies with the fundamental representation of digital data. We all know a sequence of 0 s and 1 s doesn’t make proper sense unless it is decoded properly. At the time of archiving, the browsers which could have rendered properly might have gone obsolete or upgraded way beyond to recognize old formats or the browser platforms could have been upgraded to recognize old formats. We studied various data preservation, web archiving related works and proposed a new framework that could store the exact client browser details (user-agent) in the WARC record and use it to load corresponding browser @ client side and render the archived content.

**Keywords** Personal data · Web archiving · Guaranteed retrieval

## 1 Introduction

All the history, culture, inventions, philosophies we have today are documented sometime in the past and preserved till date by someone. It is the foremost duty of any government to ensure the safety of these artefacts. People might think that taking a picture or video of the item and putting on internet should be sufficient to retain it. But it is even more dangerous than maintaining it in original medium. What

---

A. Devendran (✉)

Dr. M.G.R. Education and Research Institute, Chennai, India

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

K. Arunkumar

Technical Architect, Ppltech, Chennai, India

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

if the storage (Hard disk—SSD or DVD or something else) becomes inaccessible in 10 years from now. Even bigger problem is, what is the guarantee that the data-formats and softwares rendering them will be supported after 10 years. There lies the digital data preservations core problem.

### 1.1 Digital Data Preservation

In his interview on BBC @ 13 February 2015—Vint Cerf, one among 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 strategies and innovation, future citizens could be locked out of accessing today’s digital content enter a ‘digital dark age’. Figure 1 plots the life expectancy of the media storage types and amount of data they can store and its lifetime. Clearly it is alarming. Even if with digital representation we can store huge amount of information in small space—what is the use if we cannot retrieve them?

Also, the sustainable maintenance cost of digital data is considerably high compared to physical representation. For a while digitally born medias(example digital born movies) were converted to physical media and archived.

The fundamental problem with digital data is ‘data is not just data. It is data + metadata’. The archival problem is storing data + metadata of data + metadata of metadata of data + metadata of metadata of metadata and so on. It is more of a deeper recursive problem if you want to solve it completely. Nevertheless we can have some assumption about some metadata interpretation is very standardized and do our archival around it. It is really a complex problem and many research work happens in this area [1]. Popular archival systems available today fall under emulation or migration strategy. There is other category of people, who believe in the internet existence forever and archive all the data with internet technology. Their argument is that there will be browsers or browser-like systems which can interpret HTML, Javascript, CSS and will always exist to view all the archived data.

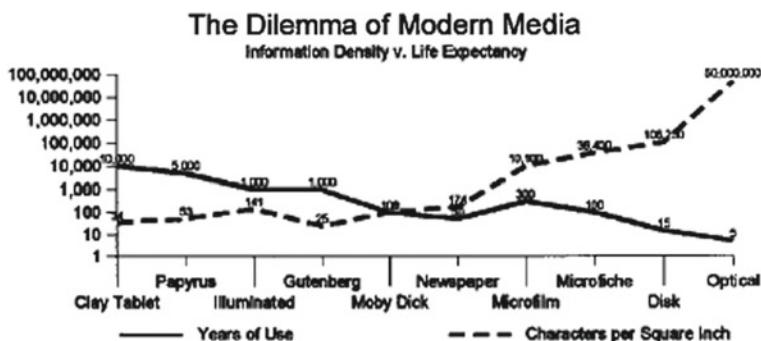


Fig. 1 Modern media—issue

## ***1.2 Web Archive***

People thought about preserving the data in web at the time of birth of WWW sometime around 1996. It is definitely beyond just bookmarking the URL, the issue about missing URLs had been studied thoroughly. Technically all the static files rendered by browser have to be archived. The process of crawling the website and storing the raw HTML, CSS, JS data is called ‘Web Archiving’. Also, the data is indexed based on URL and content. There are lot of web archiving systems around the world and the largest among them is ‘web.archive.org’ (previously internetarchive.org) claim to have archived 338 billion web pages as of today (30 September 2018).

International Internet Preservation Consortium (IIPC), formed in 2003, has greatly facilitated international collaboration in developing standards and open source tools for the creation of web archives. These developments, and growth of web make it inevitable that more and more libraries and archives will have to face the challenges of web archiving. Commercial web archiving software and services are also available to organizations who need to archive their own web content for corporate heritage, regulatory or legal purposes.

## ***1.3 Wayback Machine***

From 1996 Internet Archive organization has been archiving web pages of websites. After 5 years (in 2001) the data was opened to public with an interface named as ‘Wayback Machine’. It is named after an animated cartoon show called ‘The Rocky and Bullwinkle Show’. To put it formally, the process of retrieval of information from the archived data is associated with component of web archiving called ‘Wayback Machine’. Retrieval can be done based on the URL or the content. Work on offline mode to reconstruct the archived data are happening.

## ***1.4 WARC Format***

At the time of creating ‘Internet Archive’ created a standard for representation of the archived web page—ARC format. Later it has been extended to form a new format WARC (Web ARChive) by adding some more metadata specific to HTTP protocol.

## ***1.5 Mememto (RFC 7089)***

If you are interested in science fiction movies and novels, the concept of time travel is something always fascinating. For me one of the biggest amazing change done to the

archiving system since its inception is the introduction of ‘Memento’ a time travel kind of setup in web world. There was a HTTP standard RFC-7089 for this HTTP-based Memento framework for recording the timestamp of request being served. Every resource (URL/URI) and their states are archived with some TimeMaps. This framework enables the enumeration of all the past states and allows to go to any of them which are frozen at that time. It is more like travelling to that time and viewing.

## 2 Problem Statement

There are lot of solutions for web archiving (public and private web pages) existing today. The solutions range from open source to closed source, community-based or commercial-based solutions. The de facto and biggest archive system as of today is ‘web.archive.org’ with its ‘Wayback Machine’ for replay. What we feel is still the ‘web.archive.org’ is not completely solving the problem.

### 2.1 Core Problem

Archival is missing crucial piece of metadata, which is the client browser rendering the archived data. In the web ‘User-Agent’ HTTP header carries this information.

**What impact It will have?** The browsers which could have rendered data properly might have gone obsolete or upgraded way beyond to recognize old formats or the browser platforms could have been upgraded to recognize old formats.

#### **For example: Website with vernacular language data: (non-Unicode)**

Before Unicode standardization, there were lot of encoding formats (extension of ASCII). Many websites use these encoded formats for their data. Such websites cannot be viewed without the supported encoding font available in the operating system. Also, simply reading and rendering data from ‘txt’ file will print junk characters and no one can make any meaning out of it. For example, consider the Non-Unicode encoding TSCII of Tamil language. The Unicode standardization for Tamil language came around 2006 (The Unicode 5.0 Standard (5 ed.), Upper Saddle River, NJ: Addison-Wesley, 2006, ISBN 0-321-48091-0 at p. 324). And its adoption took a while. Hence, all the Tamil websites archived before that will only show junk characters.

### 3 Related Work

#### 3.1 Survey on Web Archives

The authors studied state of art web archiving systems across globe. One of the key metrics from their study is that the no. of web archiving initiatives significantly increased from 2003 onwards. Most of these systems are archiving data of developed countries.

They propose the need to increase web archive system not just archiving but lot of processing and indexing to retrieve data to match modern day web search engines. Their results are listed in wikipedia page and being maintained by community.

#### 3.2 A Survey of Web Archive Search Architectures

Initial days of ‘Internet Archive’ indexed based on URLs, i.e. search is based on URLs only. A full text search is needed very soon when the archive goes beyond certain level. There have been a lot of systems out there to support the full text search feature. The authors proposed a web archive search architectures to ‘time-travel search’ to do a search within time frame given.

#### 3.3 A Method for Identifying Personalized Representations in Web Archives

Later part of 2000s sees a huge increase and adoption of WWW and more and more solutions are getting delivered through web. Naturally to give better experience websites or web applications started giving various personalized content, i.e a given URL could render 2 different data for 2 different users. The personalization could be based on user or at different abstract level (geographical region, device type, etc.). The authors proposed a working prototype to analyse WARC (Web ARChive) format files, inserts extra metadata establishing user-resource relationships and later modified Wayback Machine implementation to fetch the corresponding resource.

#### 3.4 A Quantitative Approach to Evaluate Website Archivability Using the CLEAR+ Method

With the adoption of ‘Website Archiving’ across the globe there are works happening to ensure the standards for archivability. The authors proposed a system CLEAR+

(Credible Live Evaluation Method for Archive Readiness) to systematically evaluate archivability of a given website. The parameters like issues HTML and CSS (which can be found by static analysis), HTTP performance, Media files formats archivability, Sitemap. They have analysed ‘archiveready.com’ (implementation of CLEAR+) performance and reports.

### ***3.5 A Framework for Aggregating Private and Public Web Archives***

Internet Archive works great for public web resources. But the private resources (like banking application) and personalized resources (like facebook) specific to users are not straight forward to archive, even if data is archived there are lot of privacy issues related in retrieving them. The authors propose a framework to aggregate all private, personal and public Web archives without compromising potential sensitive information. They proposed changes to Memento syntax and semantics to TimeMap maps’ private Web archive captures.

### ***3.6 OldWeb***

There is a open-source implementation ([oldweb.today](http://oldweb.today)) of a Remote Browser System. They have multiple implementations of opening old web pages in old browsers. The technology of giving the old browsers varies from VM (Virtual Machine), Containerized to Browser extensions.

## **4 Our Contribution**

We are proposing a framework that stores additional metadata to uniquely identify client browser. During retrieval time this information is used to load corresponding browser and render the original data.

### ***4.1 Archival System Changes***

We propose a change in capturing additional metadata of the client browser. What to capture? Our intent is that the additional metadata should uniquely identify the client browser at the time of archival request.



#### 4.1.1 How About ‘User-Agent’? What Is It?

Any HTTP request comes with ‘User-Agent’ header. This uniquely identifies browser. This ‘User-Agent’ string is used by web server to selectively serve content for appropriate browser. HTTP standard RFC 1945 mainly talk about this tailored content delivery.

#### 4.1.2 Where to Capture?

WARC record currently has a field ‘**HTTP-header-user-agent**’. Per the specification: this field is usually sent by the harvester along with each request. Currently sent to differentiate some personalized data. Ideally this seems to be the right place for capturing client user-agent.

We have captured the sample WARC record with ‘**HTTP-header-user-agent**’ currently showing the harvester or crawler capturing the website. If the crawler user-agent is important we might need to add a new field in WARC record ‘**client-user-agent**’

#### B.I.Example of ‘warcinfo’ record

WARC/1.1

WARC-Type: warcinfo

WARC-Date: 2016-09-19T17:20:14Z

WARC-Record-ID: <urn:uuid:d7ae5c10-e6b3-4d27-967d-34780c58ba39>

Content-Type: application/warc-fields

Content-Length: 381

software: Heritrix 1.12.0 <http://crawler.archive.org>

```

hostname: crawling017.archive.org
ip: 207.241.227.234
isPartOf: testcrawl-20050708
description: testcrawl with WARC output
operator: IA_Admin
http-header-user-agent:
Mozilla/5.0 (compatible; heritrix/1.4.0 +http://crawler.
archive.org)
format: WARC file version 1.1
conformsTo:
http://bibnum.bnf.fr/warc/WARC\_ISO\_28500\_version1-1\_latestdraft.pdf

```

#### B.4.Example of ‘request’ record

```

WARC/1.1
WARC-Type: request
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Warcinfo-ID: <urn:uuid:d7ae5c10-e6b3-4d27-967
d-34780c58ba39>
WARC-Date: 2016-09-19T17:20:24Z
Content-Length: 236
WARC-Record-ID: <urn:uuid:4885803b-eebd-4
b27-a090-144450c11594>
Content-Type: application/http; msgtype=request
WARC-Concurrent-To:<urn:uuid:92283950-ef2f-4
d72-b224-f54c6ec90bb0>
GET /images/logo.jpg HTTP/1.0
User-Agent: Mozilla/5.0 (compatible; heritrix/1.10.0)
From: stack@example.org
Connection: close
Referer: http://www.archive.org/
Host: www.archive.org

```

## 4.2 Retrieval System Changes: (Wayback Machine)

We believe emulation of the original environment to reproduce the original experience of data is the right way to guaranteed retrieval. The current ‘Wayback Machine’ part of ‘web.archive.org’ plainly relies on the browser(on which it is working) to render HTML, CSS, JS to give data back.

### What change do we propose?

We believe a system like OldWeb should be used to render the data in its original browser. The original browser information is available in the WARC record. We need to render the UI with the old browser engine @ server side and send the UI to

‘Wayback Machine’. Currently we used a commercial product (<https://turbo.net/>) to render old browsers with data.

### 4.3 *Library of Browsers*

One of the important aspects in our framework is to have the entire list of browser and its VM or containerized browser apps available for everyone. It is really important that it has to be maintained and kept updated. In the market we have lot of commercial products to mimic most of the browser versions on different platforms. (like <https://turbo.net/>, <https://www.browserstack.com/>, <https://www.browserling.com/>). Figure 2 lists browsers a commercial product supports for cloud testing.

## 5 Conclusion and Future Work

We have worked on the open source ‘Core Python Web Archiving Toolkit’ a.k.a. pywb (Python Wayback) tool for the proposed framework and observed the proposed emulation based solution works as expected. The rendering of non Unicode data is really an encouraging factor and unique to the proposed framework. This could not have been possible with existing ‘Wayback Machine’ technology.

The proof of concept is done with the rendering part from a commercial product (<https://turbo.net/>) to make it work. This component has to be built on top of open source solutions like OldWeb container solution. Also, the code changes need to be reviewed and merged with main branch of pywb tool. We wanted to extend the OldWeb work especially the container based bower library to match commercial product list and make them available for everyone. We wanted to try other possibilities of having the containerized rendering in the client-side either with a thick-client approach or a standalone desktop application connecting to archive system.

 iOS	v8.3 iPad Mini 2 iPad Air	iPhone 6 iPhone 6 Plus	v7 iPhone 5S iPad Mini 4th	iPhone 4S iPhone 5	iPad 3	v5.1 iPhone 4S iPad 3	v5 iPad 2
 Android	Samsung Galaxy S5 Galaxy S4 Galaxy S3 Galaxy S2 Galaxy Tab 2 Galaxy S5 Mini	Galaxy Tab 4 Galaxy Note 3 Galaxy Note 2 Galaxy Note 10.1 Galaxy Note	Amazon Kindle Fire 2 Kindle Fire HD 8.9 Kindle Fire HDX	Razr Droid Razr Razr Maxx HD	Google Nexus 9 Nexus 7 Nexus 6 Nexus 5 Nexus 4 Nexus	HTC One M8 One X Wildfire	Sony Xperia Tipo
 Windows 10	e 0 15		50 49 48 47 46 45 44 43 42 41 40 39 38 37	45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30			
 Windows 8.1	e 11		50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22	45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16		12.16 12.15 12.14 12.10 12	
 Windows 8	e 10		50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22	45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16		12.16 12.15 12.14 12.10 12	
 Windows 7	e 11 10 9 8		50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14	45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3.6		12.16 12.15 12.14 12.10 11.6	
 Windows XP	e 7 6		49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16	45 44 43 42 41 40 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 11 10 9 8 7 6 5 4 3.6 3		12.16 12.15 12.14 12.10 11.6	
 Mac OS X El Capitan	e 9.1		50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 14	45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4		12.15 12.14 12.12	
 Mac OS X Yosemite	e 8		50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 14	45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4		12.15 12.14 12.12	
 Mac OS X Mavericks	e 7.1		50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 14	45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4		12.15 12.14 12.12 12 11.6	
 Mac OS X Mountain Lion	e 6.2		49 48 47 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 27 26 25 24 23 22 21 20 19 18 17 16 14	45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4		12.15 12.14 12.12 12 11.6	
 Mac OS X Lion	e 6		49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 14	43 42 41 40 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3.6		12.15 12.14 12.12 12 11.6	
 Mac OS X Snow Leopard	e 5.1*		49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23	42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10		12.15 12.14 12.12 12 11.6	

Fig. 2 List of popular web browsers

## References

1. Arunkumar, K., & Devendran, A. (2019). Digital data preservation—a viable solution. In V. Balas, N. Sharma, & A. Chakrabarti (Eds.), *Data management, analytics and innovation. Advances in intelligent systems and computing* (Vol. 808). Singapore: Springer.
2. Ainsworth, S. G., Nelson, M. L., & Van de Sompel, H. (2015). Only one out of five archived web pages existed as presented. In *HT 2015 Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 257–266).
3. Alam, S., Kelly, M., Weigle, M. C., & Nelson, M. L. (2017). Client-side reconstruction of composite mementos using serviceworker. In *JCDL 2017 Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (pp. 237–240).
4. Gomes, D., Miranda, J., & Costa M. (2011). A survey on web archiving initiatives. In S. Gradmann, F. Borri, C. Meghini, & H. Schuldt (Eds.), *Research and advanced technology for digital libraries. TPDL 2011*. Lecture Notes in Computer Science (Vol. 6966). Berlin, Heidelberg: Springer.
5. [https://en.wikipedia.org/wiki/List\\_of\\_Web\\_archiving\\_initiatives](https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives).
6. Costa, M., Gomes, D., Couto, F. M., & Silva, M. J. (2013). A survey of web archive search architectures. In *WWW 2013 Companion Proceedings of the 22nd International Conference on World Wide Web* (pp. 1045–1050).
7. Kelly, M., Brunelle, J. F., Weigle, M. C., & Nelson, M. L. (2013). A method for identifying personalized representations in web archives. In *D-Lib magazine November/December 2013* (Vol. 19, No. 11/12).
8. Banos, V., & Manolopoulos, Y. (2015). A quantitative approach to evaluate Website Archivability using the CLEAR+ method. *International Journal on Digital Libraries*. <https://doi.org/10.1007/s00799-015-0144-4>.
9. Kelly, M., & Nelson, M. & Weigle, M. (2018). *A framework for aggregating private and public web archives* (pp. 273–282). <https://doi.org/10.1145/3197026.3197045>.
10. Old browsers—a open source tool with remote & containerized browser system by oldweb-today. <https://github.com/oldweb-today/browsers>.
11. WebRecorder pywb 2.0—core python web archiving toolkit for replay and recording of web archives. <https://github.com/webrecorder/pywb>.
12. Turbo.net—a Cloud infrastructure to run instantly on all your desktops, mobile devices applications remotely. <https://turbo.net/>.
13. WARC format 1.1—WARC (Web ARChive) file format for archiving websites and web data. <https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/>.
14. RFC 7089—HTTP framework for time-based access to resource states—Memento. <https://tools.ietf.org/html/rfc7089>.
15. RFC 1945—HTTP with user-agent specification. <https://tools.ietf.org/html/rfc1945>.

# Business Intelligence Through Big Data Analytics, Data Mining and Machine Learning



**Wael M. S. Yafooz, Zainab Binti Abu Bakar, S. K. Ahammad Fahad and Ahmed M. Mithon**

**Abstract** There is a huge amount of data creating during the fourth industry revaluation and the data are generating explosively by various fields of the Internet of Things (IoT). The organizations are producing and storing the huge amount of data into the data servers every moment. This data comes from social media, sensors, tracking, website, and online news articles. The Google, Facebook, Walmart, and Taobao are the most remarkable organizations are generating most of the data in the web servers. Data comes into three forms as structured (text/numeric), semi structured (audio, video, and image) and unstructured (XML and RSS feeds). A business makes revenue from the analysis of 20% of such data, which is a structured form while 80% of data is unstructured. Therefore, unstructured data contains valuable information that can help the organization to improve the business productive, better decision-making, extract the insights, new products and services and understand the market conditions in various fields such as shopping, finance, education, manufacturing, and healthcare. The unstructured data are needed to be analyzed and distribute in a structured manner, that is required information's are to be gathered through the data mining techniques are used to mining the data. In this paper, expose the importance of data analytics and data management for beneficial usage of business intelligence, big data, data mining and machine and data management. In addition, the different techniques that can be used to discover the knowledge and useful information from such data been analyzed. This can be beneficial for numerous users concern on text mining and convert complex data into meaningful information for researchers, analyst, data scientist, and business decision makers as well.

---

W. M. S. Yafooz (✉) · Z. B. A. Bakar · S. K. A. Fahad · A. M. Mithon  
Faculty of Computer and Information Technology, Al-Madinah International University, Kuala Lumpur, Malaysia  
e-mail: [wael.mohamed@mediu.edu.my](mailto:wael.mohamed@mediu.edu.my); [waelmohammed@hotmail.com](mailto:waelmohammed@hotmail.com)

Z. B. A. Bakar  
e-mail: [zainab.abubakar@mediu.edu.my](mailto:zainab.abubakar@mediu.edu.my)

S. K. A. Fahad  
e-mail: [fahad.wasd@gmail.com](mailto:fahad.wasd@gmail.com)

A. M. Mithon  
e-mail: [mithun\\_lonedies@yahoo.com](mailto:mithun_lonedies@yahoo.com)

**Keywords** Data mining · Business intelligence · Unstructured data · Structured information

## 1 Introduction

There is a huge amount of data creating during the fourth industry revaluation. Such data generates by human through social media (facebook, twitter, linked-in or Instagram) or by computer machine such as sensors, GPS, website or application systems [1]. The organizations are producing and storing the huge amount of data into the data servers every moment. This data comes from social media, sensors, tracking, website, and online news articles. The Google, Facebook, Walmart, and Taobao are the most remarkable organizations are generating most of the data in the web servers. Data comes into three forms as structured (text/numeric), semi structured (audio, video, and image) and unstructured (XML and RSS feeds). A business makes revenue from the analysis of 20% of such data which is a structured form while 80% of data is unstructured. Therefore, unstructured data contains valuable information that can help the organization to improve the business productive as well as significant for security, education, manufacturing, and healthcare as well. This can be achieved through big data analytics and data management in order to achieve the business intelligence.

The Business intelligence (BI) plays a vital role to help the decision maker to see the insights to improve productive or fast and better decision. In addition, BI can assist enhance the effectiveness of operational rules and its impression on superintendence systems, corporate-level decision-making, budgeting, financial and administration recording, making strategic choices in a dynamic business environment [2]. BI is the technologies, applications, and systems for the compilation, combination, analysis, and exhibition of the business report to help immeasurable with active business decision executing way for enforced to gain, learn and control their data to further decision-making in a plan to develop business procedures [3].

On the other hand, the big data management from diverse data formats is the main competition in business and as well as for management. The data management consists of serious management problem where current tools are not adequate to manage such massive data volumes [4]. The importance of typical big data management related to storage, pre-processing, processing and security, where new challenges in terms of storage capacity, data integration complexity, analytical tools and lack of governance. The big data management is a complex process, particularly in abundant data originated from heterogeneous sources that are to be used for BI and decision-making. A report stated on managing big data that 75% of organizations manage some form of big data. The aim of big data management is to ensure the effectiveness of big data storage, analytics applications and security [5].

This paper exposes the importance of data analytics and data management for beneficial usage of big data, and data mining and machine learning for BI and decision-making of management. In addition, the different techniques that can be

used to discover the knowledge and useful information from such data also been analyzed.

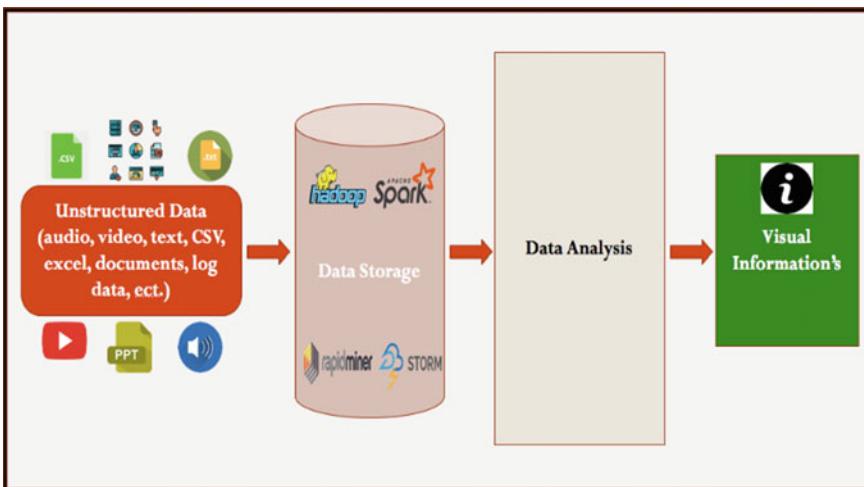
This paper organized as Sect. 2 demonstrates the big data architecture, while importance of BI highlighted in Sect. 3. Section 4 explains the big data analytics. The data mining techniques are described in Sects. 5 and 6 shows the steps of data mining. The conclusion of this paper in last section.

## 2 Big Data Architecture

Big data is used to describe the exponential growth of structured and unstructured data. The greatest big data challenge is that a large portion of it is not structured, often in the form of unstructured text. Therefore, there are several steps in order to handle such data. Big data architecture is the overarching system that a business uses to steer its data analytics work. The big data architecture is shown in Fig. 1. There are four logical layers that exist in big data architecture [6] as follows:

**Big Data Sources Layer:** This is the initial layer where data comes across organization records, social media, customer records, servers, sensors, internet logs, mobile devices etc. This accepts all structured and unstructured data types.

**Data Storage Layer:** From gathered sources the data are lived in this layer and converts to specific formats for tools accessibility. The huge amount of data can be accessible where the structured data stored in RDBMS and the unstructured in HDFS or no SQL database.



**Fig. 1** Big Data architecture

**Data Analysis/Processing Layer:** This layer includes the analysis or processing to get the data to be useful which interacts for the BI. Few tools are used in this layer to analyze them into a format such as MapReduce.

**Consumption/Output Layer:** After the analysis or processing data are prepared to visualize its information. The output can be charts, reports and figures as well depends on the requirement.

### 3 Importance of Business Intelligence

To manage and develop business from earlier data, each business requires to receive remarkable judgment from anything they have done previously. If management takes critique by BI than it will support the managerial group to consider relevant declarations. BI is the technologies, applications, and systems for the compilation, combination, analysis, and exhibition of the business report to help immeasurable with active business decision executing way for enforced to gain, learn and control their data to further decision-making in a plan to develop business procedures [3]. Therefore, BI can assist enhance the effectiveness of operational rules and its impression on superintendence systems, corporate-level decision-making, budgeting, financial and administration recording, making strategic choices in a dynamic business environment [2].

Modern businesses continue to use a strategy to leverage data (especially Big Data) and achieve a sustainable contentious advantage. By transforming raw data into presentable information and understandable knowledge through the utilization of the latest information technology that can be applied at a managerial level in the decision building. Businesses produce huge investments in BI systems to accomplish goal-oriented, modern, and sustainable competing for advantage and take possibly huge advantages as a result of certain expenses [7]. BI successfully ruling Retail Industry, Insurance, Banking, Finance & securities, Telecommunications, Manufacturing industry for appropriate data mining operation on remain data on different companies. Most important features are Analysis that supports cross selling and up selling, Customer segmentation and profiling, Analysis of Parameters Importance, Survival time analysis, Analysis of consumer loyalty and consumer switching to competition, Credit scoring, Fraud detection, Fraud Detection, Web-Farming (investigation of the Internet content) [8]. The business intelligence can be archive it's objectives using the big data analytics techniques.

### 4 Big Data Analytics

The big data analytics which is the machine learning techniques are needed due to datasets are often distributed and their size and privacy considerations evidence distributed techniques. The data resides on platforms with varying computational

and network capabilities. The benefits of big-data analytics and the diversity of application pose challenges. For example, Walmart servers handle every hour more than one million customer transactions, and these information's are inserted into databases with more than 2.5 petabytes of data, and this is the equivalent of 167 times the number of books in the Library of Congress. Herein, the Large Hadron Collider at CERN produces around 15 petabytes of data annually, this is enough to fill more than 1.7 million dual-layer DVDs per year [7]. The big data analytics are used for education, healthcare, media, insurance, manufacturing, and the government. Big data analytics has been evolved from business intelligence and decision support systems that enable healthcare organizations to analyze an immense volume, variety, and velocity of data across a wide range of healthcare networks to support evidence-based decision-making and action taking [9]. Therefore, from the discussion it's evident, the big data analytics and data management [10] is important in business intelligence for four reasons are:

First, **better decision-making**: Big data analytics can analyze past data to make predictions about the future. Thus, businesses can not only make better present decisions but also prepare for the future. Second, **Cost reduction**: Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data. In addition, provide insights on the impact of different variables. Third, **new products and services**: With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Big data analytics, more companies are creating new products to meet customers' needs. Fourth, **understand the market conditions**: By analyzing big data, you can get a better understanding of current market conditions.

In order to retrieve the significant information's, there are few challenges and features to be considered in big data analytics tools and techniques, and they are including, scalability, and fault tolerance as well [11–13]. Table 1 represented few widely used tools with the features they provide for big data analytics.

## 5 Data Mining Techniques

Picking the suitable data mining technique is one of the most significant responsibilities in the Data Mining Process. Nature of business and the kind of object or difficulty suffered in business provides relevant direction to determine the fittest technique [14]. Applying data mining techniques there is some generalized approach, it can be referring to enhance the efficiency and cost-effectiveness. Several core techniques that are performing in the data mining process, specify the character of the mining operation and reclamation option of data. The mining technique is highly productive on the result [15]. There are lot of techniques but among them, Association Rule, Classification, Clustering, Decisions Tree, and the Neural Networks are profoundly practical and successful.

**Association Rule:** Association (relation) is the usual simple and straightforward data mining procedure. It is too powerful and well-researched systems of data min-

**Table 1** Most common techniques of Big Data

Features/tools	Scalability	Fault tolerance	Visualization	Policy	Citations
Hadoop	Yes	Yes	Graphical format	Processing of big data	[29, 22]
Apache Spark	No	Yes	Through charts	Filters large scale data	[30]
Cassandra	Yes	Yes	Used tableau	NoSQL techniques	[31]
Apache SAMOA	Yes	No	Snapshots views	Mines repositories and presents	[32]
RapidMiner	Yes	Yes	Various output formats	Supports all mining process	[24]
NodeXL	Yes	No	Graph	Maps networks using excel	[33]
Apache Storm	Yes	Yes	Graph, Maps, Charts	Real-time computation	[34]
Hive	Yes	Yes	Graphical	Static data analyzes	[23]

ing. By extraction of interesting similarities, connections, common formations within collections of items in Database. Association is also particular to dependently associated variables. Because or running by determining the correlation among items in the related transaction, the association is more comprehended as relationship procedure [16]. By using association technique, retailers can perform analysis on customer's habits. Retailers might obtain an explanation from history of sales data, consumers who order drinks while they purchase fast-food. They can put them beside each other to conserve time for the client and increase sales. Association rule depends on two significant information. Those are; Help and Confidence. "How frequently is the rule implemented?" Is represent support and Confidence is "How frequently the rule is true?" [17].

$$\text{Support}(\text{FASTFOOD} \Rightarrow \text{DRINKS}) = \frac{\text{Number of time order DRINKS when purchase FASTFOOD}}{\text{Total number of Transection.}}$$

If Support (FASTFOOD  $\Rightarrow$  DRINKS) is 5%, it indicates that Consumer demands fast-food and drinks together 5% of total purchase in the history.

$$\text{Confidence}(\text{FASTFOOD} \Rightarrow \text{DRINKS}) = \frac{\text{Support}(\text{FASTFOOD} \cup \text{DRINKS})}{\text{Support}(\text{FASTFOOD})}$$

If in association rule Confidence (FASTFOOD  $\Rightarrow$  DRINKS) = 75%. That means, 75% customer order drinks when they purchase fast-food.

**Classification:** Classification is the most deliberate and generally practiced supervised learning data mining task. Given an object, allowing it to one of the predefined

target sections and classify individually in a set of data inside a predefined set of categories or collections described as classification. Classification is a complicated data mining procedure that overcomes to assemble multiple properties mutually into discernable divisions, which can apply to carry additional outcomes to accurately predict the target class for individual state. Classification algorithms attempt to define relations among properties in a training set to incorporate new observations and there have two principal rules in this method [14]. Learning—Data are examined by the analysis algorithm and model is created from the practice examples. Classification—In this rule, the data is applied to estimate the accuracy of the classification precepts and allow a label to an unlabeled analysis situation [18].

Naive Bayes classifier, Random Forest and AdaBoost are successful and rapidly utilized in classification data mining technique. Naive Bayes classifier is a great example of a classifier that estimates unknown conditional probabilities to recognize classes. It delivers successful results in medical diagnosis, banking, document categorization, marketing, and pattern recognition. Random Forest is an excellent training model including the goal of decreasing variance and building efficiency of performance. By ensemble learning approach with several training sets to classify input parameters for every tree in the forest. AdaBoost is the most suitable binary classification solution that connects a number of soft learners to perform better detachment between classes [19].

**Clustering:** Clustering is the concept to unite objects in clusters according to their similarity. Clustering is very similar to classification, but clustering is an unsupervised learning technique that grouping chunks of data together in meaningful clusters based on their similarities [20, 21]. As like classification, clustering groups those are named cluster and those are not described previously. It was majored by the specialty of data points and the relationships between the individual data points based on their properties. Clustering is a blinded unsupervised learning rule of the data mining process that can determine the correlation within data points based on the qualities it can understand. Sometimes, clustering called segmentation and it helps to understand, the changes happened in the database. Clustering algorithms are divide into meaningful groups.

There are several kinds of clustering methods including thousands of algorithm for different object. Most significant are; Partitioning, Hierarchical, Density-Based, Grid-Based, and Model-Based Methods. Clustering Algorithm is divided base on different types of clustering method [22]. (a) Partitioning Based: K-means, K-modes, K-medoids, PAM, CLARANS, CLARA, and FCM. (b) Hierarchical Based:BIRCh, CURE, ROCK, Chameleon, and Echidna. (c) Density-Based: DBSCAN, OPTICS, DBLASD, and DENCLUE. (d) Grid-Based: Wave-cluster, STING, CLIQUE, and OptiGrid (e) Model-Based: EM, COBWEB, CLASSIT, andSOMs.

**Decisions Tree:** Decision tree technique model is simple to learn for users. The decision tree can be utilized both as a component of the adoption patterns to establish the suitability and preference of particular data begin with a simple question that has two and more replies [23]. Each solution guides to an additional question to improve recognizing the data. This prediction can be performed based on any response determine the data, that can obtain the terminal determination. Several predictions might

be based on the historical practice that supports the structure of the decision tree frequently practiced with classification systems to associate standard information, and including predictive methods [14].

Decision trees produce a hierarchical partitioning of the data. Those several partitions at the leaf level to the several classes that produced by the application of a split basis. The separation principle stated on an individual attribute, or on multiple attributes. Once it connected as a univariate split and applied as multivariate split. The approach is to recursively break training data to maximize the difference in various classes over several nodes. The perception is many classes are maximized on various classes when the delivered node is maximized. The tree-shaped formation that describes collections of arrangements, tree nodes describe property value, the branch describes the result of the test and subsequently, tree leaves represent class relationships. The group instance begins at the source node and, depending on the results, regarding the proper parts till the leaf [23].

**Neural Networks:** Neural Network is an extensive technique applied in the starting stages of the data mining technology. Neural networks are automated to a remarkable extent and because the user is not required to have much knowledge regarding the database. Node and the Link are the two principal elements of the Neural Network technique [24]. Node, which coordinates to the neuron in the human brain and the Link, suits the connections among the neurons in the human brain. To execute neural network efficiently, three factors need to reflect. Wherewith the nodes are correlated? When should the training rule be suspended? And Number of processing units to be applied? [14].

The formation of neurons and their interconnections have described the architecture of the network and those interconnected are in the single or multiple layers. Every neural network model has distinctive architectures and those architectures use separate learning methods with their individual benefits and limitations. Neural networks is a forbidding modeling technique and some complicated models are impossible to understand completely. Therefore, to know the Neural network technique, there have two explications is recommended. Neural network must pack up and let to be practiced for a single application and bonded with skillful advising co-operation [25].

Forward and Backpropagation, Neural Networks hold by these two states. Continuously ultimate output activation function is frequently applied to produce the inputs to meet the class description in the forwarding condition. Final output at the output layer produces an error value and backpropagation states stat operation to updating of the weights in the prior layers are determined as a function of the errors and weights in the course before of it.

## 6 Steps of Data Mining

Data Mining is regarding interpreting the immense volume of data and extracting of knowledge from its several objects. Fundamentally, the chance of losing the rich and influential message carried by the massive databases was standing and this demands

the adoption of sufficient systems to gain beneficial data so that the scope of data mining had been developed in 1980s and is still advancing. The individual approach leads and produces the data mining assignments and its utilization [26]. For some businesses, the purposes of data mining recognize developing marketing abilities, identifying different trims, and predicting the prospect based on earlier observations and modern inclinations. As databases become extensive, it turns more challenging to maintain enterprise preparation. There is an audible demand to examine the data to sustain devising and additional purposes of an entrepreneur. Data mining could further continue practiced to recognize unusual performance. An intelligence agency could define a strange behavior of its representatives practicing some aforementioned technology [27].

There are extensive amounts of current and historical data existing and stored. Different standard models for data mining are proposed and some are established. All those models are described in subsequent steps. These steps support performing the data mining responsibilities. Three models are mostly followed by the data mining experts and researchers for data mining process and these types are; Knowledge Discovery Databases (KDD) process model, CRISP-DM and SEMMA. The Knowledge Discovery Databases (KDD) model to gaining knowledge in data and emphasizes the important level of particular data with nine steps. Cross Industry Standard Process for Data Mining (CRISP-DM) launched by Daimler with six steps or phases and improves over the years [28]. With five distinct phases identified as Sample, Explore, Modify, Model, Assess (SEMMA), this model was developed by SAS Institute Inc. Data Mining sometimes named to knowledge discovery database (KDD) due to, it is the method of examining data from different sources and comprehensions, and condensing it into knowledge that can be presentable, that knowledge can minimize loss and improve return or both [17]. See Table 2.

## 7 Conclusion

Business Intelligence is the technologies, applications, and systems for the compilation, combination, analysis, and exhibition of the business report. BI help immeasurable with active business decision executing way for enforced to gain, learn and control their data to further decision-making in a plan to develop business procedures. A business makes revenue from the analysis of 20% of such data which is a structured form while 80% of data is unstructured. Therefore, unstructured data contains valuable information that can help the organization to improve the business productive as well as significant for security, education, manufacturing, and health-care as well. So, the data management, data mining and machine learning techniques are required in order to extract the insights from huge amount of data. By using such techniques, business intelligence gets better decision-making, Cost reduction, new products and services and understand the market conditions.

**Table 2** Steps and key features of KDD

Steps	Key features		
KDD	CRISP-DM	SEMMA	
1. Learning and understanding of the application domain	<ul style="list-style-type: none"> <li>Defined objects based on the client's point</li> <li>State and assuming the purpose and principles</li> </ul>	<ul style="list-style-type: none"> <li>Uncovers factors like success patterns, enterprise, and data mining aspirations</li> <li>Learn the fundamentals of business terminologies and technical phases</li> </ul>	<ul style="list-style-type: none"> <li>Sampling data</li> <li>Portion took from a huge dataset to obtain meaningful knowledge</li> <li>Small enough to handle instantly</li> </ul>
2. Creating a target dataset	<ul style="list-style-type: none"> <li>Create a target dataset including the subset of data units which process will do performed</li> </ul>	<ul style="list-style-type: none"> <li>Data gathering, monitoring quality and examining of information form hypotheses for unexplained information</li> </ul>	<ul style="list-style-type: none"> <li>Exploration of data</li> <li>Expanding the understanding and conceptions</li> <li>Improving the development rule by combing for trends and irregularities</li> </ul>
3. Data cleaning and pre-processing	<ul style="list-style-type: none"> <li>Targeted data cleansing and pre-processing for data externally any noise and inequalities</li> <li>DBMS points are selected such as data schema, type of data, and mapping of dropping and unfamiliar values in the database</li> </ul>		
4. Data reduction and projection or data transformation	<ul style="list-style-type: none"> <li>Determining valuable characteristics and images to describe the data</li> <li>Prepossessed and transformed into a conventional format</li> <li>Transformation of data from one form to another so that data mining algorithms can be performed efficiently</li> </ul>	<ul style="list-style-type: none"> <li>Collection and development of the ultimate dataset</li> <li>Records, table and attributes assortment</li> <li>Cleaning and transformation of data</li> </ul>	<ul style="list-style-type: none"> <li>Modification of data by creating, selecting and transformation of variables</li> <li>Focus on the model selection process</li> <li>Seems for outliers and decreasing the number of variables</li> </ul>

(continued)

**Table 2** (continued)

Steps	Key features	KDD	CRISP-DM	SEMMA
5. Choosing the function suitable data mining task	<ul style="list-style-type: none"> <li>Appropriate data mining task is decided based on distinct intentions</li> <li>Determining the scope of the model assumed by the data mining algorithm</li> </ul>	<ul style="list-style-type: none"> <li>Determination and utilization of different modeling procedures</li> <li>Various models are constructed for the same data mining predicament by valuing separate parameters</li> </ul>	<ul style="list-style-type: none"> <li>Automatically explores for a sequence of data</li> <li>Various modeling methods are present and several types of the model have its individual concentration</li> <li>Relevant for the particular condition on the data for data mining</li> </ul>	
6. Choosing the suitable data mining algorithm(s)	<ul style="list-style-type: none"> <li>Appropriate data mining algorithms are picked for exploring various patterns from data</li> <li>Fitting algorithms are decided based on balancing the overall standards for data mining</li> </ul>			
7. Employing data mining algorithm		<ul style="list-style-type: none"> <li>Decided algorithms are performed on preprocessed and transformed data</li> <li>Exploring patterns of interest in a set of records or a particular representable structure</li> </ul>		(continued)

**Table 2** (continued)

Steps	Key features	CRISP-DM	SEMMA
KDD			
8. Interpreting mined patterns	<ul style="list-style-type: none"> <li>Interpretation and evaluation of mining patterns</li> <li>Associate in selected patterns visualization</li> <li>Eliminating redundant or irrelevant patterns, and transposing the useful ones into terms comprehensible by users</li> </ul>	<ul style="list-style-type: none"> <li>Evaluation of recovered models by determining the application of results</li> <li>Explanation of each models depends on the algorithm going to implement</li> </ul>	<ul style="list-style-type: none"> <li>Evaluation of the reliability and application of findings</li> <li>Evaluates the performance</li> </ul>
9. Using discovered knowledge	<ul style="list-style-type: none"> <li>Discovered knowledge is applied to different goals</li> <li>Discovered knowledge can apply involved individuals or can be integrated with another system for additional progress</li> <li>It to affected individuals, as well as monitoring for and determining possible conflicts with previously obtained knowledge</li> </ul>	<ul style="list-style-type: none"> <li>Determining the technique of gathering knowledge and decisions</li> <li>Organizing, reporting and presenting the obtained knowledge when required</li> </ul>	

## References

1. Yafooz, W. M. S., Abidin, S. Z., & Omar, N. (2011, November). Challenges and issues on online news management. In *2011 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 482–487). IEEE.
2. Richards, G., Yeoh, W., Chong, A. Y. L., & Popović, A. (2017). Business intelligence effectiveness and corporate performance management. An empirical analysis. *Journal of Computer Information Systems*, 1–9.
3. Balachandran, B. M., & Prasad, S. (2017). Challenges and benefits of deploying big data analytics in the cloud for business intelligence. *Procedia Computer Science*, 112, 1112–1122.
4. Yafooz, W. M. S., Abidin, S. Z., Omar, N. & Hilles, S. (2016, September). Interactive big data visualization model based on hot issues (online news articles). In *International Conference on Soft Computing in Data Science* (pp. 89–99). Singapore: Springer.
5. Siddiqa, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., et al. (2016). A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications*, 71, 151–166.
6. Borodo, S. M., Shamsuddin, S. M., & Hasan, S. (2016). Big data platforms and techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, 1(1), 191–200.
7. Sparks, B. H., & McCann, J. T. (2015). Factors influencing business intelligence system use in decision making and organisational performance. *International Journal of Sustainable Strategic Management*, 5(1), 31–54.
8. Qureshi, N. A., Khan, B. A., & Saif, J. A. (2017). Business intelligence systems in the holistic infrastructure development supporting decision-making in organisations.
9. Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126, 3–13.
10. Yafooz, W. M. S., Abidin, S. Z., Omar, N., & Idrus, Z. (2013, December). Managing unstructured data in relational databases. In *2013 IEEE Conference on Systems, Process & Control (ICSPC)* (pp. 198–203). IEEE.
11. Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences (HICSS)* (pp. 995–1004). IEEE.
12. Zhou, Z. H., Chawla, N. V., Jin, Y., & Williams, G. J. (2014). Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]. *IEEE Computational Intelligence Magazine*, 9(4), 62–74.
13. Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263–286.
14. Fahad, S. A., & Alam, M. M. (2016). A modified K-means algorithm for big data clustering. *International Journal of Computer Science Engineering and Technology*, 6(4), 129–132.
15. Fahad, S. A., & Yafooz, W. M. (2017). Design and develop semantic textual document clustering model. *Journal of Computer Science and information technology*, 5(2), 26–39. <https://doi.org/10.15640/jcsit.v5n2a4>.
16. Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: Literature review and challenges. *International Journal of Distributed Sensor Networks*, 11(8), 431047.
17. Birant, D., & Yıldırım, P. (2016). A Framework for data mining and knowledge discovery in cloud computing. In Z. Mahmood (Ed.), *Data science and big data computing: frameworks and methodologies* (pp. 245–267). Cham: Springer. [https://doi.org/10.1007/978-3-319-31861-5\\_11](https://doi.org/10.1007/978-3-319-31861-5_11).
18. Aggarwal, C. C. (2015). *Data classification: Algorithms and applications*. Boca Raton: CRC Press, Taylor & Francis Group.
19. Yafooz, W. M. S., Abidin, S. Z., & Omar, N. (2011, November). Towards automatic column-based data object clustering for multilingual databases. In *2011 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 415–420). IEEE

20. Yafooz, W. M. S., Abidin, S. Z., Omar, N., & Halim, R. A. (2014). Model for automatic textual data clustering in relational databases schema. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (pp. 31–40). Singapore: Springer.
21. Ali, S. M., & Tuteja, M. R. (2014). Data mining techniques. *International Journal of Computer Science and Mobile Computing*, 3(4), 879–883.
22. Kotwal, A., Fulari, P., Jadhav, D., & Kad, R. (2016). Improvement in sentiment analysis of twitter data using hadoop. *Imperial Journal of Interdisciplinary Research*, 2(7).
23. Bhawnani, D., Sanwlani, A., Ahuja, H., & Bohra, D. (2015). Big Data analytics on cab company's customer dataset using Hive and Tableau. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, 1(2), ISSN, 2395-3470.
24. Rangra, K., & Bansal, K. L. (2014). Comparative study of data mining tools. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(6).
25. Alton, L. (2017, December 22). The 7 most important data mining techniques. Retrieved October 2, 2018, from <https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques>.
26. Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 12(1), 217–222.
27. Thuraisingham, B. (2014). *Data mining technologies, techniques, tools, and trends*. CRC press.
28. Alhendawi, K. M., & Baharudin, A. S. (2014). A classification model for predicting web users satisfaction with information systems success using data mining techniques. *Journal of Software Engineering*.
29. Park, D., Wang, J., & Kee, Y. S. (2016). In-storage computing for Hadoop MapReduce framework: Challenges and possibilities. *IEEE Transactions on Computers*.
30. Pirozzi, D., Scarano, V., Begg, S., De Sercey, G., Fish, A., & Harvey, A. (2016, July). Filter large-scale engine data using apache spark. In *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)* (pp. 1300–1305). IEEE.
31. Santos, M. Y., e Sá, J. O., Andrade, C., Lima, F. V., Costa, E., Costa, C., ... & Galvão, J. (2017). A Big Data system supporting Bosch Braga Industry 4.0 strategy. *International Journal of Information Management*, 37(6), 750–760.
32. Minelli, R., & Lanza, M. (2013, September). SAMOA—A visual software analytics platform for mobile applications. In *2013 29th IEEE International Conference on Software Maintenance (ICSM)* (pp. 476–479). IEEE.
33. Yep, J., & Shulman, J. (2014). Analyzing the library's Twitter network: using NodeXL to visualize impact. *College & Research Libraries News*, 75(4), 177–186.
34. Raina, I., Gujar, S., Shah, P., Desai, A., & Bodkhe, B. (2014). Twitter sentiment analysis using apache storm. *International Journal of Recent Technology and Engineering*, 3(5), 23–26.

# The Effect of Big Data on the Quality of Decision-Making in Abu Dhabi Government Organisations



**Yazeed Alkatheeri, Ali Ameen, Osama Isaac, Mohammed Nusari, Balaganesh Duraisamy and Gamal S. A. Khalifa**

**Abstract** One of the important research topics and areas that is attracting significant interest and attention globally is ‘big data’. While big data contribute towards the quality of decision-making, it also assists in the development of and extending the knowledge in this area by harnessing available technology. This research presents and discusses the literature related to the quality of big data and its impact on the quality of decision-making. A descriptive methodology approach was also adopted by reviewing the literature of published and unpublished scientific research along with a survey in the form of a questionnaire involving participants from Abu Dhabi Police Agencies to collect their views and opinions in this area. The results from the literature review and survey led to proposing a theoretical, conceptual model according to the quantitative and numerical methodology. The findings of this research have revealed that the quality of big data predicts the quality of decision-making and that the quality of big data in Abu Dhabi Governmental Organisations (ADGO) plays a significant role in the quality of decision-making.

**Keywords** Big data · Information systems · Mediating factor, ADGO

---

Y. Alkatheeri · A. Ameen (✉) · O. Isaac · M. Nusari · B. Duraisamy · G. S. A. Khalifa  
Lincoln University College, Selangor, Malaysia  
e-mail: [ali.ameen@aol.com](mailto:ali.ameen@aol.com)

Y. Alkatheeri  
e-mail: [yazeed.alkatheeri@gmail.com](mailto:yazeed.alkatheeri@gmail.com)

O. Isaac  
e-mail: [osama2isaac@gmail.com](mailto:osama2isaac@gmail.com)

M. Nusari  
e-mail: [nusari@lincoln.edu.my](mailto:nusari@lincoln.edu.my)

B. Duraisamy  
e-mail: [balaganesh@lincoln.edu.my](mailto:balaganesh@lincoln.edu.my)

G. S. A. Khalifa  
e-mail: [gamal@lincoln.edu.my](mailto:gamal@lincoln.edu.my)

## 1 Introduction

### 1.1 *Background to the Problem*

Big Data consists of extremely large datasets which are analysed computationally to reveal significant patterns, trends, and associations, especially relating to human behaviour and other interactions. Big data is a term used to refer to the study and application of data sets that are so large and complex that traditional data-processing application software is unable to appropriately deal with them.

In recent years, the quality of data, and its use in effective decision-making has become a critical factor in driving the sustainability and growth of contemporary organisations. The decision-making process is typically reliant upon the quality and accuracy of the data, information, and knowledge [34, 36]. Data can be defined as a raw material and the input of any process whereas, information is the result or outcome of the processing process in the form of outputs that offer useful meaning and value. In 2011, according to the McKinsey Global Institute, big data can be defined as a collection of data, big in size, exceeding the capabilities of traditional data through the collection, storage, processing, management and analysis of databases [22, 31]. Big data also contains a small fraction of organised information, although a large proportion is unorganised. The concept of big data and its benefits has quickly become a global reality.

There is no single definition of big data, as some definitions describe it as an outcome from the use of computers, mobile phones, communication devices, and Internet applications. The majority of people globally nowadays use mobile phones to make voice calls, send text messages, e-mail, and browse the Internet to purchase goods and services and pay via mobile payment systems via their credit cards. Notwithstanding, many people are continually browsing social networking sites, updating content, and posting messages (i.e., Twitter, Facebook, etc. [1, 9, 33]). All of this generates data and increases the volume of digital content being hosted and stored, which ultimately results in significant data revenues and challenges. The size, speed, and variety of data (i.e., data characteristics) are likewise increasing. The value of data is when it comes to searching for information and/or to extract data [2, 4, 5, 19]. However, big data is complex, overlapping, and it cannot be processed using a single tool like a database as the data is unstructured.

### 1.2 *Problem Statement*

Recently, big data has evolved to become a widespread [universal] and desirable phenomenon in administrative, economic and political fields (Maier and Markus 2013). Furthermore, given the rapid and modern advances and changes witnessed over the last few years, especially in Arab countries, it has become necessary to investigate the ways that countries have invested in harnessing the opportunities

associated with big data. Moreover, it is important to identify the anticipated benefits and to identify the challenges of this new technology. While the United Arab Emirates (UAE) is one of the leading countries that produce large amounts of data, the ability to harness this [big] data remains a challenge [37]. Therefore, the problem associated with big data is represented in several ways:

- (a) Through the need to investigate the impact of big data on the quality of decision-making.
- (b) To determine the most important and influential factors associated with the role of big data.
- (c) To explore the influence of big data (in management information systems (MIS) as an intermediary), towards the quality of effective decision-making.
- (d) To understand the benefits of the UAEs strategy to invest in big data.

The problem of this research can be surmised based on the scarcity of research in this field which has focused on optimising the investment of big data as an influencing factor with effective decision-making. The literature review undertaken in this study has revealed that many studies have confirmed there is a weakness in the optimal investment of big data by many companies, to support effective decision-making [6–10]. For example, in 2015, the study entitled, “Big data and Trial” reported that no clear model determines the relationship between big data and the impact on the quality of decision-making.

### ***1.3 Research Contribution***

The contribution and benefits of this study will not simply revisit the importance of big data and the extent of its complexity in different fields and applications that have been undertaken by previous scholars and researchers. Instead, this study explores the uniqueness of big data by studying its use and the investment made by other countries and resultant benefits. More importantly, this study will highlight the relationship between the impact of big data and decision-makers and the important factors that must be present in order to use the data effectively.

Accordingly, this study examines the application of big data in the context of government departments and in other fields in the United Arab Emirates (UAE). Moreover, to determine the various ways and approaches that the UAE has adopted as part of its investment strategy to use big data, by measuring the benefits and comparing to the experiences of other countries.

## ***1.4 Research Objectives***

This research aims to identify the factors that have led towards investing in big data by examining the impact and influence of big data on the quality of decision-making in government organisations in the UAE.

## **2 Literature Review**

Hadoop is a distributed computing platform which has become quite synonymous with big data, and due to its high availability, expandability, fault tolerance, and low costs have become a standard for big data systems. However, Hadoop's Distributed File System HDFS storage system makes it quite challenging to face end-user applications (such as using a user's browser history to recommend news articles or products). Therefore, the more common practice is to send offline computing results to user-facing storage systems (UFSS) such as Redis and HBase. In 2018, Hadoop conducted a study to examine the organisational structure of big data by analysing several models for structuring and organising big data [3, 12, 37]. The goal of the study was also to conceptualise the available technology applications and programmes to serve the needs of big data and to organise these programmes based on their components and functions in organisational models. However, the reference structure in the document is illogical for analytically oriented storage areas, because search and performance capabilities rely heavily on the aggregation model and are not flexible enough for specific queries. Although, it can be an option for semi-structured data planning due to its potential performance and scalability in the online flow area.

Also, to store the flow of semi-structured events or data from internet weblogs a consistent model is required. For example, depending on how much storage is available, there may be a write/ write conflict, or there may be data platly loss if the master node in the master-slave model breaks down before publishing the data to any slave. This is ideal for the flow of workloads that are appended only, which usually occur in event logs. Also, it may be an option to store semi-structured data in the raw data archive. In the following sections, big data and the decision-making factors are discussed along with describing the various sub-factors that are associated with each.

### ***2.1 Big Data Factor***

Big data constitutes a wide range of large and complex data which is difficult to manage using conventional information systems (IS), given their database structure that processes data using traditional applications and programmes. Many of the challenges facing operators is the ability to access information, and the time required

relating to portability, storage, searching and transportation of data. Although, given the development of information technology (IT) over the last few decades, and the rapid emergence of the Internet, the demand for data applications has increased leading and similarly the need to analyse a broad range of data and their associated relationships. However, compared with smaller and separate groups of data, dealing with them has become quite complicated. Nowadays, big data is one of the most important sources of information for government and non-government organisations and has also become an important economic and valued source for countries as a catalyst for innovation. It is anticipated that big data will continue to become not only a vital source of information but the information will increasingly become more sensitive towards the security of countries. On the other hand, big data has enabled the discovery of commercial and legal linkages in which the applications that support big data help to combat crime and terrorism. Moreover, to determine the flow of security data in a timely and appropriate manner.

Big data includes clusters of data of vast sizes and storage areas that exceed the capacity of traditional software and systems in their ability to search and capture information in a very short time. Data management and processing within an acceptable time are pre-requisites for handling big data given the vast volumes of data that are moving continuously, which often makes it difficult to search and engage with the data. In most cases, the magnitude and size of big data requires enormous storage capacity which traditional storage media handling cannot cater for. Given this issue, the development of new systems and special tools to handle and manage big data will help to provide the means by which users can access and obtain the data quickly, accurately and with high efficiency (Snijders et al. 2012).

### **2.1.1 Data Quality**

Big data has driven the demand for highly talented information management professionals in software development companies, such as Oracle Corporation, IBM, Microsoft, SAP, EMC, HP, and Dell. These companies have spent more than US\$15 billion on software and data management software. In 2010, the software industry was valued at more than US\$100 billion and rapidly growing by almost 10 per cent per annum; about twice the speed of the software.

According to one estimate, a third of the world's stored information is in the form of alphanumeric and static image data, the most useful form for most big data applications. This also shows the degree and magnitude of unused data (i.e., in the form of video and audio content). While many vendors offer ready-made solutions to cater for big data, industry experts recommend the development of internal solutions designed to manage and solve the company's existing problem(s), if the company has adequate technical capabilities.

The application and adoption of big data in government allows for cost-efficiency, productivity, and innovation, but not without its drawbacks or disadvantages. Data analysis often requires multiple parts of government (central and local) to work collectively to create new and innovative processes to achieve the desired results.

For example, in government, one of the important applications for big data in the scientific field is in the recording data for large helium collisions. There are about 150 million sensors that deliver data at a frequency of 40 million times per second. Further, there are approximately 600 million collisions of atoms that occur per second that require recording and analysis after filtering and sorting the data which must be 99.99995% accurate. Accordingly, big data technology plays a significant role in the accuracy of data and results.

### **2.1.2 Data Relevance**

Big data is usually quite unstructured and messy, of varying quality and distributed to innumerable servers located worldwide. Regarding, big data, there is always a sense or general perception of the vastness of the data itself rather than its precision and detail. Before the advent of big data, analysis was limited to testing a limited number of hypotheses that were formulated before collecting the data. Leaving the data to talk, relationships that were not previously envisioned or evident gradually evolved. For instance, Twitter can be used to predict the performance of the stock market, and likewise, Amazon and Netflix offer products to consumers based on the feedback and ratings of tens of thousands of users. Likewise, Twitter, LinkedIn, Facebook creates a “social graph” of user relationships to show what users prefer.

Although big data will be based on the values developed and preserved globally in many cases as evidence, the data is not simply a re-enactment of old rules applied to new conditions, and therefore an awareness of the urgent need for entirely new quality of principles is needed. Big data has become an important element in the treatment and recognition of problems such as climate change, disease eradication, the creation of efficient governments and economic growth. However, through the emergence of big data, there are also many challenges imposed on organisations in order to be better prepared in harnessing the technology that will inevitably transform society, institutions and ourselves. Strategies have lost three main components that have long been used in privacy protection, namely; observation and approval of individuals, withdrawal, and ignorance. In fact, one of the most significant drawbacks of big data is that many users feel that their privacy has been violated.

Moreover, many organisations have dealt with the difficulty afforded by big data; interacting with data as an unfortunate reality, rather than viewing it for its real value. In fact, many people tend to view big data as an artificial [human-made] constraint developed by techniques over time. However, nowadays, the technical or technology environment has turned around 180 degrees. Although, there are still, and always will be, limitations on how much information can be efficiently managed, these will be less limited and restricted over time. Therefore, the criticism of the big data model can be seen from two respects. The first of which stems from those who doubt the implication of big data based on conventional or the same approaches, and second, is from those who doubt the way it is presently implemented.

### 2.1.3 Spread Data

Big data delivery is an exceptional technique to handle large volumes of data stored and received in a timely manner. A report by McKinsey (2011) proposed that to deal with big data requires a number of factors. These include information system operators, appropriate learning, authentication rule techniques, data classification, the practice of cluster analysis of the data stored, and data integration. Also, algorithms, machine learning technology, knowledge of sorting and natural language processing which handles the user. Also, the identification of data patterns, the rapid detection of abnormal evidence, digital signal processors, learning subject and non-subject to control, digital simulation and is automatic (Manyika et al. 2011). The challenge of handling significant amounts of data is not new. Historically, society has worked with limited amounts of data as the tools, organisation, storage, and analysis have been limited. Moreover, the information was filtered, relying only on the smallest part, in order to easily examine it.

Based on the 2013 Global Trends Study, improvements in supply planning and product quality provide the most significant benefits regarding large [big] manufacturing data. Notably, the data provides an infrastructure to facilitate manufacturing transparency, which is the ability to detect uncertainties in processes such as inconsistent components, performance, and availability. Moreover, predictive processing as a viable approach to near zero failure and transparency requires a vast amount of data and advanced forecasting tools for a systematic process of managing the data to gain useful information.

### 2.1.4 Data Storage

Most systems dealing with big data do not use relational databases (MPP) because they cannot store and manage large amounts of data. Moreover, these systems do not have the capability to monitor and load data of this magnitude, have no backup facilities, or use tables sufficient to cater for large [big] databases using RDBMS technique.

The data analysis programme DARPA (Defense Advanced Research Projects Agency) is one of the most important programmes used in evidence infrastructure rules for big data management. This technology first appeared in 2008 in the institutions of Anaj company claims where management responsible for data operations and analysis were unwilling to deal with the storage space required for the data, instead preferring slow direct storage spaces (DAS), starting with solid state hard drives (SSD), followed by Serial AT Attachment (SATA) hard drives of higher-capacity. The underlying architecture of the spaces associated with shared data storage are ineffective, slow, complex and are often expensive. Furthermore, the specifications of these devices do not cater for managing big data, which is characterised by high speed and efficiency nor do they allow for big data analysis and efficient system performance. However, the infrastructure designed specifically for big data needs to be affordable [37]. Accordingly, based on the information described above, it is evident

that the timely delivery of large amounts of data is one of the most distinctive characteristics of big data systems. Obviously, the cost of a storage area network (SAN) used in big data applications is significant compared with other storage technologies.

## ***2.2 Factors of Decision-Making***

Organisations rely on effective decision-making to achieve strategic objectives towards the growth and profitability of the organisation and shareholders and to solve the problems faced by the organisation (Nightingale 2008). Decision-making often involves the brainstorming of ideas and putting forward proposals and suggestions related to improving the operations of the organisation, in meeting its objectives. Furthermore, by identifying the necessary information needed and articulating the strengths and weaknesses of each idea or proposal, this helps to determine the most appropriate proposal and making amendments until reaching the most appropriate decision. This, therefore, enables the institution to achieve its objectives in the shortest possible time and perform its operations with the highest level of efficiency and effectiveness. The following section discusses the decision-making process and stages (Kutty 2007).

### **2.2.1 Problem Identification Accuracy**

The problem identification stage is the initial stage that is undertaken in the decision-making process. This stage is undertaken to identify the actual problem or decision that needs to be made and the work required to resolve a particular problem. Notably, the size and nature of the problem will influence the process and time to resolve the problem in deciding. All interested parties associated with the decision-making process will be involved in the process to gain their input and different perspectives on the cause(s) of the problem (including symptoms and impacts) and identifying those parties impacted by the problem [13, 22]. The decision-makers during this stage identify the nature of the problem, its dimensions (scope) and the situation or circumstance which creates the problem. The importance of the problem should not be confused with the symptoms, causes and time to resolve the problem in order to make an effective and appropriate decision. System information (SI) is integrated into this process during the facilitation, determination, and identification of the problem. This step is often complicated by the implication that the problem is the existence of something that hinders the implementation of some task(s) or the achievement of objectives.

### **2.2.2 Information Accuracy**

In understanding the problem, the accuracy of the information collected and analysed is important. This helps to identify the cause or reasons that have led to the problem and precedes the proposal phase. During this stage, suitable and practical alternatives and options are identified and examined. This requires data and information related to and linking to the problem to be available. Also, the capacity of decision-makers is crucial during this stage as they will be relying on the collected sources of information (and data) presented as information relating to the problem at hand. Accordingly, the information must be accurate and relevant in order for the decision-makers (or stakeholders) to compare, analyse and discuss any facts or figures to identify alternatives and options which will lead towards an appropriate decision. Therefore, in understanding the problem, they must also understand the reality of the situation or circumstance, by proposing alternatives to solve the problem.

### **2.2.3 Evaluation of Alternatives**

The evaluation phase of alternatives is important in the decision-making process and is often considered as the most important stage as it will determine the nature of the decision that will be chosen from a range of alternatives. Importantly, this stage considers the outcomes from the previous stage; the alternatives, supporting factors, organisational policies, philosophy of the organisation, the potential of each alternative and timing. All these factors will constitute towards an effective decision being made and shortlisting of alternatives if necessary. Additionally, logical thinking, visualisation and predicting the outcome of the decision is paramount during this stage of the process. Evaluating alternatives helps in rating and ranking each alternative and shortlisting to compare the benefits and disadvantages of each alternative. Selecting the most acceptable and most appropriate alternative is also based on available standards, risk, and objectivity. Standards being the most appropriate in many instances.

The researcher of this study at this point highlights that fake participation by subordinates and specialists in this particular field depends on forecasting innovation. Notably, the number and type of alternative solutions depend on several factors; position of the organisation, the policies it applies, its material resources, the time available to the decision maker, the director's [and management's] decision-making attitudes and his or her ability to think logically and creatively, based on imaginative thinking, perceptions, expectations, and ideas.

### **2.2.4 Decision Accuracy**

Decision-making processes need to be integrated and coherent with existing information systems in order for management to expand the level of knowledge of managers on the proposed decision(s) that will be made. The integration of information sys-

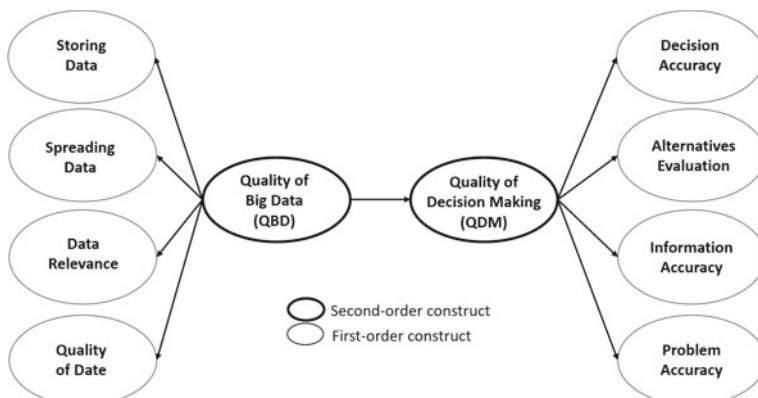
tems will importantly help to support the accuracy of decision-making [8–10, 15, 32]. Moreover, the allocated time to make and select a decision will influence the quality of the final decision in achieving the optimal result or best outcome to resolve the problem. Notably, decision-makers will also need to consider the implications of their decision based on the ability to implement the solution or change.

Also, from the analysis undertaken by the decision-makers, it will be evident in most cases as to what caused the problem, and what decision(s) need to be made to achieve the best results or outcomes. Furthermore, there are differences between the analysis of the problem and the decision made; decision-makers can solve the problem without resolution. Alternatively, a decision can be made without solving the problem as suggested by Cao et al. [16] as the analysis may reveal that the problem is moated outside the domain of the organisation's liability. In this case, decision-makers, do nothing apart from informing management of their intention. This kind of process is a trade-off between the available alternatives and choosing the most appropriate alternative according to the criteria and considerations.

### 3 Research Method

In this study, the hypothesised variables and their relationships in the model have been derived from the previous literature on the models and theories, along with the literature prescribed above. Figure 1 displays the proposed conceptual framework of the model.

An eight-item questionnaire was developed for this research, in line with existing literature in harnessing big data as a key factor in the quality of decision-making, using a multi-item Likert scale. The variables were measured using a Likert Scale; 5 = ‘Strongly Agree’ and 1 = ‘Strongly Disagree’. An online and paper-based survey



**Fig. 1** Conceptual framework

in the form of questionnaires was used for the data collection process. Participants from Abu Dhabi Police Agencies participated in the survey.

## 4 Data Analysis and Results

AMOS statistical software was used to analyse the data. AMOS stands for analysis of moment structures and is an additional SPSS module used for Structural Equation Modelling (SEM), Path Analysis (PA), and Confirmatory Factor Analysis (CFA). It is also known as analysis of covariance or causal modelling software. Structural Equation Modelling-Variance Based (SEM-VB) was also employed to examine the research model [17, 30, 35].

### 4.1 Descriptive Analysis

Table 1 presents the mean and standard deviation of each variable in the current study. Each respondent was asked to indicate their opinion in completing the survey questionnaire. The quality of big data recorded a mean score of 4.093 out of 5.0, with a standard deviation of 0.674, indicating that the respondents agreed that the quality of big data within Abu Dhabi Governmental Organisations (ADGO), is high and decision-making is based on best practices. Also, the quality afforded by big data enabled ADGO to share this knowledge with other organisations.

The quality of decision-making recorded the mean score of 4.081 out of 5.0, with a standard deviation of 0.689, indicating that the respondents agreed that restructuring of resources and altering of organisational structures was important. Moreover, employees were aware that challenges will always come along, so employees persist in order to overcome them. ADGO performance was also viewed as being very high compared to how other companies or organisations in similar industries are performing. Indeed, ADGO's position in the industry was considered admirable and of a comparably high standard.

### 4.2 Measurement Model Assessment

All the goodness-of-fit indices exceeded their respective common acceptance levels as suggested by previous research. Therefore, demonstrating that the measurement model exhibited a good fit with the data collected. Therefore, the evaluation of the psychometric properties of the measurement model regarding construct reliability, indicator reliability, convergent validity, and discriminant validity could be proceeded with.

**Table 1** Mean, standard deviation, loading, Cronbach's Alpha, CR, and AVE

Constructs	Item	Indicators	Loading (>0.5)	<i>M</i>	SD	$\alpha$ (>0.7)	CR (>0.7)	Ave. (>0.5)
Quality of big data (QBD)	QBD1	The quality of data within an organisation	0.721	4.093	0.674	0.860	0.882	0.592
	QBD2	Data relevance to the decision subject	0.736					
	QBD3	Spread data and availability	0.745					
	QBD4	Store and structure of data with firms	0.789					
Quality of decision-making (QDM)	DMQ1	Problem identification	0.843	4.081	0.689	0.831	0.891	0.584
	DMQ2	Accuracy information	0.889					
	DMQ3	Accuracy alternatives evaluation	0.678					
	DMQ4	Taking decision	0.645					

Note *M* Mean; SD Standard Deviation,  $\alpha$  Cronbach's alpha; CR Composite Reliability, Ave Average variance extracted  
Key QBD Quality of big data; QDM Quality of decision-making

**Table 2** Results of discriminant validity by Fornell-Larcker criterion

	Factors	1	2
		QBD	QDM
1	QBD	<b>0.769</b>	
2	QDM	0.490	<b>0.764</b>

Note Diagonals represent the square root of the average variance extracted while the other entries represent the correlations

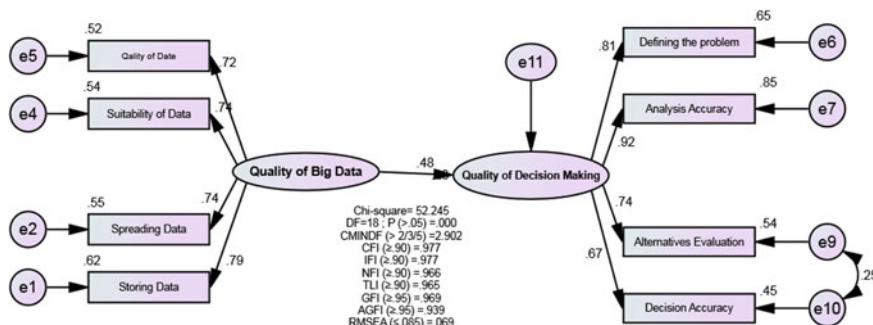
Key QBD Knowledge sharing; QBD Quality of decision-making

The values of all the individual Cronbach's alpha coefficients in this study exceeded the suggested value of 0.7 [29]. Furthermore, for testing construct reliability, the values of all the composite reliability (CR) exceeded 0.7 [21]. Furthermore, the values of all average variance extracted (AVE) exceeded the recommended value of 0.50 [23]. Table 1 shows that all items in this study had factor loadings higher than the recommended value of 0.5 [23].

Table 2 displays the results for discriminant validity using the Fornell-Larcker criterion. It was found that the square root of the AVEs on the diagonals (shown in bold) are greater than the correlations between constructs (corresponding row and column values), thereby indicating good discriminant validity [18, 20].

### 4.3 Structural Model Assessment

Figure 2 and Table 3 depict the structural model assessment. It is seen that the quality of big data significantly predicts the quality of decision-making. Hence, H1 is accepted with ( $\beta = 0.484$ ,  $t = 10.454$ ,  $p < 0.001$ ). Twenty-three per cent of the variance in the quality of decision-making is also explained by knowledge sharing.

**Fig. 2** SEM result

**Table 3** Structural path analysis result

Hypothesis	Relationship	Std. beta	Std. error	t-value	p-value
H1	QBD → QDM	0.484	0.046	10.454	0.000

Key *QBD* Knowledge sharing; *QDM* Quality of decision-making

## 5 Discussion

The fundamental nature of decision-making is influenced by everyday activities coupled with organisational behaviour. This fact was established from the outset of this study. Accurate and effective decision-making are mandatory in all forms of business, drawing reverence from both internal and external organisational domains. Accordingly, the quality of accurate decision-making has become an ongoing concern for organisations globally, thereby creating the need for a much deeper understanding of the factors that contribute and that are needed in this area. Building on the premise that decision-making by management has often ended in abysmal failure, this has attracted the interest and attention of researchers to explore and explain the use of specific theories and paradigms.

## 6 Theoretical and Practical Implications

The implications associated with key theories and how these contribute towards the concepts and theories that underlie organisational uncertainty and best practice in decision-making are important to understand. This study paves the way for better understanding of how big data has become one of the most important factors in supporting decision-makers towards achieving more predictably their goals and objectives. The findings of this study reveal that the quality of big data has a significant effect on the quality of decision-making which therefore implies that the quality of big data is embedded in the quality of decision-making. Notably, this is because, in the quality of decision-making, employees do not just go along with what data is stored in the information system(s), but are instead, drivers of how to harness big data in a deferent manner. Contrary to the hypothesised relationship, the quality of big data is an insignificant predictor of the quality of decision-making whereas, the predictive effect was not a significant one. A study by Backhaus et al. (2010) presents key arguments that may explain this in which they argue that the quality of big data is key to organisational success in general and especially in the event of change. The impact of the quality of big data may, therefore, lie more in the quality of the decision-making process itself rather than the formulation of the quality of decision-making policy. Accordingly, this points to the quality of big data as an important tool towards the quality of decision-making (Johnson et al. 2008). As mentioned earlier, both the quality of decision-making management and the quality of big data predict the quality of decision-making. Also, the quality of decision-making management

may include directives aimed at employee tensions, which may improve the overall predictive effect of the quality of big data [11, 14, 22].

This study further helps to firmly establish the quality of big data as a key variable to consider in the implementation of harnessing big data to enhance and facilitate effective decision-making. The significance of the findings in relation to the concept of big data lies in the fact that it opens the door for academic research to further study these concepts as interlinked concepts. Notwithstanding, this study has added to the theory of the quality in decision-making regarding how complexities influence can it. The findings also make room for further critique in the field of quality decision-making and serve as a precedent for the development of hybrid models that illustrate the interrelationships revealed.

The present study has profound implications for the quality of decision-making that has currently been considered towards more sustainable institutional development. ADGO need to understand and know the impact of the quality of big data on the quality of decision-making, with the main rationale to help arrive at how management policies can assist towards the implementation of big data successfully. Notably, this can help in drawing attention from other organisations to build a successful model in the region.

## 7 Limitations and Recommendations for Future Research

A limitation of this study is related to the way the data was gathered which was cross-sectional rather than longitudinal. The longitudinal method might improve the understanding of the associations and the causality between the variables [24]. Therefore, future research should investigate the relationship between the variables by conducting cross-cultural studies as recommended by previous studies [25, 26]

## 8 Conclusion

In this study, the concept surrounding the quality of big data and decision-making was discussed, supported by performing a thorough literature review and survey of participants from Abu Dhabi Police Agencies. The relationship between the quality of big data and decision-making from the perspective of the theory of complexity can help management make well-informed and accurate decisions from an organisational viewpoint. A thorough discussion was evidenced by the application of these two areas in the management of complexity in the organisational context. Increasingly, big data in information systems is continuously being converted into meaningful information which is used and analysed by decision-makers to solve organisational problems. Therefore, the information must be reliable and accurate in order to identify alternative solutions and to make informed [quality] decisions. Consequently, the diversity of data sources, the information available, and alternative solutions to

problems can benefit the organisational decision-making process. This is supported by the conceptual model, hypothesising the variables and relationships.

Importantly, the nature of information and data held in information systems needs to be timely given that the information often corresponds to the type of decisions that are taken. Moreover, the dissemination of data is often in accordance and reliant upon the diversity of data formats. Therefore, big data stored and residing in databases needs to be well structured and organised to facilitate the diagnosis of problems and solutions. With regards to the main research question which sought to investigate the relationship between the quality of big data and quality of decision-making, the findings revealed that the quality of big data predicts the quality of decision-making. This is also supported by the conclusion that the quality of big data in ADGO plays a key role in the successful quality of decision-making management. A final observation from this study related to big data implementation is that organisations should increase spending on research and development in order to increase organisational effectiveness and benefits associated with this technology [27, 28].

## References

1. Abdulrab, M., Zumrah, A.-R., Almaamari, Q., Al-tahitah, A. N., Isaac, O., & Ameen, A. (2018). The role of psychological empowerment as a mediating variable between perceived organizational support and organizational citizenship behaviour in Malaysian higher education institutions. *International Journal of Management and Human Science (IJMHS)*, 2(3), 1–14.
2. Al-Ali, W., Ameen, A., Isaac, O., Khalifa, G. S. A., & Hamoud, A. (2011). The mediating effect of job happiness on the relationship between job satisfaction and employee performance and turnover intentions: A case study on the oil and gas industry in the United Arab Emirates. *Journal of Business and Retail Management Research (JBRMR)*, 13(4), 1–15.
3. Al-Obthani, F., & Ameen, A. (2018). Towards customized smart government quality model. *International Journal of Software Engineering & Applications*, 9(2), 41–50. <https://doi.org/10.5121/ijsea.2018.9204>.
4. Al-Shamsi, R., Ameen, A., Isaac, O., Al-Shibami, A. H., & Sayed Khalifa, G. (2018). The impact of innovation and smart government on happiness: Proposing conceptual framework. *International Journal of Management and Human Science (IJMHS)*, 2(2), 10–26.
5. Alkhateri, A. S., Abuelhassan, A. E., Khalifa, G. S. A., Nusari, M., & Ameen, A. (2018). The Impact of perceived supervisor support on employees turnover intention: The mediating role of job satisfaction and affective organizational commitment. *International Business Management*, 12(7), 477–492. <https://doi.org/10.3923/ibm.2018.477.492>.
6. Ameen, A., & Ahmad, K. (2011). The Role of Finance Information Systems in anti financial corruptions: A theoretical review. In *11 International Conference on Research and Innovation in Information Systems (ICRIIS'11* (pp. 267–272). IEEE. <http://doi.org/10.1109/ICRIIS.2011.6125725>.
7. Ameen, A., & Ahmad, K. (2012). Towards harnessing financial information systems in reducing corruption: A review of strategies. *Australian Journal of Basic and Applied Sciences*, 6(8), 500–509.
8. Ameen, A., & Ahmad, K. (2013). A conceptual framework of financial information systems to reduce corruption. *Journal of Theoretical and Applied Information Technology*, 54(1), 59–72.
9. Ameen, A., & Ahmad, K. (2013b). Proposing strategy for utilizing financial information systems in reducing corruption. In *3rd International Conference on Research and Innovation in Information Systems—2013 (ICRIIS'13)* (Vol. 2013, pp. 75–80).

10. Ameen, A., & Ahmad, K. (2013c). Proposing strategy for utilizing financial information systems in reducing corruption. In *3rd International Conference on Research and Innovation in Information Systems—2013 (ICRIIS'13)* (Vol. 2013, pp. 75–80).
11. Ameen, A., Almari, H., & Isaac, O. (2018). Determining Underlying Factors that Influence Online Social Network Usage among Public Sector Employees in the UAE. In B. A. Saeed F., Gazem N., & Mohammed F. (Eds.), *3rd International Conference on Reliable Information and Communication Technology 2018 (IRICT 2018), Bangi-Putrajaya, Malaysia* (3rd ed., Vol. 843, pp. 945–954). Cham: Springer. [http://doi.org/doi.org/10.1007/978-3-319-99007-1\\_87](http://doi.org/doi.org/10.1007/978-3-319-99007-1_87).
12. Ameen, A., Almari, H., & Isaac, O. (2019). Determining Underlying factors that influence online social network usage among public sector employees in the UAE. In F. M. Faisal Saeed & N. Gazem (Eds.), *Recent trends in data science and soft computing. IRICT 2018. Advances in intelligent systems and computing* (Recent Tre, Vol. 843, pp. 945–954). Springer Nature Switzerland AG: Springer. <http://doi.org/10.1007/978-3-319-99007-1>
13. Ameen, A., Almulla, A., Maram, A., Al-Shibami, A. H., & Ghosh, A. (2018). The impact of knowledge sharing on managing organizational change within Abu Dhabi national oil organizations. *International Journal of Management and Human Science (IJMHS)*, 2(3), 27–36.
14. Ameen, A., & Kamsuriah, A. (2017). Information Systems Strategies to Reduce Financial Corruption. In S. M. & Benlamri R. (Ed.), *Springer proceedings in business and economics* (Vol. 1, pp. 731–740). Cham, Switzerland: Springer. [http://doi.org/10.1007/978-3-319-43434-6\\_65](http://doi.org/10.1007/978-3-319-43434-6_65)
15. Arefin, M. S., Hoque, M. R., & Bao, Y. (2015). The impact of business intelligence on organization's effectiveness: An empirical study. *Journal of Systems and Information Technology*, 17(3), 263–285. <https://doi.org/10.1108/JSIT-09-2014-0067>.
16. Cao, G., Duan, Y., & Li, G. (2015). Linking business analytics to decision making effectiveness: A path model analysis. *IEEE Transactions on Engineering Management*, 62(3), 384–395. <https://doi.org/10.1109/TEM.2015.2441875>.
17. Chan, Y. H. (2005). *Structural Equation Modeling*, 46(12), 675–680.
18. Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–358). New Jersey: Lawrence Erlbaum Associates, Mahwah, NJ: Lawrence Erlbaum.
19. Fahad, A.-O., & Ameen, A. (2017). Toward proposing SMART-government maturity model: Best practices, international standards, and six-sigma approach. In *1st International Conference on Management and Human Science (ICMHS 2017)* (p. 2017). Kuala Lumpur, Malaysia.
20. Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50.
21. Gefen, D., Straub, D., & Boudreau, M.-C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems*, 4(1), 1–79.
22. Haddad, A., Ameen, A., & Mukred, M. (2018). The impact of intention of use on the success of big data adoption via organization readiness factor. *International Journal of Management and Human Science (IJMHS)*, 2(1), 43–51.
23. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. New Jersey.
24. Isaac, O., Abdullah, Z., Ramayah, T., & Mutahar, A. M. (2017). Internet usage, user satisfaction, task-technology fit, and performance impact among public sector employees in Yemen. *International Journal of Information and Learning Technology*, 34(3), 210–241. <https://doi.org/10.1108/IJILT-11-2016-0051>.
25. Isaac, O., Abdullah, Z., Ramayah, T., & Mutahar, A. M. (2017). Internet usage and net benefit among employees within government institutions in Yemen: An extension of Delone and Mclean information systems success model (DMISM) with task-technology fit. *International Journal of Soft Computing*, 12(3), 178–198. <https://doi.org/10.3923/ijscmp.2017.178.198>.
26. Isaac, O., Abdullah, Z., Ramayah, T., & Mutahar, A. M. (2017). Internet usage within government institutions in Yemen: An extended technology acceptance model (TAM) with internet self-efficacy and performance impact. *Science International*, 29(4), 737–747.

27. Isaac, O., Abdullah, Z., Ramayah, T., & Mutahar, A. M. (2018). Factors determining user satisfaction of internet usage among public sector employees in Yemen. *International Journal of Technological Learning, Innovation and Development*, 10(1), 37–68. <https://doi.org/10.1504/IJTLID.2018.10012960>.
28. Isaac, O., Abdullah, Z., Ramayah, T., Mutahar, A. M., & Alrajawy, I. (2018). Integrating user satisfaction and performance impact with technology acceptance model (TAM) to examine the internet usage within organizations in Yemen. *Asian Journal of Information Technology*, 17(1), 60–78. <https://doi.org/10.3923/ajit.2018.60.78>.
29. Kannana, V. R., & Tan, K. C. (2005). Just in time, total quality management, and supply chain management: understanding their linkages and impact on business performance. *Omega: The International Journal of Management Science*, 33(2), 153–162.
30. Kline, R. B. (2008). *Principles and practice of structural equation modeling*. New York, NY, US: The Guilford Press.
31. Koltay, T. (2016). Data governance, data literacy and the management of data quality. *IFLA Journal*, 42(4), 303–312. <https://doi.org/10.1177/0340035216672238>.
32. Mohamed, N., Kaur, J., & Gian, A. P. (2012). Information management & computer security emerald article : A conceptual framework for information technology governance effectiveness in private organizations.
33. Mohsen, A.-A., Ameen, A., & Al-Gamrh, B. (2017). The impact of achievement and enablers excellence and innovation in organization: A proposed model. In *1st International Conference on Management and Human Science (ICMHS2017)* (p. 19). Kuala Lumpur, Malaysia.
34. Mugnaini, R., & Fujino, A. (2017). *Bibliometria e cientometria no Brasil: infraestrutura para avaliação da pesquisa científica na era do Big Data*. <http://doi.org/10.11606/9788572051705>
35. Ringle, C. M., Wende, S., & Becker, J.-M. (2015). *SmartPLS 3*. Bonnigstedt: SmartPLS.
36. Toivonen, M. (2015). *Big data quality challenges in the context of business analytics department of computer science*.
37. Yazeed, A., Ali, A., & Al- Shibami, H. (2018). Conceptual framework for investigating the intermediate role of information systems between big data factor and decision-making factor. *International Journal of Management and Human Science (IJMHS)*, 2(2), 39–45.

# The Impact of Technology Readiness on the Big Data Adoption Among UAE Organisations



**Adel Haddad, Ali Ameen, Osama Isaac, Ibrahim Alrajawy, Ahmed Al-Shbami and Divya Midhun Chakkavarthy**

**Abstract** Big Data is an important player in offering a highly competitive advantage, specialty, in contemporary organisations. The theory of technology readiness can be used for measuring the readiness of an organisation to adapt big data. Structural equation modelling is used in this study through AMOS to analyse 381 valid questionnaires to evaluate the proposed model built on the Theory of Technology Readiness to determine the factors that could affect big data adoption. This research concentrates on one of Abu Dhabi's public organisations (ADPO). In this model, the key independent constructs are comparable to Innovativeness, Optimism, Insecurity and Discomfort pertaining to these organisations' readiness for exploiting this massive data amounts. The dependent constructs are based on the adopted big data's readiness in ADPO. The relations between the different constructs are defined in this research. This work has enhanced our insights regarding the online social networking model. The results showed that all four independent variables considerably helped to predict the adoption of big data with different percentages. The model that was put forward explained 50% of the variance occurring in the adoption of big data.

**Keywords** Optimism · Innovativeness · Discomfort and insecurity · Big data · Adopting · Theory of technology readiness · Public sector · UAE

---

A. Haddad · A. Ameen (✉) · O. Isaac · I. Alrajawy · A. Al-Shbami · D. Midhun Chakkavarthy  
Lincoln University College, Kota Bharu, Selangor, Malaysia  
e-mail: [ali.ameen@aol.com](mailto:ali.ameen@aol.com)

A. Haddad  
e-mail: [haddadphd@gmail.com](mailto:haddadphd@gmail.com)

O. Isaac  
e-mail: [osama2isaac@gmail.com](mailto:osama2isaac@gmail.com)

I. Alrajawy  
e-mail: [ibrahim2alrajawy@gmail.com](mailto:ibrahim2alrajawy@gmail.com)

A. Al-Shbami  
e-mail: [alshibami@lincoln.edu.my](mailto:alshibami@lincoln.edu.my)

D. Midhun Chakkavarthy  
e-mail: [divya@lincoln.edu.my](mailto:divya@lincoln.edu.my)

## 1 Introduction

In a contemporary organisation, *Big Data* is an important player in offering a highly competitive advantage. Most organisations try to benefit because it provides a deeper understanding of its customers and their requirements. This helps to make appropriate and appropriate decisions within the company in a more effective manner based on information extracted from customer databases [1, 2]. Specialised IT research and consultancy defines big data as a large, fast-flowing and highly diversified information asset that requires cost-effective and innovative processing methods to develop insights and decision-making. It is also defined by the company (IBM); Big Data is created by everything around us. At all times, every digital process and every exchange in social media produces huge data, transmitted by systems, sensors, and mobile devices. Big data has multiple sources of speed, size and diversity, and to derive significant benefit from large data. “We need perfect treatment, analytical abilities, and skills” he said. The International Standards Organization (ISO) has defined big data as groups or sets of data with unique characteristics (e.g. size, speed, diversity, variability, data health, etc.), which cannot be efficiently addressed using current and traditional technology to make use of it.

Big data was defined by the International Telecommunication Union (ITU) as data sets that are super-large, fast, or versatile, compared to other types of data sets used. Speed is a crucial factor in decision-making based on these data; which is the time it takes from the moment these data arrive to when the decision is made. Previously, companies used to process small sets of data stored in a structured data image in a process database, where each dataset was analysed one by one pending the arrival of the results. Big data is an important player in offering a highly competitive advantage, specialty, in contemporary organisations. Most organisations seek it for benefits since it gives a deeper understanding regarding the customers as well as their requirements. This research will insight the impact of technology readiness on big data adoption among public organisations in Abu Dhabi, UAE.

## 2 The Status of Big Data Technology in the UAE

The United Arab Emirates (UAE) started adopting large-scale data technology since 2013. Establishment of a smart government was the first application, which was aimed at providing services to the UAE public around the clock, anywhere. The goal of this project is to take advantage of the huge data applications to serve the UAE citizens around the clock and anywhere in the world [3]. The idea of this project was based on the context of the Government’s efforts to develop government services and achieve high quality of life for UAE citizen and residents, according to the UAE Vision 2021 [4–6].

As part of its efforts to implement the Smart Government Initiative, the AE General Authority for Development has prepared the Smart Government Roadmap, which

provides a plan for the UAE to move from e-government to smart government. The map sets out a range of tasks covering the period until 2015. The scope of the road map is in line with the current federal e-government strategy 2012–2014, with emphasis on environmental improvements, enhanced user readiness and user satisfaction [7, 8].

The United Arab Emirates plans to set up Dubai Smart City in cooperation with Emirates Integrated Telecommunications Corporation. The first phase of the Dubai Smart Platform, is an interactive database that allows residents, visitors and institutions to analyse data and information that is electronically collected and collected from local government institutions to achieve the concept of satisfaction and the happiness of users. Dubai's artificial intelligence road map, Dubai Smart, in partnership with a network of private and public partners, strives to search for innovative technology solutions to enhance the quality of life in Dubai as well as make the city more efficient, safe, smooth and effective in terms of experience.

UAE government has identified areas of focus within four parallel tracks, which correspond to the Smart government, which are:

- Creating a general environment in which the smart government thrives
- Assess the capabilities of government agencies
- Establish joint resources through government agencies at a national level
- Happy citizens.

In the UAE, different entities must discuss artificial intelligence as well as the capability that allows understanding natural language, verifying and analysing large databases swiftly, and reach conclusions based on transactions, as well as recommend relevant information to aid users in selecting appropriate next steps. To install on a platform, a smart window will be created to collect the services it needs daily, which can be changed at any time, thus minimising this concept. The government's ability is considerably improved with the platform in making quick decisions with the available data, which allows city leaders to get involved in community-wide dialogues and evaluate rich city data across numerous dimensions. The platform enables additional improvement for the existing smart initiatives and services based on analysis as well as data-based innovation.

### 3 Literature Review

#### 3.1 *The Optimism*

The identification of optimism lies with the inclination to have “a constructive view regarding innovation as well as a conviction promising to give individuals expanded control, proficiency and adaptability in their lives” [9]. Self-assured individuals are suggested to embrace innovation, and in evaluation, to various buyers, are less likely be inclined to centre on the contrary parts associated with inevitable hardships as well

as disappointments in new advancements, should they occur (Kotler & Armstrong). Since innovation is seen by confident people with a perspective of conceivable outcomes, it could also be expected that hopeful purchasers may see self-scanners as both less demanding and more valuable for use versus non-idealistic buyers.

There are many studies that have shown a positive connection between big data adoption BDA and optimism. In one of the key examinations regarding BDAs, Dabholkar (1996) found that a higher level of control was permitted to the shopper, for touch-screen request in the cost-effective food industry. The impression of purchaser control has also been seen to be emphatic with the purchaser acknowledging shopper in self-requesting booths that can be found in eateries (in the same place), as well as self-registration stands at air terminals. These discoveries were in line with the findings of Dabholkar et al. (2003), who discovered self-filtering DDAs, that control and efficiency are also the major determinants for the buyer acknowledgment of BDAs. Thus, the below hypothesis is proposed:

H1: Optimism has a positive impact on big data adoption.

### **3.2 *The Innovativeness***

The identification of creativity lies with the inclination “towards being a thought pioneer and an innovation pioneer” [9], in which the creative buyer is guessed to be bold in employing innovation. Furthermore, inventive purchasers are guessed to view innovation as being simple, since they possess an abnormal state of mechanical information, as well as a certifiable enthusiasm to detect new innovation [9]. As creative purchasers perceive innovation as being interesting, it can be highly expected that imaginative customers perceive self-scanners as being more significant and less demanding to employ than others.

As opposed to the other TR-ideas, it seems that contemporary BDA researchers have not explored Innovativeness fairly. Likewise, for some of the examinations conveyed, addressing of (i) the unwavering quality of the measure, as well as (ii) the beneficial outcome has been done, as proposed by the TR-writing. In fact, experimental studies have confirmed that an absence of viability in the Innovativeness measure continues to persist by all accounts, mainly since the distinction between general innovativeness and area particular cannot be thought with the measure. Beyond a doubt, the space particular Innovativeness has been put forward as being firmly identified with the selection of innovation, while as a frail indicator to innovation acknowledgment, the general Innovativeness has been put forward. Liljander et al. (in the same place) found that the Innovativeness measure can be enhanced as a positive measure regardless of its general methodology, a measure that could possibly contribute to the aggregate informative degree. Be that as it may, the feedback for Innovativeness measure is considered to be more grounded. In this, specifically, a sharp feedback was coordinated by Roehrich (2004), and it concentrated on a non-substantial indicator to acknowledge innovation. Due to this insightful concern, as of

late, the innovation preparation record was streamlined by Parasuraman & Colby [9] who re-assessed the measure. Post the re-assessment, it was found that the measures' unwavering legitimacy and quality was of a solid help. Thus, the below hypothesis is proposed:

H2: Innovativeness has a positive impact on big data adoption.

### ***3.3 The Discomfort***

Uneasiness identifies with the inclination to have an “apparent absence of command over innovation and a sentiment of being overpowered by it” [9]. Purchasers that have a mechanical Discomfort were predicted to possess a sense of general distrustfulness towards innovative tension, development and changes, technophobia [9] and a general negative perception when linking with new or outsider innovation.

With the absence of saw value as well as saw usability for a specific innovation, the BDA writing has risen to Discomfort. Here, Kallweit et al. (2014) evaluated self-benefit data and observed advancements in terms of a reduction in saw usability that casts a critical negative effect on the adequacy of the client; hence, it appears that Discomfort and purchaser acknowledgment of BDAs possess a negative relationship. This end, however, is not questionable. Meuter et al. (2003), emphatically underscore that distress, for example, innovation nervousness, is a conceivable motivation to why customers maintain a strategic distance from innovation. Hence, at the end of the day, force a negative connection among inconvenience and the utilisation of innovation. Thus, the below hypothesis is proposed:

H3: Discomfort has a negative impact on big data adoption.

### ***3.4 The Insecurity***

Weakness identifies with the propensity to have “doubt of innovation and distrust about its capacity to work appropriately” [9]. Once in a while, shoppers with Insecurity are ready to rely on innovation. They believe that innovation comes up short during the most basic minute [9]. Accordingly, purchasers with Insecurity have been linked to both equivocalness as well as a general low utilisation of innovation. For sure, as emphasised by both Kotler and Armstrong (2012) and Parasuraman and Colby [9], customers with Insecurity are sometimes a buyer that embraces innovation enthusiastically, yet they do it when there is no more decision. With these, it can be accepted that self-scanners are observed by shaky buyers as both harder and less valuable for usability versus different purchasers.

For instance, with regards to the Innovativeness measure, unique effects were discovered by researchers which concern Insecurity. For example, Godoe and Johansen

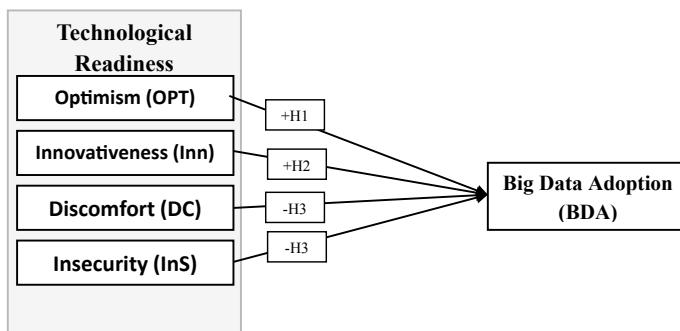
(2012) and Walczuch et al. (2007) confirmed that the identification of Insecurity is not fundamental along with a negative assessment for the saw handiness. In spite of what could be expected, “one could expect that individuals will realise fundamental estimation for a framework that pays little heed to how things are being handled”. In line with this thinking, Gelderman et al. (2011) contended that there is low effect of Insecurity, yet basically considered the measure to be inconsistent and frail. However, rather than stressing its small size, they imply that insecurity is a negative idea; yet, because of its shortcomings, should be joined with the more grounded proportion of Discomfort. Notwithstanding, in the on-going TR re-assessment, Parasuraman and Colby [9] found that Insecurity is without a doubt emphatically identified with absence of trust in innovation, from one viewpoint, and a lower inclination to utilise innovation, then again; therefore, forcing a negative connection among Insecurity and the general acknowledgment of advances. Thus, the below hypothesis is proposed:

H4: Insecurity has a negative impact on big data adoption.

## 4 Research Methodology

### 4.1 Proposed Conceptual Framework

In the conceptual framework, the hypothesised relationships between the constructs have been taken from the relevant literature. Figure 1 displays the put forward model with optimism, discomfort, innovation and insecurity to forecast the adoption of big data. These relationships are taken from [9]. The said model examines the relationship between the constructs among employees in the public organisations of Abu Dhabi in the UAE. Four hypotheses are tested with the suggested conceptual framework.



**Fig. 1** Proposed conceptual framework

## 4.2 Research Instruments

A 17-item questionnaire was used to construct the instrument for this study and a multi-item Likert scale was also applied as per the information systems in the literature [10]. A Likert scale was employed to measure the constructs, which was recommended in the earlier studies [11, 12], in which 5 referred to ‘Strongly Agree’ and 1 to ‘Strongly Disagree’. Since the respondents were Arabic speakers, translation of the questionnaires from English to Arabic was done in a precise manner. Thus, a back translation was also employed, which is an approach employed broadly in cross-cultural surveys [13–15]. In this study, the measurement of the variables was validated by employing extant research. For each construct, the number of items was determined based on the guidelines of Hayduk and Littvay (2012) who advocated to employ few optimal items.

## 4.3 Data Collection

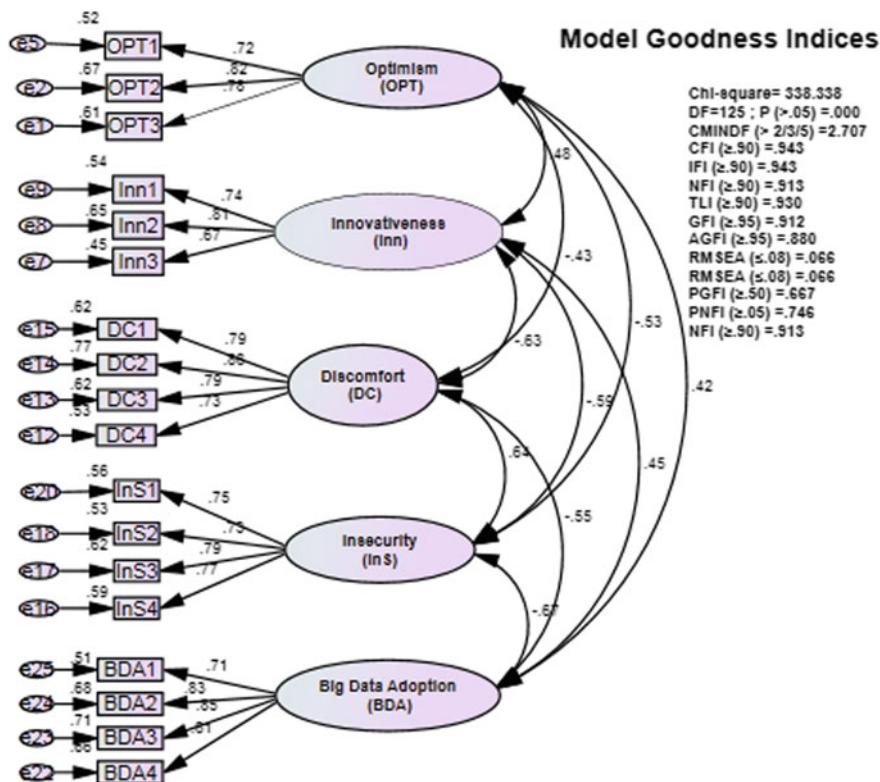
To employees within the public sector in the UAE, self-administered questionnaires were delivered personally from February to July 2018 for data collection. Out of the total of 550 distributed questionnaires, 403 were returned; for the analysis, 381 were considered appropriate. The sample size was adequate as per Krejcie and Morgan [16] and Tabachnick and Fidell [17]. In comparison to relevant literature, this study’s 69.27% response rate was considered to be highly satisfactory [18]. There were a total of 33 excluded questionnaires, including 12 cases that had missing data for more than 18% of the questions, 11 cases with straight lining and 5 cases as outliers.

## 5 Data Analysis and Results

For this study, structural equation modelling (SEM) was selected as an analytical technique since it allows simultaneous analysis to get enhanced accurate estimates [13, 14, 19–22].

### 5.1 Measurement Model Assessment and Confirmatory Factor Analysis (CFA)

The indices of goodness can be seen in Fig. 2. The SEM software is implemented practically here. Table 1 shows the acceptable outcomes as per the earlier studies. On the basis of Table 1 and Fig. 2, all indices representing *goodness-of-fit* exceeded levels



**Fig. 2** Result of confirmatory factor analysis (CFA)

of acceptance as suggested by the earlier research, thus pointing out that the model of measurement exhibited a good fit compared to the gathered data. The indices representing the total fit show that the chi-square is insignificant ( $p$ -value should be  $>0.5$ ). In spite of the insignificant chi-square, the prototype still fits since the chi-square value almost always discounts the prototype when sizeable samples are considered [23]. The fact that the chi-square is responsive for a sample size of greater than 200 is noteworthy [24], and the size of the sample for this research is 381. Thus, we can proceed to assess the psychometric attributes of the measurement prototype in terms of indicator and construct reliability, and discriminant and convergent validities.

As far as the construct reliability is concerned, the findings indicate that each of the individual alpha coefficient of Cronbach are greater than the recommended level of 0.7 [25]. Moreover, in evaluating construct reliability, all CR (composite reliability) values were larger than the suggested value of 0.7 [26, 27]. This conclusion corroborates that there has been achievement of construct reliability (Table 2). To find out indicator reliability, loadings of factor were examined [28]. The loading for every article surpassed the advised value 0.5, and hence the loadings for each article are fulfilled not counting article OPT4 and article Inn4, which had been removed due

**Table 1** Goodness-of-fit indices for the measurement model

Fit index	References	Admissibility	Result	Fit (Yes/No)
$X^2$			338.338	
DF			125	
p-value		>0.05	0.000	No
<b><math>X^2/DF</math></b>	[27]	1.00–5.00	<b>2.707</b>	<b>Yes</b>
<b>RMSEA</b>	[39]	<0.08	<b>0.066</b>	<b>Yes</b>
SRMR	[40]	<0.08	0.066	Yes
GFI	[41]	>0.90	0.919	Yes
AGFI	[41]	>0.80	0.880	Yes
NFI	[23]	>0.80	0.913	Yes
PNFI	[23]	>0.05	0.746	Yes
IFI	[42]	>0.90	0.943	Yes
TLI	[43]	>0.90	0.930	Yes
<b>CFI</b>	[24]	>0.90	<b>0.943</b>	<b>Yes</b>
PGFI	[44]	>0.50	0.746	Yes

The indexes in bold are recommended because they are frequently reported in the literature [45]. Note  $X^2$  Chi Square; DF Degree of freedom; CFI Comparative-fit-index; RMSEA Root mean square error of approximation; SRMR Standardized root mean square residual; GFI Goodness-of-fit; NFI Normed fit index; AGFI Adjusted goodness of fit index; IFI Increment fit index; TLI Tucker–Lewis coefficient index; PNFI Parsimony normed fit index

to low loading. Also, in order to observe convergent validity, AVE (average variance extracted) was utilised, and all values of AVE were bigger than the recommended value 0.50 [29]. Thus, adequate convergent validity was exhibited successfully. As far as the construct reliability is concerned, the findings indicate that each of the individual alpha coefficient of Cronbach is greater than the recommended level of 0.7 [25]. Moreover, in evaluating construct reliability, all CR (composite reliability) values were larger than the suggested value of 0.7 [26, 27]. This conclusion corroborates that there has been achievement of construct reliability (Table 3). To find out indicator reliability, loadings of factor were examined [28]. The loading for every article surpassed the advised value 0.5, and hence the loadings for each article are fulfilled not counting article OPT4 and article Inn4, which had been removed due to low loading. Also, in order to observe convergent validity, AVE (average variance extracted) was utilised, and all values of AVE were bigger than the recommended value 0.50 [29]. Thus, adequate convergent validity was exhibited successfully.

**Table 2** Measurement assessment

Constructs	Item	Loading (>0.5)	M	SD	$\alpha$ (>0.7)	CR (>0.7)	AVE (>0.5)
Optimism (OPT)	OPT1	0.718	3.405	1.025	0.914	0.82	0.60
	OPT2	0.819					
	OPT3	0.778					
Innovativeness (Inn)	Inn1	0.737	3.395	1.037	0.914	0.78	0.55
	Inn2	0.808					
	Inn3	0.668					
Discomfort (DC)	DC1	0.788	3.259	0.996	0.903	0.84	0.64
	DC2	0.877					
	DC3	0.787					
	DC4	0.731					
Insecurity (InS)	InS1	0.737	3.333	1.091	0.927	0.78	0.55
	InS2	0.808					
	InS3	0.668					
Big data adoption (BDA)	BDA1	0.711	3.201	0.969	0.886	0.87	0.64
	BDA2	0.826					
	BDA3	0.846					
	BDA3	0.813					

Key OPT Optimism; Inn Innovations; DC Discomfort; InS Insecurity; BDA Big data adoption

Note M Mean; SD Standard deviation; AVE Average variance extracted; CR Composite reliability;  $\alpha$  Cronbach's alpha

**Table 3** Discriminant validity assessment

	Factors	1	2	3	4	3
		InS	OPT	Inn	DC	BDA
1	InS	<b>0.894</b>				
2	OPT	0.731	<b>0.894</b>			
3	Inn	0.712	0.605	<b>0.874</b>		
4	DC	0.774	0.611	0.631	<b>0.848</b>	
5	BDA	0.667	0.655	0.511	0.611	<b>0.865</b>

Note Diagonals represent the square root of the average variance extracted while the other entries represent the correlations

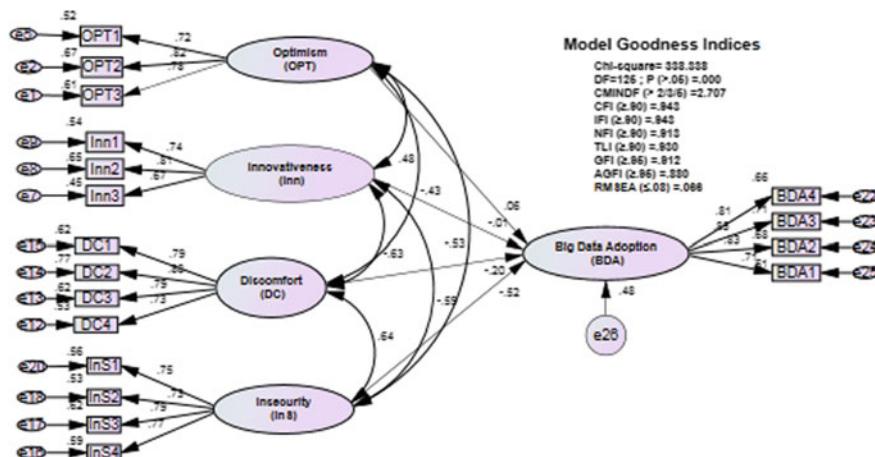
Key OPT Optimism; Inn Innovations; DC Discomfort; InS Insecurity; BDA Big data adoption

## 5.2 Structural Model Assessment

The structural prototype's goodness-of-fit can be compared to the earlier model for CFA measurement. In the case of this structural prototype, the values were documented as  $CFI = 0.943$ ,  $X^2/df = 2.707$ , and  $RMSEA = 0.066$ . These indices of fit point out to the acceptable fit amid the theorised model and experimental data [24]. Therefore, the structural prototype's path coefficients can now be investigated.

### 5.2.1 Direct Hypotheses Tests

The conjectures of this research were verified by employing SEM through AMOS (Fig. 3). The structural prototype assessment displayed in Table 4 gives indication of the experiments on the theories, with all the 4 theories of this research being backed



**Fig. 3** Structural model results

**Table 4** Structural path analysis results

Hypothesis	Dependent variables		Independent variables	Estimate B (path coefficient)	S.E	C.R (t-value)	p-value	Decision
H1	BDA	<—	OPT	0.48	0.039	2.331	0.004	Supported
H2	BDA	<—	Inn	<b>0.53</b>	0.042	2.011	0.022	Supported
H3	BDA	<—	DC	0.20	0.042	2.033	0.011	Supported
H4	BDA	<—	InS	0.52	0.035	2.422	0.017	Supported

S.E Standard error; C.R Critical ratio

Key OPT Optimism; Inn Innovations; DC Discomfort; InS Insecurity; BDA Big data adoption

**Table 5** Coefficient of determination result  $R^2$ 

Exogenous construct	Endogenous construct	$R^2$	Cohen [46]	Chin [47, 48]	Hair et al. (2013)
OPT, Inn, DC, and InS	BDA	0.48	Substantial	Moderate	Moderate

Key *OPT* Optimism; *Inn* Innovations; *DC* Discomfort; *InS* Insecurity; *BDA* Big data adoption

up by optimism ( $\beta = 0.48, p < 0.05$ ), discomfort ( $\beta = 0.20, p < 0.05$ ), innovations ( $\beta = 0.53, p < 0.05$ ), and insecurity ( $\beta = 0.52, p < 0.05$ ) factors, all having an affirmative effect on the big data. Thus, H1, H2, H3, as well as H4 are encouraged. Observe that the coefficient of the standardised path point out to the strengths of the correlation between dependent and independent variables, and so the explicit effects of the factor of innovation on big data acceptance of ADPO are more solid compared to other independent variables.

### 5.2.2 Coefficient of Determination $R^2$ : The Variance Explained

The structural model  $R^2$  value showed that all  $R^2$  values are adequately high in order that the prototype can fulfil a reasonable amount of explanatory power [30] (see Table 5).

## 6 Discussion

Using the suggested prototype, this research provides an improved insight into the role played by the theory attributes such as the readiness of technology as well as other similar aspects in terms of actuality which have a direct effect of embracing big data for providing conditions in estimating the acceptance of big data among the staff members in Abu Dhabi and stresses the relevant consequences. The analyses are given further as follows.

The research discovered that optimism parameter has an affirmative effect on the ADPO adoption of big data among participants, and this result is supported by earlier studies [31, 32]. This result can be justified by the fact that more positivity is useful for enhancements in inspiring the organisation to use few resources on big data depending on their financial situation and to get equipped with the most recent technology for competitive advantage.

Similarly, the innovation certainly affects ADPO adoption of big data among participants, and this result is supported by the earlier studies [32–34]. This result is explained by the fact that increase assisting organisation for turning to the big data system for supporting process available for big data adoption through providing most of the necessary help and resources to enable people to use big data application.

Moreover, the discomfort factor was discovered to have an optimistic effect on the ADPO adoption of big data among participants, and this result is backed up by earlier studies [33]. This outcome can be justified on the basis of the reality that it is uncomfortable when there is trouble related to the big data technology while it is being watched by the people and also when that application is not easy to use. Certain conclusive results from big data are difficult to comprehend.

At last, the factor of insecurity was proved to have an affirmative effect on the ADPO adoption of big data among participants, and this result is supported by the earlier research [35–37]. The justification for this result can be given by the fact that the constructed inputs raise insecurity in the adoption of big data and decrease the self confidence in the use of big data system by having validations and security concerns.

## 7 Implications, Limitations and Future Directions

The theory of the readiness technology (TRD) has played a vital role in understanding what affects the acceptance and adoption of different types of technology applications of big data as a main part of ionisation successes. This work successfully validates TRD in a new context, namely, in the usage of ADPO among employees in a public organisation in the UAE.

This study has inferences for better understanding of the relations among the different significant aspects concerned with the big data technology adoption in a public organisation. The findings should be relevant to researchers, policy makers, and industry players. Given the trend and status of big data and ADPO, it seems probable that they are able to promote UAE in general and in particular Abu Dhabi to appeal to visitors from the entire globe. Thus, the Abu Dhabi tourism agency and policy makers of the government are directed to integrate the adoption of big data into their operational procedures. These positive opinions are also documented in the literature [38]. This research is limited to only a single public sector organisation of the UAE, and so its findings should be considered with caution.

## 8 Conclusion

Big Data is an important player in offering a highly competitive advantage, specialty, in contemporary organisations. This study has inferences for better understanding of the relations among the different significant aspects concerned with the big data technology adoption in a public organisation. The findings should be relevant to researchers, policy makers, and industry players. Given the trend and status of big data and ADPO, it seems probable that they are able to promote UAE in general and in particular Abu Dhabi to appeal to visitors from the entire globe. Thus, the Abu Dhabi tourism agency and policy makers of the government are directed to integrate

the adoption of big data into their operational procedures. These positive opinions are also documented in the literature [38]. This research is limited to only a single public sector organisation of the UAE, and so its findings should be considered with caution.

## References

1. Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The ‘big data’ revolution in healthcare. *McKinsey Quarterly*. Retrieved January 22, 2013 from [http://www.pharmataleents.es/assets/files/Big\\_Data\\_Revolution.pdf](http://www.pharmataleents.es/assets/files/Big_Data_Revolution.pdf).
2. Groves, P., & Knott, D. (2013). The ‘big data’ revolution in healthcare (January).
3. Al-Shamsi, R., Ameen, A., & Al-Shibami, A. H. (2018). The influence of smart government on happiness: Proposing framework. *International Journal of Management and Human Science (IJMHS)*, 2(2), 10–26.
4. Ameen, A., Almulla, A., Maram, A., Al-Shibami, A. H., & Ghosh, A. (2018). The impact of knowledge sharing on managing organizational change within Abu Dhabi national oil organizations. *International Journal of Management and Human Science (IJMHS)*, 2(3), 27–36.
5. Haddad, A., Ameen, A., & Mukred, M. (2018). The impact of intention of use on the success of big data adoption via organization readiness factor. *International Journal of Management and Human Science*, 2(1), 43–51.
6. AL-khatheeri, Y., Ali Ameen, A. H. A.-S., & Lincoln. (2018). Conceptual framework for investigating the intermediate role of information systems between big data factor and decision-making factor. *International Journal of Management and Human Science (IJMHS)*, 2(2), 39–45.
7. Al-Obthani, F., Ameen, A., Nusari, M., & Alrajawy, I. (2018). Proposing SMART-government model: Theoretical framework. *International Journal of Management and Human Science (IJMHS)*, 2.
8. Fahad, A.-O., & Ameen, A. (2017). Toward proposing SMART-government maturity model: best practices, international standards, and six-sigma approach. In *ICMHS 2017 1 st International Conference on Management and Human Science* (p. 2017).
9. Parasuraman, A., & Colby, C. L. (2015). An updated and streamlined technology readiness index: TRI 2.0. *Journal of Service Research*, 18(1), 59–74. <https://doi.org/10.1177/1094670514539730>.
10. Lee, B. C., Yoon, J. O., & Lee, I. (2009). Learners’ acceptance of e-learning in South Korea: Theories and results. *Computers & Education*, 53(4), 1320–1329. <https://doi.org/10.1016/j.compedu.2009.06.014>.
11. Ameen, A., & Ahmad, K. (2014). A systematic strategy for harnessing financial information systems in fighting corruption electronically. In *Knowledge Management International Conference (KMICe) 2014, Malaysia* (pp. 12–15). August 12–15, 2014. Retrieved from <http://www.kmice.cms.net.my/>.
12. Isaac, O., Abdullah, Z., Ramayah, T., Mutahar, A. M., & Alrajawy, I. (2017). Towards a better understanding of internet technology usage by yemeni employees in the public sector: An extension of the task-technology fit (TTF) model. *Research Journal of Applied Sciences*, 12(2), 205–223. <https://doi.org/10.3923/rjasci.2017.205.223>.
13. Ameen, A., & Ahmad, K. (2013). A conceptual framework of financial information systems to reduce corruption. *Journal of Theoretical and Applied Information Technology*, 54(1), 59–72.
14. Ameen, A., & Ahmad, K. (2013). Proposing strategy for utilizing financial information systems in reducing corruption. In *3rd International Conference on Research and Innovation in Information Systems—2013 (ICRIS'13)* (Vol. 2013, pp. 75–80).
15. Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1. <http://doi.org/10.1177/135910457000100301>.

16. Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, 38, 607–610.
17. Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics. *PsycCRITIQUES*, 28, 980. <https://doi.org/10.1037/022267>.
18. Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, 61(8), 1139–1160. <https://doi.org/10.1177/0018726708094863>.
19. Al-Shibami, A. A., & Hamoud, R. A. A. (2018). The influence of smart government on happiness: Proposing framework. *International Journal of Management and Human Science (IJMHS)*, 2(2), 10–26.
20. Isaac, O., Abdullah, Z., Ramayah, T., & Mutahar, A. M. (2017). Internet usage and net benefit among employees within government institutions in Yemen: An extension of delone and mclean information systems success model (DMISM) with task-technology fit. *International Journal of Soft Computing*, 12(3), 178–198. <https://doi.org/10.3923/ijsscomp.2017.178.198>.
21. Isaac, O., Abdullah, Z., Ramayah, T., & Mutahar, A. M. (2017). Internet usage within government institutions in Yemen: An extended technology acceptance model (TAM) with internet self-efficacy and performance impact. *Science International*, 29(4), 737–747.
22. Isaac, O., Masoud, Y., Samad, S., & Abdullah, Z. (2016). The mediating effect of strategic implementation between strategy formulation and organizational performance within government institutions in Yemen. *Research Journal of Applied Sciences*, 11(10), 1002–1013. <https://doi.org/10.3923/rjasci.2016.1002.1013>.
23. Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588–606.
24. Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). Abingdon: Routledge.
25. Kannana, V. R., & Tan, K. C. (2005). Just in time, total quality management, and supply chain management: Understanding their linkages and impact on business performance. *Omega: The International Journal of Management Science*, 33(2), 153–162.
26. Gefen, D., Straub, D., & Boudreau, M.-C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems*, 4(1), 1–79.
27. Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press.
28. Hair, J. F. J., Hult, G. T. M., Ringle, C., & Sarstedt, M. A. (2014). Primer on partial least squares structural equation modeling (PLS-SEM). In *46 Long range planning* (p. 328). London, Thousand Oaks: SAGE. <http://doi.org/10.1016/j.lrp.2013.01.002>.
29. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. New Jersey.
30. Urbach, N., & Ahlemann, F. (2010). Structural equation modelling in information systems research using partial least squares. *Journal of Information Technology Theory and Application*, 11(2), 5–40.
31. Lohr, S., Einav, L., Levin, J., Lohr, S., Einav, L., Levin, J., et al. (2012). The age of big data. *New York Times*, 11(6210), 1–5. <https://doi.org/10.1126/science.1243089>.
32. Wang, Y., Kung, L. A., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126, 3–13. <https://doi.org/10.1016/j.techfore.2015.12.019>.
33. Braganza, A., Brooks, L., Nepelski, D., Ali, M., & Moro, R. (2017). Resource management in big data initiatives: Processes and dynamic capabilities. *Journal of Business Research*, 70, 328–337. <https://doi.org/10.1016/j.jbusres.2016.08.006>.
34. Gu, J., & Zhang, L. (2014). Data, DIKW, big data and data science. *Procedia Computer Science*, 31, 814–821. <https://doi.org/10.1016/j.procs.2014.05.332>.
35. Aldholay, A. H., Isaac, O., Abdullah, Z., Alrajawy, I., & Nusari, M. (2018). The role of compatibility as a moderating variable in the information system success model: The context of online learning usage. *International Journal of Management and Human Science (IJMHS)*, 2(1), 9–15.

36. Ifinedo, P. (2012). Technology acceptance by health professionals in Canada: An analysis with a modified UTAUT model. In *Proceedings of the Annual Hawaii International Conference on System Sciences* (pp. 2937–2946). <http://doi.org/10.1109/HICSS.2012.556>.
37. Sumak, B., Polancic, G., & Hericko, M. (2010). An empirical study of virtual learning environment adoption using UTAUT. In *2010 Second International Conference on Mobile, Hybrid, and On-Line Learning* (pp. 17–22). <http://doi.org/10.1109/eLmL.2010.11>.
38. Amato, F., Castiglione, A., De Santo, A., Moscato, V., Picariello, A., Persia, F., et al. (2018). Recognizing human behaviours in online social networks. *Computers and Security*, 74, 355–370. <https://doi.org/10.1016/j.cose.2017.06.002>.
39. Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25(2), 173–180.
40. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
41. Jöreskog, K., & Sörbom, D. (1998). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International Inc.
42. Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin*, 107(2), 256–259.
43. Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1–10.
44. James, L. R., Muliak, S. A., & Brett, J. M. (1982). *Causal analysis: Models, assumptions and data*. Beverly Hills, CA: SAGE.
45. Awang, Z. (2014). *Structural equation modeling using AMOS*. Shah Alam, Malaysia: University Teknologi MARA Publication Center.
46. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Associates Erlbaum.
47. Chin, W. W. (1998). Issues and opinion on structural equation modeling. *MIS Quarterly*, 22(1), 7–16.
48. Chin, W. W. (1998). *The partial least squares approach to structural equation modeling* (pp. 295–358). New Jersey: Lawrence Erlbaum.

# **Artificial Intelligence and Data Analysis**

# Sports Data Analytics: A Case Study of off-Field Behavior of Players



Malini Patil, Neha Sharma and B. R. Dinakar

**Abstract** The field of sports science is highly emerging and has made the headlines in the research and development activities, with new challenges and trends in the recent past. Sports analytics is the analysis of historical big data mainly available in the form of statistics to provide a proper insight into the entire team, an individual, or even the coach. It also facilitates decision-making for both on-field and off-field performances of all the concerned. The term “sports analytics” was popularized in 2011. It has also created a platform under the confluence of many disciplines such as data mining, machine learning, big data analytics, artificial intelligence, and predictive analytics. The paper aims at analyzing the off-field behavior of players using a statistical approach. The focus of the study is to comprehend the nature of data set, explore the relation between the attributes of the data set, and create a model to understand how the data relates to the underlying population using a real world data set. It comprises motion sensor data of 19 activities among both male and female categories. The sports data set is referred from the UCI repository. It consists of motor sensor data collected for both male and female players. The data set also clearly displays the three dimensions of big data namely, volume, variety, and veracity.

**Keywords** Sports data analytics · Predictive analytics · Big data · Players behavior · Recurrent neural network · LSTM

---

M. Patil (✉)  
J.S.S. Academy of Technical Education, Bengaluru, India  
e-mail: [drmalinipatil@gmail.com](mailto:drmalinipatil@gmail.com)

N. Sharma  
Society for Data Science, Pune, India  
e-mail: [nvsharma@rediffmail.com](mailto:nvsharma@rediffmail.com)

B. R. Dinakar  
Dayanand Sagar Academy of Technology and Management, Bengaluru, India  
e-mail: [dinakar.br96@gmail.com](mailto:dinakar.br96@gmail.com)

## 1 Introduction

Today, sports or sporting events are the important facts of day-to-day life. They have become a major content of the news feed. Though data analytics is a recent term in the field of data science, but the data which is in the form of statistics was the part of the sports industry since early 1870. The first box score was recorded at that time in the basketball match. The introduction of machine learning and data mining techniques made the sports industry to predict many dimensions in sports data and also emphasized on the collection of more fine-grained data. There exist many franchises in today's sports world that makes use of sports analytics to support decision-making, namely game-day decision-making, draft selection, and player evaluation. Other stakeholders are the leagues, media, and sports clubs. They depend on the data to bring in a potential change in rules. Enormous amount of data is collected from various sporting events like Common Wealth Games, Olympics, etc. and majorly encourages the analysis of the players' behavior off-field and on-field.

Incredible amount of data is available in today's digitized world in a variety of forms. From the literature review, it is found that in early days of sporting events, the sports data was collected or recorded in the form of text. Few references to quote: trading cards which provide the complete details about the sport, place of the sports, results about players, etc. Exclusive newspapers on sports were also.

About trading cards, the best example can be a Baseball card. The trading card in baseball game provides the player's defensive position(s), batting average, total hits, number of losses, wins, and saves earned over the past several seasons. It also contains the players shooting percentage, free throws, total points earned by the player, and fouls committed. This data covers the details of recent seasons also about the player and the play. Similar types of trading cards are found in football, hockey, cricket, etc. As now technology support is widened for the data collection method, the recent news feed says, an AI-enabled cricket bat records the player's performance from all dimensions. From the above facts, it is understood that the focus of the sports industry was either business related or performance related.

With the introduction of the buzz words like data mining or machine learning, exciting and impactful work is carried out by researchers and scientists to provide a broad understanding of the way in which the sports industry can utilize the data. These works are mainly focused on identifying the challenges in information systems and technology as well as applications of analytics in the sports industry by transforming the data into useful information and knowledge. Recently, conducted studies prove that they can be used in a variety of domains. However, the effective use of data in sports has shown moderate growth.

The latest survey report says sports analytics and big data are two major buzzwords within the sports sector, whether it be from the business perspective or performance. It is widely acknowledged and not only captures a good amount of data but it also utilizes to improve on-the-field and off-the-field performance, which is critical to success in modern sport. Sports science is an evolving field and each season presents new challenges. Some industry experts from the top leagues National Basket Ball

Association (NBA) and Premier League (Football) to college sport, encourage the use of data and analytics and also highlights the challenges in sports data analytics.

The paper presents the study of sports data set, nature of the variable, and the relationship among them to create a model to generate patterns and insights using a real-world data set. In this paper, we have attempted to use a statistical approach to analyze the off-field behavior of players. The next section reviews the available literature and work done by researchers and sports enthusiasts. Section 3 provides the preliminary study required to understand the sports data analytics domain and Sect. 4 provides the information about Recurrent Neural Network and Long-Short Term Memory Network, which is used to analyze the data. Section 5 presents the experimental results and discussions, Sect. 6 concludes the chapter, followed by references, which are mentioned in the last section.

## 2 Literature Survey

The following section provides the study of research work in the area of sports data analytics, which helps understand the existing research, to make appropriate improvement in the existing work or to suggest new strategies to analyze the sports data. Bonidiaet.al have conducted a considerable amount of survey on several parameters of sporting events [1]. The authors also describe that the use of data mining techniques for the sports data has proved the effective transformation of data into knowledge. The word sports data mining is coined by the work carried out by the authors of this paper, which have seen a breakthrough growth [1]. Shih in his survey, has emphasized on the trends, challenges, and other related fundamentals of sports data analysis related to video content analysis [2]. This method is referred to as sports video analysis as visualization of sports activities plays a major role in the field of sports [2]. Takahashi et al. have very specifically used volleyball as the sporting activity in terms of visualization of the sport [3]. Twenty-four sample players were studies to verify the efficiency of the parameters “speed” and “accuracy” for the decision-making to improve the skills of the game [3]. Knobbe et al. mainly focus on the performance of the athlete before the sporting event “speed skating” [4]. The detailed 15-year-old historical data recorded by the Elite Speed Skating team was taken for analysis. To get the combined training effect of the sports, two different techniques were proposed by the authors. One is based on the physiological model of the athlete and other is based on the sliding window model. Linear modeling and Subgroup Discovery method was used to extract meaningful models of the data [4].

Cheng et al. opted a new approach of sports analysis for handling noteworthy challenges in tracking a ball by using computer vision technology [5]. The techniques identified by the authors are accurateness of predicted 3D ball trajectory under tough conditions, irregular motions of the ball due to external forces added by players, and unpredictable conditions in the actual game. The author proposed an anti-occlusion observation model, an abrupt motion adaptive system model, and a spatial density-based automatic recovery based on particle filter [5]. Baumer et al have focused on

the comprehensive statistical analysis of the players' performance in baseball game [6]. The authors of the paper have used one measure called as WAR (wins above replacement) strategy to aggregate the contributions of the players at each stage of the game such as hitting, pitching, fielding, etc. The authors have proposed another competitive measure called as OPENWAR based on public data and nebulous concept of a "replacement" player [6]. The success of prediction of National Football League (NFL) is based on the complex drafting strategies. Mulholland et al have highlighted the analysis on the tight end position of the player on the field [7]. A special predictive model is created using linear regression and recursive partitioning decision trees to predict NFL draft and NFL career based on the performance draft which is available. It is found that the size measures BMI, weight, and height are given more weight age in the NFL draft, which contributes to the players' performance. The work carried out by the authors is found very interesting for analyzing the players' performance in NFL [7].

The introduction of machine learning algorithms focused on the performance of players and relevant predictions also. Lopez et al. in their paper have proved that statistical methods are more often efficient means of predicting the NCAA men's football tournament [8]. The authors have proposed a predictive model that merges the point spreads set by Las Vegas sports books using logistic regression analysis. They have predicted the optimistic game scenarios with the accuracy of 12% in the success of the game and 50% chance computing one of the best ten scores [8]. Becker et al. have worked on a fantasy football, an online game [9]. The work is to check the players' performance of the National Football league. The statistics available indicate that approximately 35 million people in Canada and US play online fantasy sports in major websites YAHOO, ESPN, MSN, NFL, etc. sports online. The authors have developed an optimized methodology to handle this historical data and that predicts team and player performance [9].

Kolbush et al. focus have developed a model using logistic regression/Markov Chain method to rank basketball teams in NCAA [10]. Experimental results show that ranking system F-LRMC followed by Massey's College Football Ranking Composite is among the best [10]. The athlete's performance in runner's model for a 24-h ultra-running race is developed as a trajectory model [11]. It uses a clustering approach based on the speed, age, and gender of runners. Expectation–maximization algorithm is the best fit here. Base data used by authors belong to 2013 World championship. Irrespective of age and gender, the average moving speed and their propensity to rest during the race were the vital data for analysis [11].

### **3 Sports Data Analytics (SDA)**

#### ***3.1 Background***

In recent times, sports analytics has become a major domain in the domain-specific analytics scenario of data science. The precise definition of sports analytics is stated as a process in which statistics is applied to relevant and historical sports data to offer an edge to the entire team or an individual. Sports analytics obtained the importance and popularity with the release of the film “Money Ball” in which the main focus was given to the competitive team building tactics with less budget on the basis of analytics. The main stakeholders identified for sports analytics are players, managers, coaches, and other related staffs. The results of sports analytics can be best used by all the stakeholders. Every sporting event is governed by a set of rules to be followed by the players and other stakeholders, recording of results of the performance of players, and also the recording of complete visual displays. These serve the sports events by way of providing fair competition, judgment, and fair evaluation of the winners. Each sport activity has a set of rules or regulations that governs the game. Records of performance are often kept as historical data and it will be announced and reported in the sporting events.

Sports analytics can be categorized into two parts, one is on-field and other is off-field analytics. On-field analytics focus on the on-field performance of the team and individual player and mainly about the overall tactics involved in the game and fitness of the players. Off-field analytics focuses on the business aspect of sports by helping a sport organization to increase the popularity, scope, viewers' interest, attract the other business supporters, retailers of the game, and provide an improved decision-making system leading to profit and growth. As a part of academics in the year 1998, sports engineering is emerged as a discipline. The main focus is given to the use and implementation of technology like big data, sports analytics, wearable technology etc. in sports.

#### ***3.2 Sports and Technology***

Modern sports depend on modern technology. Improvement in the performance of the sporting event among the stakeholders is the outcome. Decision-making and team building are the special features here. As mentioned earlier, academic discipline sports science can be applied to some of the key areas identified by the sporting events. They are athlete performance, video analysis to fine-tune their own technique of playing the game, improved running shoes or other suitable and comfortable wears/clothing and good swimwear, etc. The increased use of technology led the decision-making very transparent and fair. “REPLAY”, “THIRD UMPIRE” are few such technology supports extended for sporting events. In international cricket, if the decision is challenged, the third umpire is the final decision maker.

The growth of sports analytics is noticeable with respect to the growth of technology. This is possible due to in-depth data collection in one decade. This leads to the use of advanced statistical metrics and tools and led to the growth of the sports of the specific technologies. The technique of simulation and performance analysis plays a vital role here. The technological innovations also support the researchers for the proper interpretation of the data in the recent years and also provide new challenges for analysis.

## 4 Algorithms for Sports Data Analytics Models

The sports data is a sequential data like speech, time series, video, text, weather, and financial data. In this research work, the sports data analytics is performed with the help of machine learning model created using recurrent neural networks (RNN) and LSTM Recurrent Neural Networks. The algorithms are chosen as these are looped networks and has a memory, which allows the information to persist, unlike tradition neural networks where it is difficult to reason about the event of previous second to predict the event of next second. RNN was first introduced in 1980, but it displayed its optimal potential only after the availability of huge data, increase in computational power, and invention of Long-Short Term Memory (LSTM) in the 1990s. RNN can remember the input because of the internal memory and process it to precisely predict the next event.

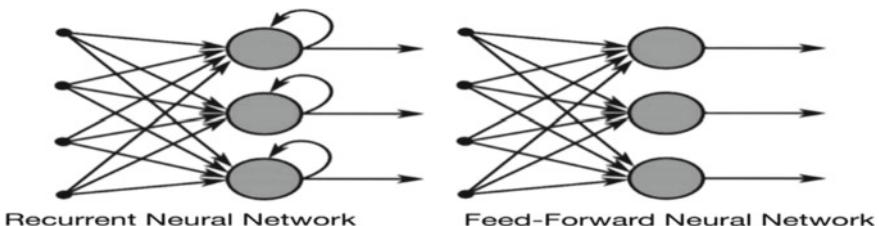
### 4.1 Recurrent Neural Network (RNN)

RNN is the modification of Feedforward Neural Network. The Feedforward Neural Network has three layers (input, hidden, and output) and the information flows straight from input to output layer through a hidden layer. The network works on current data and has no memory to store previous input and hence cannot predict next item in the sequence. Whereas, in RNN, the information flows in a loop and stores the previous learning in a memory, which can be short-term or long-term memory. The analysis is done using current as well as previous information. Figure 1 shows the difference between both the networks.

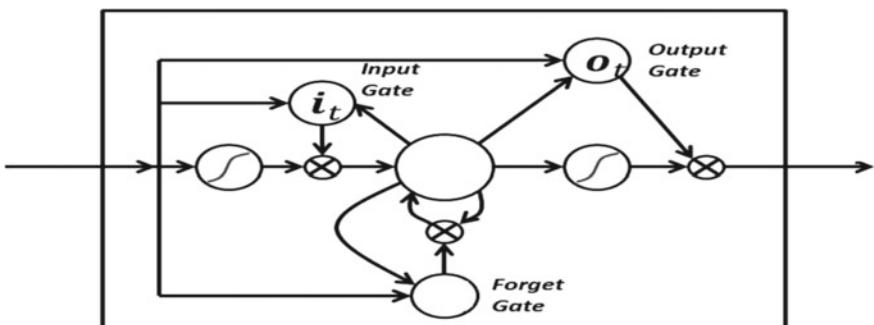
Feedforward Neural Network assigns a weight to the input before it produces the output. Whereas, RNN has two inputs (the current and the previous) and it assigns a weight to both, and then produce the output. The two major challenges with RNN are exploding gradient and vanishing gradient. The gradient is the rate of change of weight with respect to the change in error. When the algorithm assigns high weight without any purpose, it leads to exploding gradient. It can be handled by truncating the gradient. However, when the algorithm assigns too small weight then the model takes too long to learn or stops learning and is known as vanishing gradient. It can be resolved with the help of Long-Short Term Memory.

## 4.2 Long-Short Term Memory (LSTM)

LSTM networks extend the memory of recurrent neural networks, which makes it capable to learn from experiences of the past that have a long gap. LSTM is used with the layers of RNN and form a network to remember the inputs from the past for a longer period of time. LSTM is similar to the computer's memory as the information can be read, written, and deleted from it. The memory can be perceived as a gated cell that takes the decision to store or delete the information, depending on the weight assigns to the information by the algorithm. There are three types of gates: (i) Input gate—takes the decision to allow the new input, (ii) Forget gate—takes the decision about deleting the information if it is redundant, and (iii) Output gate—takes the decision about impacting the output at the current timestamp. Therefore, LSTM learns over time regarding the importance of the information. Figure 2 illustrates the RNN with all the three gates of the LSTM network.



**Fig. 1** Comparison between RNN and FFNN



**Fig. 2** Recurrent neural network with three gates of LSTM network

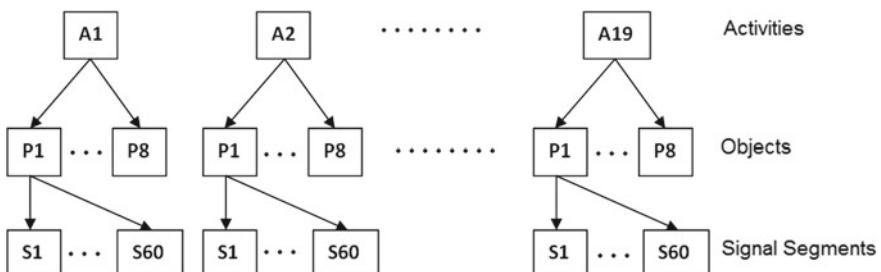
## 5 Experimental Results

### 5.1 Data Set Description

The data set used in the research is accessed from UCI machine learning repository, which is an exclusive collection of data set from various domains, created explicitly for machine learning enthusiasts to experiment with the machine learning algorithms [12]. This FTP archive is created in 1987 by David Aha at UC Irvine. The data set used for sports analytics is multivariate, three-dimensional, time series with 5 classes and 3 subclasses, consisting of 9120 instances and 5625 attributes. The structure of data set is shown in Fig. 3 and Table 1. The 3D view of the structure of data set is shown in Fig. 4. The sports data is usually a motion sensor data which records the activity. The details of the data regarding 19 activities used in this research are described in this section. Each of the 19 activities is performed by 8 players (4 female, 4 male, age group is between 20 and 30) for 5 min [12–15]. For every activity of each player, the signal duration is 5 min. The players are asked to perform the activities in their own style and were not restricted on how the activities should be performed. For this reason, there are intersubject variations in the speeds and amplitudes of some activities. Sensor units are calibrated to acquire data at 25 Hz sampling frequency. Overall observation of all 60 data sets are discrete in nature. The 5-min signals are divided into 5-s segments so that 480 (=60 × 8) signal segments are obtained for each activity. Eight subjects ( $p$ ), 60 segments ( $s$ ), 5 units on torso ( $T$ ), right arm (RA), left arm (LA), right leg (RL), left leg (LL), and 9 sensors on each unit ( $x, y, z$  accelerometers,  $x, y, z$  gyroscopes,  $x, y, z$  magnetometers).

### 5.2 Experimental Framework and Result

The sports data analytics is performed using RNN-LSTM algorithm, which is implemented on an Intel Core i7-4790 processor with 8 M Cache with 4.00 GHz speed using Keras. Keras is an API that is developed with a focus to perform experimenta-



**Fig. 3** Structure of sports data set

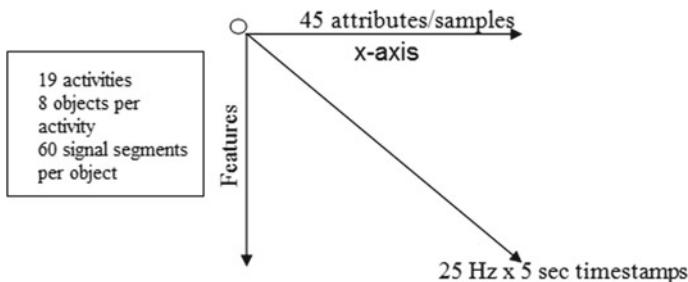
**Table 1** Description of sports data set

Number of Activities in the data set: 19  
 (Folders a01, a02, ..., a19 contain data recorded from the 19 activities)

No	Description of the activity	Notation	No	Description of the activity	Notation
1.	Sitting	A01	11.	Walking on a treadmill with a speed of 15° inclined positions	A11
2.	Standing	A02	12.	Running on a treadmill with a speed of 8 km/h	A12
3.	Llying on back	A03	13.	Exercising on a stepper	A13
4.	Lying on back on right side	A04	14.	Exercising on a cross trainer	A14
5.	Ascending stairs	A05	15.	Cycling on an exercise bike in horizontal positions	A15
6.	Descending stairs	A06	16.	Cycling on an exercise bike in vertical positions	A16
7.	Standing in an elevator still	A07	17.	Rowing	A17
8.	Moving around in an elevator	A08	18.	Jumping	A18
9	Walking in a parking lot	A09	19.	Playing basketball	A19
10.	Walking on a treadmill with a speed of 4 km/h in flat	A10			

Number of subjects in the data set: 8 with age group 20–30 years  
 (For each activity, the subfolders p1, p2, p3, p4, p5, p6, p7, p8 contain data from each of the eight subjects)

No	Male players	Notation	No	Female players	Notation
1.	M_Player 1	SM 1	5.	F_Player 1	SM 1
2.	M_Player 2	SM 2	6.	F_Player 2	SM 2
3.	M_Player 3	SM 3	7.	F_Player 3	SM 3
4.	M_Player 4	SM 4	8.	F_Player 4	SM 4
1.	Correspond to the sensors in unit 1 (T) Txacc, Tyacc, Tzacc, Txgyro, Tygyro, Tzgyro, Txmag, Tymag, Tzmag,				
2.	Columns 10–18 correspond to the sensors in unit 2 (RA) RAxacc, RAYacc, RAzacc, RAxgyro, RAYgyro, RAzgyro, RAxmag, RAYmag, RAzmag				
3.	Columns 19–27 correspond to the sensors in unit 3 (LA) LAXacc, LAyacc, LAzacc, LAXgyro, LAygyro, LAzgyro, LAXmag, LAymag, LAzmag				
4.	Columns 28–36 correspond to the sensors in unit 4 (RL) RLXacc, RLyacc, RLzacc, RLXgyro, RLgyro, RLzgyro, RLXmag, RLymag, RL_zmag				
5.	Columns 37–45 correspond to the sensors in unit 5 (LL) LLXacc, LLyacc, LLzacc, LLXgyro, LLgyro, LAzgyro, LLXmag, LLymag, LLzmag				



**Fig. 4** 3D view of structure of sports data set

tion without any delay, and is written using Python. It is basically a neural network of very high level that can run on the top of CNTK, TensorFlow or Theano. Keras is preferred due to its fast and easy prototyping feature. It supports recurrent and convolution networks, as well as the hybrid network and runs seamlessly GPU and CPU. 75% of the subjects from the sports data set (from P1 to P8) has been randomly taken as training sets and the remaining 25% are considered as test set. The same random selection process has been applied to 60 signal segments of each P1–P8, to consider S1–S45 as training set and s46–s60 is considered as test set. The implementation details indicate that the number of instances in training data set is 6840 and in test data set are 2280, which equals 9120 instances as per the given sports data set.

The experiment is made to run for three times by keeping the epoch as 6 and with varying batch sizes 45, 60, and 100. The results are tabulated in Table 2 for all the 19 activities. The graphical representation is shown in Fig. 5. The interesting fact to note here is that the behavior of players in case of activity 2(standing). As per the condition of activities, players are free to perform any activity for the given time. The analysis result shows that irrespective of subjects (male, female), it is the standing activity, which makes the player's behavior slightly different. In case of activity 2, correctly classified instances a101 and incorrectly classified instances are 19 in training phase 1. Similarly in T2, CCI = 84, ICI = 36, In T3 CCI = 78 and ICI = 42. But in all the training phases, the learning accuracy and predictive accuracy is found to be good.

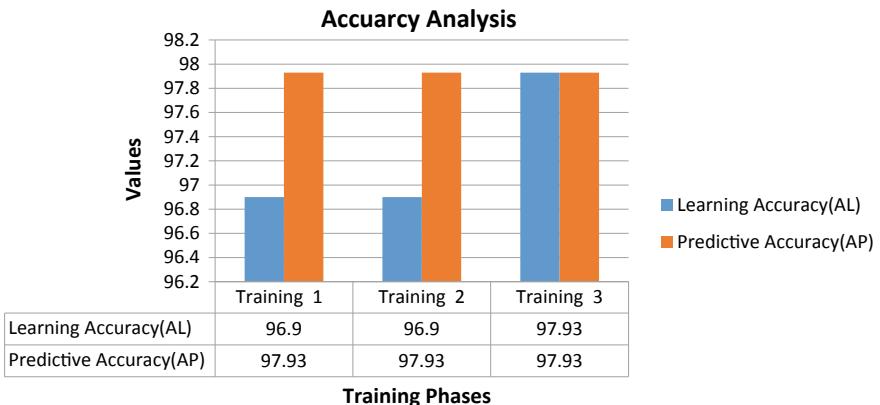
## 6 Conclusion

Sports or a sports science is a term, which has geared the wider interest of researchers across the world. Sports analytics is a successful domain specific initiative with the growth of technology. As this field is new, the primary focus is to promote success stories so that the target audience is made aware of the future goals and objectives of sports analytics. Until recently, there was a notion in sports analytics includes a lot of mathematical and statistical techniques that are difficult to understand and apply. With this motivation, the authors have made a sincere effort to conduct the

**Table 2** Results of CCI and ICI using RNN-LSTM

Training 1			Training 2			Training 3			
Learning accuracy (AL) = 96.90			Learning accuracy (AL) = 96.90			Learning accuracy (AL) = 96.90			
Predictive accuracy (AP) = 97.93			Predictive accuracy (AP) = 97.93			Predictive accuracy (AP) = 97.93			
Epoch = 6		Batch size = 45		Epoch = 6	Batch size = 60		Epoch = 6	Batch size = 100	
Activities	CCI	ICI	Activities	CCI	ICI	Activities	CCI	ICI	
A01	120	0	A01	119	1	A01	118	2	
A02	101	19	A02	84	36	A02	78	42	
A03	120	0	A03	120	0	A03	120	0	
A04	119	1	A04	118	2	A04	119	1	
A05	120	0	A05	120	0	A05	120	0	
A06	120	0	A06	120	0	A06	118	2	
A07	120	0	A07	120	0	A07	120	0	
A08	117	3	A08	118	2	A08	120	0	
A09	120	0	A09	120	0	A09	115	5	
A10	120	0	A10	120	0	A10	120	0	
A11	118	2	A11	120	0	A11	120	0	
A12	120	0	A12	120	0	A12	120	0	
A13	105	15	A13	120	0	A13	120	0	
A14	120	0	A14	119	1	A14	105	15	
A15	120	0	A15	120	0	A15	120	0	
A16	120	0	A16	120	0	A16	120	0	
A17	120	0	A17	120	0	A17	120	0	
A18	120	0	A18	106	14	A18	120	0	
A19	113	7	A19	116	4	A19	112	8	

experiment on sports data using RNN and LSTM techniques. The data set used for sports analytics is multivariate in nature, three-dimensional, time series with 5 classes and 3 subclasses, consisting of 9120 instances and 5625 attributes. The work uses the benchmark data set from the UCI repository. The experiment is made to run for three times by keeping the epoch as 6 and with varying batch sizes 45, 60, and 100. The analysis result shows that irrespective of the gender of players, it is the standing activity which makes the player's behavior slightly different. In case of activity 2, correctly classified instances a101 and incorrectly classified instances are 19 in training phase 1. Similarly in T2, CCI = 84, ICI = 36, In T3 CCI = 78 and ICI = 42. But in all the training phases, the learning accuracy and predictive accuracy is found to be good. Understanding and interpretation of data set was very unique in nature. Selection of RNN and LSTM proved that the training data and



**Fig. 5** Accuracy analysis

test data selection is very feasible. Thus, this work proves to be one of the unique works conducted by exploring the unique sports data set and to certain extent, it has overcome the notion of “How sports data set can be used”. The work forms the basis for new methodology, which is relatively simple.

**Acknowledgements** The authors wish to acknowledge the UCI Machine Learning repository, Centre for Machine Learning, and Intelligent Systems for providing the data set.

## References

1. ParmezanBonidia, R., DuilioBrancher, J., & Marques Busto, R. (2018). Data mining in sports: A systematic review. *IEEE Latin America Transactions*, 16(1), 232–239.
2. Shih, H.-C. (2017). A survey on content-aware video analysis for sports. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(5), 1212–1231.
3. Takahashi, M., Ikeya, K., Kano, M., Ookubo, H., & Mishina, T. (2016). Robust volleyball tracking system using multi-view cameras. In *2016 23rd International Conference on Pattern Recognition Pattern Recognition (ICPR)* (pp. 2740–2745).
4. Knobbe, A., Orie, J., Hofman, N., et al. (2017). Sports analytics for professional speed skating. *Data Mining and Knowledge Discovery*, 31, 1872. <https://doi.org/10.1007/s10618-017-0512-3>.
5. Cheng, X., Ikoma, N., Honda, M., & Ikenaga, T. (2017). Ball state based parallel ball tracking and event detection for volleyball game analysis. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 100(11), 2285–2294.
6. Baumer, B., Jensen, S., & Matthews, G. (2015). openWAR: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports*, 11(2), 69–84. Retrieved October 31, 2018 from <https://doi.org/10.1515/jqas-2014-0098>.
7. Mulholland, J., & Jensen, S. T. (2014). Predicting the draft and career success of tight ends in the national football league. *Journal of Quantitative Analysis in Sports*, 10(4), 381–396. <https://doi.org/10.1515/jqas-2013-0134>.

8. Lopez, M. J., & Matthews, G. (2015). Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, 11(1), 5–12.
9. Becker, A., & Sun, X. A. (2016). An analytical approach for fantasy football draft and lineup management. *Journal of Quantitative Analysis in Sports*, 12(1), 17–30.
10. Kolbush, J., & Sokol, J. H. (2017). A logistic regression/Markov chain model for American college football. *International Journal of Computer Science in Sport*, 16(3), 185–196.
11. Bartolucci, F., & Murphy, T. B. (2015). A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. *Journal of Quantitative Analysis in Sports*, 11(4), 193–203.
12. <https://archive.ics.uci.edu/ml/datasets/daily+and+sports+activities> (accessed on 10–11–2018).
13. Altun, K., Barshan, B., & Tunçel, O. (2010). Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10), 3605–3620.
14. Barshan, B., & Yüksek, M. C. (2014). Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, 57(11), 1649–1667.
15. Altun, K., & Barshan, B. (2010). Human activity recognition using inertial/magnetic sensor units. In *Proceedings First International Workshop on Human Behavior Understanding (in conjunction with the 20th International Conference on Pattern Recognition)*, August 22, 2010. Istanbul, Turkey,

# Phrase Based Information Retrieval Analysis in Various Search Engines Using Machine Learning Algorithms



S. Amudha and I. Elizabeth Shanthi

**Abstract** Query-based information retrieval is an essential part of the web search engine. Many researchers have applied different types of web mining technologies to find more relevant information based on the keyword but are not able to know the correct meaning of the term (keyword) single, multiword or phrases. In this paper we address this problem of searching phrases. In this work the phrase searching process is three-fold as whole Phrase, Sequence of term in phrase and Mingle of the term in the phrase. Here the user enters a query as phrases that is passed to various search engines and retrieves the top ' $n$ ' list of web pages. Initially preprocessing is performed on the Sequence of Keyword in phrase and Mingle of the keyword in the phrase. Then feature extraction is done based on the web pages in the various search engines using term Frequency-Inverse Document Frequency method. Following the feature extraction, grouping of the top ' $n$ ' list of web pages from various search engines based on the parameter as title-based, snippet-based, content-based, address-based, link-based, uniform resource locator-based, and co-occurrence-based calculation is done using LBG clustering algorithm. Then identified the unique link from the above grouping of web pages from the various search engines using SVM classifier and assigned the rank value to the unique link web pages are done using proposed ranking algorithm. Finally it is observed from this experiment that precision, recall,  $f$ -measure, accuracy, speed, and error rate show significant improvement than the traditional search engines.

**Keywords** Information retrieval · LBG clustering · Machine learning algorithm · Search engine · SVM classification

---

S. Amudha (✉) · I. Elizabeth Shanthi

Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India  
e-mail: [amudhajaya@gmail.com](mailto:amudhajaya@gmail.com)

I. Elizabeth Shanthi  
e-mail: [elizabeth\\_cs@avinuty.ac.in](mailto:elizabeth_cs@avinuty.ac.in)

## 1 Introduction

Nowadays search engines are more essential tools for retrieval of information from large repository and fast-growing corpuses in web. Search engine is a software that performs retrieval of a web page (document) based on the user's entered query in the form of the keyword and output a list of web pages using WebCrawler [1, 2]. An information retrieval process aims to find more relevant information to users based on keyword frequencies and also finds the similarity of the keyword [3]. The World Wide Web has turned into the biggest archive and it's more difficult to search in it. The traditional search engines provide a huge amount of search results based on the user query. User query submitted to the traditional search engine normally contains set of terms. For example, suppose if a User search as a query as a phrase "java is a programming language", it contains 3 terms like java, programming and language. The phrase query contains a number of query terms available in the web documents [4, 5]. The traditional search engine retrieval process is very difficult because if retrieval information in web the user clearly specify the information requirements. Information retrieval systems based on the phrases that assign the index of the document, searching the document, assign the ranking, and specify the document on the net [6, 7]. Those phrases are identified as good phrase or bad phrase and match those phrases to all web pages and retrieve similar web pages to the good phrase or specific words [8].

Accordingly the phrase based on retrieval of information system is identifying the phrase, index the web documents depend on the phrase, search and ranked top final web documents [9, 10]. The phrase retrieval system uses the following steps Identifying phrases and related phrases, Indexing web document based on phrase query, ranking web document with the phrase, create a description of the web document and eliminate the duplicate documents [11].

In this paper used a novel method for ranking the search information in various search engines based on the phrase query. These phrases pass to the various search engines in different way to find more relevant web pages in the repository [12, 13]. The first method passes the whole phrase without preprocessing to the multiple search engines and retrieves the top ' $n$ ' pages from each search engine. The second method is phrase query has preprocessed and finds the number of the term in the query. The numbers of term form in sequence and pass to the multiple search engines and retrieve the most relevant to the sequence term. The third method is the numbers of terms form in mingle, search a relevant web page and retrieved from multiple search engines. Then top ' $n$ ' web page results are collected from multiple search engines and filter the web pages using term frequency and inverse document frequency. Accordingly, using a LBG clustering algorithm to group the similar web pages based on the method of phrase and to optimize the unique web pages using the SVM classifier. Finally, to assign a rank value of the unique top link web pages and produced the final optimized result using a ranking algorithm.

The rest of the paper is organized as follows: in the following section, we describe the related work, while Sect. 3 details the phrase based search on various search engines, highlighting the differences with traditional search engines. A section 4

describes the experiments results and evaluation methods. Finally, Sect. 5 addresses concluding comments and remarks.

## 2 Related Work

**Aramatzis et al.** [1] have developed a IRENA (information retrieval engine based on natural language analysis) system using NLP techniques for document retrieval based on the precision and recall values. This system is mainly based on noun phrase and using co-occurrence matrix techniques to identify the similar word occur in the phrase text. They collected music text corpus (6.7 Mb) manually which contains magazine article, FAQ about article, interviews, reviews etc. Noun phrase compared in three ways are keyword, keyword with morphological, keyword with synonyms, morphological variants.

**Lang et al.** [3] have proposed prediction of the query performance for information retrieval based on covering topic score. They estimate the topic of a user's query is covered by the document retrieval from the particular retrieval system and also incorporate with the features in bag of words like phrases and proximity of terms. This experiment Covering Topic Score constantly performs better than the previous methods like high effectiveness, low complexity.

**Masłowska** [4] has proposed phrase based searching and retrieving information using the hierarchical clustering. They proposed clustering techniques for online processing of web documents based on the time, requirement of cluster text order of the reading web pages.

**Laura et al.** [6] have created a search engine for semantic illegal content hunter (SICH) and it's used to automatically identify the illegal content in the internet. This SICH system is categorized into three sections as crawler, indexer and query processor. It provides better results for searching, downloading, filtering, clustering and organizing illegal content contained in unstructured text documents.

**Adriani et al.** [8] have proposed a system for query expansion based on the term similarity with cross language an information retrieval using TREC corpus. They proposed novel dictionary-based cross-language information retrieval (CLIR) method for a query expansion is based on a similarity measure between terms to improve the effectiveness. They demonstrated the effectiveness using combining of two techniques query expansion and translation of queries with cross language namely German, Spanish, and Indonesian are improved the performance. Finally indicate that the term similarity better than more phrase in the queries.

**Bhatia et al.** [11] have developed a system for automatic classification of web search query into five classes as low frequency high entropy, low frequency low entropy, high frequency low entropy, high frequency and high entropy. This work uses the commercial search engine user queries and click entropy. Then reformulation of query achieves better results using query-based and click based.

**Stoyanchev et al.** [12] has proposed system based on the question answering to retrieved information in phrases using AQUAINT dataset. This work automatically

finding the question is similar to the search query and use the two datasets are AQUAINT of 3 gigabyte of news documents and TREC 2006. In this work improve the result of sentence retrieval calculation in WEB dataset. Finally 9.5% improved the performance of automatically generated phrase compared with manual annotated phrase.

**Fatmawati et al.** [13] have analyzed the common phrase method on information retrieval system in the English news. The initial process of research is to find the relevancy level of the documents listed. The dataset collected has 100 documents and 20 queries with English news text. The dataset passes through preprocessing steps viz tokenizing and stemming process. The tokenization process is splitting of sentence to words and stemming process uses porter stemming algorithm to removing the ion, ious, ed, ing, ect in the word. The second step is finding the common words, each term calculates the frequency of occurrence in all documents and add each term frequency to get the total frequency. The third step is term weight calculation on every term to each document and query. The next step is to find the similarity between the two documents in each query in corpus. This system calculates the performance in two manners (i) determine the relevant documents using kappa statistic (ii) determine the success rate like precision, recall and  $f$ -measure. Finally the experiment produced more relevant documents are eligible for evaluation and success rate of relevant document is low.

**Mala et al.** [14] have proposed a system for combining the two techniques semantic and keyword depends on the web searching information and retrieval information like a new search engine. Semantic search engines and keyword search engine are selected based on their precision ratio, natural language queries. The selected semantic search engines are Wikipedia, Google, yahoo and keyword search engines are Hakia, Bing, DuckDuckGo. Finally compared the performance evaluation of the relevant and non-relevant document is better than the semantic and keyword search engine features.

**Li et al.** [15] have developed a system for hierachal clustering of web snippets based on the phrases. This system has used five datasets are topology, geometry, number theory, meat, knowledge and have four parameters classes, documents, words, phrases. They have creates a novel method for hierachal clustering of web snippets with index values of document. Phrase-based all snippets and its move to the corresponding cluster with phrase.

**Yu et al.** [16] have proposed a system for phrase-based topic modeling for semantic information processing system for biomedicine. This system uses phrase based LDA model converted from bag of words to bag of key phrases using key phrase extraction method. The proposed system incorporates with open source interactive topic browser and qualitative analysis of browser with biomedical information.

### 3 Proposed Methodology

The main objective of the work is retrieving the most relevant information from various search engines. The modules are phrase query processor, searching process and resulting process. Figure 1 shows the process of phrase query.

#### 3.1 Phrase Query Processor (PQP)

The initial process of PQP is getting the phrase query from the user and transfer to the various search engines. Before search query is selecting the file type as pdf,

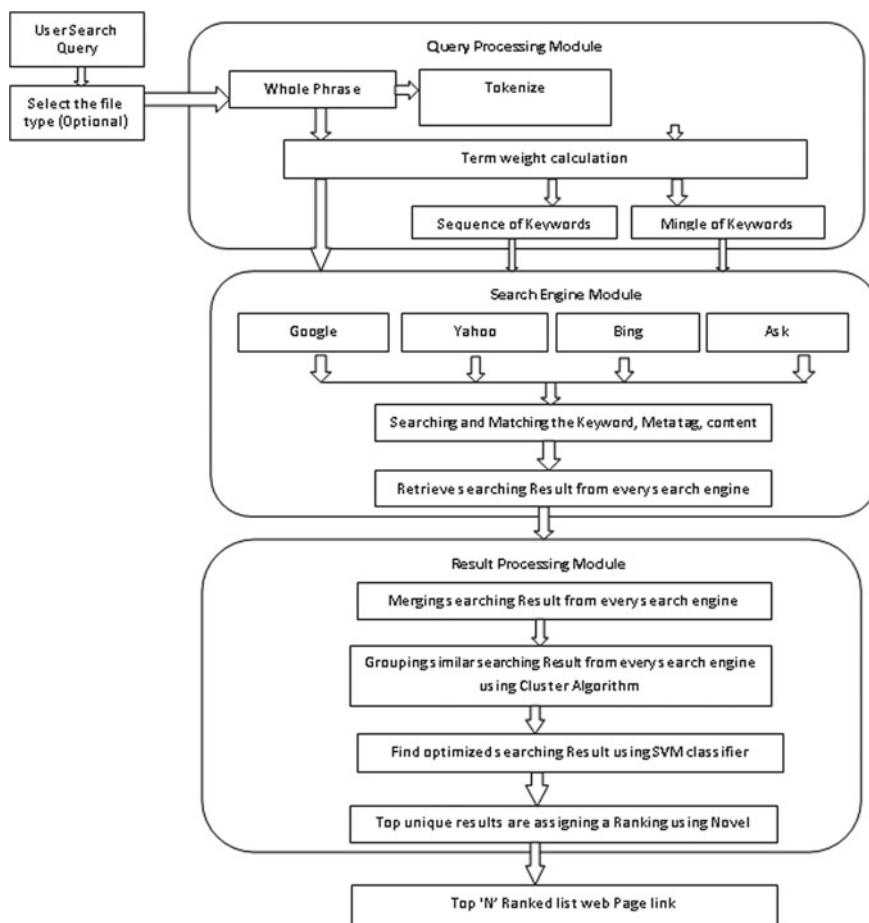


Fig. 1 PQP architecture

word, excel, text formats. These search engines match number of terms with the link content, content of the web pages and meta-tag content. There are three methods using PQP. The first method passes the whole phrase without preprocessing to the multiple search engines. The second method is phrase queries to perform preprocessing are tokenization, removal of the stop word list and finally perform the stemming. Then to find the number of the term in the phrase query and the number of terms form in sequence of terms and calculate the term weight in the number of terms in the query. Hence, based on the number of terms to form a sequence of terms and mingle of terms to transfer to the searching module.

### **3.2 Searching Process**

The above module produces the three types of the query results. There are whole phrase, sequence of term, mingle of term and these passed to the various search engines (Google, Yahoo, Bing, and Ask). Each search engines retrieved the top ‘ $n$ ’ web pages based on the type of query type and similarly matching with the parameters are meta-tag and content of the web pages. The window size is set as one; retrieve the first web pages from the search engines. For example, Google retrieved first top ‘10’ pages like similar to the remaining search engine. Entirely retrieve the 40 web pages from the four search engines. The result web pages are transferred to the next module.

### **3.3 Retrieving Result**

From the above module results use to perform the fusion from the all search engine top ‘ $n$ ’ results. The fusion results are split into the similar results and grouped together using the LBG cluster algorithm. Then find the optimized results using the SVM classifier.

#### **3.3.1 LBG Cluster Algorithm**

The LBG algorithm is similar to  $K$ -means algorithm, but creates a quantizer contain total distortion less or equal to previous cluster. The initial process is a whole training vector of data that represents as one cluster and calculates the centroid is called as code vector (CV). The mean vector of information in every step proceeds with CV and it was selected for the entire cluster with the centroid. The Euclidian distance calculated for two vectors are  $V+V+$  and  $V-V-$ . Then forming two clusters, one cluster contains minimum Euclidian distance and another contains maximum Euclidian distance. After calculating the centroids of the above two vectors and create a

new two CV. Similarly next iteration forms a four CV until reaches as the desired size was generated. The following are the step of LBG algorithm

- Create a vector codebook and calculate the centroid of the whole training vectors.
- Then double the size of the codebook by dividing into the new codebook  $V_i$   $V_i$  based on the rule.

$$V + i V - i = Vi(1 + \varepsilon) = Vi(1 - \varepsilon)Vi+ = Vi(1 + \varepsilon)Vi- = Vi(1 - \varepsilon)$$

- Then new codebook is calculating centroids.
- Every time the cluster algorithm divides the data into two sections, otherwise merged the data.
- Repeat the step to reach the finishing condition.

### 3.3.2 SVM Classifier

Support vector machine (SVM) classifier performs the classification task by creating hyperplane in a multidimensional space. The multidimensional space then separates into the different class labels like relevant document and irrelevant document in the information retrieval processing system. SVM performs both task classification and regression step. It can also handle multiple sequences of variable and categorical variables or dummy variables. The dummy variables contain either 0 or 1 and dependent variable contains three states viz.,  $A$ ,  $B$  and  $C$  denoted by a set of 3 dummy variables.

$$A : \{1 \ 0 \ 0\}, B : \{0 \ 1 \ 0\}, C : \{0 \ 0 \ 1\}$$

To build an ideal hyperplane, SVM utilizes an iterative training algorithm, which has the advantage of reducing the error function. For this, the training algorithm for reducing the error function with constraints is given below:

$$\frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i$$

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0, i = 1, \dots, N$$

where the variable  $C$  is the capacity constant, constant is represented in variable  $b$ , vector of coefficients is represented in variable  $w$ , and input variable for handling nonseparable data in  $\xi_i$ . The  $N$  training has index  $i$  labels, independent variable represented as  $x_i$ , class label is represented as variable  $y \in \pm 1$ . The kernel  $\phi$  is used to change information from the independent input to the feature space and it ought to be noticed that the bigger the  $C$ , the more the mistake is punished. Along these lines,  $C$  ought to be picked with care to maintain a strategic distance from overfitting.

## 4 Experimental Results and Evaluation

This experiment has been carried out for retrieving information from the web. The user searches the information in various search engines based on the phrase. These phrases pass the various search engines in different ways to find more relevant web pages in the repository. The first method passes the whole phrase without preprocessing to the multiple search engines and retrieves the top ‘ $n$ ’ pages from each search engine. The second method is phrase query has preprocessed and finds the number of the term in the query. The number of terms to form sequence of terms and passed to the multiple search engines and retrieve the most relevant to the sequence term. The third method is the numbers of terms to form in mingle of terms, search a relevant web page and retrieved from multiple search engines. Then top ‘ $n$ ’ web page result is collected from multiple search engines and filtered using term frequency and inverse document frequency. Using the LBG clustering algorithm to group the similar web pages based on the method of phrase and to optimize the unique web pages using the SVM classifier. Finally, using a ranking algorithm rank value of the unique top link from the final optimized result is assigned. Finally the performance metrics precision, recall,  $f$ -measure, accuracy, speed and error rate are calculated.

This experiment using the evaluation metrics calculated based on the confusion matrix. The focuses on the parameters are precision, recall,  $F$ -measure, speed and error rate. Table 1 represents a confusion matrix for ranking web pages based on the phrases. This confusion matrix counts the relevant web pages and irrelevant web pages from the various search engines. True positives (TP) are relevant web pages in the ranking and false positives (FP) are irrelevant web pages in the ranking. Then, true negatives are irrelevant web pages missing from the ranking and false negatives are relevant web pages missing from the ranking.

Now, Eq. 1 for precision, Eq. 2 for recall and Eq. 3 for  $F$ -measure calculation in terms of confusion matrix are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$F - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

An information retrieval based on the phrase system by its accuracy, which classifies the correct relevant web pages. In terms of the confusion matrix Table 1, the

**Table 1** Confusion matrix for information retrieval

	Relevant	Non-relevant
Retrieved	TP	FP
Non Retrieved	FN	TN

accuracy and error rate calculation can be done by using Eqs. 4 and 5 respectively.

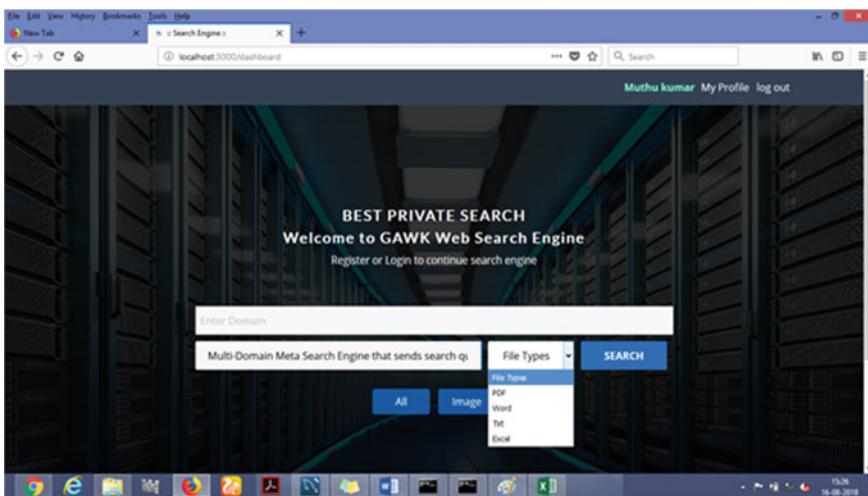
$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (4)$$

$$\text{Error Rate} = \frac{(FP + FN)}{(TP + FP + FN + TN)} \quad (5)$$

Figure 2 represents the users enter the phrase in our proposed search engine. The phrase is classified into three categories as whole, mingle, sequence of terms.

The Phrase Query as “Multi-Domain Meta Search Engine that sends search queries to various search engines and to retrieve results from them” and pass the various search engines. Then, used to overall fusion results using LBG clustering algorithm to group together the similar result links shown in Table 2.

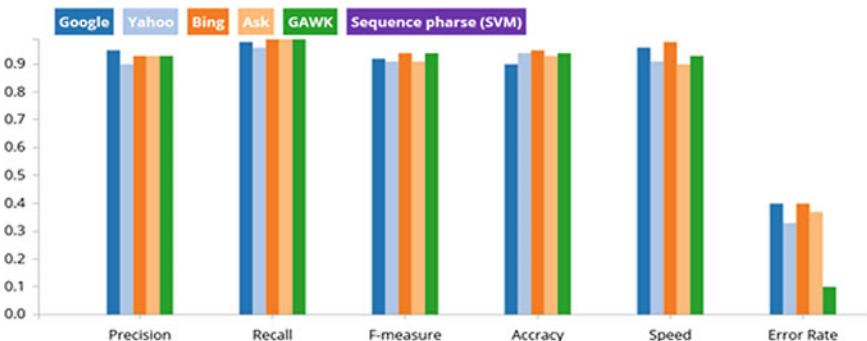
The above LBG clustering algorithm results are transferred to the SVM classifier to identifying the unique link based on the three types of phrase searching categories like whole Phrase, Sequence of Keyword in phrase and Mingle of the keyword in the phrase. Figure 3 shows the performance calculation using in the form of sequence terms using SVM classifier. Figure 4 shows the performance calculation using in the whole phrase using SVM classifier. Figure 5 shows the performance calculations using in the form of mingle terms using SVM classifier. Figure 6 shows the performance calculation are combined the above three types using SVM classifier. Figure 7 shows the top ‘n’ unique link from the various search engine.

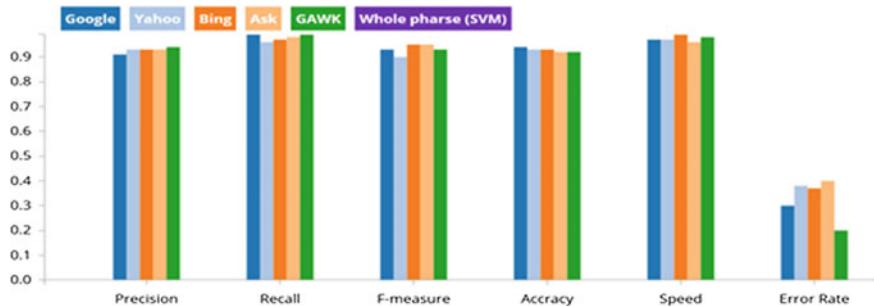


**Fig. 2** Phrase based searching

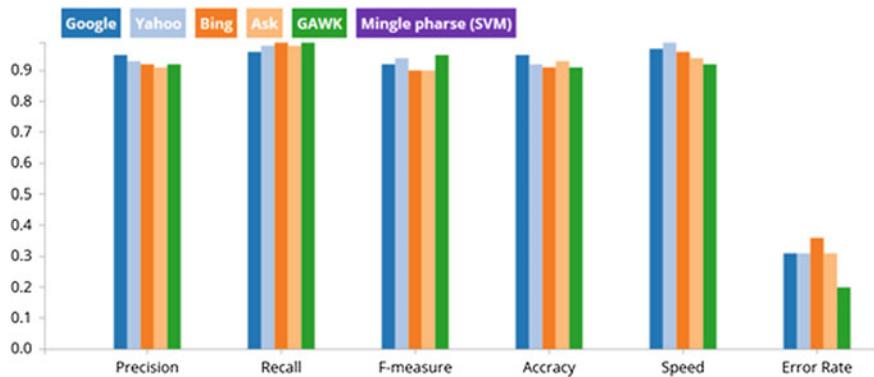
**Table 2** Similar links grouped together using LBG clustering algorithm

Cluster	URL	LBG clustering value
Cluster 0	<a href="http://airccse.org/journal/ijcseit/papers/1011ijcseit04.pdf">http://airccse.org/journal/ijcseit/papers/1011ijcseit04.pdf</a>	0.94
Cluster 1	<a href="https://www.researchgate.net/publication/254052336_EGG_Enhanced_Guided_Google_-_A_meta_search_engine_for_combinatorial_keyword_search">https://www.researchgate.net/publication/254052336_EGG_Enhanced_Guided_Google_-_A_meta_search_engine_for_combinatorial_keyword_search</a>	0.75
Cluster 2	<a href="https://link.springer.com/content/pdf/10.1007%2F978-3-642-24043-0_13.pdf">https://link.springer.com/content/pdf/10.1007%2F978-3-642-24043-0_13.pdf</a>	0.7
Cluster 3	<a href="https://www.google.com/search/howsearchworks/">https://www.google.com/search/howsearchworks/</a>	0.94
Cluster 4	<a href="https://www.scribd.com/doc/72370424/Text-Supplement">https://www.scribd.com/doc/72370424/Text-Supplement</a>	0.94
Cluster 5	<a href="https://www.researchgate.net/publication/221194963_Evaluation_of_Result_Merging_Strategies_for_Metasearch_Engines">https://www.researchgate.net/publication/221194963_Evaluation_of_Result_Merging_Strategies_for_Metasearch_Engines</a>	0.87
Cluster 6	<a href="https://www.researchgate.net/publication/281632206_The_Essence_of_the_Essence_from_the_WebThe_Metasearch_Engine">https://www.researchgate.net/publication/281632206_The_Essence_of_the_Essence_from_the_WebThe_Metasearch_Engine</a>	0.88
Cluster 7	<a href="http://www.allwebdevhelp.com/php/help-tutorials.php?i=31041">http://www.allwebdevhelp.com/php/help-tutorials.php?i=31041</a>	0.93
Cluster 8	<a href="https://www.startpage.com/">https://www.startpage.com/</a>	0.81
Cluster 9	<a href="http://www.dogpile.com/">http://www.dogpile.com/</a>	0.95
Cluster 10	<a href="https://www.researchgate.net/publication/221194963_Evaluation_of_Result_Merging_Strategies_for_Metasearch_Engines">https://www.researchgate.net/publication/221194963_Evaluation_of_Result_Merging_Strategies_for_Metasearch_Engines</a>	0.87

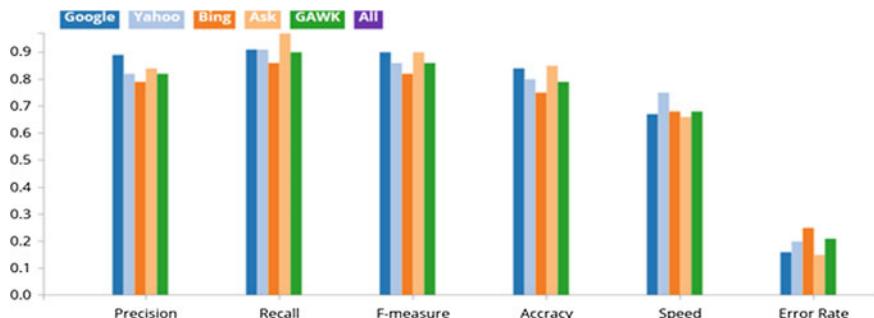
**Fig. 3** Sequence of terms based on evaluation of performance metrics



**Fig. 4** Whole phrase based on evaluation of performance metrics



**Fig. 5** Mingle of terms based on evaluation of performance metrics



**Fig. 6** Overall evaluation of performance metrics

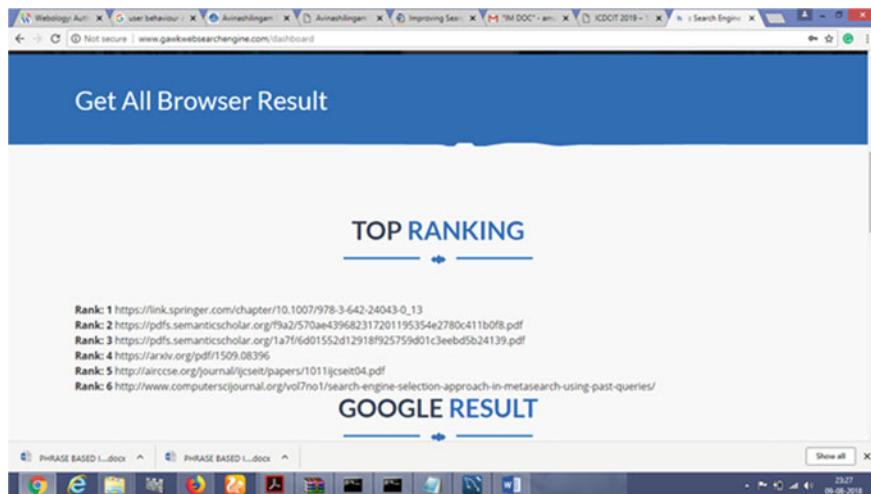


Fig. 7 Top 'n' ranking results from proposed work

## 5 Conclusion

In this paper, we have presented a phrase-based model for information retrieval from the web using machine learning techniques. The phrase model builds and maintains the retrieval dynamically during the query processing the type of file and its categories the phrase type as whole phrase, sequence of keyword, mingle of the keyword are identified and preprocessed. Our proposed work searching phrase with special character and it provides most relevant result compared with the traditional search engines. The proposed work outperforms the existing search engine in terms of precision, Recall, F-measure, accuracy, speed and error rate are calculated. These performance metrics show significant improvement with the accuracy 15% and comparatively lower error rate 10% of the proposed framework.

## References

1. Arampatzis, A. T., Tsoris, T., Koster, C. H. A., & Van Der Weide, Th. P. (1998). Phase-based information retrieval. *Information Processing & Management (Elsevier Science Ltd)*, 34(6), 693–707.
2. Zanaty, E. A. (2012). Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal*, 13, 177–183.
3. Lang, H., Wang, B., Jones, G., Li, J.-T., Ding, F., & Yi-Xuan, L. (2008). Query performance prediction for information retrieval based on covering topic score. *Journal of Computer Science and Technology*, 23(4), 590–601.
4. Masłowska, I. (2003). Phrase-based hierarchical clustering of web search results. In *European Conference on Information Retrieval, ECIR 2003: Advances in Information Retrieval* (pp. 555–562).

5. Remesh Babu, K. R., Samuel, P. (2015). Concept networks for personalized web search using genetic algorithm. In *International Conference on Information and Communication Technologies (ICICT 2014)*, *Procedia Computer Science* 46 (pp. 566–573).
6. Laura, L., & Me, G. (2017). Searching the web for illegal content: The anatomy of a semantic search engine, methodologies and application. *Soft Computing*, 21, 1245–1252.
7. Arias, M., Cantera, J. M., Vegas, J. (2008). Context-based personalization for mobile web search. In *ACM. VLDB'08*. August 24–30, 2008.
8. Adriani, M., van Rijsbergen, C. J. (1999). Term similarity-based query expansion for cross-language information retrieval. In *International Conference on Theory and Practice of Digital Libraries ECDL 1999: Research and Advanced Technology for Digital Libraries* (pp. 311–322).
9. Mangla, N., Jain, V. (2014). Context based indexing in information retrieval system using BST. *International Journal of Scientific and Research Publications*, 4(6), ISSN:2250-3153.
10. Patterson, A. L. (2006). Phrase identification in an information retrieval system. In *Google*.
11. Bhatia, S., Brunk, C., & Mitra, P. (2012). Analysis and automatic classification of web search queries for diversification requirements. In *ASIST 2012*, October 28–31, 2012.
12. Stoyanchev, S., Song, Y. C., Lahti, W. (2008). Exact phrases in information retrieval for question answering. In *Proceedings of the 2nd workshop on Information Retrieval for Question Answering (IR4QA)*, pp. 9–16.
13. Fatmawati, T., Zaman, B., Werdiningsih, I. Implementation of the common phrase index method on the phrase query for information retrieval. In *International Conference on Mathematics: Pure, Applied and Computation, AIP Conf. Proc.* 1867, 020027-1-020027-9.
14. Mala, V., Lobiyal, D. K. (2016). Semantic and keyword based web techniques in information retrieval. In *IEEE 2016 International Conference on Computing, Communication and Automation (ICCCA)*. ISBN:978-1-5090-1666-2.
15. Li, Z., & Wu, X. A phrase-based method for hierarchical clustering of web snippets. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*.
16. Yu, Z., Johnson, T. R., & Kavuluru, R. (2014). Phrase based topic modeling for semantic information processing in biomedicine. In *IEEE Explore, 2013 12th International Conference on Machine Learning and Applications*. ISBN:978-0-7695-5144-9.

# The Politics of Artificial Intelligence Behaviour and Human Rights Violation Issues in the 2016 US Presidential Elections: An Appraisal



**Patrick A. Assibong, Ikedinachi Ayodele Power Wogu,  
Muyiwa Adeniyi Sholarin, Sanjay Misra, Robertast Damasevičius  
and Neha Sharma**

**Abstract** The outcry from organizations and concerned bodies responsible for regulating and protecting the security and privacy of citizens' data have suddenly been on the increase since the unearthing of the recent scandal regarding data gathered from the users of Facebook by Cambridge Analytica (CA) without prior consent. Data belonging to over 87 million Facebook users were admittedly passed on to third-party organizations who used them to manipulate the views and opinions of American citizens, a move largely believed influenced and altered the results and outcome of the 2016 US presidential elections in favour of certain individuals, an action largely believed undermines the Fundamental Human Rights (FHR) of the citizens affected. The Marxian Alienation Theory and Derrida's critical and qualitative analytical method for evaluating arguments and data gathered on the subject matter of the paper were adopted for the study, with the view to highlight the inimical influences of AI politicking on the FHR of Americans and her institution of democracy. The paper observed that the Facebook and CA scandal has given room for questioning the integrity and credibility of future election outcomes in America. There is an alarming dearth of viable regulations and ethical codes governing the

---

P. A. Assibong · I. A. P. Wogu · M. A. Sholarin · S. Misra (✉)  
Covenant University, Ota Nigeria, Ota, Ogun State, Nigeria  
e-mail: [sanjay.misra@covenantuniversity.edu.ng](mailto:sanjay.misra@covenantuniversity.edu.ng)

P. A. Assibong  
e-mail: [patrick.assibong@covenantuniversity.edu.ng](mailto:patrick.assibong@covenantuniversity.edu.ng)

I. A. P. Wogu  
e-mail: [ike.wogu@rhemauniversity.edu.ng](mailto:ike.wogu@rhemauniversity.edu.ng)

M. A. Sholarin  
e-mail: [solarinadeniyi@gmail.com](mailto:solarinadeniyi@gmail.com)

R. Damasevičius  
Kaunas University of Technology, Kaunas, Lithuania  
e-mail: [robertas.damasevicius@ktu.lt](mailto:robertas.damasevicius@ktu.lt)

N. Sharma  
Society for Data Science, Pune, India  
e-mail: [nvsharma1975@gmail.com](mailto:nvsharma1975@gmail.com)

reckless use of AI politicking platforms in the public domain. The paper suggests measures of improving and securing the private rights and data of citizens generally. It also suggests measures for preserving the sanctity and credibility of future election results in America.

**Keywords** Artificial intelligence behaviour · Artificial intelligence politicking · American elections · American politics · Cambridge Analytica · Facebook · Fundamental human rights · Human rights violations · Marxian alienation theory · Presidential elections

## 1 Introduction

In search of appropriate words to capture the very complex nature of man and how he relates with his fellow man and his environment, Aristotle inferred that *man is a political animal* by nature. Hence, he always consciously or unconsciously participates passively or actively in politics. Politics in this wise, is largely referred to as the procedure by which individuals, clusters or group of persons make decisions on matters regarding governance, the rule of law and the wellbeing of people in the state [1, 2]. From Aristotle's point of view, it is an activity that men are naturally drawn to. Hence, they would go to any length to acquire power, sustain power or take over power wherever the need arises.

With the twenty-first century ushering an era of massive adoption and implementation of Artificial Intelligent (AI) innovations into all sectors of human endeavour [3], one of the sectors that commenced the massive deployment of AI innovations and technology for achieving its goals and aspirations is the political sector [4, 5]. In view of this, Lindsie Polhemus among other scholars, observed that humans and their counterparts (AI machines), have since the past decade—as a result of rising participation and involvement of man in political matters—put mechanisms in place to facilitate the accurate predictions of outcomes of elections for a while now [4]. Recent studies reveal that they (Man and Machines) have largely succeeded in achieving this feat [5, 6].

Following the whistle-blowing efforts and testimonies of a former employees of one of the AI servicing firms that perpetrated acts believed to undermine the privacy and fundamental human rights of certain US citizen's data on the internet, an action also believed to have occasioned scholars [7, 8] to questioned the sanctity and credibility of US elections, the world has witnessed series of litigations and public outcry over the callous and reckless violation of individuals' rights and privacy. In the light of the above, scholars [5, 6] and activist now fear for the sanctity of future election results in America. They also fear for the safety and privacy of individual's data on the internet [4–6]. It is on this premise that this paper interrogates the use of AI politicking approaches for political campaigns in America's democracy. It also evaluates the attendant consequences of this mode of politicking on the privacy

and fundamental human rights of US electorates who participated in the 2016 US presidential elections.

**The Problematic:** The under listed are some of the specific issues that prompted the writing of this paper:

- i. Evidence from literature reveals that the reckless adoption and implementation of AI politicking approaches for political campaigning purposes in the 2016 US presidential elections, poses a rising threat to the sanctity and credibility of election results in America.
- ii. The rising adoptions of AI technology for political purposes in America by companies like Cambridge Analytica (CA), largely undermines the privacy and rights of unsuspecting citizens who subscribe to various internet and social media platforms.
- iii. The testimonies of Christopher Wylie about the meddling of the political processes that brought Donald Trump into power, largely questions the justification and legitimacy of his presidency and his position in the White House.

**Objectives:** In the light of the above issues raised, the paper specifically strives to:

- i. highlight and evaluate the threats, which AI politicking approaches for political campaigning in the 2016 US presidential election exerted on American citizens and on its democracy.
- ii. assesses and evaluates the degree of privacy and fundamental human right violation committed on the American citizens by the malicious and unlawful use of private data from unsuspecting internet users by AI firms like CA.
- iii. evaluates the degree of merits in the claim that the 2016 US presidential election results, which brought Donald Trump into the White House and into power was largely flawed and manipulated.

**Theory and Methodology:** The theory adopted for this paper largely relies on the Marxian Alienation Theory [9] because it provides adequate theoretical foundations for interrogating the degree of human right violations perceived to emanate from adopting AI politicking approaches presently at play in America's polity. Creswell's mixed method approach [10] for analyzing and evaluating qualitative data and arguments in the social sciences, was also adopted for the study. The *ex-postfacto* research method [11] was also utilized since the study largely relied on previously analysed data from other studies. Derrida's deconstructive and reconstructive method [12] of analyzing deeper meanings of concepts and arguments was also utilized for this study.

## 2 Advances in Artificial Intelligence in the Twenty-First Century

### 2.1 Current Research in AI

Consequent on these new developments and advances in AI research, scientist, scholars, businessmen and industrialists have sought for ways of exploring and maximizing the new abilities in the intelligent machine in every sector of life's endeavour. The medical profession/sector, for instance, have witnessed a geometric advancement in the ability of Doctors/Surgeons to utilize Artificial Intelligent Machines (AIM) technology for running very accurate and precise medical diagnosis [13] for patients in matter of seconds, thus improving the chances of saving lives and enhancing the medical experience generally. The business sector and the stock markets—through the adoption of High-Level Machine Intelligence (HLMI) software—are now able to make very precise predictions and up-to-date stock market evaluations, which has helped businessmen know when to sell and when to hold back their shares [14]. In fact, the past decade has witnessed tremendous successes as a result of the implementation of HLMI systems in virtually every sector of human endeavour. Rising from the above premise, Drum in [15] made bold to say that 'AIM will in no distant future, take over completely the jobs of man' irrespective of what the job is.

### 2.2 The Behaviour and Capabilities of HLMI for Politics

One very significant sector where HLMI systems are having a tremendous impact on the trend of things is in the political sector [4–6]. In the past, political candidates running for elections had little or no tools to work with regard to knowing the temperament of the electorates. Hence, more often than not, they relied only on instincts and not calculated insights before making a decision concerning their campaign tactics for running for office. But with the advent of innovations in HLMI systems, Machine learning, Deep Learning, Big data analytics and Artificial Intelligent systems [4–6], founded on statistical techniques which automatically helps analysts to identify special patterns in the behaviour of electorates. This feat is made possible by a set of data collected in advance about the preferences of the electorate. These HLMI systems have already been used to predict which United States Congress bills were passed into law by simply analyzing 'the algorithmic assessments of the content of the bill' and other essential variables like: what time of the year the bill is being proposed, what number of sponsors the bill has had in the past and the exact time and season the bill is being passed to members of Congress [6].

In the same vein, recent studies [7, 16] revealed that these advances in HLMI systems are now being carefully deployed for use even in election campaigns where they have been tactically used to manipulate the psyche of the electorates such that votes eventually casted by unsuspecting citizens, are largely influenced by series of adverts

and information provided to targeted members of the electorates in accordance with their individual political, social, psychological or spiritual preferences as the case may be [6]. A process of this nature utilized in a standard democracy like America', raises critical and ethical questions about the sanctity and credibility of the results obtained from such elections. It also raises questions about the moral justifications for manipulating unsuspecting minds and psyches of electorates, whose data were obtained without their consent and used largely to their disadvantage. Some studies [4–6] revealed that where the minds of the electorates had not been subjected to any kind of manipulation by these special ads on HLMI systems, the results of the votes casted during the elections could have turned out differently. Recent evidence from scholars [7, 16] suggests that this same HLMI technologies were deployed by Donald Trump's campaign team, during the 2016 US presidential elections in America. These scholars noted that the adoption of these HLMI systems during the period in question, helped to sway the votes of the electorates in a manner that eventually favoured the Trump's Republican Party. Others experts like [6] believed that the deployment of the HLMI tools by Trump's campaign team was decisive in the outcome of the results that were eventually released to the general public, after the 2016 presidential elections.

### ***2.3 Cambridge Analytica and HLMI Technologies***

The recent scandal about the gross violations of the privacy and fundamental human rights of over 87 Million US citizens during the 2016 US presidential elections, had in its centre a data science firm simply known as Cambridge Analytica (CA). The company, on its home page, on the internet, describes herself thus: 'We are the global leader in data-driven campaigning with over 25 years of experience, supporting more than 100 campaigns across five continents'. They claimed to have also played pivotal roles in winning congressional, state and presidential elections [17].

Largely an advertising agency, the company's political arm, claimed to have redefined the relationship that initially existed between data and campaigns. The company believed that a precise knowledge of the electorate gives them profound knowledge and influence over the electorate, thus bringing desired goals and objectives to a reality for any individual or group hiring and utilizing their services. In the light of the garrulous statements made by CA, the recent whistle-blowing acts of a former employee of the company: Christopher Wylie—who exposed in some details, the nefarious acts of gross human right violations perpetrated by the company—have landed the company in a critical situation with rising legal suites from human right organizations and law makers. They have also had summons from the Senate and Congressional Houses in the Capitol, forcing them to short down officially, all it operations and offices in the US, UK and Canada.

Before the scandal began, Cambridge Analytica had in 2016, largely rolled out a widespread advertising campaign designed to target fluid electorates based on their psychological dispositions [6, 7]. Advertisements from CA were designed in such a

way that different individuals received different messages according to their various dispositions and susceptibility to different arguments proposed by each political party and their candidates. Those considered to be paranoid with the basic ideology of certain political parties were given advertisements that were associated with their fears [16]. Those discovered to have strong conservative predispositions were made to receive advertisements which emphasize on themes relating to community and tradition. All these were made possible by the availability of real-time data gathered on the electorates' behaviours and preferences from social media platforms. The internet footprints arising from the analysis and evaluations of data generated from these electorates were then used to build psychological and behavioural profiles that were unique to each individual [6]. This platform created by CA is what Christopher Wylie in [8] referred to as a 'political ad targeting technology' or 'an arsenal of weapons for fighting a culture war on the electorates'.

### 3 Politicking in Twenty-First Century American Politics

#### 3.1 *A Review of AI Technology's Involvement in 2016 US Presidential Elections*

The scenario describes in the previous section on the nature of politicking going on in twenty-first century American politics perhaps, explains why there have been serious endeavours by political parties and organizations concerned with US politics, to naturally display the behaviour for which Aristotle described man as 'a political animal'. An evaluation of the presence of the political animalistic behaviour among American politicians, revealed that the 2016 US presidential elections witnessed in full scale, man's reckless and animalistic tendencies deployed in the bid to acquire political power, that is, in the case of those who were not yet in positions of power. All these behaviours were vividly displayed in the recently concluded 2016 US presidential election [8].

The testimony provided by Christopher Wylie before the Senate Judiciary Committee on 16 May 2018, revealed that certain members of key political parties that took part in the 2016 US presidential elections, held nothing back at taking steps to ensure that the outcome of the results of the said elections, turned out in their favour. One of the testimony provided by the 28-year-old revealed how Steve Bannon, a former White House senior strategist, connived with billionaire Robert Mercer to seek and obtain the assistance for, what the whistle blower described as: 'Cambridge Analytica's political ad targeting technology'. A technology latter discovered to have the capacity to unleash an 'arsenal of weapons to fight a culture war' [8], since according to Wylie, Steve Bannon believed that politics is downstream from culture. Hence, they sought companies that would build weapons targeted at fighting the culture war that would break or manipulate unsuspecting electorates' psychological dispositions

and their abilities to make a clear subjective political decision via individual ads, which the company made available to the electorates.

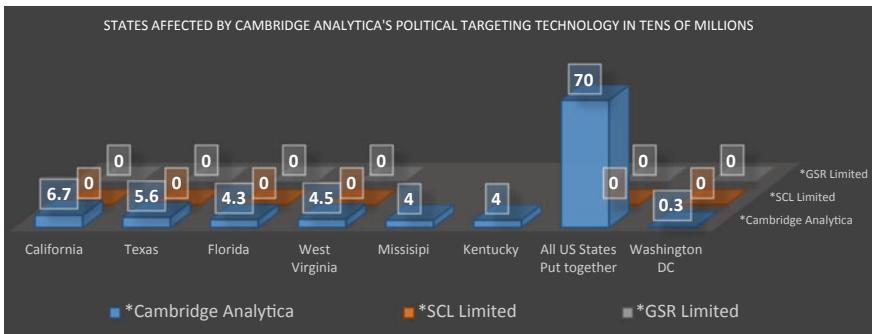
The scenario described in the above passage—the scenario where specially designed algometric tools by AI companies were used to confuse, misinform, manipulate and malign electorates into doing otherwise—records indicate [6–8, 13, 16], had become the order of the day, some years and months leading into the 2016 US presidential elections. Ted Cruz's presidential campaign, for instance, was in 2015, observed to have used the services of Cambridge Analytica who provided for his campaign team, the psychological profiles of tens of millions of unsuspecting citizens generated from Facebook to enhance his presidential campaign plans and strategies [7]. Other big names in this category include Ben Carson, John Bolton and Tom Hills. In the same vein, Donald Trump's campaign executives without knowing it, revealed to undercover reporters that they too subscribed to Cambridge Analytica. They affirmed that the company's expertise was pivotal and instrumental to their victory in the 2016 presidential elections [6, 7]. The instances discussed above is typical of the nature of the kind of politicking which took place during the 2016 US presidential elections. This paper notes that all the actors in this era, played their politics without recourse to violating the rights and privacy of citizens whose data were stolen and used to manipulate the electorates. In Wylie's own words: 'We exploited Facebook to harvest millions of people's profiles. And built models to exploit what we knew about them to target their inner demons. That was the basis on which the entire company (Cambridge Analytica) was built on' [7].

### ***3.2 Matters Arising from the 2016 US Presidential Elections***

One established fact of the 2016 US presidential election is that AI was used to manipulate the electorates who participated in the series of elections that were conducted that year. All these were made possible by the availability of real-time data on voters' behaviour and preferences on the social media. These data were largely used to build unique psychographic and behavioural profiles on the electorates.

Candidates like Donald Trump who is so flexible with his campaign promises is an example of the kind of person who could benefit from these ads tactics. The campaign ads sent different tailored messages that focused on a different perspective of the candidate. In other words, each voter got a different version of Donald Trump's advertisement sent to them. The key to this approach is to simply find the right soft spot in each individual capable of spurring them into acting otherwise [5, 6, 13].

The whistle-blowing efforts of Christopher Wylie revealed that firms like Cambridge Analytica had masterminded the designing of robots and systems which were responsible for aggressively spreading, as it were, a one-sided political information which invariable manufactured illusory public support that never really existed. These tactics was known to have been increasingly prevalent during the 2016 US elections. The false propaganda arising from these ads were often used to highlight false social media messages about opponents to a targeted population or group of persons, such



**Fig. 1** The breakdown of Facebook users affected by Cambridge Analytica. *Source* Created by the authors from available data on Brexit Referendum (2016)

that they electorates were discouraged from going out on election day to vote for their candidate [6]. Vyacheslav Polonski in a report offered a good instance of the scenario where pro-Trump bots in 2016, infiltrated Twitter hashtags and Facebook pages used by Hillary Clinton supporters to spread automated contents' [6].

Regarding the degree of impact which CA had on most states in the US, the data in Fig. 1 indicates that California topped the chart with 6.7 million users who were affected. This is followed by Texas 5.6 million and Florida 4.3 million. Where the comparison is done on the basis of state percentage, Washington DC topped the list with over 50% of its population affected by the breach of data privacy. This record is followed by West Virginia, Mississippi and Kentucky, all states that largely voted for Trump [18].

## 4 New Perspectives on Human Right Violations in America

From the very first week after Donald Trump was sworn in into office as the 45th president of the United States of America, all through his first year in office (December 2016–December 2017), the country's strong civil society and democratic institutions were tested across a wide range of issues by Human Rights Watch organizations. Scholars who assessed the country's human right initiatives believed the country scored so poorly on the local and on the international scale [19]. Some of the areas that scored so poorly on the human right scale include areas of Immigration, Rights of non-citizen, Rights to affordable health care, Rights of privacy and Human Right violations generally, to mention but a few.

#### **4.1 *Human Rights Violations in Donald Trump's Administration***

Scholars argued that the high degree of nonchalance and lackadaisical attitude of the Trump's administration towards certain policies were largely responsible for the poor civil society's reports recorded and the very flaccid and porous sense of human right laws in the United States of America [20, 21]. Natasha Lomas referred to it as 'the precariously placed EU-US Privacy Shield'. Thus, most civil liberty and civil rights movements like the Libe Committee, continually argued that the mechanisms guarding the privacy of individuals' data online and human rights generally, in its current state, does not have the capacity to provide the much needed protection for American and EU citizens alike, especially in the light of the recent scandal which has rocked Facebook and Cambridge Analytica to its very foundations [20, 21]. The existence of these porous and fluid human right laws and provisos, largely explains why companies like Cambridge Analytica could boldly advertise its treacherous and callous business on the internet for all to see without the fear of being reprimanded or having any form of litigation levelled against her by concerned authorities.

The secret video documentary released by Channel 4 [22, 23], made on some of the executives of CA who boasted about the abilities and resources their company had at its disposal for manipulations and for winning elections, are clear indications of instances where CA and their associate companies jointly violated America's electoral laws and regulations. The FHR of the electorates whose minds and psyche was manipulated to get targeted election outcomes, was also violated. Yet, Mr. Alexander Nix and the executives of CA on various advertisement platforms on their website, claimed to 'hold the electorates in very high esteem' [22]. 'This kind of deceit won't be necessary if the nation had a central location where data is stored, as it is done in China', argues Maya Wang, a researcher on China's mass surveillance systems [20]. Some scholars contend that were this scenarios becomes the order of the day, it has the capacity to distort the democratic processes of countries [24].

#### **4.2 *The Genesis of Facebook and Cambridge Analytica Scandal***

The whole ploy to develop and foster excellent political propaganda campaigns for the Republican Party began with a request from GSR to collect data from Facebook for purely academic research purposes. This was, however, not to be so since GSR passed the same data to CA. The company obtained the information with the aid of a digital application designed by a Psychologist and a university professor named Aleksandr Kogan. He created a personality quiz which required subscribers on the Facebook platform to login in with their credentials before they could partake in the quiz. A total number of 270,000 individuals on Facebook participated in the quiz, but that was after they had installed the apps on their accounts. By so doing, all 270,000

persons had given their data to the University professor, who forwarded the same to CA. The application was design in such a way that it—without any prior consent—allowed the University professor and CA an unhindered access to the personal files and details of over 87 million Facebook fans in the UK and in America who were in one way or the other, friends linked or connected to the 270,000 persons who initially participated in the quiz arranged by Aleksandr Kogan from Cambridge University [25].

To affirm that there was a strong sense of corruption and human rights violation involved in the whole ploy, Aleksandr Kogan, the professor from Cambridge who designed the initial quiz, was implicated in the lawsuit that followed the scandal, since he was found to be one of the founding directors of GSR Limited [25]. Steve Bannon, a former White House Adviser to Donald Trump, was discovered to have chaired the Board of CA in 2014 when the initial data was obtained [23, 25]. Another Rebekah Mercer, who was a prominent donor to the Trump's campaign group [23, 25, 26], was also discovered to have been a core member of CA at the time when the initial 270,000 data was gathered.

Going by this revelation, the Republican Party, CA and all the names mentioned above had already started planning this scheme some 2 years before it was set into motion. This is what the number one Democratic nations of the world had degenerated to, all because of the desire to satisfy that animalistic tendency which Aristotle made reference to when he declared that 'man is a political animal'. Most scholars on this grounds believe that the individuals and the groups mentioned above are guilty of grossly violating the FHRs of all 87 million persons and more, who were affected by this scandal [24, 26, 27]. There are scholars who also believed that Facebook should also take the fall for the grievous part it played in allowing such a magnitude of scandal to take place, especially after getting wind of the conspiracy, in 2014 [23, 27].

## 5 Further Discussion and Summary of Findings

### 5.1 *The Fate of Elections in America in the Wake of Rising AI Technology*

Consequent on the new trend (AI politicking) today, the game of political campaigning have been redefined across the globe by AI ad firms who have perfected the act of mining data profusely, what Cheng in [22] describes as 'the mining of a profusion of fine-grained data about voters and their habits'. While the political consultant of the 1960s and 1970s sought to look for ways of projecting the images of their candidates on television in their private homes, today's campaign masters utilize big data to manipulate the deepest hopes and fears of the electorates via cutting edge 'psychographic profiling systems' believed to have the capability of assessing the personalities of voters even better than their friends could attest about them. The end

of which is to sway the votes of the electorates to the advantage of the campaign professionals.

While this paper notes that the actual process of conducting research on the opposition with the view to providing messages that would be appealing to the electorate, it is not an illegal or necessarily wrong thing to do. This is because marketing and advertising companies do it all the time when they desire to sway the consumer from buying one product over another. There are however standards and regulations which guards against subliminal modes of advertising and campaigns targeted at consciously manipulating consumers, in this case, the American electorates into altering the choice of votes cast at the end of the day. Further reflections on the documentary aired by Channel 4 [23], clearly indicates the extent to which these companies and individuals that subscribe to them for political assistance, are willing to go to, in other to achieving whatever set goals their political animalistic minds and ambitions set out to achieve.

A polity so adversely taken over by the animalistic drive to acquire and retain political power, by all means, necessary (via human and non-human means), irrespective of the damaging effects these political actions may have on the electorates' privacy and their FHR, this paper believes, is tantamount to having a negative impact on the sanctity and the credibility of the results obtained from elections conducted in such states. Such polity would have problems with establishing and strengthening democratic institutions in the country. In the light of massive adoption of AI platforms like Facebook and CA for political campaigns purposes and the scandal arising from the 2016 US presidential elections, the future of America's politics and democracy, is bleak.

## ***5.2 Summary of Findings***

In the light of the three objectives set aside for the study, the authors made the following pertinent findings:

- On the first objective, the authors of this paper identified that a lacking of the sense of ethical and moral values in America's political arena and among the leadership and ruling elite class are largely aiding the emergence and use of AI ad firms like CA whose sole aim, among other things, is to utilize whatever means at its disposal to provide ad campaigns which eventually alters the outcome of election results in favour of those who subscribe to their firms.
- On the second objective, the authors identified the existence of none viable legislations and regulations governing the conduct of Internet-based firms who largely process peoples data online. Consequent on the weak or non-implementation of these regulations, many organizations continue to get away with fraudulent practices, some of which have been known to have inimical and colossal damaging effect on the citizen's affected.

- Regarding the third objective, the authors found overwhelming evidence from literature, etc., which points to the fact that the AI ad firms threw caution to the wind and used every means at their disposal to achieve victory for their client, Donald Trump. Thus, the legitimacy of Donald Trump's Presidency in the White House and the sanctity of future US elections—in the light of AI politicking campaign approaches rigorously adopted during the 2016 US presidential elections—gave reasons for many to question the legitimacy and the sanctity of the result obtained.

### ***5.3 Recommendations***

In the light of the evidence and data gathered in view of the objectives outlined for this study, the authors of this make the following recommendations:

- While striving to enforce existing laws, the government needs to urgently implement the UN guiding principles on businesses and human rights since it offers better logical platforms for discussing the responsibilities of companies as regarding human right and privacy issues.
- Companies like Facebook should assure its users that those who she decides to give her data to would not misuse it and that the rights of the users are protected from abuse. However, where such rights are breached, they must be remedied.
- AI can better be used to deploy micro-targeting campaigns techniques which will help inform the electorates about different political issues at stake in a polity and the various views shared by each political organization. Such information will help the electorate make informed decision about which party to affiliate themselves with.
- From the way it stands, the use of AI technology in politics has come to stay, considering the valuable role it now plays in political campaigns. Politicians and all concerned are however advised to judiciously and ethically commit to using AI technology in ways that will not undermine the rights of its citizens and that of democracy at large.

### ***5.4 Contribution to Knowledge***

From the foregoing studies and discussions carried out on the three specific research problems identified for this study, the authors of this paper identified the following as the main contribution this paper would make to the body of knowledge generally.

- The study for the first time, affirms Aristotle's animalistic political nature of man as one of the overarching factors influencing the reckless adoption of AI politicking platforms for politics in America, and more recently, in all her political campaign processes.

- The colossal damage done to the privacy of individual's data on online and the violation of their FHR when they failed to get the consent of the users of these platforms, would have all been averted if necessary attention were given to strengthening legislations and enforcing compliance and stiff penalty to defaulters.
- While Facebook and Cambridge Analytica did not directly vote Donald Trump into the White House, their online platform and targeted messaging structure provided the avenue through which the manipulation which eventually swayed the votes of the electorates to vote the way they did, to the advantage of those who hired the services of the CA. The electorates, therefore, are largely responsible for who presently occupies the White House. They may have to live with it.

## 6 Conclusion

The massive adoption of AI politicking in America's polity and more recently, in the political campaign processes which took place in 2016 US presidential elections, was found to be inimical to the sanctity and integrity of elections conducted in America. The overarching factor propelling the reckless implementation of AI politicking approaches in the US is rooted in the Aristotelian animalistic political theory and nature of man, which was found to be a feature prevalent amongst the political elites and leadership class of Americans. This political nature, in turn, makes it easy for all concerned parties to easily disregard existing regulations guarding individuals' and group's data online, thereby violating the privacy and FHR of all those who entrusted the care of their private data to these online platforms for safe keeping. Consequently, the authors of this paper recommend a discontinuing of the present AI politicking approach now prevalent in the US polity. Doing so will prevent further colossal damage to America's electoral processes and its democracy. In its place, a morally oriented version of AI politicking approach is recommended.

## References

1. Wogu, I. A. (2010). *A preface to logic, philosophy and human existence* (p. 634). Lagos, Nigeria: Pumack Educational Publishers. ISBN:978-978-50060-0-1.
2. Wogu, I. A. P. (2018). *Plato's republic and the crisis of leadership in the 21st century: Implications for political development in Nigeria*. Unpublished Ph.D. thesis from Covenant University, Ota Ogun State Nigeria.
3. Wogu, I. A. P., Olu-Owolabi, F. E., Assibong, P. A., Apeh, H. A., Agoha, B. C., Sholarin, M, A., Elegbeleye, A., & Igbokwe, D. (2017). Artificial intelligence, alienation and ontological problems of other minds: A critical investigation into the future of man and machines. In *Proceedings of the IEEE International Conference on Computing, Networking and Informatics (ICCDI 2017)*, October 29–31, 2017. Covenant University, Ota. <https://doi.org/10.1109/iccdi.2017.8123792>, <http://ieeexplore.ieee.org/document/8123792/?part=1>, <https://www.scopus.com/record/display.uri?eid=2-s2.0-85047079881&origin=resultslist&sort=plf-f&src=s&st1=>

- ICCN1+&nlo=&nlr=&nls=&sid=02fc90c5e2fe9f02e52574629f67b0d&sot=b&sdt=b&sl=12&s=CONF%28ICCN1+%29&relpos=51&citeCnt=0&searchTerm=.
- 4. Polhemus, L. (2016). Artificial intelligence and politics. In *An Online Publication of Futurism*. Retrieved from <https://futurism.media/artificial-intelligence-and-politics>.
  - 5. Patterson, D. (2018). *Campaigns are catching up to the consumer: How AI is shaping the world of politics*.
  - 6. Polonski, V. (2018). How artificial intelligence conquered democracy: Who is really running elections this days. In *The Conversation. An Online publication*. Retrieved from <https://theconversation.com/how-artificial-intelligence-conquered-democracy-77675>.
  - 7. Solon, O., & Graham-Harrison, E. (2018). The six weeks that brought Cambridge Analytica down: How revelations by the observer and others brought an unknown firm into the light and to its demise, amid international fury. In *The Guardian Online*. Retrieved on the June 13 from <https://www.theguardian.com/uk-news/2018/may/03/cambridge-analytica-closing-what-happened-trump-brexit>.
  - 8. Solon, O. (2018). Cambridge Analytica whistleblower says Bannon wanted to suppress voters. In *The Guardian Online*. Retrieved from June 13, 2018. <https://www.theguardian.com/uk-news/2018/may/16/steve-bannon-cambridge-analytica-whistleblower-suppress-voters-testimony>.
  - 9. Cox, J. (1998). An introduction to Marx's theory of alienation. *International Socialism: Quarterly Journal of the Socialist Workers Party (Britain)*, 79(5). Published July 1998 Copyright © International Socialism.
  - 10. Creswell, J. W. (2003). *Research design: Qualitative and quantitative approaches*. Thousand Oaks, CA: SAGE.
  - 11. Marilyn, K. (2013). *Ex-post facto research: Dissertation and scholarly research, Recipes for success*. Seattle, WA: Dissertation Success LLC. <http://www.dissertationrecipes.com/wp-content/uploads/2011/04/Ex-Post-Facto-research.pdf>.
  - 12. Balkin, J. M. (1987). Deconstructive practice and legal theory. *Yale LJ* 96(15).
  - 13. Radowitz, J. (2017). Intelligent machines will replace teachers within 10 years, leading public school head teacher predicts. In *Independent News Online*. Retrieved on February 16, 2018 from <http://www.independent.co.uk/news/education/education-news/intelligent>.
  - 14. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When will AI exceed human performance? In *Evidence From AI Experts*. arXiv:1705.08807v2 [cs.AI] 30 May 2017. arXiv: 1705.08807v2[cs.AI].
  - 15. Drum, K. (2018). You will lose your job to a robot—Sooner than you think! In *Mother Jones. Online Blogger*. Retrieved on February 26, 2018, <https://www.motherjones.com/politics/2017/10/you-will-lose-your-job-to-a-robot-and-sooner-than-you-think/>.
  - 16. Greenfield, P. (2018). The Cambridge Analytica files: The story so far. In *An Online Publication of the Guardian.com*. Retrieved June 13, 2018. <https://www.theguardian.com/news/2018/mar/26/the-cambridge-analytica-files-the-story-so-far>.
  - 17. Cambridge Analytica. (2017). CA Political: Data driven campaigns. In *An online Publications of Cambridge Analytica*. Retrieved on June 22, 2018 from <https://ca-political.com/>.
  - 18. Bhardwaj, P., & Lee, S. (2018). Here's a state-by-state breakdown of Facebook users impacted by the Cambridge Analytica scandal. In *An Online Publication of business insider.com*. Retrieved on June 13 2018 from <http://www.businessinsider.com/facebook-cambridge-analytica-affected-us-states-graphic-2018-6?IR=T>.
  - 19. Human Rights Watch. (2018). Human rights issues and events of 2017. In *A Publications of Human Rights Watch Online*. Retrieved on June 19, 2018 from <https://www.hrw.org/world-report/2018/country-chapters/united-states>.
  - 20. Wang, M. (2018). Cambridge Analytica, big data and China. In *A Publications of Human Rights Watch Online*. Retrieved on June 19, 2018 from <https://www.hrw.org/news/2018/04/18/cambridge-analytica-big-data-and-china>.
  - 21. Lomas, N. (2018). Pressure mounts on EU-US privacy shield after Facebook-Cambridge Analytica data scandal. In *Online Publication of techcrunch.com*. <https://techcrunch.com/2018/06/12/pressure-mounts-on-eu-us-privacy-shield-after-facebook-cambridge-analytica-data-scandal/>.

22. Chen, A. (2018). Cambridge Analytica and our lives inside the surveillance machine. In *Online Publications of The New Yorker*. Retrieved on June 19 2018 from <https://www.newyorker.com/tech/elements/cambridge-analytica-and-our-lives-inside-the-surveillance-machine>.
23. Tripathi, T. (2018). Facebook and Cambridge Analytica—Where lies privacy? In *Online Publications*. Retrieved on June 19, 2018 from <https://www.ihrb.org/focus-areas/information-communication-technology/commentary-facebook-cambridge-analytica>.
24. Collins, J. (2018). Where does the Cambridge Analytica scandal leave our privacy and democracy? In *Online Publication of Human Right News, Views & Info*. Retrieved on June 19, 2018 from <https://rightsinfo.org/cambridge-analytica-scandal-leave-privacy-democracy/>.
25. CISION Newswire. (2018). Class action lawsuit filed against Facebook and Cambridge Analytica for stealing and improperly using more than 71 Million users' data. In *Online Publication of CISION PR Newswire*. Retrieved on June 19, 2018 from <https://www.prnewswire.com/news-releases/class-action-lawsuit-filed-against-facebook-and-cambridge-analytica-for-stealing-and-improperly-using-more-than-71-million-users-data-300627281.html>.
26. Remnick, D. (2018). Cambridge Analytica and a moral reckoning in silicon valley. In *An Online Publications of the New Yorker*. Retrieved on June 19, (2018) from <https://www.newyorker.com/magazine/2018/04/02/cambridge-analytica-and-a-moral-reckoning-in-silicon-valley>.
27. White, J. B. (2018). Shareholder accuses Facebook of human rights violation at tense meeting. In *An Online Publication of Independent News Online*. Retrieved on June 19, 2018 from <https://www.independent.co.uk/news/business/news/facebook-privacy-human-rights-shareholders-cambridge-analytica-a8378296.html>.

# Crop Prediction Using Artificial Neural Network and Support Vector Machine



Tanuja K. Fegade and B. V. Pawar

**Abstract** In India, the general problem faced by farmers is the selection of proper crop for farming. There are many factors which influence the yield of crop like rainfall, temperature humidity, soil, etc. Crop prediction helps farmers in selecting proper crop for plantation to maximize their earning. Prediction of crops can be accurately done with the help of machine learning techniques and considering the environmental parameters. In this work, the classifiers used are support vector machine and artificial neural networks. Prediction of crop is done by considering parameters like amount of rainfall, minimum and maximum temperature, soil type, humidity, and soil pH value. The data is collected from the agricultural website of Maharashtra. The data is divided into nine agricultural zones. An interface is been designed through which farmers can enter the required information to predict the crop. Neural network gives 86.80% of prediction accuracy.

**Keywords** Artificial neural networks · Support vector machine · Crop prediction

## 1 Introduction

Agricultural sector is an integral part of Indian economic system. Farmers find it difficult to cultivate the total crop yield because of uncertain climate and the changes occurred in the climatic process. In recent time, Modern Artificial Intelligence is emerging as a dominating technology in worldwide development. Machine learning techniques can improve the productivity in the field of agriculture; thus, most of the researchers are using this technique to provide positive results. The agricultural sector is noticed as the quick adopter of artificial intelligence and machine learning both in terms of agriculture products and in field farming techniques.

---

T. K. Fegade (✉)

KCES's Institute of Management & Research, Jalgaon, Maharashtra, India

e-mail: [tanujamahajan18@gmail.com](mailto:tanujamahajan18@gmail.com)

B. V. Pawar

School of Computer Sciences, KBC NMU, Jalgaon, Maharashtra, India

e-mail: [bvpawar@hotmail.com](mailto:bvpawar@hotmail.com)

In this research work, we are using two machine techniques: ANN and SVM, to develop our system model. This system can be used to improve the productivity of farmers in the overall agriculture field through exact prediction of crop with the help of environmental and regional parameters as a crucial input. Through this system, farmers can also get to know the seasonal crops they can take with maximum yield and profit.

Through this innovative system, we can see overall progress in the lives of humans with the use of modern-day technology. Indian farmers are way behind from the actual use and advantages of these technologies, due to their incapable financial position and lack of knowledge and education for using these technologies.

Sellam et al. [1] described in detail the different environmental factors which are cultivation area, total rainfall (annual), and price of food. Mark up exaggerating the crop yield using regression analysis. They conclude that  $R^2$  value states that TAR, AUC, and FPM have an average 70% impact in the crop yield.

Hemageetha [2] formulated the survey of analysis parameters of soil like Nitrogen, moisture, and pH for the prediction of crop yield and assumptions theory using numerous data mining techniques.

Sujatha et al. [3] explained the motives of different techniques of classification like random forests, Naïve Bayes, support vector machine, J-48, and Artificial Neural Networks (ANN) that can be applied for the yield of crop prediction.

Ankalaki et al. [4] explain the comparison of DBSCAN and AGNES algorithms. MLR (Multiple Linear Regression) was used for forecasting the yield of crop.

In Gayatri et al. [5] in order to control a massive amount of data, IOT and web services were used. The data had been gathered with the help of sensors. GPS was used to capture images of agriculture field and stored the data along with their mapping position into repositories.

Kushwaha et al. [6] presented a whole new algorithm which is used to anticipate the crop suitability according to the specific type of soil and enriching the overall quality of agriculture.

Bendre et al. [7] gathered the data from various resources like GPS, GIS, VRT systems, etc., to predict the database of weather. For this, they studied Map Reduce and linear regression algorithm.

Fathima et al. [8] recited various data mining techniques for knowledge uncovering in the agriculture sector.

Kaur et al. [9] studied the various techniques of data mining in order to reach accuracy for speculation of price.

Veenadhari et al. [10] discussed the use of techniques of data mining in crop yield prediction using climatic parameters and achieved 75% accuracy for forecasting of crops.

Raorane et al. [11] discussed a few machine learning techniques, which are decision tree, ANN, Bayesian network, SVM, and K-means, for the betterment of crop cultivation in agriculture.

Rub et al. [12] studied K-means and SVM techniques to be used in the fields of agriculture.

We have studied and compared some data to get the detail information about agriculture based study type along with techniques analyzed with the dataset parameters and results through statistical analysis. Following is Table 1 representing our comparative study over the crop yield presented and studied by various authors and researchers.

Our study shows that most of the researchers have worked on the use of artificial intelligence technologies for the upliftment of the farmer's life. In this study, we have used two classifiers, namely, artificial neural network and support vector machine for the prediction of crop. For training and testing, this model large dataset is collected preprocessed and normalized.

## 2 Proposed Architecture

Figure 1 explains the architecture of speculated system, which is formed into two major phases such as training phase and testing phase.

## 3 Methodology

An interface is designed which enables to access the necessary information for selecting the proper crop. This proposed system looks like a recommender system and assists the farmer to choose suitable crop within the constraints of their farming parameters (Fig. 2).

The phases that are used in this work are discussed in the following text.

### 3.1 Data Collection/Preparation

The required attributed data for crop prediction is collected from the agricultural website of Maharashtra state. The crop knowledge base consists of parameters such as crop varieties, soil varieties, soil pH values, and seasonal structured such as kharif, rabi, and summer crops. The dataset also includes zonal segmentations as well as district database, environmental parameter such as high- and low-temperature counts and average rainfall in the monsoon. The datasets of ten parameters are merged using the Microsoft Excel application as shown in Fig. 3.

The raw data collected from public resources requires extensive preprocessing for handling missing values and other data inconsistencies. The data needs to be normalized for classification.

To prepare preprocessed dataset, we used the following scale:

**Table 1** Comparative study of crop yield prediction

Year	Agriculture type	Techniques used	Dataset parameters	Statistical analysis
2014 [10]	Forecasting crop yield	C4.5, Decision tree	Temperature (min and max), yield, evapotranspiration, cloud cover, frequency of wet day	Crops Soybean Paddy Maize Wheat
2015 [13]	Crop yield prediction for wheat crop	Regression (multiple linear) fuzzy logic, neural fuzzy inference system	Extractable soil water, biomass, rain, and radiation	RMSE values MLR      FL 9.252    6.4251

(continued)

**Table 1** (continued)

Year	Agriculture type	Techniques used	Dataset parameters	Statistical analysis												
2015 [14]	Crop selection	Crop selection method	Weather, soil type, crop type water density, sowing time, plantation days	Accuracy and performance depends on predicted value of affecting parameters												
2015 [15]	Analysis of soil behavior and prediction of crop	Naïve Bayes, K-nearest neighbor	Soil testing parameters like nutrients and micronutrients	<table border="1"> <thead> <tr> <th>Classifier</th> <th>Poor yielding</th> <th>Good yielding</th> <th>Moderate yielding</th> </tr> </thead> <tbody> <tr> <td>K-NN</td> <td>30 lands</td> <td>45 lands</td> <td>25 lands</td> </tr> <tr> <td>Naïve Bayes</td> <td>15 lands</td> <td>40 lands</td> <td>45 lands</td> </tr> </tbody> </table>	Classifier	Poor yielding	Good yielding	Moderate yielding	K-NN	30 lands	45 lands	25 lands	Naïve Bayes	15 lands	40 lands	45 lands
Classifier	Poor yielding	Good yielding	Moderate yielding													
K-NN	30 lands	45 lands	25 lands													
Naïve Bayes	15 lands	40 lands	45 lands													
2016 [1]	Crop yield prediction	Regression analysis	Cultivation area, rainfall, food price index	70% influence in crop yield												

(continued)

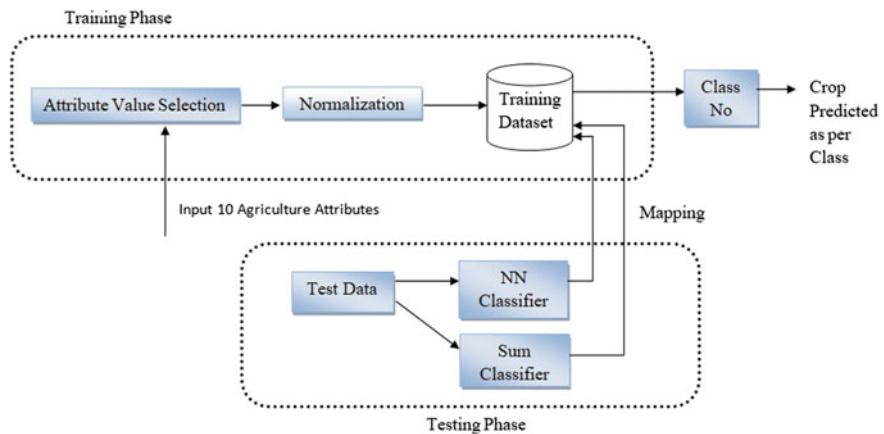
**Table 1** (continued)

Year	Agriculture type	Techniques used	Dataset parameters	Statistical analysis
2016 [16]	Forecast annual yield	DBSCAN, AGNES, regression analysis	Area, Production, Rainfall, Temperature, PH soil minerals	Regression analysis gives highly dependency on dataset
2016 [17]	Selection of maximum yield crop	Fuzzy logic	Historical soil parameters, whether parameters, and cost	Achieved <i>F</i> -score 54% with precision 77%
2016 [18]	Indian rice crop prediction	Decision support system	Min, max, avg. temperature, area, production, yield, precipitation, crop evapotranspiration	High variation is observed in the production of rice with DSS tools

(continued)

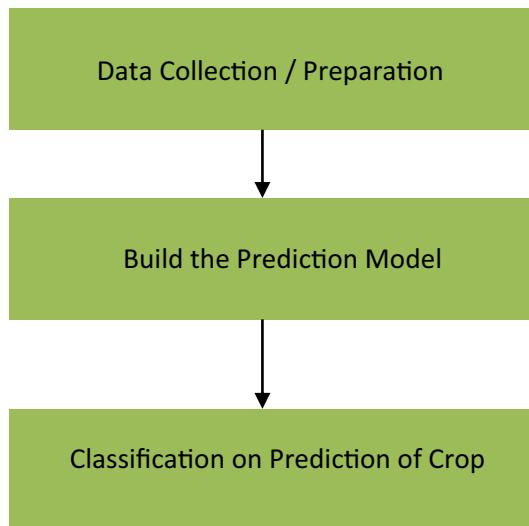
**Table 1** (continued)

Year	Agriculture type	Techniques used	Dataset parameters	Statistical analysis															
2016 [19]	Rice crop prediction	Artificial neural network	Min, max, avg. temperature, area, production, yield for kharif season precipitation, crop evapotranspiration	Accuracy of 97.5% and sensitivity of 96.3%															
2016 [20]	Prediction of tea yield	Multiple linear regression	Temperature, humidity, rainfall sunshine, evaporation	<table border="1"> <tr> <td>Type</td> <td>South Bank</td> <td>Upper Assam</td> <td>North Bank</td> <td>Cacher</td> </tr> <tr> <td>Coefficient of correlation</td> <td>0.82</td> <td>0.8000</td> <td>0.7739</td> <td>0.6555</td> </tr> <tr> <td>Determination</td> <td>0.68</td> <td>0.64</td> <td>0.59</td> <td>0.42</td> </tr> </table>	Type	South Bank	Upper Assam	North Bank	Cacher	Coefficient of correlation	0.82	0.8000	0.7739	0.6555	Determination	0.68	0.64	0.59	0.42
Type	South Bank	Upper Assam	North Bank	Cacher															
Coefficient of correlation	0.82	0.8000	0.7739	0.6555															
Determination	0.68	0.64	0.59	0.42															



**Fig. 1** Proposed architecture

**Fig. 2** Design flowchart



- (1) **Zone:** There are nine zones identified with number from 0 to 8 in Maharashtra state.
- (2) **District:** 29 Districts are considered under Maharashtra state which is represented by values 0.1 to 0.29.
- (3) **Kharif Season, Rabi Season, and Summer Season:** If the value is true, then it is represented by 1; otherwise, it is represented by 0.
- (4) **Soil Type:** There are eight soil types which are represented by values 0.1 to 0.8.
- (5) **Max Temperature and Minimum Temperature:** The temperature values vary district-wise in different ranges. For example, for Bhuldhana district, it is 7–7.78,

Agriculture Crop Dataset													
Sr. No	Zone	District	Crop Type	rainif	Seas	stab	Seas	Summer Sea	Soil Type	Soil Ph	Temperature	temperature	Rainfall[Avg]
1	Central Maharashtra Pl	Jalgaon	Rice	Yes	No	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
2	Central Maharashtra Pl	Jowar	Jowar	Yes	Yes	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
3	Central Maharashtra Pl	Jalgaon	Bajra	Yes	No	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
4	Central Maharashtra Pl	Jalgaon	Tur	Yes	No	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
5	Central Maharashtra Pl	Jalgaon	Mung	Yes	No	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
6	Central Maharashtra Pl	Jalgaon	Udil	Yes	No	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
7	Central Maharashtra Pl	Jalgaon	Maize	Yes	Yes	Yes	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
8	Central Maharashtra Pl	Jalgaon	Gram	No	Yes	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
9	Central Maharashtra Pl	Jalgaon	Groundnut	Yes	Yes	Yes	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
10	Central Maharashtra Pl	Jalgaon	Sesamum	Yes	No	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
11	Central Maharashtra Pl	Jalgaon	Soyabean	Yes	No	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
12	Central Maharashtra Pl	Jalgaon	Sunflower	Yes	Yes	Yes	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
13	Central Maharashtra Pl	Jalgaon	Cotton	Yes	Yes	No	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
14	Central Maharashtra Pl	Jalgaon	Banana	Yes	Yes	Yes	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
15	Central Maharashtra Pl	Jalgaon	Sugarcane	Yes	Yes	Yes	Colour changes frt	7-7.5	34.6C	19.2C	750 mm		
16	Central Maharashtra Pl	Dhule	Rice	Yes	No	No	Colour changes frt	7-7.5	41C	21C	700 to 900mm		
17	Central Maharashtra Pl	Dhule	Jowar	Yes	Yes	No	Colour changes frt	7-7.5	41C	21C	700 to 900mm		
18	Central Maharashtra Pl	Dhule	Bajra	Yes	No	No	Colour changes frt	7-7.5	41C	21C	700 to 900mm		
19	Central Maharashtra Pl	Dhule	Maize	Yes	Yes	Yes	Colour changes frt	7-7.5	41C	21C	700 to 900mm		
20	Central Maharashtra Pl	Dhule	Tur	Yes	No	No	Colour changes frt	7-7.5	41C	21C	700 to 900mm		

**Fig. 3** Sample of agriculture crop dataset

and for Bhandara, it is from 32 to 37. For normalization, these values are multiplied by 0.01. So that it can be converted between 0 and 1 values.

- (6) **Rainfall:** Rainfall values are multiplied by 0.0001.

The above scale of crop dataset parameters are normalized in the range of 0–1 for building the prediction model, which is explained in the upcoming segment.

### 3.2 Building the Prediction System

In order to build the model to predict the crop, we have used two classifiers—neural network and support vector machine, which are described as follows:

#### (A) Neural Network for Crop Prediction

Pattern recognition networks are feedforward networks that can be trained to categorize values in accordance with target classes. We have used an existing algorithm **NN\_Train** using neural network to predict the crop.

---

**Algorithm: NN\_Train (input\_param, Classes)**

---

Input\_param: Preprocessed Normalized dataset with 10 Parameters  
 Classes: 44 crop types

**Step 1: [Select a crop dataset.]**  
 $[x,t] = \text{crop\_dataset};$

**Step 2: [Design of training model.]**  
 $\text{net} = \text{patternnet}(10);$

**Step 3: [Train the neural network and visualized it.]**  
 $\text{net} = \text{train}(\text{net}, x, t);$   
 $\text{view}(\text{net})$

**Step 4: [Test the network.]**  
 $y = \text{net}(x);$

**Step 5: [Check the performance of the system.]**  
 $\text{perf} = \text{perform}(\text{net}, t, y);$

**Step 6: [Display the network result in the form of crop type.]**  
 $\text{classes} = \text{vec2ind}(y);$

End Algorithm.

**(B) Support Vector Machine for Crop Prediction**

Support Vector Machine (SVM) is a supervised machine learning technique, which could be used for both classification and regression. The algorithm SVM\_Train is used to predict the crop.

---

**Algorithm: SVM\_Train (input\_param, Classes)**

---

Input\_param: Preprocessed Normalized dataset with 10 Parameters  
 Classes: 44 crop types

**Step 1: [Select unique crop types.]**  
 $u = \text{unique}(\text{out}) \text{ where}$   
 $\text{out} = \text{crop type}$

**Step 2: [find number of classes for crop type.]**  
 $\text{Num\_class} = \text{length}(u);$

**Step 3: [Prepared Support Vector Machine training model]**

```

For i= 1 to Num_class
    GlvAll=find(strcmp(out,u(k));
// used for vectorized the class
    nnn=zeros(size(out));
    nnn(GlvAll)=1;
    Models(i)=svmtrain(Input_param, nnn);
End for
```

**Step 4: [Test the model.]**

```

For j=1 to size (testdata, 1)
    For k=1 to Num_class
        If (svmclassify (Models (k), testdata (j))
            Break;
        End if
    End for
    Result (i)=k;
```

End for

**Step 5: [predict the crop using SVM.]**

```
Crop_predicted_svm (i)=u (result (i));
```

**End**

## 4 Applications

If the farmers get the information in advance, they can easily predict which crop is beneficial for them. This predicted crop maximizes the crop yield, and hence, it is nearly combined to the overall Agricultural Development and well-being of the rural masses. Prompt information transfer between the researchers and the farmers has specific advantages. Henceforth this research study is concentrated on the issues of developing an agricultural framework for supporting farmers in plectrum of the crop based on versatile parameters. With the help of current model, the farmers can access interactive, interlinked, and convenient information for selecting the crop according to the farm's climatic conditions.

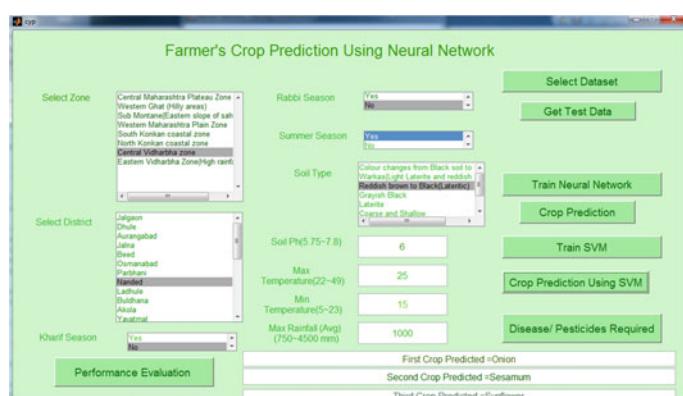
The model will help the farmers in the expansion of their productivity by choosing the appropriate crop.

## 5 Results and Discussions

In this work, MATLAB is used to build the prediction model. The results are divided into two parts using SVM and ANN:

### A. Crop Prediction Using SVM

Crop is predicted using SVM by giving the various input parameters. The initially trained database can be selected by using button “select database” and “train neural network”. Then test data value is provided using “get test data” button. Farmers can provide input parameters by selecting zone, district, season type, and soil type. The proposed model predicted three crops as per their ranking. The first crop predicted by SVM is onion, which is shown in Fig. 4.



**Fig. 4** Crop prediction using SVM

## B. Crop Prediction Using Artificial Neural Network

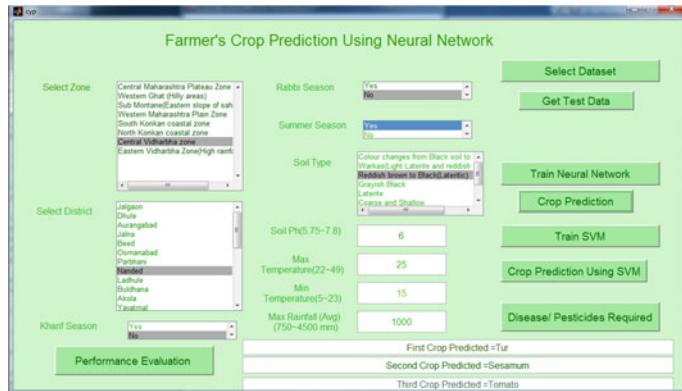
Crop is predicted by ANN by providing various input parameters. Here the first predicted crop by ANN is Tur, which is shown in Fig. 5.

### C. Accuracy

For evaluating the proposed model, accuracy is measured using Eq. 1.

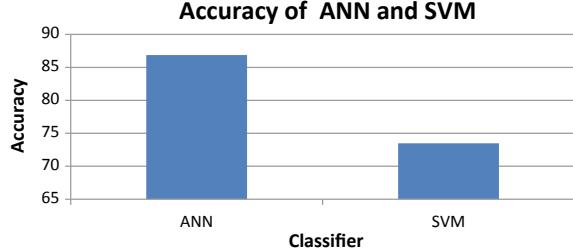
$$\text{Accuracy} = \frac{\text{Number of crop predicted}}{\text{Total number of crop}} \times 100 \quad (1)$$

By using SVM classifier, we have received accuracy of 73.48% which is further improved with the help of artificial neural network which gives the accuracy of 86.80%. The accuracy in graphical form is shown in Fig. 6.



**Fig. 5** Crop prediction using ANN

**Fig. 6** Comparison of ANN and SVM



## 6 Conclusion

Crop prediction is important in the agriculture community. In this work, crop prediction is done by considering various parameters like rainfall, soil type, soil pH, temperature, etc., using ANN and SVM. A comparative study of the results obtained from SVM and NN is also performed. The accuracy of ANN models on the test set was found to be 86.80%. The results indicate that the ANN model had better accuracy and better prediction rate as compared to SVM. This reveals that the ANN techniques could speculate the crop type better than SVM for the given dataset. The outcome of this work assists farmers for proper selection of the crop. In the future, a generalized prediction model for various crops along with crop yield and profit analysis can be developed.

## References

1. Sellam, V., & Poovammal, E. (2016). Prediction of crop yield using regression analysis. *Indian Journal of Science and Technology*, 9(38), 1–5.
2. Hemaneeetha, N. (2016). A survey on application of data mining techniques to analyze the soil for agricultural purpose. In *3rd International Conference on Computing for Sustainable Global Development (INDIA-Com)* (pp. 3112–3117).
3. Sujatha, R., & Isakki, P. (2016). A study on crop yield forecasting using classification techniques. In *International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE)* (pp. 1–4).
4. Ankalaki, S., Chandra, N., & Majumdar, J. (2016). Applying data mining approach and regression model to forecast annual yield of major crops in different district of Karnataka. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(2), 25–29.
5. Gayatri, M. K., Jayasakthi, J., & Anandha Mala, G. S. (2015). Providing smart agricultural solutions to farmers for better yielding using IOT. In *IEEE Technological Innovation in ICT for Agriculture and Rural Development (TIAR)* (pp. 40–43).
6. Kushwaha, A. K., & Bhattacharya, S. (2015). Crop yield prediction using Agro Algorithm in Hadoop". *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 5(2), 271–274.
7. Bendre, M. R., Thool, R. C., & Thool, V. R. (2015). Big data in precision agriculture: Weather forecasting for future farming. In *1st International Conference on Next Generation Computing Technologies* (pp. 744–750).
8. Fathima, G. N., & Geetha, R. (2014). Agriculture crop pattern using data mining techniques. *International Journal of Advanced Research in Computer Science and Engineering*, 4(5), 781–786.
9. Kaur, M., Gulati, H., & Kundra, H. (2014). Data mining in agriculture on crop price prediction: Techniques and applications. *International Journal of Computer Applications*, 99(12), 1–3.
10. Veenadhari, S., Misra, B., & Singh, C. D. (2014). Machine learning approach for forecasting crop yield based on climatic parameters. In *International Conference on Computer Communication and Informatics (ICCCI-2014)*, Coimbatore, India (pp. 1–5).
11. Raorane, A. A., & Kulkarni, R. V. (2012). Data mining: An effective tool for yield estimation in the agricultural sector. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 1(2), 75–79.
12. Rub, G. (2009). Data mining of agricultural yield data: A comparison of regression models. In *9th Industrial Conference* (Vol. 5633, pp. 24–37).

13. Shastry, A., Sanjay, H. A., & Hegde, M. (2015). A parameter based ANFIS model for crop yield prediction. In *IEEE International Advance Computing Conference (IACC)*.
14. Kumar, R., Singh, M. P., Kumar, P., & Singh, J. P. (2015, May). Crop selection method to maximize crop yield rate using machine learning techniques. In *International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)* (pp. 138–145).
15. Paul, M., Vishwakarma, S. K., & Verma, A., (2015). Analysis of soil behaviour and prediction of crop yield using data mining approach. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*.
16. Chandra, S., Chandra, N., & Majumdar, J. (2016, October). Applying data mining approach and regression model to forecast annual yield of major crops in different district of Karnataka. *International Journals of Advanced Research in Computer and Communication Engineering*, 5.
17. Aadithya, U., Anushya, S., Bala Lakshmi, N., & Sridhar, R. (2016). Fuzzy logic based hybrid recommender of maximum yield crop using soil, whether and cost. *ICTACT Journal on Soft Computing*, 6(4).
18. Gandhi, N., Armstrong, L. J., & Petkar, O. (2016). Proposed decision support system (DSS) for Indian rice crop yield prediction. In *IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*.
19. Gandhi, N., Petkar, O., & Armstrong, L. J. (2016). Rice crop yield prediction using artificial neural networks. In *IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*.
20. Rupanjali, D., Baruah, S. R., Bhagat, R. M., & Sethi, L. N. (2016). Use of data mining technique for prediction of tea yield in the face of climate change of Assam, India. In *International Conference on Information Technology (ICIT)*.

# A Hybrid and Adaptive Approach for Classification of Indian Stock Market-Related Tweets



Sourav Malakar, Saptarsi Goswami, Amlan Chakrabarti and Basabi Chakraborty

**Abstract** Twitter generates an enormous amount of data daily. Various studies over the years have concluded that tweets have a significant impact in predicting and understanding the stock price movement. Designing a system to store relevant tweets and extracting information for specific stocks and industry is a relevant and unattempted problem for Indian stock market, which is the eighth largest in terms of market capitalization. As people with diverse backgrounds are tweeting about many topics simultaneously, it is nontrivial to identify tweets which are relevant for the stock market. Therefore, a critical component of the aforesaid system should contain one module for the extraction and storage of the tweets and another module for text classification. In the current study, we have proposed a hybrid approach for text classification which combines lexicon-based and machine learning-based techniques. The proposed scheme handles class imbalance problems effectively and has an adaptive characteristic, where it automatically grows the lexicon both through WordNet and by using a machine learning techniques. This system achieves F1-score over 98% of the relevant class, as compared to 60% achieved using the baseline method over a corpus of 10,000 tweets. The coverage of tweets by lexicons also improves by 8%.

**Keywords** Cross-validation · Stock market · Twitter · Text classification

---

S. Malakar (✉) · S. Goswami · A. Chakrabarti  
A.K. Choudhury School of Information Technology, University of Calcutta,  
Kolkata, India  
e-mail: [sourav.xaviers@gmail.com](mailto:sourav.xaviers@gmail.com)

S. Goswami  
e-mail: [saptarsi007@gmail.com](mailto:saptarsi007@gmail.com)

A. Chakrabarti  
e-mail: [achakra12@yahoo.com](mailto:achakra12@yahoo.com)

B. Chakraborty  
Faculty of Software and Information Science, Iwate Prefectural University,  
Takizawa, Japan  
e-mail: [basabi@iwate-pu.ac.jp](mailto:basabi@iwate-pu.ac.jp)

## 1 Introduction

Twitter is one of the most popular social networking sites, where each day almost 500 million tweets are generated on diverse topics. Tweets represent information as well as opinion about various topics. Though the tweets lack structures, if properly messaged, they can generate meaningful insights into diverse communities. Liu et al. [1] have clearly analyzed good, bad, and the ugliness of social media data. They have also discussed new research opportunities to design novel algorithms and tools for data mining. In some recent works, Edger et al. [2] used tweets to create psychological landscapes. Ashktorab et al. [3] developed a system called Tweedr to extract actionable information for disaster relief workers for better informed relief operation. Abboute et al. [4] have developed an approach to identify people with suicidal tendency based on their tweets. There are many more papers which study the effect of twitters after a disaster [5] or the effects of a drug. In a recent paper, Jain et al. [6] have proposed a novel intelligent surveillance process based on Twitter and RSS Feeds to give better insights into government agencies for better management of healthcare emergencies.

While the abovementioned papers are related to social goods, companies are also using Twitter to improve their earnings, by analyzing public perception about their products and brands. Ghiassi et al. [7] take a particularly notable effort in this direction. Bollen et al. [8] must be credited for their first significant work to establish the relationship between Twitter mood and the stocks. They have extracted moods of the tweets and established the correlation of the moods with the movement of Dow Jones Industrial Average for United States of America (USA). There is another interesting study where the effect of announcement of CEO Succession on the stock market and its effect on the return of the stock have been studied for the USA and the United Kingdom (UK). Rao et al. [9] also studied the correlation between Twitter sentiment and 13 high cap stocks in the USA. Study of tweets in developing counties has been relatively limited [10]. Zhang et al. [10] have studied the relationship between sentiment of retail investors in China, using a special social network named Xueqiu. Some other studies, like [11, 12] also upheld the view of positive correlation and causation between the sentiment expressed by tweets and stock market movement.

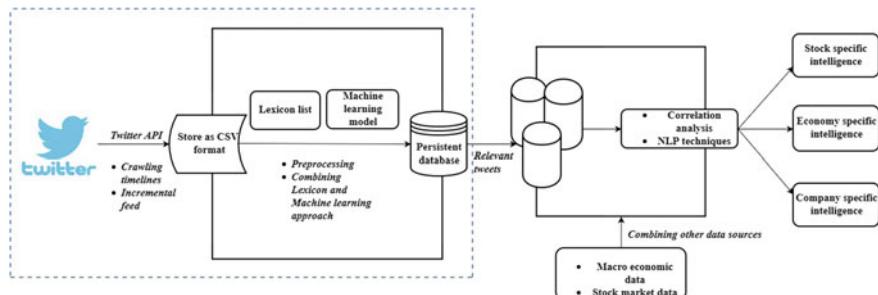
However, studies are limited for India, and to effectively use Twitter-based sentiment for the prediction of the stock market, probably it needs a far more rigorous and systematic approach. The usage pattern of Twitter in India is very different [13]. There is no trivial way to filter tweets which are related to Indian stock market, as, first, not all tweets have geocodes, and second, with the diversity of Indian stock market, there is no established set of hashtags which can be used to extract tweets in India. In this context, an approach to crawl timelines of economic and business news agencies and brokerage houses was adopted. These need to be stored in a database for systematic analysis. Even then, we observed that only 20% of the tweets are relevant to the stock market. If the rest 80% of the tweets are fed into the system and analyzed, then the insights will not be correct. This motivates us to build a strategy to

filter/classify relevant tweets for Indian stock market as an important part of building a Twitter-based intelligence system for the same. A schematic representation of such a system is provided in Fig. 1, which illustrates different layers of the system. The first layer concentrates on acquiring the data, cleaning and filtering of the irrelevant tweets. The second layer focuses on finding features using Natural Language Processing (NLP) techniques and integration of other sources of data, for example, stock market data. In the final layer, economy specific, industry-specific or stock-specific intelligence is extracted as applicable. The scope of the current paper is the extraction and classification of tweets and then necessary preprocessing for storage, which is highlighted in Fig. 1.

The filtering can either be done using a lexicon-based approach or a machine learning-based approach. A lexicon-based approach will have a dictionary of words like “bullish”, “dividend”, “BSE”, etc., and eventually, any tweet having such words will be classified as relevant. Lexical approaches work very fast, but it is very difficult to build an all-comprehensive lexicon. As a result, a lexicon-based approach will tend to have a low recall [14]. On the other hand, a machine learning-based approach can be used, by labeling the tweets as “relevant” and “nonrelevant” by a domain expert and then by training a suitable model. Machine learning methods are more adaptive; however, getting the tweets trained by a domain expert is expensive. There are some efforts, which have combined lexical-based approaches for text classification. Many of them have been applied for sentiment analysis, where lexicons have helped to extract the aspects and then sentiment analysis has been applied [14, 15]. However, in contrast to the existing approaches, we aim to build a combined system, where the output of the machine learning model can be exploited to grow the lexicon. WordNet [16] has been used to expand the lexicon. Our proposed system also successfully handles the class imbalance problem.

The main contribution of our paper is summarized below:

1. The corpus has been generated using a Twitter “list” and has been labeled by domain expert.
2. A lexicon has been prepared with the help of a domain expert and expanded using WordNet.



**Fig. 1** Architecture diagram for Twitter-based intelligence system for the stock market

3. The lexicon has also been updated using a feedback path from tweets classified as relevant with high confidence.
4. Class imbalance problem has been suitably addressed.

The rest of the paper is organized as follows. In Sect. 2, some related works in this area are briefly discussed. In Sect. 3, the proposed system and the necessary algorithm have been discussed. In Sect. 4, the experimental setup has been detailed. In Sect. 5, the results and analysis of the experiment are discussed in detail. Section 6 sums up the conclusion of the paper.

## 2 Related Works

Text classification involves classifying documents, tweets, emails or even URLs in some predefined categories. Some of the popular tasks in this domain involve sentiment analysis [14, 15], spam detection [17], authorship attribution [18], and topic classification [19]. Textual data are usually preprocessed by removing punctuation, stop words and then by applying stemming or lemmatization. Subsequently, it can be represented using the bag of words model (can vary from uni-gram to N-gram) in the traditional manner. For assigning weights to the terms, different techniques like tf, tf-idf, or BM25 can be used. Veningston et al. [20] proposed two approaches named Term Rank-based Approach (TRA) and Path Traversal-based Approaches (PTA1, PTA2, and PTA3) to improve the document re-ranking task based on a Term Association Graph Model. Part-of-Speech (POS) tagging [21] is one of the popular tasks in NLP, used to classify a word in a corpus with respect to different POS tags based on the context of the corpus. Sirsat et al. [22] reviewed several methods to differentiate the applicability of Information Extraction (IE) either for key-phrase extraction, building concept dictionaries for annotating a corpus directly or can be used to extract structured data from unstructured text. They have also proposed that most of the IE systems are based on supervised approach. Topics modeling-based techniques assume that a document is a mixture of few topics, takes the bag of word matrix as an input and produces lower rank approximations using techniques like LSI (Latent Semantic Indexing) or LDA (Latent Dirichlet Allocation). There are newer neural network-based techniques like word2vec and paragraph vector, which have been successfully used for text representation.

In this paper, a combined approach using both lexical and machine learning-based methods have been proposed. There have been some works where a combined approach has been employed. Melville et al. [23] have successfully used background lexical information along with the learning paradigm for analyzing the sentiment of the blog. Yenala et al. [24] have used a deep learning architecture named Convolutional Bi-Directional LSTM (C-BiLSTM) to identify inappropriate query suggestions of search engines, where C-BiLSTM outperforms traditional pattern and keyword-based filtering as well as SVM and gradient boosted decision trees-based classifiers. They have also proposed LSTM and BiLSTM sequential modules for

filtering out inappropriate conversations in messengers. Lu et al. [25] have used a large lexicon with machine learning techniques and reported superior performance for subjectivity analysis. Zahoor et al. [26] have proposed a lexical-based approach where some specific feeling related words are used to extract tweets for sentiment analysis. Pagolu et al. [27] have been proposed a Word2vec-based feature extraction model for performing sentiment analysis of a large dataset using random forest. Oliveira et al. [28] used of several well-known lexicon resources have been used to classify tweets in two or three sentiments by two approaches: the first one is daily positive and negative words count based on different lexicons. Another one is by selecting each tweet to see the proportion of negative and positive words. However, most of these methods were specially for sentiment analysis, which has quite a few established lexicons and a balanced class distribution. In this paper, we have built the lexicon with the help of domain experts. Khan et al. [14] also reported improvement when using a combined approach for sentiment analysis. We have applied WordNet to expand the same, handled the imbalanced class problem, and subsequently used a feedback path to grow the lexicons.

As mentioned earlier, some of the researches which studied the relationship of stock market with tweets are as follows:

- Bollen et al. [8] must be credited for their first significant work to establish the relationship between Twitter mood and the stocks in the year 2011. They have extracted moods of the tweets and illustrated strong correlation of the moods with the movement of Dow Jones Industrial Average for US for individual stocks.
- There is another interesting study where the effect of announcement of CEO Succession on the stock market and its effect on the return of the stocks have been analyzed for USA and UK by Rao et al. [29], where again the authors found a significant correlation between them.

Yet in another study, the correlation between Twitter sentiment and 13 high cap stocks in the USA was investigated by Zhang et al. [10]. Chang et al. [30] have used a sentiment list provided by Hu and Lu to map tweets to classify into two sentiment classes and then have employed correlation-based analysis to associate a sudden increase or decrease of stock prices with sentiment scores over time. Nisar et al. [12] studied the effect of politics-related sentiment and FTSE 100, during election time and found a positive correlation between the tweets and the stock market movement. Yuexin et al. [31] performed a correlation analysis to investigate the relationship between Twitter volume spike with multiple factors and used a Bayesian classifier to analyze the importance of stock. Simsek et al. [32] used, some common sentiment terms of Turkish language to decompose the dataset into two sentiment classes and correlate the average sentiment score of tweets with stock price movement over the times.

- Ruan et al. [11] in a very recent work proposed a user-to-user trust-based network, based on the reputation of the users, and how the reputation can be used to justify abnormal returns of some stocks.
- Smailović et al. [33] have proposed a predictive sentiment classifier which annotates tweets by using positive and negative words provided by Stanford University and

dataset belongs to two sentiment classes. Finally, Support Vector Machine (SVM) with linear kernel is used to classify tweets into different sentiments. Subsequently, they have also performed casualty analysis to correlate Twitter sentiment with stock price movement. Khan et al. [14] also reported improvement when using a combined approach for sentiment analysis.

It can be found that, though most of the studies confirm a positive correlation between the stock market movement and sentiment expressed from tweets, there has been a limited number of studies for Indian stock market which is one of the top 10 stock markets as far as market capitalization is concerned.

### 3 Proposed Methodology

In this section, first, the proposed architecture of the system is discussed, followed by the algorithm.

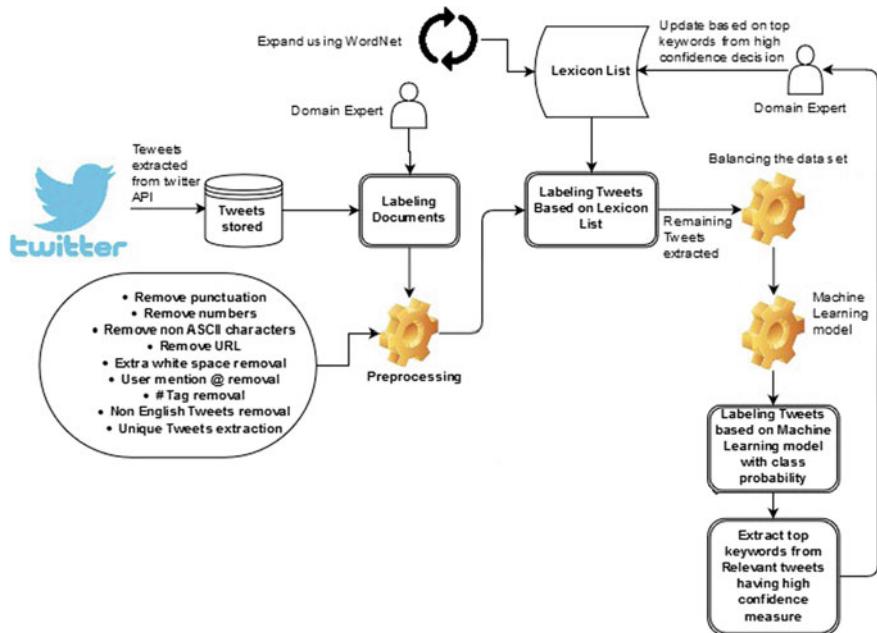
#### 3.1 *Proposed Architecture*

The architecture is explained using Fig. 2.

In our paper, we have proposed a hybrid approach for text classification that will improve the existing machine learning approach. We have combined both lexicon-based and machine learning approach to design the classifier. By applying different preprocessing tasks on our dataset, we clean our text. Then, based on the lexicon list, initially, we select a subset of tweets as relevant. Our motive is to extract maximum possible tweets depending on the lexicon list because it will optimize the processing time. Next, the dataset is evaluated for class imbalance, if it is imbalanced then standard techniques have been applied to rebalance them. Now, for the remaining tweets, we have applied a machine learning algorithm to classify them into two classes relevant and nonrelevant. Our lexicon list is fully dynamic in nature; in real time, it can grow using a well-known database called WordNet. Another strategy to enrich the lexicon list is to extract top keywords from relevant tweets with the high confidence score, which are not existent in the current lexicon. Any new word, thus added is always expanded using WordNet.

#### 3.2 *Algorithm: Adaptive Text Classifier (ATC)*

The proposed algorithm can be designated as Adaptive Text Classifier or ATC 1, which would scrutinize all aspects of current scenario at the time of execution for identifying the next machine learning model.



**Fig. 2** A detail architecture of adaptive text miner

The execution flow of the ATC algorithm is divided into three stages:

- Lexicon-based tweets extraction.
- Classify remaining tweets using a machine learning algorithm.
- Extraction of stock-specific keywords.

The detailed approach is discussed below:

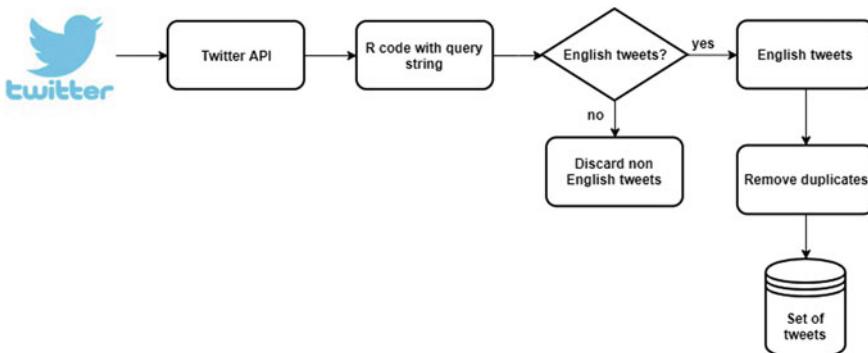
1. **Lexicon-based tweets extraction:** At this stage, we assume a lexicon list named  $L$  and a set of tweets named  $T_{\text{total}}$ . Based on how many numbers of lexicons are appearing in a tweet, compared to total words, we have assigned a score to each tweet in the corpus. Then, all tweets having a score greater than or equal to a dynamic threshold “ $f$ ” are selected as relevant and stored in a set called  $R_{\text{lexicon}}$ ; otherwise, we store them in  $T_{\text{rm}}$ .
2. **Classification with a machine learning algorithm:** Before applying any traditional machine learning algorithm, we first check whether the remaining dataset is imbalanced or not. For this, we have calculated the percentage of relevant tweets with respect to the whole dataset and if the percentage is lower than or equal to a threshold  $\lambda$ , we have balanced the dataset using a rebalancing technique. This explains the adaptive nature of our algorithm. Next, we have classified the dataset  $T_{\text{rm}}$  into two classes by keeping both responses and class probabilities for each prediction.

3. **Top keyword extraction:** Here, we have selected only those tweets classified as relevant with a high confidence (greater than equal to  $\gamma$ ) and store them as  $R_{\text{conf}}$ . Now, all tweets belonging to the set  $R_{\text{conf}}$ , we extracted top “ $n$ ” keywords from them. A domain expert helped us to identify only those terms relevant to stock. Finally, we have added the newly generated keywords to our existing lexicon list named  $L$ .

## 4 Materials and Methods

### 4.1 Data Collection and Domain Keyword List Generation

**Tweet Collection:** A total of 10,009 tweets have been extracted from Twitter using Twitter API between the periods of May 2017 to June 2017 on a daily basis. We have crawled different popular Indian business news channels like ET NOW, Financial Xpress, NDTV Profit, Times of India Business, etc. As well as we are monitoring some very popular stock market brokers timeline on daily basis from Twitter, and we store all tweets in a database. After removing redundant tweets, they are labeled by the domain of experts into two categories relevant and nonrelevant. After removal of redundant tweets and multilingual tweets, we have 8250 tweets out of 10,009. We further notice that out of 8520 unique tweets 1873 tweets are relevant and 6647 tweets are nonrelevant. In Fig. 3, we have drawn the whole pipeline of tweet collection.



**Fig. 3** A pipeline of tweets collection

**Algorithm 1** Adaptive Text Classifier(ATC)

---

```

1: Input: Set of tweets( $T_{total}$ ) ,Lexicon list(L),f # is minimum cutoff score for
   lexicon based tweet selection ,n # is number of top keywords, $\gamma$  # is the
   confidence cutoff score used to extract tweets with high confidence,  $\lambda$  # is the
   cutoff percentage of relevant tweets over total tweets
2: Output: Class labels of each tweet(c), New lexicons( $L_w$ )
3: procedure ATC( $T_{total}$ , L,f,n, $\gamma$ , $\lambda$ )
4:    $R_{lexicon} \leftarrow \{\}$  # store tweets as relevant outputted by Lexicon approach
5:    $T_{rm} \leftarrow \{\}$ # store remaining tweets to be classified by machine learning
   algorithm
6:    $R_{conf} \leftarrow \{\}$ # store relevant tweets having high confidence measure
7:    $L_{old} \leftarrow L$ 
8:   #assign a score to each tweet based on lexicon list
9:   for Lexicon words $\in L$  do
10:     $score(x) \leftarrow score_{lexicon}$  where  $x \in T_{total}$ 
11:    Merge(x, $score(x)$ )
12:    # select tweets as relevant which have lexicon based score greater than a
   real time threshold k
13:    for Each  $x \in T_{total}$  do
14:      if  $score(x) \geq f$  then
15:         $R_{lexicon} \leftarrow R_{lexicon} \cup x$ 
16:      else
17:         $T_{rm} \leftarrow T_{rm} \cup x$ 
18:      # examine class imbalance problem or not
19:      if  $\frac{Totalobservation(T_{rm})}{Totalobservation(T_{total})} * 100 \leq \lambda$  then
20:         $T_{rm} \leftarrow Balance\ data\ set(T_{rm})$ 
21:      # classified by machine learning algorithm with both probability distribu-
   tion and response
22:       $Prediction_{response,probability\ distribution} \leftarrow M_L(T_{rm})$ 
23:      # confidence greater than a threshold  $\gamma$  considered as tweet classified rel-
   evant with high confidence
24:      for Each  $x \in T_{rm}$  do
25:        if predicted response(x) = "relevant" then
26:          if ( $probability(x_{relevant}) - probability(x_{non-relevant}) \geq \gamma$ ) then
27:             $R_{conf} \leftarrow R_{conf} \cup x$ 
28:          # extracting top 'n' lexicons
29:           $L_w \leftarrow Top\ Keyword(sort(t_f),n)$ 
30:          #update lexicon list
31:           $L \leftarrow (L_w \cup L_{old})$ 
32:          # predicted class labels for all tweets
33:           $c \leftarrow Response(R_{lexicon}) \cup Response(Prediction)$ 
34:        return ( $L_w, c$ )

```

---

**Domain-specific lexicon list generation:** In our algorithm, we are motivated to build a robust lexicon list. To build the lexicon list, we have taken the help of a domain

expert. In our lexicon list, all words related to three different categories named stock, economy, and company.

- At the initial stage, we have a total of 561 numbers of unique keywords belonging to three different domains, but the list is not static in nature.
- We have used, WordNet 2.1 in our experiment to expand the lexicon list with new keywords. In R [34], we have used the “wordnet” [35] package to extract all synonyms of type noun or adjective from the WordNet database, and finally, we have found that total lexicons in our lexicon list become 762.

## 4.2 Data Cleaning and Feature Extraction

- English language specific and domain-specific stop words are removed using the “tm” [36] package in R.
- We have used different customized regular expressions like “https\ s+\s\* | http\ s+\s\*” for URL removal, “@\w+” for user mention removal, “#\w+” for hash tag removal, and so on.
- For Lemmatization, we have used package “textstem” [37], and we have lemmatized all words based on an English dictionary.
- For creating the document-term matrix, tf-idf weighting has been used. Other techniques like Skip-gram, CBOW could have been used; however, it may be noted for small corpora the result might be unstable [38]. The continued research in tf-idf [39] is also a testament of its viability as an embedding method.

## 4.3 Keyword-Based Classification Strategy

Our final objective is to detect as many relevant tweets as possible based on lexicon with high accuracy. For achieving that, initially, we have selected all tweets having at least one domain-specific keyword in it. Then we have assigned a score with respect to all tweets collected from the previous step. To calculate the score, we use a collection of formulas they are defined as follows:

$$X = \frac{WC_{\text{feature}}}{T_{\text{count}}} \quad (1)$$

$$\text{score}_{\text{relevant}} = \frac{f_{\text{relevant}}}{n_{\text{relevant}}} + X \quad (2)$$

$$\text{score}_{\text{non\_relevant}} = \frac{f_{\text{non\_relevant}}}{n_{\text{non\_relevant}}} + X \quad (3)$$

$$\text{score}_{\text{mod}}^{\text{relevant}} = \text{score}_{\text{relevant}} + b_t \quad (4)$$

In Eq. (1),  $\text{WC}_{\text{feature}}$  implies total feature keyword count with respect to the lexicon list available in our database in each tweet, where  $T_{\text{count}}$  means the total number of words in each tweet. Now for each tweet belonging to the relevant class we have assigned a score  $\text{score}_{\text{relevant}}$ , where  $f_{\text{relevant}}$  is total number of relevant tweets containing at least two keywords and  $n_{\text{relevant}}$  is a count of the total number of tweets belonging to the class relevant. So the idea of  $\text{score}_{\text{non\_relevant}}$  is same but it is used for the nonrelevant class. After that, we have compared between Eqs. (2) and (3) to decide which one is greater. Interestingly, we found that first one is greater because it is quite obvious that out of all tweets where at least one keyword is present, the value of  $f_{\text{relevant}}$  and  $n_{\text{relevant}}$  is closer than the nonrelevant class. So, we increased the score of all relevant tweets with a bias term  $b_t$  in Eq. (4) as  $\text{score}_{\text{relevant}} + b_t$ . We have selected a dynamic threshold value “ $f$ ” as 0.7. Finally, select tweets as relevant if the score is greater than “ $f$ ”.

#### **4.4 Balancing the Dataset**

After extracting a subset of tweets based on the lexical approach, we have found that our remaining dataset is not balanced because the percentage of relevant observations compared to the whole dataset is only 3.8% that is far below of  $\lambda$  taken as 15% as a cutoff. We used  $\lambda$  as 15% because, traditionally, researchers are considered the percentage of minority class as from 15 to 20%, but the dataset can be more imbalanced in nature.<sup>1</sup>

- So, we use SMOTE from the “DMwR” [39] package of “R” [40] to balance the dataset.
- We have seen that in our dataset, the total proportion of relevant tweets with respect to nonrelevant tweets is almost 5%. So, for each observation from the relevant class, we generate five synthetic new relevant observations and append them to our dataset. In SMOTE as parameters, we have used the percentage of oversampling as 500 and the percentage of undersampling as 100. For the  $K$ -nearest neighbors used in SMOTE, we have used the value “ $k$ ” as 5.

#### **4.5 Cross-Validation Using SVM**

In this approach, the support vector machine has been used as the classifier:

---

<sup>1</sup>Natalie Hockham makes this point in her talk Machine learning with imbalanced data sets, which focuses on imbalance in the context of credit card fraud detection.

- A tenfold cross-validation has been used to protect against overfitting. For cross-validation, we have used package “caret” [41] in R.
- We have applied  $K$ -means clustering algorithm to our dataset to check the purity of our results. We have seen that purity is less than 95%. This justifies our choice of a non-learn kernel namely a radial kernel.
- For evaluating the classification model, we have performed ROC and AUC measure using package “pROC” [42] in R.

## 4.6 Top Keyword Extraction

- After classification, we have selected a subset of tweets predicted as relevant having confidence score greater than equal to  $\gamma$  as 0.93.
- After that, top “ $n$ ” keywords are extracted from the set of high confidence tweets. We have taken the value of “ $n$ ” as 50 in our experiment. Then added them to our existing lexicon list. The newly added keywords may be validated by a domain expert and again expanded using WordNet.

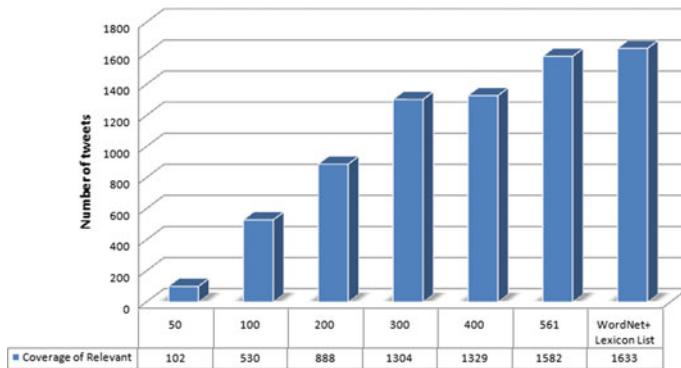
## 5 Results and Discussion

In this section, the results of our experiment are critically discussed. This has four main subsections. The first subsection discussed the improvement of the lexicon in terms of coverage of relevant tweets. In the second subsection, the improvement due to Synthetic Minority Oversampling (SMOTE) is elaborated. In the third subsection, change of recall and specificity with variation of different parameters of SMOTE have been elaborated. In the fourth subsection, a comparison has been done in terms of improvement over a baseline text classification method.

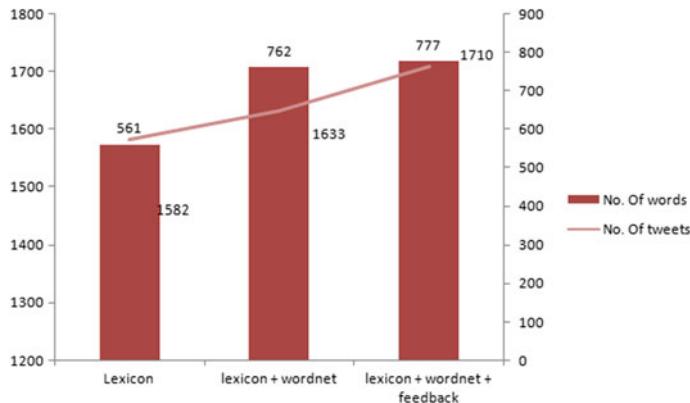
### 5.1 Percentage of Coverage by the Lexicon List

In Fig. 4, the effect of size of the lexicon with percentage of coverage of the tweet is examined. As expected, the coverage increases with lexicon size. The lexicons added by WordNet are also relevant and brings an improvement in coverage of relevant tweets by 3.2%.

The results of improvement of the lexicon list by WordNet and new keywords from high confidence relevant tweets are elucidated in Fig. 5. Interestingly, in the above figure, we see that by using only the lexicon list we covered a total of 1582 relevant tweets, using lexicon list plus WordNet, we can cover total 1633 tweets and finally using existing lexicon list, WordNet and feedback keywords we can



**Fig. 4** Improvement by lexicon



**Fig. 5** Coverage by the lexicon list

cover ultimately 1710 number of relevant tweets. Therefore, we can see that the percentage of improvement of Lexicon and WordNet approach over lexicon is 3% and the percentage of improvement of final approach over the second one is over 4.7%. In Table 1, the sample of added words in the lexicons is shown from both the result of WordNet and our feedback process. Getting such lexicons from WordNet and the feedback process implies that we have got very good lexicons relevant to stock from both approaches.

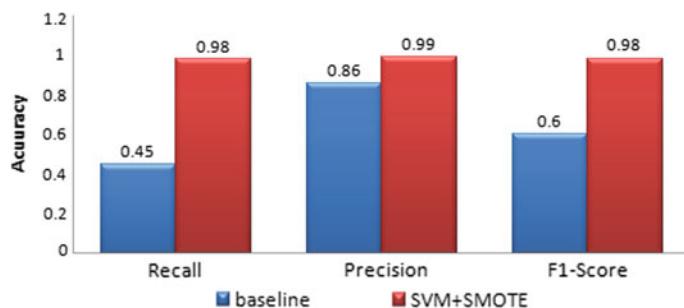
**Table 1** New lexicons

Lexicons from WordNet	Lexicons from feedback
Downfall, gross, tax, trade, impression, return, sale	Bullish, rise, target, investor, buy, look, stock

## 5.2 Improving Using SMOTE

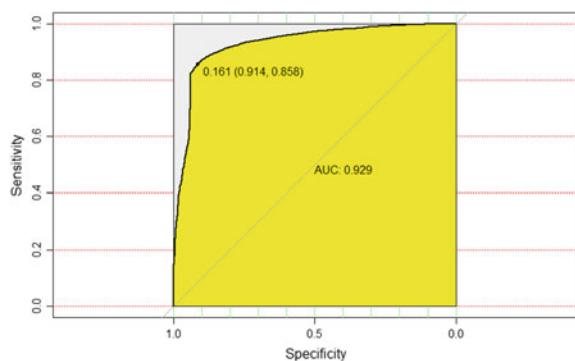
In Fig. 6, we can observe that the performance of SMOTE over the baseline approach is very significant. Where the baseline produces a recall of 45% for the relevant class, SMOTE achieves a 98% recall. Precision for the baseline approach is 86% where after applying SMOTE we have been achieved 99% of precision for the relevant class. F1-score after SMOTE has been increased by over 30%. In the above figure, we can see that although the precision of baseline and SVM with SMOTE is almost near to each other, the improvement over recall is very clear.

In Figs. 7 and 8, we observed that SVM + SMOTE for best threshold value 0.483 produced a sensitivity of 98% and specificity of 97%, which implies that SVM + SMOTE having better discriminating power over classes than baseline. Another improvement is SVM + SMOTE having area under the curve (AUC) of 99%, which is also outperforming the baseline approach.

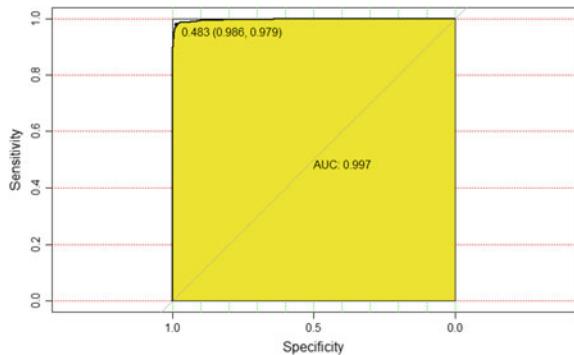


**Fig. 6** Accuracy measure of SMOTE a rebalancing technique over baseline

**Fig. 7** ROC-AUC measure for baseline



**Fig. 8** ROC-AUC measure for SVM + SMOTE

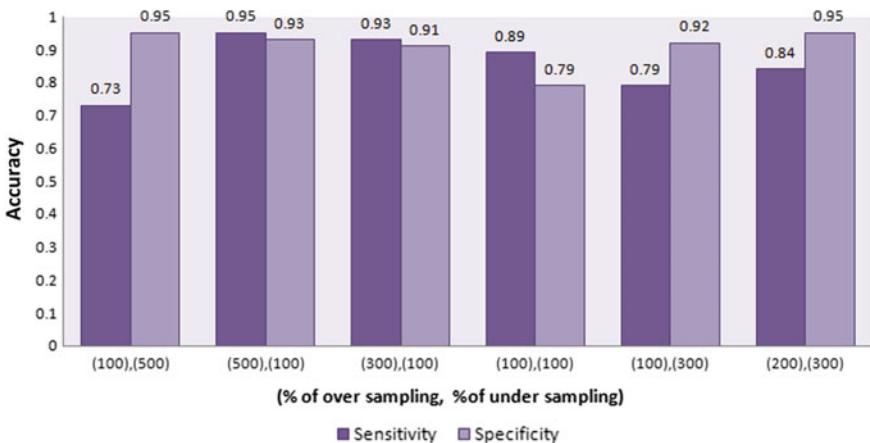


### 5.3 Variation with SMOTE

In Fig. 9, we have shown the tuning of SMOTE with various parameters for oversampling and undersampling. The oversampling at 500 and undersampling at 100 give best sensitivity or recall and specificity or true negative rate pair.

### 5.4 Performance Comparison Between Baseline and Proposed Methodology

In Table 2, we have shown the performance comparison between the baseline and the proposed approach. Using our hybrid approach the recall of relevant class signif-



**Fig. 9** Tune SMOTE with various parameters

**Table 2** Performance measure between baseline and proposed methodology

Recall + precision + F1-score	Baseline		Proposed approach	
	Relevant	Nonrelevant	Relevant	Nonrelevant
Recall	0.45	0.98	0.99	0.99
Precision	0.90	0.98	0.98	0.99
F1-score	0.60		0.98	

icantly improves from 45 to 99%. To compare with the overall performance of two approaches, for our proposed approach, F1-score improves by 38%.

## 6 Conclusion

In recent studies, it was observed that there is a strong correlation between sentiments expressed in Twitter and other social media and stock market movements w.r.t. indices as well as individual stocks. However, most of these studies have been done in the USA and UK, and our study focuses on Indian stock market only, where we intended to build a system, which will extract tweets on daily basis and then filter out the irrelevant tweets and store them in a database. Finally, we will use named entity recognition and dependency parsing to associate the sentiment of the tweets to a proper target which may be a stock, a particular industry or the economy in general.

**Keyword list:** The keyword list has been first expanded using WordNet and then frequently occurring terms which are not present in the lexicons that are added, with suitable expansion by WordNet. The lexicon list has grown from 561 to 777 words, i.e., by 38.5%, and the coverage of relevant tweets by keywords has increased by 8%. The lexicon-based strategy can finally cover 91.3% of the relevant tweets, which is highly significant, and with few more tweets, this coverage can be further improved.

**Balancing the dataset:** SMOTE has been used to balance the dataset, which has improved the class-specific recall, precision, and *F*-Score from 0.45, 0.86, and 0.6 to 0.98, 0.99, and 0.98, respectively.

**Improvement in classifier performance in combined approach:** It may be noted that recall is the most important parameter in this context, as we want to get as many relevant tweets as possible with better precision. The proposed approach shows remarkable improvement by more than doubling the recall rate. The system is also unique, as it has the capability to grow the lexicons on its own.

## References

1. Liu, H., et al. (2016). The good, the bad, and the ugly: Uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics*, 1(3–4), 137–143.

2. Ediger, D., Jiang, K., Riedy, J., Bader, D.A., & Corley, C. (2010, September). Massive social network analysis: Mining Twitter for social good. In *2010 39th International Conference on Parallel Processing (ICPP)* (pp. 583–593). IEEE.
3. Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014, May). Tweedr: Mining Twitter to inform disaster response. In *ISCRAM*.
4. Abouze, A., Boudjериou, Y., Entringer, G., Az, J., Bringay, S., & Poncelet, P. (2014, June). Mining Twitter for suicide prevention. In *International Conference on Applications of Natural Language to Data Bases/Information Systems* (pp. 250–253). Cham: Springer.
5. Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A., & Chakraborty, B. (2016). A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal*, 9(3), 362–378.
6. Jain, V. K., & Kumar, S. (2017). Effective surveillance and predictive mapping of mosquito-borne diseases using social media. *Journal of Computational Science*, 25, 406–415.
7. Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16), 6266–6282.
8. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
9. Rao, T., & Srivastava, S. (2012, August). Analyzing stock market movements using Twitter sentiment analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (pp. 119–123). IEEE Computer Society.
10. Zhang, X., Shi, J., Wang, D., & Fang, B. (2017). Exploiting investors social network for stock prediction in China's market. *Journal of Computational Science*, 28, 294–303.
11. Ruan, Y., Durresi, A., & Alfantoukh, L. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems*, 1(145), 207–218.
12. Nisar, T. M., & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science*, 4(2), 101–119.
13. Rajput, H. (2014). Social media and politics in India: A study on Twitter usage among Indian Political Leaders. *Asian Journal of Multidisciplinary Studies*, 2(1), 63–69.
14. Khan, A. Z., Atique, M., & Thakare, V. M. (2015). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science and Engineering (IJECSCE)*, 89.
15. Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining* (p. 5). ACM.
16. Christiane, F. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
17. Rothwell, A. C., Jagger, L. D., Dennis, W. R., & Clarke, D. R. (2004). Networks Associates Technology Inc, 2004. Intelligent SPAM detection system using an updateable neural analysis engine. *U.S. Patent 6,769,016*.
18. Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–334.
19. Kumar, M., & Rangan, V. (2011). Clearwell Systems Inc, 2011. Methods and systems for e-mail topic classification. *U.S. Patent 7,899,871*.
20. Veningston, K., Shanmugalakshmi, R., & Nirmala, V. (2015). Semantic association ranking schemes for information retrieval applications using term association graph representation. *Sadhana*, 40(6), 1793–1819.
21. Rani, P., Pudi, V., & Sharma, D. M. (2016). A semi-supervised associative classification method for POS tagging. *International Journal of Data Science and Analytics*, 1(2), 123–136.
22. Lpez, V., et al. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250, 113–141.
23. Melville, P., Gryc, W., & Lawrence, R. D. (2009, June). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1275–1284). ACM.

24. Yenala, H., et al. (2017). Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*, 6(4), 273–286.
25. Lu, B., & Tsou, B. K. (2010, July). Combining a large sentiment lexicon and machine learning for subjectivity classification. In *2010 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 6, pp. 3311–3316). IEEE.
26. Zhao, S., et al. (2016). Correlating Twitter with the stock market through non-Gaussian SVAR. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE.
27. Pagolu, V. S., et al. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. IEEE.
28. Oliveira, N., Paulo C., & Nelson, A. (2013). Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*. ACM.
29. Leitch, D., & Sherif, M. (2017). Twitter mood, CEO succession announcements and stock returns. *Journal of Computational Science*, 21, 1–10.
30. Chung, S., & Sandy, L. (2011). *Predicting stock market fluctuations from Twitter*. Berkeley, California.
31. Mao, Y., Wei, W., & Bing, W. (2013). Twitter volume spikes: analysis and application in stock trading. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM.
32. Simsek, M. U., & Suat, Z. (2012). Analysis of the relation between Turkish Twitter messages and stock market index. In *2012 6th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE.
33. Smailovi, J., et al. (2013). Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 77–88). Berlin, Heidelberg: Springer.
34. R Core Team. (2017). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*, Vienna, Austria, <https://www.R-project.org/>.
35. Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Bradford Books.
36. Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5), 1–54.
37. Rinker, T. W. (2018). *Textstem: Tools for stemming and lemmatizing text version 0.1.4*. New York: Buffalo.
38. Faruqui, M., et al. (2016). *Problems with evaluation of word embeddings using word similarity tasks*. arXiv preprint [arXiv:1605.02276](https://arxiv.org/abs/1605.02276).
39. Torgo, L. (2010). *Data mining with R, learning with case studies*. Boca Rotan: Chapman and Hall/CRC.
40. R Development Core Team. (2008). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*, Vienna, Austria. ISBN:3-900051-07-0.
41. Kuhn, M. (2018). Caret: classification and regression training. Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., The R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Tyler Hunt. In *R Package Version 6.0-79*.
42. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77.

# Generative Adversarial Networks as an Advancement in 2D to 3D Reconstruction Techniques



Amol Dhondse, Siddhivinayak Kulkarni, Kunal Khadilkar, Indrajeet Kane,  
Sumit Chavan and Rahul Barhate

**Abstract** Synthesizing three-dimensional objects from single or multiple two-dimensional views has been a challenging task. To combat this, several techniques involving Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and Recurrent Neural Network (RNN) have been proposed. Since its advent in 2014, there has been a tremendous amount of research done in the area of Generative Adversarial Networks (GANs). Among the various applications of GANs, image synthesis has shown great potential due to the power of two deep neural networks—generator and discriminator, trained in a competitive way, which are able to produce reasonably realistic images. Formulation of 3D-GANs—which are able to generate three-dimensional objects from multiple two-dimensional views with impressive accuracy—has emerged as a promising solution to the aforementioned issue. This paper provides a comprehensive analysis of deep learning methods used in generating three-dimensional objects, reviews the different models and frameworks for three-dimensional object generation, and discusses some evaluation

---

Kunal Khadilkar, Indrajeet Kane, Sumit Chavan and Rahul Barhate: Indicates equal contribution.

---

A. Dhondse

IBM Master Inventor, IBM, Pune, India

e-mail: [amol.dhondse@in.ibm.com](mailto:amol.dhondse@in.ibm.com)

S. Kulkarni

Department of Computer Engineering, MIT-WPU, Pune, India

e-mail: [siddhivinayak.kulkarni@mitcoe.edu.in](mailto:siddhivinayak.kulkarni@mitcoe.edu.in)

K. Khadilkar (✉) · I. Kane · S. Chavan

Department of Computer Engineering, MITCOE, Pune, India

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

I. Kane

e-mail: [kaneindrajeet1202@gmail.com](mailto:kaneindrajeet1202@gmail.com)

S. Chavan

e-mail: [chavansumit13@gmail.com](mailto:chavansumit13@gmail.com)

R. Barhate

Department of Information Technology, MITCOE, Pune, India

e-mail: [rahulbarhate97@gmail.com](mailto:rahulbarhate97@gmail.com)

metrics and future research direction in using GANs as an alternative for simultaneous localization and environment mapping as well as leveraging the power of GANs to revolutionize the field of education and medicine.

**Keywords** Generative adversarial networks · Convolutional neural network · Deep learning · Three-dimensional object reconstruction · Image synthesis

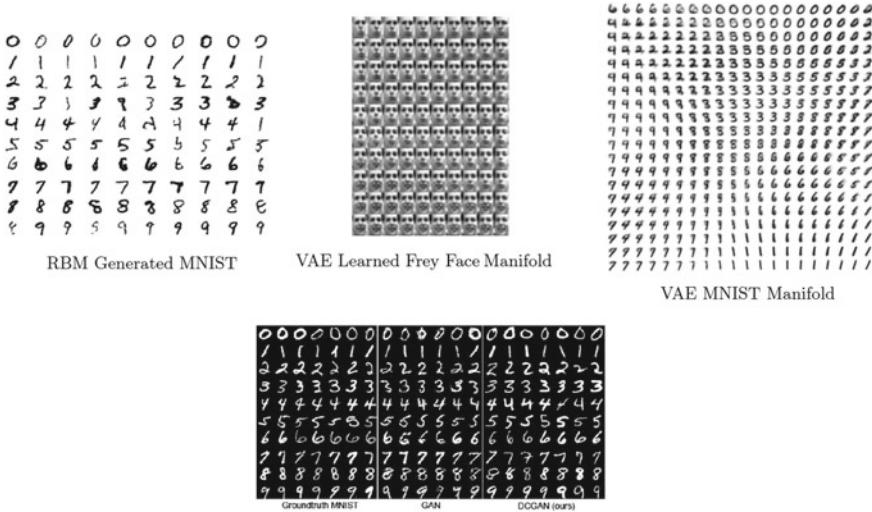
## 1 Introduction

With advancements in the field of deep learning, algorithms are now able to outperform humans in some tasks such as classification of images on ImageNet [1], playing Go [2], and Texas Hold’em Poker [3]. However, it is not possible to come to the conclusion that these algorithms have true “intelligence”. This is because knowing how to perform a task does not necessarily mean that the agent has a complete understanding of the task. As said by the renowned physicist Richard Feynman, “What I cannot create, I cannot understand”, it is pivotal for an agent to understand its task. In case of machines, to understand the input data they need to learn to generate the data. A promising approach is to use generative models which learn specific structural and semantic properties using which the model can synthesize new samples. Generative algorithms attempt to estimate attributes given a specific label. Discriminative algorithms, on the other hand, try to classify the input data. That is, given the features of a data, the label or category to which a dataset belongs is predicted [4]. Using a learned generative model, it is even possible to draw samples which are not in the training set but follow the same distribution. Other popular generative models such as Restricted Boltzmann Machines (RBMs) [5], Variational Autoencoders (VAEs) [6] use latent variables as a hidden representation of data samples. They specify an explicit parameterized log-likelihood function which represents the data. Integrating over the entire space of latent variables is needed for estimating maximum likelihood of parameters, which is difficult to deal with. Hence, the approximation methods do not always yield good results [7]. On the other hand, Generative Adversarial Network (GAN) [8] is an implicit probabilistic model which defines a stochastic procedure capable of generating data from a latent variable belonging to a low-dimensional space. Deep Convolutional GANs (DCGANs) [9] are proved to be an improvement over GANs. Figure 1 illustrates the images generated by RBMs, VAEs, and DCGANs models when trained on the MNIST dataset.

Having a crude implementation, the images generated using GANs are significantly sharper compared to RBMs and VAEs. For example, when trained on datasets such as CIFAR [10] and SVHN [11], RBMs and VAEs fail to produce complex images. On the other hand, GANs generate far better images on these datasets too [7].

### Generative Adversarial Networks (GANs)

GAN, invented by a team of researchers headed by Goodfellow [8], is a state-of-the-art neural network architecture having wide-ranging applications in the field of 2D



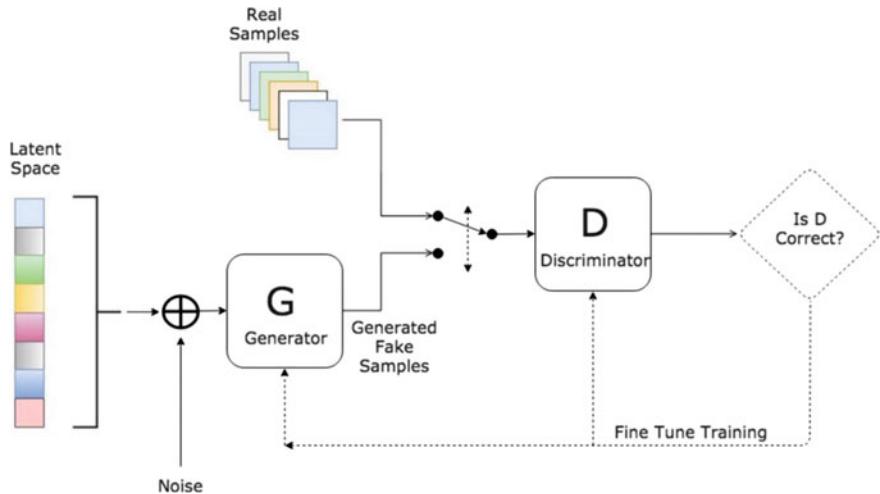
**Fig. 1** Generative results on the MNIST produced by RBM, VAE, and DCGANs [7]

to 3D reconstructions and visualizations. Initially, GANs were applied extensively in the field of image generation [7]. It was found that the primitive GANs were quite prone to noise and the images generated were unclear. As a result, different varieties of GANs like the DCGANs and Laplacian GANs (LAPGANs) [9] have been introduced to overcome the shortcomings of traditional GANs. GANs are a kind of a network in which two networks compete with each other. The aim is to generate data which is very much similar to the training data [12].

Figure 2 describes the flow and working of a GAN, highlighting important modules like the generator and the discriminator. The aim of the generator is to generate random data mostly in the form of a latent random variable represented as a matrix. This random data obtained from the generator is fed as input to the discriminator. The other input to the generator is the data of real-world images or authentic images. Discriminator does the important task of differentiating between real and fake. Loss is calculated and the evaluation function is manipulated such that the loss is minimized.

### Structure and Working of GANs

GANs belong to a set of algorithmic approaches known as generative models, mainly in the field of unsupervised learning. Generative models have the ability to learn intrinsic distribution function of the input dataset  $p(x)$  for a single class, or  $p(x, y)$  for multiple classes. This allows the generative models to create synthetic inputs  $x_1$  and synthetic outputs  $y_1$ . A neural network  $G(z, \theta_1)$  is used for modeling a generator with the aim of mapping input noise variables (say  $z$ ) to the desired data space. At the same time, a second neural network  $D(x, \theta_2)$  models the discriminator and outputs the probability that the data came from the real dataset. Here,  $\theta_i$  represents weights or parameters for defining a neural network. The discriminator is trained correctly to differentiate between authentic or fake images. The aim of the generator is to



**Fig. 2** Architecture of GANs [12]

fool the discriminator by generating data as realistic as possible. After several steps of training, if the generator and discriminator are powerful enough, the system will reach a point where both of them will not be able to improve anymore. When this occurs, the discriminator will fail to differentiate between real data and synthetic data. A GAN can be thought of as playing a min-max game, represented by Eq. (2), where the generator is trying to maximize its probability of having its outputs recognized as real, while the discriminator is trying to minimize this same value [12].

Equation (1) represents the min-max function used by the generator and the discriminator.

$$\min \max V(D, G) = E_{x \sim p_{\text{data}}} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

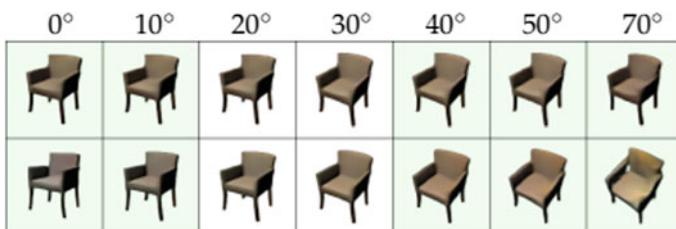
The paper is organized as follows. Section 2 describes 3D image regeneration techniques focusing on two major areas: Object and Face Regeneration. In Sect. 3 of the paper, a comprehensive study of existing datasets has been carried out. Along with this, future scope of improvements as well as prospective fields of dataset creation has been described. In Sect. 4, possible applications in the field of health and education using such GANs architectures have been described. Prospective future advancements in this field have been discussed.

## 2 Related Work

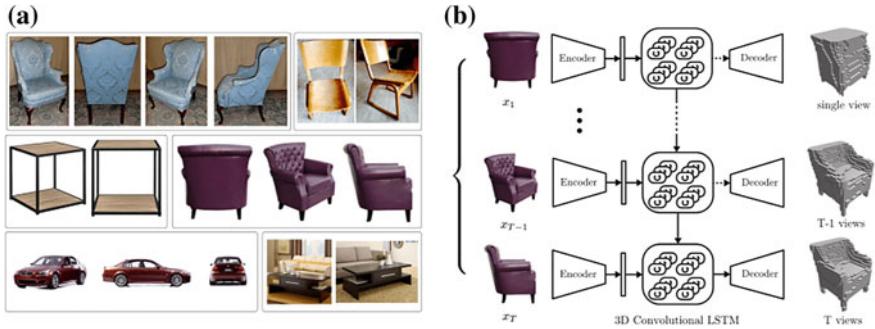
Since the advent of Computer Vision, 3D object understanding and generation has been a core problem. While different approaches to this problem have been tried and tested over time, attempts based on Deep CNNs and GANs have been found to be able to overcome the traditional drawbacks faced. Substantial progress has been made by the researchers in the field of 3D object modeling and synthesis [13–15]. This work was mainly focused on meshes, skeletons, and multi-view images. The synthesized objects created using these methods were realistic but did not qualify as conceptually novel. In the field of 3D object recognition, various techniques like combined embedding of 3D shapes and blended pictures [16–18], 3D object synthesis from in-the-wild images have been proposed [19, 20]. More recently, DC-GAN [9] achieved impressive performance by adopting GANs with convolutional networks for image synthesis.

Dosovitskiy et al. [21] took advantage of the fact that CNNs work efficiently given a large enough dataset. They used the concept of generating multi-view of objects given detailed information about the object which included the model number that defined the style, viewpoints, and other key points such as zoom, brightness, color, saturation, etc. As seen in Fig. 3, the network allows itself to smoothly transform between different object views or object instances with every transitional picture being important, and thus, the network could perform knowledge transfer within a class and between two different classes as well as perform feature arithmetic leading to acceptable results.

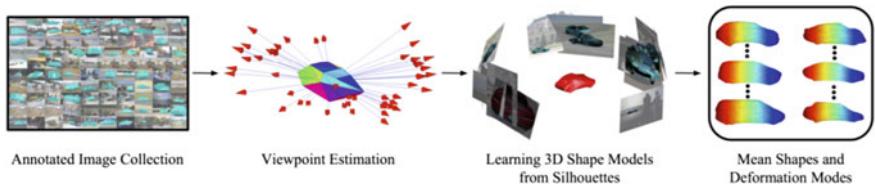
The success of Long Short-Term Memory (LSTM) [22] networks, along with the progress made in single-view 3D reconstruction using CNNs has enabled us to overcome many difficulties in 3D object regeneration. Inspired by efficient methods that make use of shape priors to accomplish strong 3D recreations, Choy et al. [20] proposed a novel recurrent neural network architecture called as the 3D Recurrent Reconstruction Neural Network (3D-R2N2). The architecture is shown in Fig. 4b and the objects that are aimed to reconstruct is shown in Fig. 4a. They have assumed that the earlier information about the object appearance and shape is accessible and used this to their benefit. Thus, shape prior-based strategies can work with fewer pictures and fewer presumptions on the object appearance. The system accepts as



**Fig. 3** Evaluation angle knowledge transfer [21]



**Fig. 4** **a** Objects aiming to reconstruct [20]. **b** Overview of the network [20]



**Fig. 5** Architecture of training pipeline [23]

input manifold images of an object example from distinct viewing angles. The output forms a 3D occupancy grid that shows a representation of the object.

Kar et al. [23] has used the architecture of NRSfM [24] to accurately evaluate the camera viewpoints (turning, translation, and scale) for each and every preparation examples in each class. They have proposed a simple addition to NRSfM algorithm that inculcates outline data along with key point correspondences to vigorously recuperate cameras and shape bases. This is the first attempt to execute fully autonomous object regeneration using a single image on a huge and realistic dataset. The architecture of the training pipeline used in [23] is shown in Fig. 5.

For all experimental evaluations, the work in [23] has considered pictures from the PASCAL VOC 2012 dataset [25] and has additionally utilized openly accessible ground truth class-particular key points [26]. An evaluation of the learned 3D models on the PASCAL 3D+ dataset [27] which has up to 10 3D computer-aided design models for the inflexible classifications in PASCAL VOC has been performed.

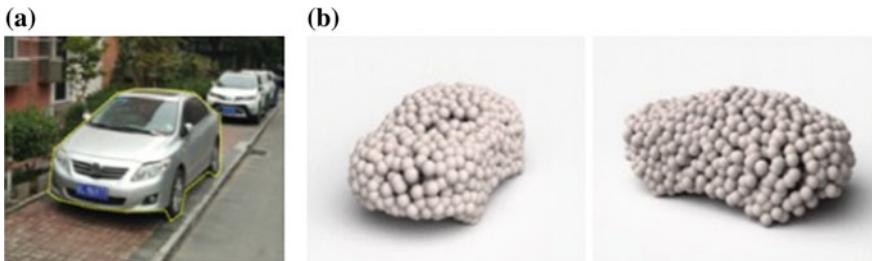
The problem to naturally reconstruct novel pictures after applying inherent transitions (e.g., 3D rotation and disfigurement) to an input image is a tough task. This is because of the loss of data apparent in representing a 3D object into the image space. Yang et al. [28] tried to solve this problem by using a model based on deep convolutional encoder-decoder network trained to perform trajectories of multiple transformations. The Multi-PIE [29] dataset along with a dataset of 3D chair models was used for checking the quality of their predictions. The network demonstrated its potency for generating 3D views of unknown object instances. Zhou et al. [30]

also attempted to solve the multi-view task by using the fact that any two different perspectives of the same object are highly connected. They used CNN to predict vectors determining which pixels in the input view could be used to reproduce the objective view. Tatarchenko et al. [31] have also contributed to this domain by using a network that is trained end-to-end on renderings of 3D models from the ShapeNet dataset [32]. The network can anticipate an RGB image and a depth map of the object as seen from a random view. Kulkarni et al. [33] prepared a variant of a VAE, where particular measurements offer responses to the diverse components of differences in the input information such as perspective and illumination. This technique is appreciable and permits to create previously unobserved perspectives of objects; however, the prediction standard of work in [31] is superior than that in [33].

For generating complex 3D shapes, voxels or 3D parts are utilized as underlying portrayals. In the past, people have tested different sampling novel voxelized 3D shapes by modeling them by the means of volumetric CNNs [32, 34, 35]. The disadvantages of these methods are high memory requirement and low resolutions which occur due to dimensionality.

The majority of existing work that has been done use the normal portrayals such as volumetric frameworks or set of pictures. The drawbacks of these portrayals are hiding the natural invariance of 3D shapes under geometric transitions, and furthermore experiencing other different issues. Another drawback is that convolutions on the 3D space are computationally expensive and grow cubically with resolution, thus typically limiting the 3D reconstruction to exceedingly coarse representations. Fan et al. [36] have tried to solve the problem of constructing 3D objects from single images using a particular form of representation: Point cloud coordinates as seen in Fig. 6. A point cloud is a basic, consistent structure that represents geometric shapes as a set of 3D locations in a Euclidean frame. In 3D, these locations are defined by their  $x$ ,  $y$ ,  $z$  coordinates. Thus, the point cloud representation of an object or scene is a  $N \times 3$  matrix, where  $N$  is the number of points, referred to as the point cloud resolution.

Table 1 shows how the work in [31] fares compared to 3D-R2N2 which is currently the avant-garde in 3D object recognition. As can be seen in the single-view reconstruction setting [36] attained greater Information over Union (IoU) in all sections and in eight out of thirteen for other sections as well.



**Fig. 6** **a** Input [36] and **b** reconstructed 3D point cloud [36]

**Table 1** 3D reconstruction comparison (per category)

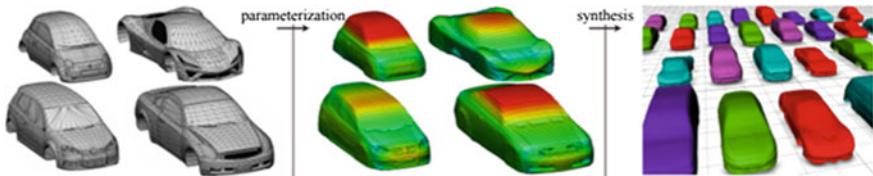
Category	[31]	3D-R2N2		
	1 view	1 view	3 views	5 views
Plane	0.601	0.513	0.549	0.561
Bench	0.55	0.421	0.502	0.527
Cabinet	0.771	0.716	0.763	0.772
Car	0.831	0.798	0.829	0.836
Chair	0.544	0.466	0.533	0.55
Monitor	0.552	0.468	0.545	0.565
<i>Mean</i>	0.6415	0.563	0.62	0.635

The new PointNet introduced by Qi et al. [37] has inspired the architecture of the network [38] which is based on a patch-based learning method. The generic surface properties, for example, normals or curvature is difficult to extract and techniques have utilized the way to deal with the extraction of these features from smooth surfaces corresponding to neighborhood patches of the point cloud. But these approaches were found to be sensitive to various parameter settings such as the neighborhood size, or the degree of the fitted surface.

They have undertaken an approach that is developed upon a deep neural network trained on a comparatively small collection of shapes that can accommodate an extensive variety of conditions with similar parameter settings.

Machine learning algorithms require a steady portrayal of input and output information in the form of orthogonally aligned grid. Hence, for the appropriate representation, the following approach in Fig. 7 is presented. From the unstructured triangle mesh (left), the approach can effectively generate a quad mesh with a uniform structure (middle). The autoencoder generates a low-dimensional portrayal of a collection of shapes to construct new shapes (right).

Umetani [39], has leveraged depth map coupled with Shrink Wrapping Parameterization in order to create a concise and effective parameterization of 3D shapes. Further, an autoencoder constructs numerous 3D shapes. This acts as an input to the direct manipulation interface which explores generative shapes. The approach used in [39] handles difficulties like inverted faces, holes, and self-intersections. The

**Fig. 7** Pipeline for exploring generative 3D shapes [39]

use of autoencoder network helps to synthesize different shapes by extracting many forms of a category.

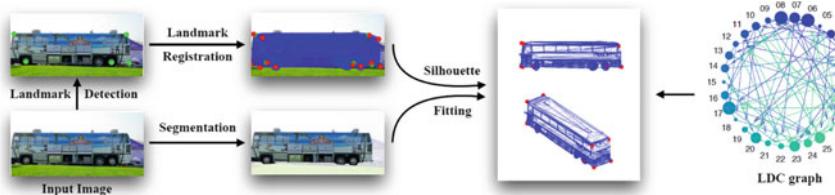
Although volumetric representations have shown commendable results in 3D object reconstruction, they have their own issues. These issues include low spatial resolution resulting due to the computational complexity involved with training a network with large 3D signals. 3D mesh depiction of CAD models preserves significant 3D details as opposed to volumetric methods. Kong et al. [40] have first proposed a graph embedding model and shows that a shape dictionary can be formed from every subgraph. Second, they propose a two-step process to select a subgraph using landmark registration and then generate a compact model using landmarks and silhouette. Eventually, they estimate fine geometry for different object categories.

In Fig. 8, the authors have created a dense correspondence graph as seen in the rightmost figure from a single image having noticeable landmarks and outliers. This is followed by landmark registration of CAD models by a rough estimate of camera position. This eventually creates a dense deformable model which fits both silhouettes and landmarks.

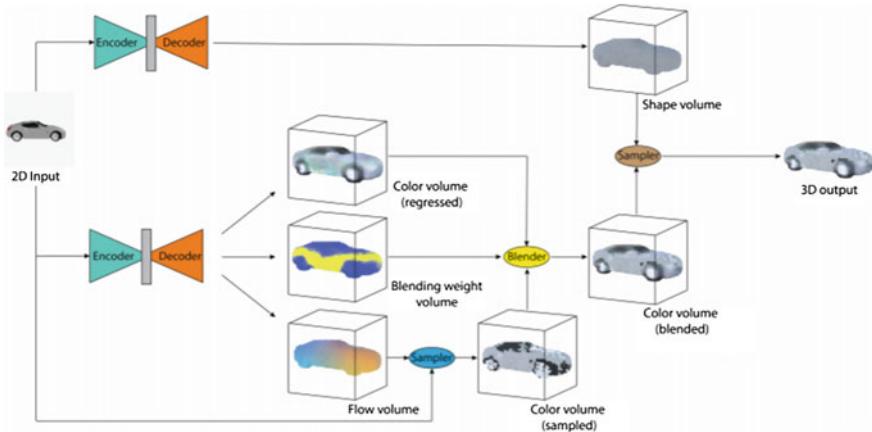
Pontes et al. [41] have first proposed to represent 3D mesh models by utilizing Free-Form Deformation (FFD) control points and the weight of sporadic linear combinations. Second, they developed an FFD algorithm to choose a 3D model that fits an image perfectly and evaluate the displacements of FFD control points. Finally, they experimentally show how the proposed framework is beneficial for denser realistic from an image. An idea of colorful 3D reconstruction is put forth by Sun et al. [42] which mean to extract both the 3D shape and the surface color from an input image. An end-to-end trainable system, Colorful Voxel Networks (CVN), which combines the strength of appearance and the geometric projections is also used.

In Fig. 9, the architecture has two branches one for Shape Construction and the other for Color Estimating. It is an end-to-end trainable framework to introduce a concept of Colorful Voxel Network (CVN) which combines the strength of appearances and geometric projections. On a single 2D input, the CVN decomposes the shape and color information of a 3D object into a shape branch and a surface color branch. To handle sparse representation, a loss function, Mean Squared False Cross-Entropy Loss (MSFCEL), is designed.

Liu et al. [43] use GANs as an interactive 3D editing tool. They propose to input a random input into the latent space of a GAN to balance the likelihood of the input space and the authenticity of the output space. In addition to this, they have provided



**Fig. 8** Overview of method used in [40]

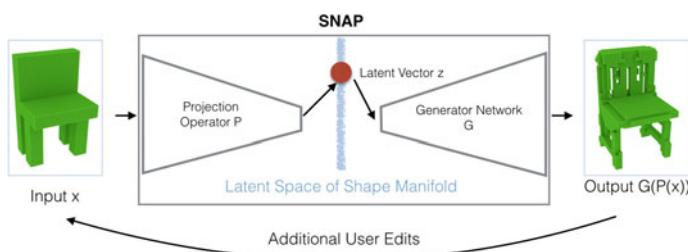


**Fig. 9** The model architecture used for colorful 3D reconstruction [42]

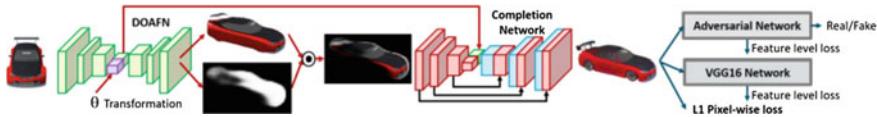
a dataset for 3D polynomials of 101 objects. In all, a basic interactive modeling tool for beginners is put forward.

With a Minecraft-like interface, a user can select a voxel grid to regenerate. The selected voxel grid is replaced by a similar voxel grid generated by a 3D-GAN. Figure 10 shows that 3D GANs are a useful interactive tool for beginners. An exclusive dataset was created for the use of GANs. An augmented dataset of 3D models consists of an augmentation of ShapeNet Core55 dataset [44], but inculcated from ModelNet40 [32], SHREC 2014 [45], and Yobi3D [46].

Recently introduced Appearance Flow Network (AFN) [30] trains a convolutional encoder-decoder network to figure out to move pixels without needing direct access to the basic 3D geometry. Given an input image and an objective change, Park et al. [47] have proposed two methods; the first method—Disocclusion-aware Appearance Flow Network (DOAFN)—changes the input view by moving pixels that can be seen in both the input and objective view along with generating a visibility map. The image completion network, at that point, visualizes the missing parts and furthermore processes regions that experience ill effects of bending or unreasonable specifications



**Fig. 10** Interactive 3D modeling with a GAN [43]



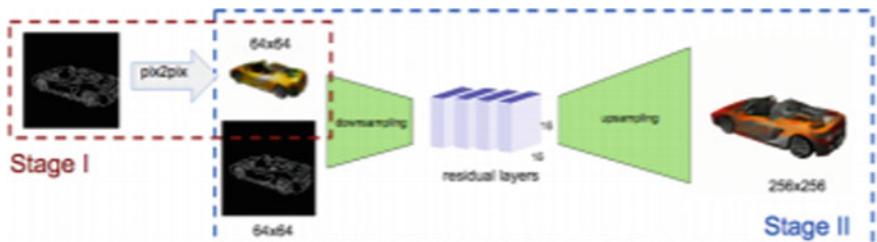
**Fig. 11** Architecture of Transformation-grounded View Synthesis Network (TVSN) [47]

because of incorrect transformation prediction with a combination of adversarial and feature-reconstruction loss. This architecture can be seen in Fig. 11.

The VGG16 network [48] has been utilized for computing the component reconstruction losses from various layers, called as the perceptual loss. Adversarial training [49] has been effective for training the loss network simultaneously with training the image generation network. In this way, they undertook the idea of feature matching presented in [50] to increase the stability of the training process. In this strategy, two uses of GANs—semi-supervised learning along with generation of realistic images have successfully passed the visual Turing test [51]. The essential objective of this methodology is to enhance the effectiveness of GANs for semi-supervised learning. The attributes of produced images with the two types of loss networks, perceptual and adversarial, are complementary. Hence, they have consolidated them together with the standard image reconstruction loss (L1) to boost execution.

Revinskaya and Feng [52] have applied different conditional Generative Adversarial Networks (cGANs) to sketches in order to generate colored images having a 3D looking shading. A pix2pix model is trained to output 3D-looking images, which acts as a baseline and further refinement is applied to this baseline with the help of StackGAN [53]. The generator consists of eight encoder and decoder convolutional layers, a batch normalization layer, and a Leaky ReLU activation. The discriminator consists of six convolutional layers with an additional layer at the end to map to a single output value. Stage-I GAN reuses a pix2pix model while Stage-II GAN conditions on both low-resolution pictures created by the past stage, and furthermore the edge input again to redress defects in Stage-I results and include the more convincing points of interest.

The model as shown in Fig. 12 achieved impressive results for edge inputs. The model was able to add 3D-looking shading, color, and detail to input edges. One drawback obtained in this model was that it could not generalize well on sketches.

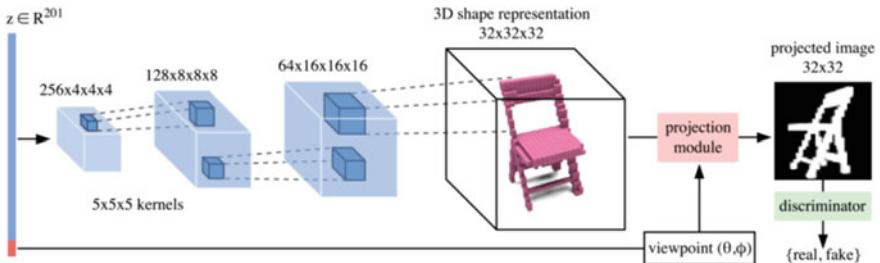


**Fig. 12** Stage-I and Stage-II architecture [52]

Gadelha et al. [54] have proposed Projective GANs (Pr-GANs) that trains a deep generative model of 3D shapes where projections correspond to the distributions of the input 2D projections. Given a collection of images of an unseen set of objects taken from an unseen set of views, the generative model performs its training process. There were various issues in the method of [54] such as shading signals are no longer accessible; the occurrence used to generate image; the perspective from which the image was produced; and the quantity of basic occurrences was not given. These problems are the reasons why existing approaches of constructing 3D geometry such as structure-from-motion [55] and visual hulls [56] could not be used. The architecture of Pr-GAN is shown in Fig. 13. The proposed Pr-GAN model in [54] as compared to 3D-GANs cannot discover structures which are not visible due to obstruction from all the remaining views.

For example, it fails to identify that some chairs have concave shapes and the generator fills these, as it does not modify the silhouette from any direction. This can be seen in Fig. 14.

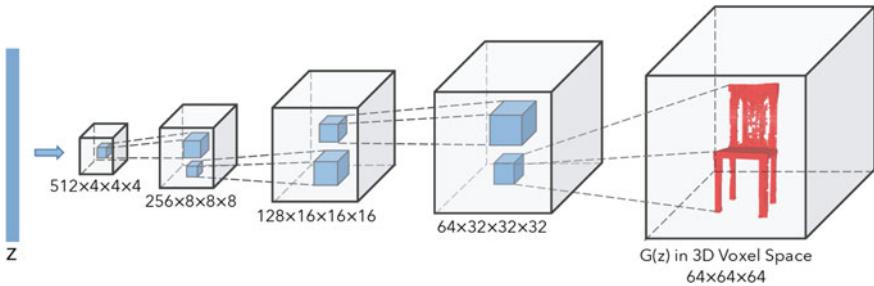
Wu et al. [57] have put up a comprehensive study on 3D object generation. An advanced framework capable of generating 3D objects from a probabilistic space using volumetric CNN [12] and GANs [8] has been proposed. This model benefits due to three particular factors. First, due to the utilization of an adversarial criterion, the object structure is implicitly captured by the generator in order to produce



**Fig. 13** Architecture of Pr-GAN [54]



**Fig. 14** Drawbacks of Pr-GAN [54]



**Fig. 15** 3D-GANs architecture [57]

high-quality 3D objects. Second, the generator is able to map a low-dimensional probabilistic space to the space of 3D objects. Due to this, it is possible to sample objects without the need of a reference image. Third, a 3D shape descriptor is provided by the discriminator which is learned using unsupervised learning and has a variety of applications in 3D object recognition. Figure 15 represents the architecture of 3D-GANs.

In the proposed architecture, a 200-dimensional latent vector  $z$  is mapped by the generator  $G$ . This vector is sampled arbitrarily from a probabilistic latent space to a  $64 \times 64 \times 64$  cube which represents  $G(z)$ , an object in a 3D voxel space. The discriminator  $D$  produces a confidence value  $D(x)$  which denotes if a 3D object input  $x$  is genuine or manufactured.

Inspired from [8], the classification loss is given by the binary cross-entropy, and the overall adversarial loss function is defined in Eq. (2).

Equation (2) gives adversarial loss function

$$L_{\text{3D-GAN}} = \log D(x) + \log(1 - D(G(x))) \quad (2)$$

Here,  $x$  is a real object in a  $64 \times 64 \times 64$  space.  $z$  is an arbitrarily sampled noise vector from a distribution  $p(z)$  having each of its dimension in a uniform distribution over  $[0, 1]$ . In order to generate 3D objects, an all-convolutional neural network has been proposed inspired by Radford et al. [9]. As seen in Fig. 5,  $G$  consists of five volumetric fully convolutional layers of kernel sizes  $4 \times 4 \times 4$  and strides 2 having batch normalization and ReLU layers in between and a Sigmoid layer at the end. The discriminator mirrors the generator which uses Leaky ReLU instead of ReLU layers [58]. This method generates superior quality 3D objects while the features learned using unsupervised learning achieve high performance on 3D object recognition on par with supervised learning method.

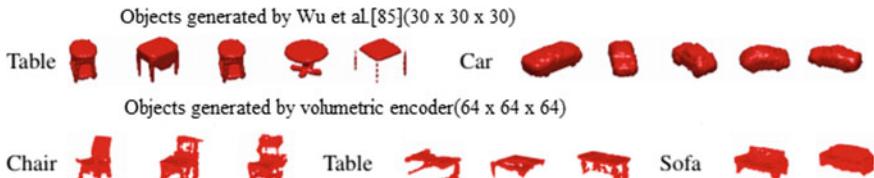
3D-VAE-GANs have been used as an expansion to 3D-GAN and consolidated VAE and GAN by imparting the decoder of VAE to the generator of GAN. The 3D-VAE-GAN performs well on single Image 3D-reconstruction as can be seen from the tests on IKEA datasets [59]. The 3D-VAE-GANS contains an image encoder  $E$ , generator  $G$ , and a discriminator  $D$ . The image encoder takes a 2D image  $x$  as

an input and outputs the latent representation vector  $z$ . Both 2D images and their corresponding 3D models are needed to train the 3D-VAE-GAN. The 3D shapes are rendered in front of background images (16,913 indoor images from the SUN database [60] in 72 views (from 24 angles and 3 elevations). The model has been tested on six categories of objects. The objects generated using this method are shown in Fig. 16 while the statistical results are shown in Table 2.

In Fig. 17, the images on the left show that the resultant “arm” vector can be included in different chairs. The ones to the right show the “layer” vector can be appended to other tables.

The results obtained from [57], as shown in Table 3, prove that 3D-GANs excels other unsupervised learning methods significantly. The proposed methodology gives an accuracy of 83.3% for the ModelNet40 Dataset [28], while an impressive 91% accuracy is obtained for the ModelNet10 Dataset.

The techniques for 3D object regeneration from images can be classified as active and passive techniques, where some form of temporal or spatial modulation of the



**Fig. 16** Object generated from 3D-GAN

**Table 2** Average prediction on the IKEA dataset [59]

Method	Bed	Bookcase	Chair	Desk	Sofa	Table	Mean
AlexNet-fc8 [58]	29.5	17.3	20.4	19.7	38.8	16	23.6
AlexNet-conv4 [58]	38.2	26.6	31.4	26.6	69.3	19.1	35.2
T-L network [58]	56.3	30.2	32.9	25.8	71.7	23.3	40
3D-VAE-GAN (jointly trained) [61]	49.1	31.9	42.6	34.8	79.8	33.1	45.2
3D-VAE-GAN (separately trained) [61]	63.2	46.3	47.2	40.7	78.8	42.3	53.1



**Fig. 17** Shape arithmetic for chairs and tables [57]

**Table 3** Accuracy and results table [57]

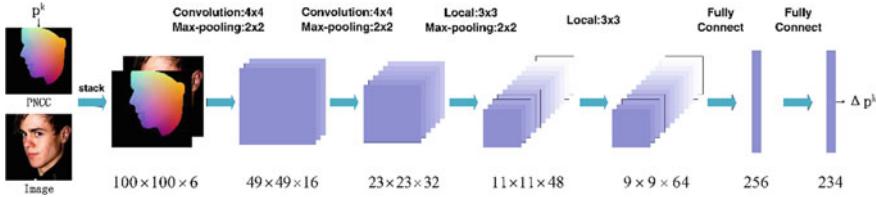
Supervision	Pretraining	Method	Classification (accuracy)	
			ModelNet40 (%)	ModelNet10 (%)
Category Labels	ImageNet	MVCNN	90.10	
		MVCNN-MultiRes [62]	91.40	
	None	3D ShapeNets [28]	77.30	83.50
		DeepPano [19]	77.60	85.50
		VoxNet [63]	83.00	92.00
		ORION [64]		93.80
	Unsupervised	SPH [65]	68.20	79.80
		LFD [66]	75.50	79.90
		T-L Network [58]	74.40	–
		VConv-DAE [67]	75.50	80.50
		3D-GAN [61]	83.30	91.00

illumination occurs in the active ones and there is no control over the amount of light in the passive ones. The systems may employ single or multiple vantage points. For multiple vantage points, the process of triangulation for the extraction of depth information is utilized. This forms the fundamental core of techniques like the self-calibrating Structure-from-Motion [24] (SfM) methods. Different methods such as shape-from-silhouettes, shape-from-occlusions, and passive stereo can be employed depending on the geometry and material features of the object or scene.

Bregler et al. [24] address the major issue of 3D nonrigid shape model's recovery from image sequences. That is, given a video of a person talking, the aim is to estimate a 3D model of the lips, eyes, and face. Most of the previous techniques used the assumption of a rigid face. This paper proposes a new methodology taking into consideration nonrigidness of the data elements. Here, the 3D shape is considered as a combination of  $K$  basic shapes. For better illustration, a shaded smooth surface was fitted to the 3D shape points. A shortcoming obtained while considering this particular approach was that the model could not handle missing tracks. To overcome such shortcomings, a 3DDFA structure [68] was introduced. This framework was proposed to solve three major issues:

The traditional face models assume that all the landmarks like “eye corner”, “nose tip”, and “chin center” are available and therefore is not usable for profile images. This is due to the fact that profile images may not have all the specified landmarks. The framework was proposed to solve the following major issues. First, the appearance of the face varies more dramatically for poses ranging from frontal view to profile view. Second, labeling the landmarks in large pose images is challenging as the invisible landmarks have to be guessed.

Figure 18 describes an overview of 3DDFA where a  $100 \times 100 \times 3$  color image stacked by PNCC is taken as input. The network is made up of four convolution



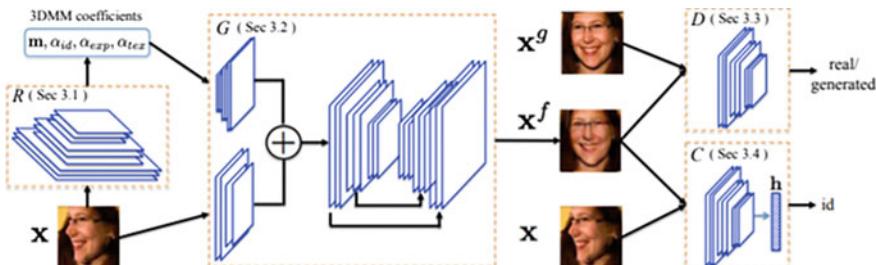
**Fig. 18** An overview of 3DDFA [68]

layers, three pooling layers, and two fully connected layers. With the help of Multi-Feature Framework (MFF) [69], a 3DMM is fit on the facial region. 3D meshing methods help in meshing beyond the face regions. 3D triangulation completes the process of 3D meshing [70]. When the depth information is estimated, the 3D model can be rotated in the 3D space to generate new instances from different viewpoints. The yaw angle of the depth image is increased from 5° to 90° [70].

A more recent development in this field of 2D to 3D visualization is the introduction of the FF-GAN framework as shown in Fig. 19.

Frontalization is the way toward delivering frontal confronting perspectives of faces showing up in single unconstrained photographs [71]. The process starts by recognizing faces seen from unconstrained perspectives to generate the recognized faces in constrained and front oriented postures. In [71], the model proposed by the authors consists of four major components, namely, a regeneration module for 3DMM coefficients, a generator  $G$ , a discriminator  $D$ , and a face recognition engine  $C$  that standardizes the generator output to preserve the various identity features. A frontal output is generated when an image which is non-frontal is given as input to  $G$ . The job of discriminator  $D$  is to try to classify the images as genuine frontal or a synthesized frontal one. Additionally, a face recognition engine  $C$  helps to regularize and keep tabs on the generator output to help preserve identity features. This step of regularizing helps to keep the generator in check.

Figure 20 shows (b) 3DMM estimation from (a) and the ideal threshold from [68]. The estimation for some images is better than the ground truth. FFGANs make use of 3DMM to serve as a reference for the frontalization process. The method proposed



**Fig. 19** Proposed FFGANs model [71]



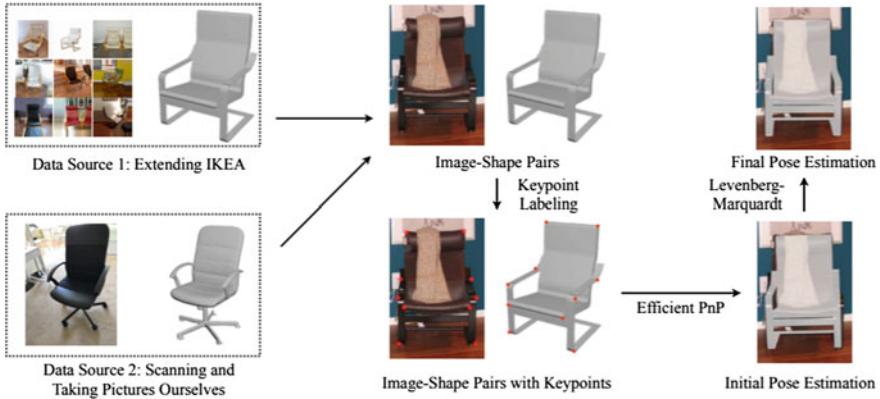
**Fig. 20** Comparison with ground rule [71]

by the authors locates and localizes key points accurately and then helps generate authentic frontal faces even for furthermore angle viewpoints and extreme profile inputs. The discriminator  $D$  treats every pixel equitably that results in loss of critical power for the identity attributes. As a result, the utilization of recognition module  $C$  (Fig. 20) to convey correct identity to the produced images becomes necessary.

### 3 Analysis of Datasets

Since the introduction of Augmented Reality, Virtual Reality as well as handheld console devices like Microsoft Kinect, there has been a significant rise in the creation of datasets mainly for the purpose of object recreation, object pose estimation, and faces. In [72], a description of current datasets as well as future scope of RGB-D datasets has been reported. With the introduction of 3D-GANs, introduction of good quality 2D to 3D datasets has become a field of primary importance in the field of deep learning. Multiple clearly visible views of the same object from different angles are considered in the RGB-D dataset. The 2011 RGB-D dataset [72] is a dataset of 300 objects. The paper [59] provides a new dataset, IKEA, consisting of fine-aligned objects with the perfectly matched 3D models. The drawbacks of these datasets were rectified by recent datasets like BigBIRD [49]. The Princeton Shape Benchmark contains a database of 3D models and software tools needed for their analysis. Version 1 of this dataset contains 1814 models. The main aim of ImageNet dataset [1] was to create a large-scale network of images. The WordNet acts as a base for the ImageNet dataset. ImageNet targets to contain around 50 million cleanly labeled high-resolution images (500–1000 per synset). ImageNet contains more than 12 subtrees.

The PASCAL 3D+ dataset [73] is a comprehensive dataset, containing many data elements from ImageNet. PASCAL 3D+ exhibits much more variability compared to existing 3D datasets. PASCAL 3D+ Dataset contains the different viewpoints of a particular data element in the dataset. Existing datasets like ShapeNet, ImageNet, and PASCAL 3D+ have some drawbacks like lack of precise alignment between 2D images and 3D shapes or they contain only synthetic data. As a result, a new dataset Pix3D [74] was introduced by authors Xingyuan Sun et al. to overcome these drawbacks. Pix3D has 395 3D shapes in nine different object categories. Figure 21 explains the construction steps needed to create the Pix3D dataset.



**Fig. 21** Construction of Pix3D dataset [74]

Scene understanding is a basic task for numerous uses of computer vision, scene modeling as well as 2D to 3D visualization. Some drawbacks of present RGB-D datasets are a smaller number of images, restricted scene description as well as blurred imagery. The Matterport 3D dataset [75] overcomes these drawbacks. The dataset comprises a collection of 194,400 RGB-D images obtained in 10,800 panorama cameras in home environments. The paper [9] explains in depth the modifications that were done to current datasets for 3D face regeneration. The dataset 300 W-LP consists of 122,450 images that are extracted from 300 W [76, 77] using the face profiling technique. The Multi-PIE dataset contains 754,200 images from 337 subjects having big differences in pose, lighting, as well as expression. Apart from these, some other important 3D datasets include SHREC 2014 [78], ModelNet [32], and ShapeNet.

## 4 Conclusion and Future Scope

In this paper, an extensive survey of the different approaches undertaken for 2D to 3D reconstruction models has been put forth. Along with this, an in-depth review of why GANs will help in such reconstruction techniques has been discussed. While GANs are slightly difficult to optimize due to unstable training dynamics, the following advantages supporting the framework of GANs have been identified. First, GANs currently generate the sharpest images compared to contemporary neural frameworks. Second, GANs are comparatively easy to train, as no statistical inference is required, and only back-propagation is needed to obtain the gradients. Introduction of new GANs frameworks like 3D-GANs and FFGANs has given a new insight into the field of 2D to 3D visualizations and reconstructions. Optimized 3D face visualizations are obtained using such advanced frameworks.

In today's era, we are moving toward smart classrooms and digital education environments. Using advanced GANs, 3D visualizations of complex chemical compounds will transform the way students learn chemistry. One more possible application lies in the field of medicine. Relative depth in complex microscopic organisms can be perceived, making the study of living characteristics, detection of harmful microorganisms more feasible and easier as compared to current techniques. Traditional scene recreation and localization techniques include physical representations and sketches. Using such advanced GANs architectures, automatic drone-based scene creation would be possible from simple 2D snapshots of the particular elements.

## References

1. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211–252.
2. Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484–489. <https://doi.org/10.1038/nature16961>.
3. Noam, B., & Sandholm, T. (2017). *Safe and nested subgame solving for imperfect-information games*. NIPS.
4. Ng, A. Y., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems* (Vol. 2).
5. Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–1554. <https://doi.org/10.1162/neco.2006.18.7.1527>. URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6796673&isnumber=6795851>.
6. Kingma, D. P. (2014). *Stochastic gradient VB and the variational auto-encoder*.
7. Manisha, P., & Gujar, S. (2018). Generative adversarial networks (GANs): What it can generate and what it cannot? CoRR abs/1804.00140: n. pag.
8. Goodfellow, I., Jean, P.-A., Mehdi, M., Bing, X., David, W.-F., Sherjil, O., et al. (2014). Generative adversarial networks. In *Advances in neural information processing systems* (Vol. 3).
9. Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR, abs/1511.06434.
10. The CIFAR-10 dataset. Retrieved from: <https://www.cs.toronto.edu/~kriz/cifar.html>, on September 30, 2018.
11. Adversarially Learned Inference—Scientific Figure on ResearchGate. Available from: [https://www.researchgate.net/Samples-and-reconstructions-on-the-SVHN-dataset-For-the-reconstructions-odd-columns-are\\_fig2\\_303755744](https://www.researchgate.net/Samples-and-reconstructions-on-the-SVHN-dataset-For-the-reconstructions-odd-columns-are_fig2_303755744). Accessed September 30, 2018.
12. Introduction to GANs, retrieved: From <https://medium.com/ai-society/gans-from-scratch>, on September 24, 2018.
13. Carlson, W. E. (1982). An algorithm and data structure for 3D object synthesis using surface patch intersections. In *SIGGRAPH*.
14. Tangelder, J. W. H., & Veltkamp, R. C. (2008). A survey of content based 3D shape retrieval methods. *Multimedia Tools and Applications*, 39(3), 441–471.
15. Van Kaick, O., Zhang, H., Hamarneh, G., & Cohen-Or, D. (2011). *A survey on shape correspondence*. CGF.
16. Li, Y., Su, H., Qi, C. R., Fish, N., Cohen-Or, D., & Guibas, L. J. (2015). Joint embeddings of shapes and images via cnn image purification. *ACM TOG*, 34(6), 234.
17. Su, H., Qi, C. R., Li, Y., & Guibas, L. (2015). Render for CNN: Viewpoint estimation in images using CNNS trained with rendered 3D model views. In *ICCV*.

18. Girdhar, R., Fouhey, D. F., Rodriguez, M., & Gupta, A. (2016). Learning a predictable and generative vector representation for objects. In *ECCV*.
19. Shi, B., Bai, S., Zhou, Z., & Bai, X. (2015). Deeppano: Deep panoramic representation for 3-d shape recognition. *IEEE SPL*, 22(12), 2339–2343.
20. Choy, C. B., et al. (2016). 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*.
21. Dosovitskiy, A., et al. (2017). Learning to generate chairs, tables and cars with convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 692–705.
22. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
23. Kar, A., et al. (2015). Category-specific object reconstruction from a single image. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1966–1974).
24. Bregler, C., et al. (2000). Recovering non-rigid 3D shape from image streams. In *CVPR*.
25. Everingham, M., et al. (2014). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111, 98–136.
26. Bourdev, L. D., et al. (2010). Detecting people using mutually consistent Poselet activations. In *ECCV*.
27. Yu, X., Roozbeh, M., & Silvio, S. (2014). *Beyond PASCAL: A benchmark for 3D object detection in the wild* (pp. 75–82). <https://doi.org/10.1109/wacv.2014.6836101>.
28. Yang, J., et al. (2015). *Weakly-supervised disentangling with recurrent transformations for 3D view synthesis*. NIPS.
29. Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-PIE. *Image and Vision Computing*, 28(5), 807–813.
30. Zhou, T., et al. (2016). View synthesis by appearance flow. In *ECCV*.
31. Tatarchenko, M., et al. (2016). Multi-view 3D models from single images with a convolutional network. In *ECCV*.
32. Wu, Z., et al. (2015). 3D ShapeNets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1912–1920).
33. Kulkarni, T. D., et al. (2015). *Deep convolutional inverse graphics network*. NIPS.
34. Kitani, K. (2016). *Learning a predictable and generative vector representation for objects*.
35. Qi, C. R., et al. (2016). Volumetric and multi-view CNNs for object classification on 3D data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp 5648–5656).
36. Fan, H., et al. (2017). A point set generation network for 3D object reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2463–2471.
37. Qi, C. R., et al. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 77–85).
38. Guerrero, P., et al. (2018). *Learning local shape properties from raw point clouds*.
39. Umetani, N. (2017). *Exploring generative 3D shapes using autoencoder networks*. SIGGRAPH Asia Technical Briefs.
40. Kong, C., Lin, C.-H., & Lucey, S. (2017). Using locally corresponding CAD models for dense 3D reconstructions from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
41. Pontes, J. K., Kong, C., Eriksson, A. P., Fookes, C., Sridharan, S., & Lucey, S. (2017). Compact model representation for 3D reconstruction. In *2017 International Conference on 3D Vision (3DV)* (pp. 88–96).
42. Sun, Y., Liu, Z., Wang, Y., & Sarma, S. E. (2018). *Im2Avatar: Colorful 3D Reconstruction from a single image*. CoRR, abs/1804.06375.
43. Liu, J., Yu, F., & Funkhouser, T. A. (2017). Interactive 3D modeling with a generative adversarial network. In *2017 International Conference on 3D Vision (3DV)* (pp. 126–134).
44. Chang, A. X., Funkhouser, T. A., Guibas, L. J., Hanrahan, P., Huang, Q., Li, Z., et al. (2015). *Shapenet: An information-rich 3D model repository*. CoRR, abs/1512.03012.

45. Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., et al. (2014). Large scale comprehensive 3D shape retrieval. In *Proceedings of the 7th Eurographics Workshop on 3D Object Retrieval, 3DOR'15* (pp. 131–140). Aire-la-Ville, Switzerland, Switzerland: Eurographics Association.
46. Lee, J. (2014). *Yobi3d*.
47. Park, E., et al. (2017). Transformation-grounded image generation network for novel 3D view synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
48. Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. CoRR abs/1409.1556: n. pag.
49. Singh, A., Sha, J., Narayan, K., Achim, T., & Abbeel, P. (2014). BigBIRD: A large-scale 3D database of object instances. In *International Conference on Robotics and Automation (ICRA)*. <http://rll.berkeley.edu/bigbird/>.
50. Salimans, T., et al. (2016). *Improved techniques for training GANs*. NIPS.
51. Visual Turing Test. Retrieved from <http://visualturingtest.org/>, on September 30, 2018.
52. Revinskaia, A., & Feng, Y. *From 2D Sketch to 3D shading and multi-view images*. Stanford University.
53. Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., & Belongie, S. (2016). *Stacked generative adversarial networks*.
54. Gadelha, M., et al. (2017). *3D shape induction from 2D views of multiple objects*. In *2017 International Conference on 3D Vision (3DV)* (pp. 402–411).
55. Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *SIGGRAPH*.
56. Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 150–162.
57. Wu, J., et al. (2016). *Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling*. NIPS .
58. Maas, A. L. (2013). *Rectifier nonlinearities improve neural network acoustic models*.
59. Lim, J. J., Pirsiavash, H., & Torralba, A. (2013). Parsing IKEA objects: Fine pose estimation. *IEEE International Conference on Computer Vision, 2013*, 2992–2999.
60. Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, pp. 3485–3492. <https://doi.org/10.1109/CVPR.2010.5539970>.
61. Wu, J., Xue, T., Lim, J. J., Tian, Y., Tenenbaum, J. B., Torralba, A., et al. (2016). Single image 3D interpreter network. In *ECCV*.
62. Qi, C. R., Su, H., Niessner, M., Dai, A., Yan, M., & Guibas, L. J. (2016). Volumetric and multi-view CNNs for object classification on 3D data. In *CVPR*.
63. Maturana, D., & Scherer, S. (2015). Voxnet: A 3D convolutional neural network for real-time object recognition. In *IROS*.
64. Sedaghat, N., Zolfaghari, M., Amiri, E., & Brox, T. (2016). Orientation-boosted Voxel Nets for 3D Object Recognition. [arXiv:1604.03351](https://arxiv.org/abs/1604.03351).
65. Kazhdan, M.M., Funkhouser, T.A., & Rusinkiewicz, S. (2003). Rotation invariant spherical harmonic representation of 3D shape descriptors. *Symposium on Geometry Processing*.
66. Chen, D., Tian, X., Shen, E. Y., & Ouhyoung, M. (2003). On visual similarity based 3D Model Retrieval. *Comput. Graph. Forum*, 22, 223–232.
67. Sharma, A., Grau, O., & Fritz, M. (2016). Vconv-dae: Deep volumetric shape learning without object labels. arXiv preprint, [arXiv:1604.03755](https://arxiv.org/abs/1604.03755).
68. Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016). Face alignment across large poses: A 3D solution. In *CVPR*.
69. Blanz, V., & Vetter, T. (1999). A morphable model for the synthesize of 3D faces. In *SIGGRAPH*.
70. Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 787–796).
71. Yin, X., Yu, X., Sohn, K., Liu, X., & Chandraker, M. K. (2017). Towards large-pose face frontalization in the wild. *IEEE International Conference on Computer Vision (ICCV), 2017*, 4010–4019.

72. Firman, M. (2016). *RGBD datasets: past, present and future* (pp. 661–673). <https://doi.org/10.1109/cvprw.2016.88>.
73. Kar, A., Tulsiani, S., Carreira, J., & Malik, J. (2015). Category-specific object reconstruction from a single image. In *CVPR*.
74. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., et al. (2018). Pix3D: *Dataset and methods for single-image 3D Shape modeling*.
75. Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., et al. (2017). *Matterport3D: Learning from RGB-D data in indoor environments*.
76. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing (IMAVIS)*, Special issue on facial landmark localisation. In *In-the-wild*.
77. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of IEEE International Conference on Computer Vision (ICCV-W), 300 Faces in-the-Wild Challenge (300-W)*. Sydney, Australia. December, 2013.
78. Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., et al. (2014). Large scale comprehensive 3D shape retrieval. In *Proceedings of the 7th Eurographics Workshop on 3D Object Retrieval, 3DOR'15* (pp. 131–140). Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

# Impact of Artificial Intelligence on Human Resources



**Sapna Khatri, Devendra Kumar Pandey, Daniel Penkar  
and Jaiprakash Ramani**

**Abstract** In the age of technology advancement and development, the recent and latest in-technology is artificial intelligence known as (AI). AI is an advanced level of technology, developed with the intention of economic growth, high productivity and to help humans get over their repetitive task. AI is based on Big Data, and a set of algorithms sense, study, analyse and perform the task, as a human would normally do. Artificial intelligence is a buzzword and everywhere there is the talk of AI; however like every new technology, AI also comes with its pros and cons. The challenge is of its usage, implementation and its impact on human resources to survive and sustain in the competitive world. Artificial intelligence is an intrinsic part of the Industrial Revolution 4.0. Every revolution comes with the demand of major change in the existing system and environment. Until it settles, all the aspects of new technology with the required setup and outcome with reference to the willingness of an employee to learn and adopt it remain intriguing. How are human resources ready to adopt this disruptive technology and its usage? How is the readiness of employers to implement AI technology? The interplay of both these questions is crucial considering the overall management of human resources and organization. On one side, AI requires specialized technical knowledge to develop and operate it, which is a clear indicator of increasing the technical employment. However, this very requirement possess a huge challenge for skill upgradation, employability of middle management, older

---

S. Khatri (✉)  
Amity University, Gwalior, Madhya Pradesh, India  
e-mail: [ramani.sapna@gmail.com](mailto:ramani.sapna@gmail.com)

D. K. Pandey  
Amity Business School, Amity University, Gwalior, Madhya Pradesh, India  
e-mail: [dkpandey@gwa.amity.edu](mailto:dkpandey@gwa.amity.edu)

D. Penkar  
SB Patil Institute of Management, Savitribai Phule  
Pune University, Pune, Maharashtra, India  
e-mail: [drdanielpenkar@rediffmail.com](mailto:drdanielpenkar@rediffmail.com)

J. Ramani  
Faurecia Interior Systems, Pune, India  
e-mail: [jaiprakashramani@hotmail.com](mailto:jaiprakashramani@hotmail.com)

employees and all human resources of the organization. This paper focuses on the infusion of artificial intelligence-based systems in an organization and the emerging challenges and opportunities in human resources management considering both technical and nontechnical resources of the organizations.

**Keywords** Artificial intelligence · Competitive age · Employee and employment · Niche skills · Human resources · Human interface

## 1 Introduction

Artificial intelligence (AI), also referred to as machine intelligence, is the area of science and technology, which is growing dramatically and rapidly in the current age of technological advancement and globalization. It would impact the organizations across the spectrum and has already started to run deep into organization structures in certain countries. Obviously, being a key disruption in modern times, besides everything else, it also would enter the domain of human resources of the organization, rather it already has initiated the baby-steps. Artificial intelligence is challenging human resources by threatening to replace them in routine jobs and cognitive tasks, pushing them to develop a bigger appetite to acquire newer skill. Indeed the potential of this threat is so significant that it might get tough, to maintain the energy levels of the organization, at all levels. Once the organizations begin to unveil AI based plans and strategies, there would be a palpable sense of desperation across the ranks and the imperative to retain the morale of the human resources, will be too high. Employee sustainability and growth planning would become crucial question marks. We, therefore, will deliberate on two important concepts, i.e. (i) artificial intelligence and (ii) human resources, which forms the basis for this paper. In the following paragraphs, the authors attempt to deal with both of these concepts in detail. Technology, innovation and human resources complement each other, being key ingredients to the recipe for organizational growth. This dissertation is an attempt to understand and analyse this very recipe.

### 1.1 Artificial Intelligence

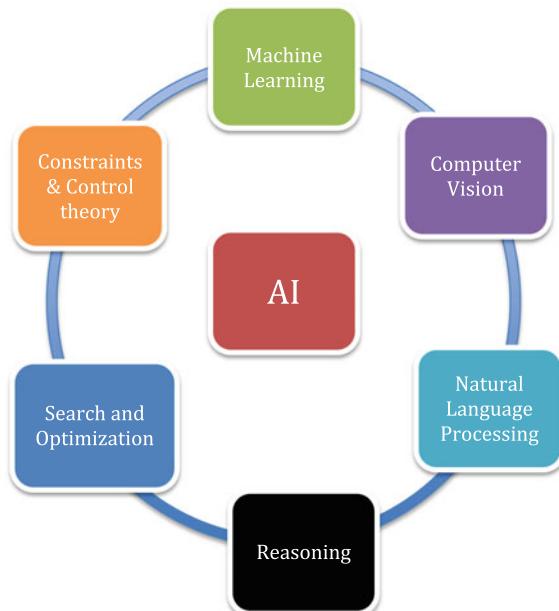
Artificial Intelligence is a specialized stream of computer science which enables or builds intelligence in the machines and systems, harnessing the available Big Data around. Humans had created computers, networks, Internet and cloud servers, enabling them to start collecting the huge transactional data. Unknowingly, we created an asset base, which till a few years back was not even known as asset class. AI leverages this asset class, with the help of complex algorithms, harnessing Big Data to create predictive solutions for human problems. All these solutions have always remained the exclusive domain of humankind, due to developed mind our

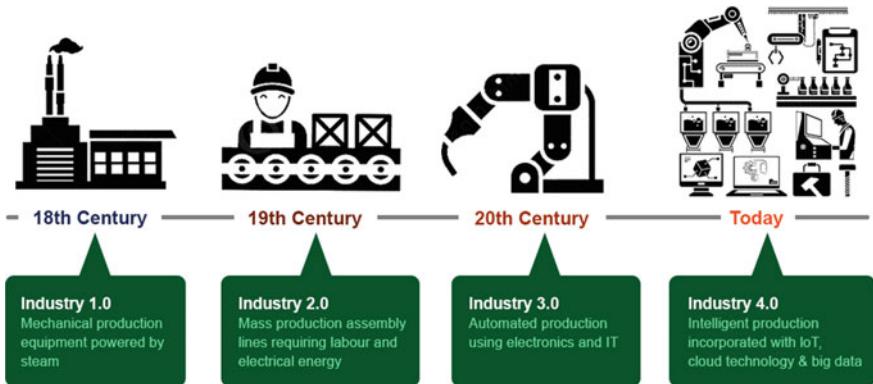
species possesses. AI is now attempting to provide the machines with traits similar to the human brain, so that they can become intelligent. We can see abundant examples scattered all around us, from digital assistants to robots, voice recognition to digital customer care agents. Machines are programmed to demonstrate human-like skills such as reasoning, knowledge, problem-solving, learning, perception, planning, manipulation, etc., by using algorithms and programming. AI is indeed an ensemble of many technologies, viz. machine learning, computer vision, etc., as illustrated in Fig. 1.

AI is the fourth stage of automation in technological development. It is an established fact that the third stage of technology development, i.e. automation has already contributed to job losses of the workers at the shop floor in the manufacturing industry. AI aims to develop cognitive skills and power, to perform the human exclusive tasks. AI already has their presence in voice assistance, face recognition, science fiction games and is faster than the humans in things like using apps, sensing the environment and act accordingly. Few examples of AI, which is forcing us to think about future employment challenges, are self-driven cars, cashier and salesman free stores, etc. In fact, AI has its application in almost all areas such as healthcare, automotive, finance, economics, infrastructure, manufacturing, etc.

Industrial Revolution 4.0 has heralded itself silently—moving through the stages of mechanized production (1.0), mass production (2.0) and automation (3.0)—connectivity (4.0) will drive the future, where prediction will be the force behind analytics and machine intelligence (refer to Fig. 2).

**Fig. 1** AI topography





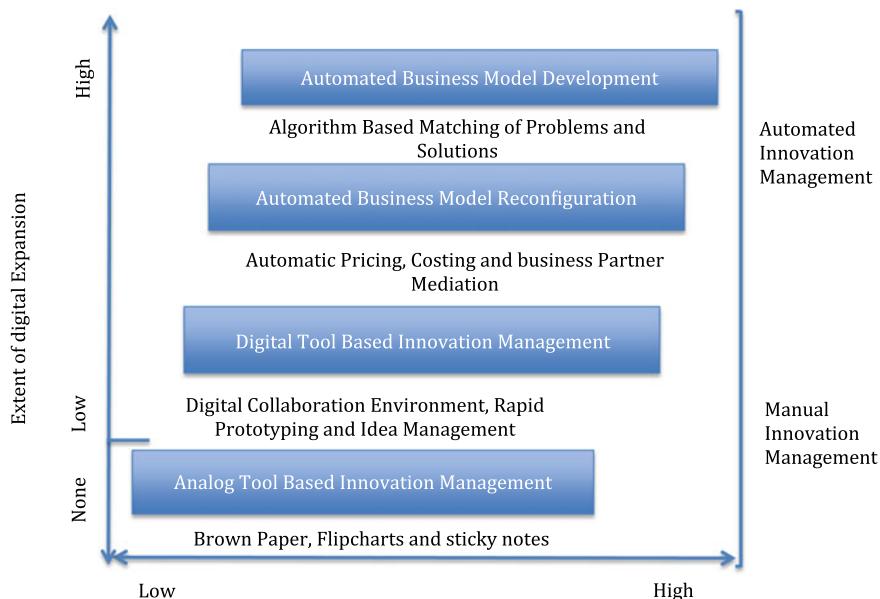
**Fig. 2** Industrial evolution

## 1.2 Human Resources

Human resources are the blood flow of the organization. Managing human resources and industry is an integral part of the economy and technological development. It relates to the life cycle of employment from recruitment to retirement of the employee. It demands continuous learning, creativity and innovation. Human resources deal with the product life cycle—its design, manufacturing planning, process outcome, service, distribution and reclamation. It is a combination and coordination of product, supply, services and human resources, which yields desired economic results.

Human resources management can be measured qualitatively and quantitatively considering the vision, mission, planning, strategy, policy, culture, structure and communication of an organization. But most important is the adoption of technological advancement for the betterment of employees, their motivation and retention. AI is the best example of science and technology aiding the employee productivity and human resources management in a non-intrusive way. Figure 3 shows the degree of impact of innovation and human resources management with digital penetration, which clearly indicates the arrival of the era of automated business model.

AI is impacting and challenging humans for some of its very important characteristics, viz. trust building, interpersonal skills and interface, huge investment and higher productivity and returns. Simultaneously, it is demanding the upgradation of skills of human resources to cope up with the changing technology. AI throws many questions and creates dilemma on human resources in the organization about; what may or would be the benefits, repercussions/consequences and impact of AI technology on human resources at all level in the organization.



**Fig. 3** Extent of AI on innovation management

### 1.3 Impact of AI on HR

The open questions and problem statements to analyse the interdependence or rather the impact of AI on human resources are as follows:

- Will artificial intelligence replace human resources?
- Is artificial intelligence creating job opportunities?
- Can AI and human resources be complementary to each other?
- Will AI create demand for human resources the newer and niche skills?
- Are organizations ready to face the threats and challenges of AI?
- What kind of strategies would have to be adopted by organizations to retain, motivate and develop their human resources?

## 2 Literature Review

Artificial intelligence is relatively new on the horizon. However, several authors have gone on to write odes about artificial intelligence and its relevance.

**Bloomberg News, The Times of India (23 September 2018)** outlines the changes face of the Bowery Farming workspace in New Jersey in US. *Morich and her fellow human farmers do what computer tells them to do. Bowery says machines keep*

*learning how to grow crop more effectively and are more than a match for the intuition of a seasoned farmer.* The article throws up the important point of intelligent machines taking over from humans and replacing them eventually [1].

**Crystal Miller Ray**, CEO, Branded Technologies in her article titled **Lions, Tigers & AI Oh My! Exploring HR Tech Fairy tales (October 2016)**, exalts about her experiences at **2016 h Technology Conference at Chicago**, ‘like I was listening to an HR Tech version of a fairy-tale.... none more so, perhaps, than A.I.’ In her article, she goes on to explain the various *AI concepts and argues about the noise around the AI technology being promoted or sold to the HR community* [2].

**Dupress.com** in its forecast of **‘Future of Artificial Intelligence—Government 2020’** cites **Bloomberg study**, which states ‘*by—2014–2024, mobile robots and artificial intelligence make it likely that occupations employing about half of today’s US workers could be automated to some degree*’. This article further goes on to stress that A.I. would allow the machines to almost mimic a process of human thinking, ‘*These are intelligent systems which improve the predictive accuracy, speed up the problem solving and administrative function bringing in an age of automation*’. The article further goes on to emphasize what would be key areas of AI development by 2020, including Cognitive Analytics, Deep Learning and Intelligent Automation [3].

**Andrew Ng**, VP & Chief Scientist of Baidu (Chinese equivalent of Google), an Adjunct Professor at Stanford University, in his critical analytics published in **Harvard Business Review (November 2016)** titled **What Artificial Intelligence Can and Can’t Do Right Now**, has had a very different take from other scholars. Andrew, who has been a founder lead of Google Brain Team and has been a former Director at Stanford Artificial Intelligence Laboratory, claims that *media sometimes paints very fancy picture of AI sometimes, as if it is going to take over the world very soon*. He feels that despite the breadth of the impact of AI, its deployment is very limited [4].

**Matthias Breunig, Matthias Kässer, Heinz Klein, and Jan Paul Stein (January 2016)**, in their study for **Mckinsey & Company**, titled **‘Disruptive trends that will transform the auto industry’** (Article published under title ‘Building smarter cars with smarter factories: How AI will change the auto business’ on McKinsey.com) clearly opine that ‘*Over the coming 20 years, artificial intelligence (AI) will enable autonomous vehicles to become mainstream. At the same time, most of the aspects of the auto-manufacturing process will be transformed by AI, which include research, designing, business support functions and project management. These changes are constantly approaching. Industrialists should understand the foundations and causes of value really are and then start developing the essential analytical competences and inaugurating an AI-ready culture*’. They further move to identify six key ways in which AI will improve automotive industry, viz. less equipment failure, more productive employees through robot–human collaboration, fewer quality problems, leaner supply chains, smarter project management and improved business support functions [5].

**Daniel Faggella (2016, March 21)**, in his contribution, titled **‘Exploring the risks of artificial intelligence’** highlights the flip side as well. He writes, ‘*Dives in AI are previously being made in the area of workstations automation and machine*

*learning aptitudes are speedily increasing to our energy and other business purposes and functions, including mobile and automotive. The next industrial revolution may be the last one that humans guide in by their own direct doing, with AI as a future collaborator and – dare we say – a potential leader’ [6].*

### **3 Methodology**

Descriptive qualitative review research methodology is used for this research paper. The paper is based on review of secondary data through articles, periodical, books and journals—both online and printed, as well as the experience, learning and observations of the researchers.

#### ***3.1 Scope of the Research***

In general, all profit-driven Human Capital Organizations can use the study analysis and conclusion of this research paper. The outcome of the study will push human resources in learning and understanding the technical advancement. It helps in creating a conducive environment in the professional surroundings. The study can also help in boosting the employees’ morale and high productivity, so that both organization and individual would be able to drive maximum advantage in terms of human resources management and required skill sets development. Tech giants are setting up research laboratories for the AI. Universities and government are also taking keen and focused initiative and investing in AI labs, putting efforts in understanding its impact on human resources management and organizational growth. Several research papers have been published on AI and its use as a tool of successful development of human resources and adoption of technological advancement.

#### ***3.2 Objectives of the Research***

- To study the readiness of the organization to adopt the AI technology.
- To study the strategies of the organization to keep the morale high of its human resources, whilst adopting new technology.
- To study the willingness of human resources to accept the changing technological environment.

## 4 Significance of the Research

The aim of artificial intelligence is to create an accurate and error-free execution; cost and time saving; increase in productivity and seamlessly perform the job as any human would do, i.e. intelligently. AI is assisting human resources in their repetitive or risky jobs, providing the platform and opportunity to perform, and use their skills in more challenging judgement-oriented tasks rather than engaged in a routine job. The composition of AI is algorithm, Big Data and machine learning through which AI will perform and reduce the human interface in mundane or risk-oriented activities. However, it is very important to know whether it is reducing the human interface, making them redundant OR really assisting human resources in reducing their efforts, thus driving them to understand and learn niche skills [7].

## 5 Expected Outcome

Researcher wishes to establish the correlation between the artificial intelligence and challenges/opportunities presented by AI on human resources. Focus is on how organizations are ready to accept the technological advancement for the betterment of organization as a whole. This would include skill development, improving employees' efficiency, operational productivity, ensuring organizational growth and sustainability. Also, researchers aim to understand the degree of impact of this innovation on human resources.

Researcher wishes to use the outcome of the research to further strengthen the field of organization behaviour, industrial relations and human resources management in the age of AI. This could well be a very potent field of study and advancements in the future human-machine interactions, human motivation, human behaviour, human satisfaction and overall human resources management in the digitized world.

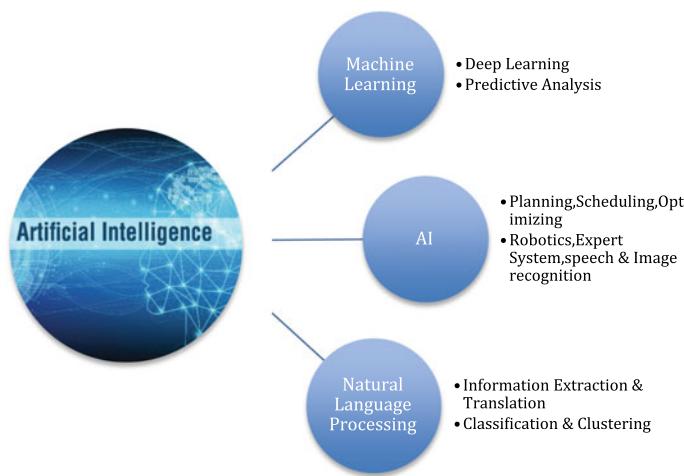
### 5.1 *Proposed Model*

Considering the research idea and expected outcome, a model has been proposed to get the desired collaboration between technology and human resources, so that win-win situation can be created and dilemma can be reduced, if not eliminated.

Researchers have attempted to build two models for interlaying the usage of technology with human resources management. Developing the skills of human resources in alignment to AI is the need of the hour. The first model shows the progress of machine self-learning; something human used to do and perform by using cognitive skills. The second model shows the overall human resource management strategies, which consists of knowledge enrichment, competitive environment, training and niche skill development, motivation and retention through reskilling, and assistance

to adoption of newer technology. The second model is directly impacted by the first model of technological advancement and has a relationship with the human resources, innovation management, organization sustainability and economic growth (Figs. 4, 5 and 6).

From the above proposed model of alignment between artificial intelligence and human resources management, it is can be understood that considering the existing practices of the organization, it is not easy to adopt and bring about the change without creating awareness among employees and maintaining the conducive environment in the organization. Creating awareness about the pros and cons of technology,

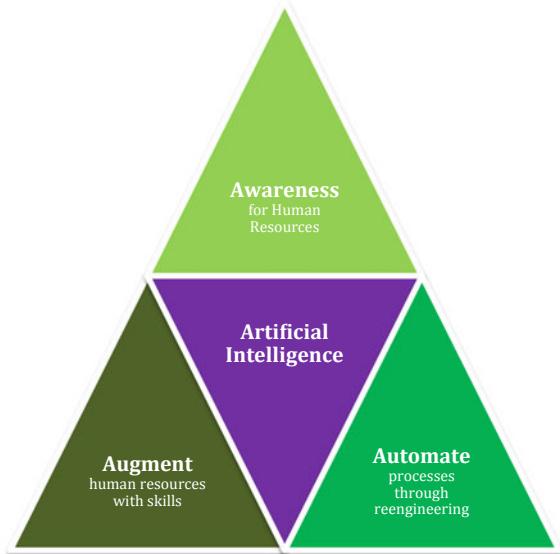


**Fig. 4** Artificial intelligence zones



**Fig. 5** Human resources management strategies

**Fig. 6** Interlay between artificial intelligence and human resources management



augmenting skill sets to keep their morale high and automating their performance through reengineering are the best ways to assimilate AI within the organization.

## 5.2 Discussions and Analysis

Let us understand and dissect the dimensions of impact of AI on human resources through a very recent example. CARLSBERG, the famous Danish brewer, has been working with Microsoft and a couple of universities in developing an artificial intelligence-based process of develop new beers and improve the quality control. Traditionally, the beer development process is tedious and involves human tasting and sense for aromas and flavours to bring newer products to market. This new AI-based approach, where it is investing millions of dollars, would help create the new beers using advanced sensors and Big Data. It will also help reduce the time to market. Broadly speaking, the success of this endeavour will not only revolutionize Carlsberg but also the whole food, pharmaceutical and other manufacturing organization. The unique dimension provided by human senses will probably be replicated scientifically, bringing far more accuracy. This is surely going to disrupt the human dynamics within the industry, causing a potential displacement of humans. Such future workplace would mean human resource professionals would have to work on creating new chemistry in the organizational culture, where both human intelligence and artificial intelligence, not only coexist but also compete and collaborate. However, isn't it predictable that such a transition would mean human resources would

become more insecure? Won't the workplace start facing newer leadership and power battles? Won't the newer political dynamics emerge? [8].

Andrew Ng, VP & Chief Scientist of Baidu (Chinese equivalent of Google), an Adjunct Professor at Stanford University, in his critical analytics published in Harvard Business Review (November 2016) titled 'What Artificial Intelligence Can and Can't Do Right Now', urges the economists and business leaders to be careful and ensure to understand AI well before making it a part of their strategy, in a manner that it benefits all, provides opportunity for everyone to thrive. Undoubtedly, he opens up the debate on the key area related to AI, which stress on how will it impact the HR and its processes and what kind of strategy the CHROs will have to develop in near future? [4].

This churns out an area of research that what kind of strategies will be required by the employers and what kind of skills will be required by human resources. Overall human resources management has to an alignment between organization and its human resources chartering growth path, with the adoption of technological artificial intelligence and working together with it. It is all about change and desire to learn, develop and harness the niche skill sets for the workspaces of tomorrow. The researchers also wish to draw attention towards a very critical point—technology and human are complementary to each other, as proposed in the interlay model no. 3.

## 6 Conclusion

Artificial intelligence is an area of research to develop learnt systems, which execute functions that usually needs human cognitive intelligence. Industrial AI is more concerned with the application of technologies to tackle industrial pain-points for value creation of customer and productivity improvement. However, the technology never creates any business value alone if the problems in the industry are not well understood and studied. Human and technology are complementary to each other in operating and progressing the organizational growth. The only requirement is for humans to upgrade their existing skills and demonstrate keenness to learn newer ones, viz. knowledge-based techniques to cope and compete with AI. Artificial intelligence promises to be the biggest technological shift in our lifetimes. Every industry will have to fundamentally reassess how it operates in order to incorporate AI and coexist with machines that will become invaluable partners in solving real problems. So we can say that aims and objectives of the organization should be to align with the change and its adoption in a structured way, considering the peripherals, and make employees ready through the required training, reskilling and redeployment. The skills need to be updated and upgraded, as they are also perishable if not mapped with technology. That is the key to successful human resources management.

In the above discussion, AI has thrown few questions and here we can understand through Carlsberg example that future belongs not only to the organizations who adapt and invest in intelligent systems, but it will also belong to the industry which

prepares their human resources to productively harness the power of the Big Data and artificial intelligence for the competitive advantage. The history of the human race is a fitting example of human grit and instincts overcoming all the disruptions; Industry 4.0 will not be an aberration. Humans will prevail and remain at the roost.

## References

1. Bloomberg. (2014). *Your job taught to machines puts half U.S. work at risk*. Retrieved from <http://www.bloomberg.com/news/2014-03-12/yourjob-taught-to-machines-puts-half-u-s-work-at-risk.html>.
2. Lay, C. M., CEO Branded Strategies. (2016). *Lions, tigers & AI oh my! Exploring HR tech fairytales*. LinkedIn, Retrieved from <https://www.linkedin.com/pulse/lions-tigers-ai-oh-my-exploring-hr-tech-fairytales-crystal-miller?trk=prof-post>.
3. Dupress.com. (2014). *Future of artificial intelligence—Government 2020*. Retrieved from <http://government-2020.dupress.com/driver/artificial-intelligence/>.
4. Andrew Ng, V. P., & Chief Scientist of Baidu, Co-Chairman and Co-Founder of Coursera, an Adjunct Professor at Stanford University. (2016). *What artificial intelligence can and can't do right now*. Harvard Business Review, November, 2016. Retrieved from <https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now>.
5. Breunig, M., Kässer, M., Klein, H., & Stein, J. P. (2017). *Building smarter cars with smarter factories: How AI will change the auto business*. Digital McKinsey, October 2017. Retrieved from <https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/building-smarter-cars>.
6. Faggella, D. (2016). *Exploring the risks of artificial intelligence*, March 21, 2016. Retrieved from <https://techcrunch.com/2016/03/21/exploring-the-risks-of-artificial-intelligence/>.
7. Allegis Group. (2017). *AI and World of Work: Embracing the promises and realities*. A White Paper, Retrieved from [https://www.allegisgroup.com/insights/ai?ecid=ag\\_ag\\_gen\\_ai-2017\\_20170530\\_bad63f06](https://www.allegisgroup.com/insights/ai?ecid=ag_ag_gen_ai-2017_20170530_bad63f06).
8. Milne, R., Nordic Correspondent, Financial Times. (2017). *Carlsberg turns to AI to help develop beers*, December, 26 , 2017. Retrieved from: <https://www.ft.com/content/be042eb2-e4cf-11e7-97e2-916d4fbac0da>.

# Role of Activation Functions and Order of Input Sequences in Question Answering



B. S. Chenna Keshava, P. K. Sumukha, K. Chandrasekaran and D. Usha

**Abstract** This paper describes a solution for the Question Answering problem in Natural Language Processing using LSTMs. We perform an analysis on the effect of choice of activation functions in the final layer of LSTM cell on the accuracy. Facebook Research's bAbI dataset is used for our experiments. We also propose an alternative solution, which exploits the language structure and order of words in the English language, i.e. reversing the order of paragraph will introduce many short-term dependencies between the textual data and the initial tokens of a question. This method improves the accuracy in more than half of the tasks by more than 30% over the current state of the art. Our contributions in this paper are improving the accuracy of most of the Q&A tasks by reversing the order of words in the query and the story sections. Also, we have provided a comparison of different activation functions and their respective accuracies with respect to all the 20 different NLP tasks.

**Keywords** RNN—Recurrent neural networks · LSTM—Long-short term memory · QA—Question answer · NLP—Natural language processing · Seq2seq models—Sequence-to-sequence models

---

B. S. Chenna Keshava (✉) · P. K. Sumukha · K. Chandrasekaran

Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, Mangalore, India

e-mail: [16co108.keshava@nitk.edu.in](mailto:16co108.keshava@nitk.edu.in)

P. K. Sumukha

e-mail: [sumukhapk46@gmail.com](mailto:sumukhapk46@gmail.com)

K. Chandrasekaran

e-mail: [kchnitk@ieee.org](mailto:kchnitk@ieee.org)

D. Usha

St. Joseph Engineering College, Vamanjoor, Mangalore, India

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

## 1 Introduction

The Question Answering problem has been studied in Natural Language Processing since the early 1970s. Entity extraction and information retrieval have been one of the main areas of applications of Question Answering. The Question answering problem requires to keep track of a long sequence of texts that are all necessary for retrieving an Answer to the given Question. So in order to solve such a problem, we need a mechanism to keep track of relevant data in certain sequences that lead us to the answer of the query.

Section 2 contains the Literature Review. Section 3 contains a description of the problem statement. A brief description of the dataset is also provided. In the penultimate section, the solution approach is described. We conclude by providing experimental results.

## 2 Literature Review

**Question Answering (QA)** can be dealt with in two major ways. The first one involves semantically parsing the query and converting it to a database query on a knowledge base. The second method involves many intermediate steps. First, the question type is identified, and relevant documents are fetched. Then sentences that might contain the answer are chosen from this set of documents. Then finally, the answer is extracted out of these documents. Finally, the relevance of the answer is evaluated based on the semantic matching of the parse trees [1].

**Generating Synthetic Datasets for Supervised Learning in NLP:** [2] Using supervised learning for machine reading and understanding is a difficult task, because we do not have a large label dataset, and it is hard to develop flexible statistical models that exploit the structure of documents.

Researchers have explored creating synthetic narratives/query pairs. This method allows for almost unlimited amount of supervised training data. But such transitions have often been unsuccessful in the history of Computational Linguistics [2].

This paper aims to build new datasets to facilitate supervised learning in document reading and comprehension by using paraphrase and summary sentences of a document. We can convert it to a context-query-answer triple, and using entity detection and anonymization algorithms, obtain new datasets for machine reading and comprehension [2–4].

**Microsoft Research Question Answering System** The Microsoft Research question answering system presented in the Text REtrieval Conference 9 (TREC-9) was a modified version of Microsoft's Natural Language Processing system (NLPWin) and the Okapi retrieval engine.

The NLPWin analyses the question to produce a logical form of it, whilst obtaining a set of query terms. Most relevant words that will link to the answer will be extracted

here. These query terms will be used by the Okapi IR engine. This produces a list of documents based on the BM25 weighting scheme. These documents are converted into sentences by segmentation by NLPWin's linguistic analysis capabilities [5].

### 3 Problem Description

Question Answering is a field in the intersection of information retrieval and natural language processing. This is related to building systems that are capable of automatically answering questions which are in the natural language of humans. A QA implementation, usually a computer program, might be done by constructing a structured database and querying it for knowledge or information, usually a knowledge base. More commonly, QA systems can detect and produce answers from an unstructured collection of natural language documents.

The question answering problem is unique because the answers here must have a semantic meaning to them. The analysis on which the answer is based is a major issue to solve. The machine must learn how to figure out the context of the question and then provide an answer to the question.

There are multiple types of question answering problem. One of the categories is, **factoid and non-factoid question answering**. In Factoid QA, there exists a factual correct answer for every question. On the other hand, non-factoid QA is those which are subjective and could have multiple correct answers for a question. The other classification in QA is **open and closed question answering**. In closed question answering, the answer for all questions is present within a sentence in the given dataset. On the other hand, to answer open questions, the machine needs to develop a *common-sense intuition* about the workings of the world. The answers will not necessarily be embedded in any sentence in the dataset [6].

The database we are using is the bAbI dataset, which is available in languages English and Hindi. bAbI is a set of 20 tasks, where each consists of numerous context-question-answer triplets, generated, prepared and released by Facebook Artificial Intelligence Research (FAIR Lab). Every task in it aims to test a particular aspect of reasoning and is, therefore, aimed towards testing a specific capability of QA learning models. The bAbI is a closed dataset, i.e. all the queries related to the dataset has solutions or answers inside the same dataset. It is also advantageous as it is a synthetic dataset and doesn't include any real-life situations which are usually very noisy, and contain information that is not particularly relevant to our problem. The vocabulary is very limited and the sentence formation is constrained making it an ideal dataset to work on. By working on such synthetic dataset, we can develop different models and test the accuracy of those models before they can be deployed to actually process human speech filled with irrelevant noise.

## 4 Solution Approach

RNNs are a class of artificial neural networks where connections between different units form a directed graph along a sequence. Different variants of the basic RNN are the Gated Recurrent Units were introduced in 2014, and the LSTMs.

In this section, we first have touched upon the working of the Long Short-Term Memory Unit (LSTM). We have also provided a concise explanation of the activation functions that have been considered for comparison.

Figure 1 depicts the unfolding of the RNN layer across time steps.

### Solution

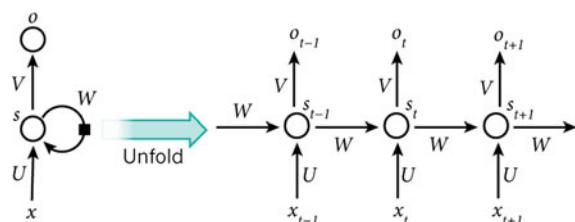
For every question posed at the algorithm, it must generate an answer from the dataset available. Since the answer has to be extracted from the dataset, we need to store the data in a special manner, so that whenever we provide a question to the algorithm it searches from the dataset, processes it based on that data and provides relevant data as an answer for the question. We store the question as with a start and an end token to signify the start and end of the question. The input to the RNN is always in a sequential manner.

We have used three different sets of RNNs for obtaining a solution for the Question Answering problem. The following are how each RNN is used:

- Input for the first RNN is the question for which the answer is to be found. This takes in the question with the tokens in its start and end and processes them into tokens. The sequential input is taken in and the RNN converts them into a set of tokens of the input question.
- The second RNN does the same job as the first RNN but this takes in the story or the dataset as the input and provides output as the tokens for the entire dataset.
- The third RNN is the one which provides the final output for the question. This takes in the tokens created from the previous two RNNs and combines them to create a vector. This vector is compared with the question and the answer is generated.

Specifically, we choose an answer sentence, from previously unseen candidate sentences also. Moreover, the number of candidate sentences may vary depending on the question. We model the problem of choosing an appropriate answer sentence as a binary classification problem. Every QA is posed in the form of ( $y = \text{Is the answer correct or not?}$ ). So, the task then becomes classifying if  $(q, a, y)$  is a correct tuple or not.

**Fig. 1** Recurrent neural network [7]



In order to verify the relevance of the answer, we generate a new question  $q'$ , from the obtained answer  $a$ . Then we perform a semantic similarity on the generated question  $q'$  and the original question  $q$ . This similarity score is used to judge the relevance of the answer  $a$ . This model can also be extended to tasks like Textual entailment, Paraphrase detection, etc.

A detailed explanation of a few variants has been proposed in some of the previous works like [8] (Chap. 28).

There are many different types of activation functions that can be used with LSTMs. These different activation functions operate on the outputs of the RNN in different ways and correspondingly provide different interpretations. We have used the following activation functions to interpret our data:

- **softmax**: The **softmax function**, squashes an  $N$ -dimensional vector  $z$  to an  $N$ -dimensional vector  $\sigma(z)$  that are in the range of 0–1 and add up to 1.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^N e^{z_k}} \quad (1)$$

Softmax function is used to convert values into action probabilities in many of the problems of Reinforcement Learning [9]. More recently, adaptive softmax has been developed to eliminate the problem of very large vocabularies. This softmax does not calculate the probabilities over the entire vocabulary. It uses an approximation to increase the speed of inference and training by a huge order of magnitude. This speedup is achieved at the cost of a small value of accuracy.

- **relu**: Stands for Rectified Linear Unit. It returns zero if the input is less than zero, else, returns the positive value [10].

$$f(x) = \max(x, 0) \quad (2)$$

Above is the description of the relu function.

- **tanh**: The hyperbolic tangent function rescales of the logistic sigmoid into the range from  $-1$  to  $1$  [11].

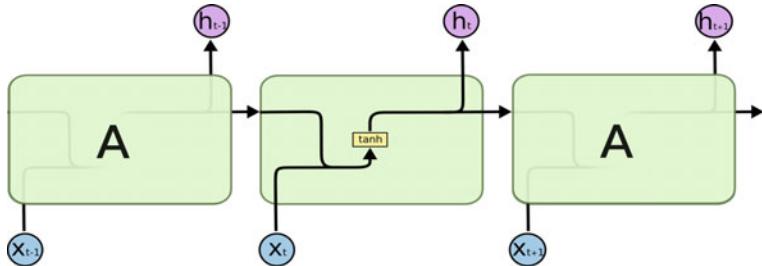
To overcome the vanishing gradient problem, we adopt other methods like elu, relu and selu so that we get a definite answer mapped to every query given to the neural network.

- **elu**: Exponential Linear Unit function checks the vanishing gradient problem. The other mentioned *activation functions* are prone to reach a point from where the gradient of the *activation functions* does not change at all.

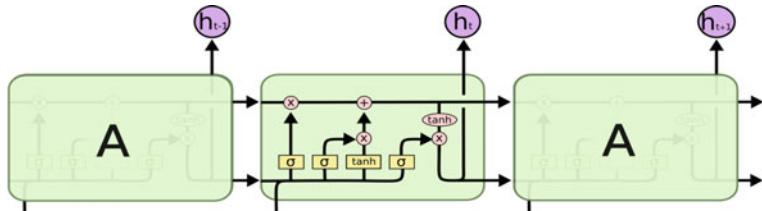
Above is the description of the selu function (Figs. 2 and 3).

- **selu**

$$\text{selu}(x) = \lambda \begin{cases} x, & x > 0 \\ \alpha e^x - \alpha, & x \leq 0 \end{cases} \quad (3)$$



**Fig. 2** Recurrent neural network [7]



**Fig. 3** LSTM [12]

## 5 Dataset Description

### *The bAbI Dataset*

The bAbI dataset was created to cater to the needs of the project that aimed at automatic question answering. This is a synthetically generated dataset which aims to provide a kickstart to the basic testing and running of different algorithms in the domain of automatic text understanding and reasoning. The dataset is basically providing an environment for the development of new algorithms by isolating from the real-life situations and noise which are very difficult to control and filter using computer algorithms. Thus, this simple clean dataset has small sentences with questions that all have answers inside the same dataset. It ensures that the algorithm doesn't have any difficulty like in real life, to figure out emotions, sarcasm, etc., and just had to plainly sift through the words, figure out the grammatical meaning and provide a grammatically correct answer for the question asked [13].

The bAbI is a set of 20 tasks, where each task consists of several context-question-answer triplets, prepared and released by Facebook. Each task has a unique aim of testing certain aspects of reasoning and is, therefore, aimed towards testing specific capabilities of different QA learning models. We are attempting to solve the task 1 and task 2 of this dataset. The dataset has the same tasks in both languages, first in readable form, then in a shuffled form which is not readable by the humans so that the learner is more forced to rely on given training data which mimics the learner learning something from scratch.

Some of the important reasons why we chose the bAbI dataset are as follows:

1. This dataset is a synthetically generated one. This helps in eliminate a lot of the noise and errors, which is the norm in real-world data.
2. This is a closed dataset, unlike the open-ended datasets, which require the knowledge of the world to answer certain questions.
3. There are well-defined baselines and tasks, upon which our model can be evaluated.

**MCTest** is a dataset created by Microsoft. Similar to the bAbI dataset, it provides information about context, question and answer. MCTest has different fictional stories which are created solely for testing different QA models using Mechanical Turk and aims at the reading comprehension level of 7-year-old children. MCTest is a multiple-choice question answering task whereas the bAbI is not. Two MCTest datasets were gathered using a slightly different methodology, together consisting of around 660 stories with an excess of 2000 questions. MCTest is a very small dataset which, renders it tricky for deep learning methods. Since other datasets didn't provide the functionalities as features like that bAbI dataset, bAbI was chosen rather than the rest. We have only used the bAbI dataset for analysis. But similar results can be reproduced for other datasets also.

## 6 Experimental Results and Analysis

Table 2 contains the results of our experiment on question-1 with 1000 training samples. The accuracy values will increase with increasing the number of iterations/epochs. The only concern with this type of model is that it is hard to predict when the model is **overfitting** the dataset [14, 15].

An interesting observation is the corresponding values for test accuracy for the 1000 and the 10,000 samples' dataset are of stark contrast to one another. For example, after 10 iterations, the model trained on 1000 dataset achieves 19% accuracy on the test dataset, whilst the model trained on 10,000 dataset achieves 51% accuracy. The divide in the accuracy values goes on until up to 40 iterations. This further suggests that a larger dataset provides a larger degree of diversity in the data for the deep learning model.

The above tables are shown in order to prove the increase in accuracy with respect to the epochs. The baseline for the above experiments shown in Tables 1 and 2 is 50 as quoted in Table 7, Task 1(single supporting facts) the bAbI dataset, LSTM and for Tables 3 and 4, the other tasks (two supporting facts) is 20, as shown in Table 7, LSTM.

This table shows the output for different activation functions. These activation functions are used on the last layer of the LSTM. Although these functions only affect the last layer, they play a crucial role in making the predictions.

The following tables show the accuracies for each activation function. All the experiments have been run for 50 iterations.

**Table 1** Task-1, dataset with 10,000 samples (training on 9500 samples, validation on 500 samples)

Epoch number	Test loss	Test accuracy (out of 1)
10	1.8079	0.1870
20	1.7984	0.2120
30	1.7921	0.2300
40	1.6908	0.3200
50	1.6765	0.3090

**Table 2** Task-1, dataset with 1000 samples (training on 950 samples, validation on 50 samples)

Epoch number	Test loss	Test accuracy (out of 1)
10	1.0893	0.5070
20	1.0553	0.5310
30	0.9355	0.6310
40	0.0108	0.9990

**Table 3** Task-2, dataset with 1000 samples (training on 950 samples, validation on 50 samples)

Epoch number	Test loss	Test accuracy (out of 1)
10	1.8162	0.1870
20	1.7923	0.2030
30	1.7387	0.2380
40	1.6858	0.3010

**Table 4** Task-2, dataset with 10,000 samples (training on 9500 samples, validation on 500 samples)

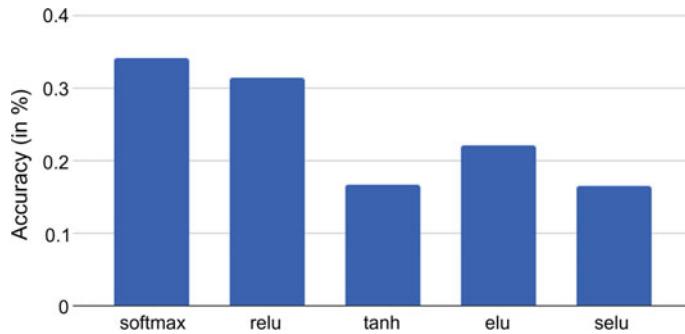
Epoch number	Test loss (cross-entropy)	Test accuracy (out of 1)
10	1.4049	0.4280
20	1.3497	0.4180
30	1.2964	0.4530
40	1.2880	0.4410

## 1. Experiments with a dataset size of 1000 training data

See Table 5 and Fig. 4.

**Table 5** Results from training on 1000 samples

Activation function	Accuracy on test data (out of 1)
softmax	0.3420
relu	0.3140
tanh	0.1670
elu	0.2210
selu	0.1650

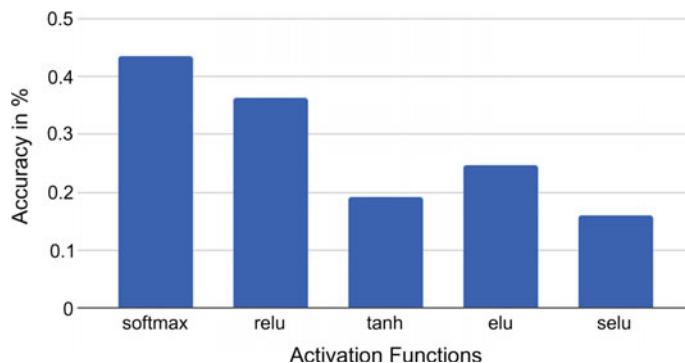


**Fig. 4** Experiments with a dataset size of 1000 training data

## 2. Experiments with a dataset size of 10,000 training data

The softmax activation function gives us the best possible results over other activation functions. It has been the de facto activation function for many of the deep learning solutions for a variety of problems. This further proves the stance of the work by Dunne et al. [16] (Fig. 5).

It is quite evident that ReLU comes to a close second when compared to the softmax activation function. Recent research on ReLU and its variants like LReLU and PReLU [17] indicate that usage of ReLU activation function can have a significant advantage in the performance of the model. For instance, some of the existing literature in Sentiment Analysis using tweets [18] have used the ReLU activation function to achieve impressive results. Moreover, many of the existing implementations of CNN use the ReLU Activation function. But it is important to note that the deep neural network models used in NLP, such as those in [18–21], also have used CNN as a feature extractor for preprocessing tasks, and have consequently used ReLU as an activation function.



**Fig. 5** Experiments with a dataset size of 10,000 training data

**Table 6** Results after training on 10,000 data points

Activation function	Accuracy on test data (out of 1)
softmax	0.4360
relu**	0.3640
tanh	0.1910
elu	0.2460
selu	0.1600

From the data shown in Table 6, one of the causes for the disparity in accuracies could be the vanishing gradient problem.

But, a similar trend doesn't seem to hold for the case of NLP tasks, possibly because of the different requirements with the problem of Question Answering. The model produces grammatically and semantically correct sentences, without having a prior knowledge of the rules of the language system.

\*\* The experiment with ReLu stopped after 27 iterations. The accuracy values resulted in NaN, so we had to stop the experiment, and proceed with the test data directly.

#### Analysis of Reversing the Order of Words in Query and Story

All of these experiments have been run with 1000 samples. The hyperparameters used are as follows:

BATCH\_SIZE = 32

EPOCHS = 50

Hidden-States of LSTM for processing queries = 100

Hidden-States of LSTM for processing sentences = 100.

We have used Categorical Cross-Entropy (built-in function in Keras library) loss for all the experiments and Adam optimizer.

Using the techniques from Sutskever et al. [1], we ran two experiments with respect to reversal of the input sequences. Initially, only the queries were reversed, and the story was fed in order to the model. In the next step, we also reversed the order of the stories (along with reversing the query) and obtained the accuracy values. We have not come across any research paper, that used this technique for a weakly supervised solution.

The results from this have been compared with the LSTM Baseline from Facebook Research. We have used the results from [22] for the accuracy values from the baseline LSTM model and the Structured SVM model.

Tasks like 3, 7, 8, 16 and 18 are worthy of special mention, as the Model-2 outperforms the strongly supervised model structured SVM by a substantial margin. These tasks where our models perform better than the LSTM baseline or the structured SVM model have been highlighted in Table 7.

In Model-1, the words of the input story and the query have been reversed. In Model-2, only the words in the query have been reversed.

**Table 7** Comparison of accuracies of different models (highlighted are ones where improvements are seen). All accuracy values are given in percentages

Sl no.	Task	LSTM baseline facebook	Structured SVM	Model-1 (reversing story and query)	Model-2 (reversing- query-only)
1.	Single supporting fact	50	99	28.6	49.10
2.	Two supporting facts	20	74	16.3	<b>34.8</b>
3.	Three supporting facts	20	17	<b>21.60</b>	<b>19.0</b>
4.	Two arg. relations	61	98	<b>76.50</b>	<b>68.0</b>
5.	Three arg. relations	70	83	33.2	51.0
6.	Yes/no questions	48	99	47.60	<b>48.3</b>
7.	Counting	49	69	46.7	<b>79.5</b>
8.	Lists/sets	45	70	31.2	<b>76.0</b>
9.	Simple negation	64	100	63.8	63.8
10.	Indefinite knowledge	44	99	<b>44.5</b>	<b>48.6</b>
11.	Basic coreference	72	100	29.9	68.4
12.	Conjunction	74	96	16.8	71.2
13.	Compound Coref.	94	99	14.2	93.2
14.	Time reasoning	27	99	20.8	23.1
15.	Basic deduction	21	96	<b>52.9</b>	<b>23.9</b>
16.	Basic induction	23	24	<b>23.7</b>	<b>48.9</b>
17.	Positional reasoning	51	61	48.0	<b>52.0</b>
18.	Size reasoning	52	62	<b>59.4</b>	<b>91.6</b>
19.	Path finding	8	49	<b>8.2</b>	<b>8.3</b>
20.	Agent's motivations	91	95	74.6	90.9

We are not aware of any other research paper that proposes a technique under Weakly Supervised and achieves a better accuracy. But under the strongly supervised techniques, the model proposed by [23] achieves more than 95% accuracy on most of the tasks.

We understand that this technique of reversing the input query will work only for certain languages. Specifically, it works for languages where the factual content might present at the end of the sentence. So, by reversing the sentence, *the model might not work for other languages with a different linguistic structure*. We leave it as future work to explore this technique further with other languages. This property could also serve as a measure of linguistic similarity between languages.

This explanation has already been provided by Sutskever et al. [1]. But from our experiments above, we can see that reversing the queries will improve the accuracies on a variety of tasks. This improvement is not specific to any particular task, and is observed across multiple tasks. This could mean that in the English language, most of the factual questions could be better-answered by increasing the short-term dependency to the last part of the sentence. This also implies that the answers to these queries are more likely to be found in the later parts of a sentence.

From the data above, we observed that it is not beneficial to reverse the input data also. A possible explanation could be that a story comprises of multiple sentences. By reversing the story, we could be focusing less on the initial parts of the story. This could be detrimental if many questions are asked using the facts in the initial parts of the paragraph. As a part of future work, we can explore the usage of Bidirectional LSTMs for these tasks.

## 7 Conclusion

We have attempted to draw parallels between the usage of different activation functions, for the task of Question Answering in Natural Language Processing, using Deep Learning models, specifically RNNs and its variants. We believe that no one-size-fits-all activation function exists across various tasks, even for a narrow subdomain like Question answering in NLP.

We felt a comparison of different activation functions with the same modalities and identical tasks is important. This assumes more significance today because NLP encompasses a lot of different types of tasks. Different activation functions may prove to shed light on different aspects of Natural Language Understanding. We believe that techniques like reversing the order of words play an important role in languages with a similar grammatical structure to English.

We have also shown that reversing the order of input words potentially increases the number of short-term dependencies and hence improves the accuracy across a wide array of tasks. This feature could be specific to the structure of the English

language. Future work could explore the existence of such structures in different languages. As a part of future work, we intend to explore the merits of more recent activation functions like adaptive softmax across various NLP tasks.

## References

1. Sutskever, I., et al. (2014). Sequence to sequence learning with neural networks.
2. Zhou, G.-B., et al. (2016). Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing*, 13(3), 226–234.
3. Jurafsky, D., & James, H. M. (2017). *Speech and language processing* [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/28.pdf>, August 7, 2017.
4. Szegedy, C., et al. (2013). *Intriguing properties of neural networks*. ArXiv preprint, arXiv: 1312.6199 .
5. Severyn, A., & Moschitti, A. (2015). Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.
6. Stroh, E., & Mathur, P. (2016). *Question answering using deep learning* [Online]. Available: <https://cs224d.stanford.edu/reports/StrohMathur.pdf>.
7. WildML. (2017). *Introduction to RNNs* [Online]. Available: <http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/>, September 17, 2017.
8. Fan, E. (2000). Extended tanh-function method and its applications to nonlinear equations. *Physics Letters A*, 277(4-5), 212–218.
9. Elworthy, D. (2000). Question answering using a large NLP system. In *TREC*.
10. Dunne, R. A., & Campbell, N. A. (1997). On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. In *Proceedings of the 8th Australian Conference on the Neural Networks* (Vol. 181). Melbourne.
11. Rajpurkar, P., et al. (2016). SQuAD: 100,000 + questions for machine comprehension of text. In *EMNLP*.
12. Colah's Blog. (2015). *Understanding LSTM networks* [Online]. Available: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>, August 27, 2015.
13. Facebook Research. (2015). *Babi dataset* [Online]. Available: <https://research.fb.com/downloads/babi/>.
14. Brownlee, J. (2016). *Over and underfitting in ML* [Online]. Available: <https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/>, March 21, 2016.
15. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436.
16. Jin, X., et al. (2016). Deep learning with S-shaped rectified linear activation units. In *AAAI*.
17. Yu, L., et al. (2014). Deep learning for answer sentence selection. ArXiv preprint, arXiv:1412.1632 .
18. Razvan, P., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*.
19. Martin, J. H., & Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall.
20. Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

21. Chung, J., et al. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling*. ArXiv preprint, [arXiv:1412.3555](https://arxiv.org/abs/1412.3555).
22. Weston, J., et al. (2015). *Towards AI-complete question answering: A set of prerequisite toy tasks*.
23. Sukhbaatar, S., et al. (2015). End-to-end memory networks.

# GestTalk—Real-Time Gesture to Speech Conversion Glove



**Varun Shanbhag, Ashish Prabhune, Sabyasachi Roy Choudhury and Harsh Jain**

**Abstract** It is often observed that people who are speech or hearing impaired find it difficult to communicate with others. According to WHO [5], over 5% of the global population has disabling hearing loss, while many more are speech impaired. These people mainly rely on sign languages for their daily communication. Sign language is quite complicated for an average person to understand, which makes the world less accessible to a person who has acquired this disability. Hence, to solve this existing problem and make the world more accessible for such people, we propose GestTalk, a smart glove specifically designed to enable speech-impaired people to communicate with others by translating their performed gesture to speech with the help of machine learning and IoT. We are using hardware-based glove loaded with home-made flex sensor to capture data points and machine learning algorithm to map these gestures to speech dynamically in real time. Usage of home-made flex sensor, each costing under 10 INR, plays the vital role in capturing finger movement during the gesture, and certainly makes GestTalk a cost-effective gadget. We were able to achieve ~95% of prediction accuracy on the trained sample with our minimal prototype. We present this case study to show how GestTalk can aid the hearing/speech impaired by enabling them to communicate with the world via speech cheaply and economically.

**Keywords** Artificial intelligence · Dynamic time warping · Home-made flex sensors · Internet of things · Machine learning · Virtual reality

---

V. Shanbhag (✉)

Information Technology, RAIT, Nerul 400706, Maharashtra, India  
e-mail: [shanbhagvarun55@gmail.com](mailto:shanbhagvarun55@gmail.com)

A. Prabhune

Information Technology, DMCE, Airoli 400708, Maharashtra, India  
e-mail: [ashuprabhune@gmail.com](mailto:ashuprabhune@gmail.com)

S. R. Choudhury

Data Science, Indiana University, Bloomington 47405, IN, USA  
e-mail: [sabysachi087@gmail.com](mailto:sabysachi087@gmail.com)

H. Jain

Computer Science, IIIT, Delhi, New Delhi 110020, India  
e-mail: [harsh18006@iiitd.ac.in](mailto:harsh18006@iiitd.ac.in)

## 1 Introduction

Verbal communication plays the vital role in expressing ideas and feelings of an individual, while at the same time helps us to understand the emotions and thoughts of the others quickly. To achieve any goal, may it be personal or professional, verbal communication plays a significant role. Sadly, people who are speech or hearing impaired cannot use this useful tool to their benefit. They majorly rely on sign language as an alternative.

There are various sign languages one can learn, namely American Sign Language (ASL), French Sign Language (LSF), and so on. Although these sign languages have proved useful for speech/hearing impaired over the years, they take a considerable amount of time to be learned. Moreover, it is unlikely to think that majority of us would be aware of such sign language to make sense of it, thus making it difficult for them to communicate with others.

For the above-mentioned reason, majority of such population are dependent on others who can understand this language. They also face various challenges due to the same in daily life. So to make their life easy and the world more accessible, we propose GestTalk, which leverages recent advancement in machine learning and internet of things to aid hearing/speech impaired people to communicate via speech with the world.

Not only GestTalk will help to communicate dynamically in real time but it is also quite cheap and one can self-train and configure with gestures as one desires; it may not necessarily be ASL or LSF.

In a nutshell, GestTalk captures the data from glove as you make the gesture, find the closest match from the available training set using machine learning algorithm, called dynamic time warping, and convert the identified gesture label to speech via a mobile app in real time.

However, a good deal of people may argue that image processing (IP) could be used instead of the proposed hardware-based approach. Similar objective can be achieved using image processing too, but we opted for hardware-based approach majorly because the disadvantages of image processing outweigh the benefits it provides. The only advantage for the image processing-based approach is data can be seamlessly captured without dependency on an external device using the smart phone camera which is readily available. On the other hand, the disadvantages are as follows.

First, the images captured by the camera may be affected by many physical interventions, such as lighting, background, the position of the camera, and so on, which can hamper result processing. Second, image recognition algorithm is considerably slow, as they need to eliminate noise (since everything gets captured as part of footage except for gesture) and identify the gesture from the footage, which takes a large amount of computational power and time, making it infeasible for efficient daily usage. Third, person is always dependent on other to capture the footage and translate it, while our approach provides the user with flexibility to use the technology at his/her will.

Our hardware-based glove approach consists of home-made low-cost flex sensors, placed on each finger which provides us with data upon bending them. We also used the gyroscope and accelerometer to provide the data on orientation and positioning of the hand. All the sensors' information are processed using DTW (dynamic time warping) algorithm, which is a time series classification algorithm that measures similarities between two gestures (performed gesture and entire training set), and may vary in speed. DTW is applied separately on flex sensors, gyroscope and accelerometer data separately using pipelines. Each of them predicts three closest gestures performed, and based on these varied outputs, a collaborative decision is made based on the voting.

This identified gesture is sent across to a mobile application, which then converts text to speech and speaks out the gesture label.

The advantage of using the sensor-based approach and dynamic time warping for gesture recognition is that the technique works irrespective of the surrounding or the user who performs the gesture and provides accurate results, which makes it robust, fast, economical and convenient. On another hand, the limitation is the dependency on hardware/device.

## 2 Working

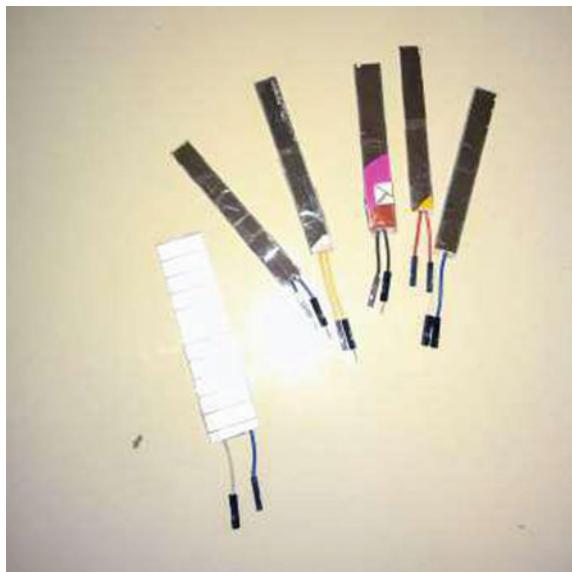
### 2.1 Home-Made Flex Sensor

We took 4½-in regular paper and scratched with pencil on it (depositing graphite layer on the same in the process). Then we coupled it with strips of aluminium tape on both sides. Lastly, we inserted two wires between graphite and aluminium, and we got our flex sensor ready. We tested this using multimeter and could capture voltage difference as we flex/bend the device. Whole production cost is under 10 INR each, and the results were pretty impressive (Fig. 1).

### 2.2 The Setup

We used Arduino Mega to integrate all the hardware components. This includes home-made flex sensor, accelerometer and gyroscope plugged onto a regular glove, as shown in Fig. 2. There are various solutions proposed for a similar setup, which costs around USD 250 [4]; our total setup costs around USD 20.

**Fig. 1** Home-made flex sensors



**Fig. 2** Setup of GestTalk with flex sensors and Aurdino



## 2.3 Data Generation

The vital part for any machine learning problem is gathering relevant training data. In this case, we are relying on data generated from the glove we built, which generates the series stream of data. We calibrated our Arduino to emit data points from the glove at the rate of four times per second which provides us better control, detail and accuracy over the model. These captured data points are then broken down into time series data [AccX, AccY, AccZ, GysX, GysY, GysZ, FS1, FS2, FS3, FS4, FS5].

Accelerometer data: AccX, AccY, AccZ

Gyroscope data: GysX, GysY, GysZ

Flex sensor: FS1, FS2, FS3, FS4, FS5.

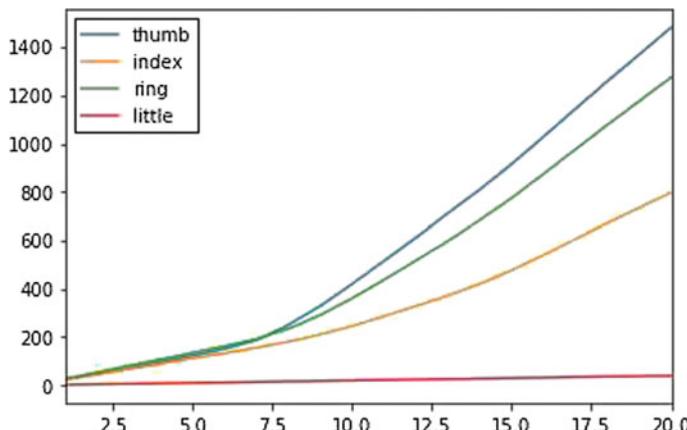
This data generation is managed using python script, which has the functionality to start and stop capturing the data points from the hardware device.

Almost ten samples for every gesture varying in the temporal sequence are collected and stored in a file. Data collected may contain a certain amount of noise, which is cleaned to get the accurate training data.

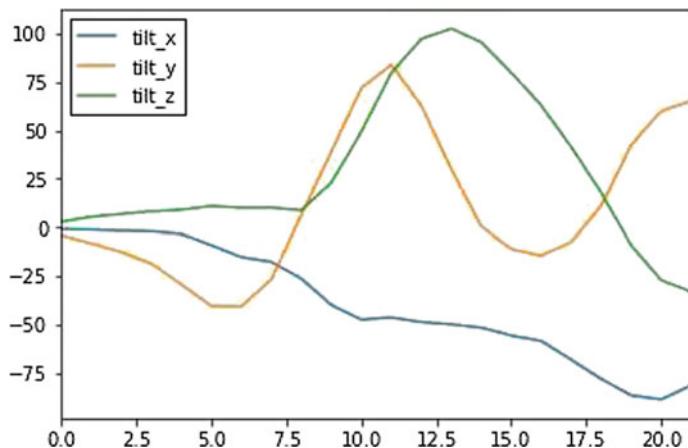
Figures 3, 4 and 5 are the graphical representation of sensor data for gesture ‘X’.

We have used three different transformers to normalise the data:

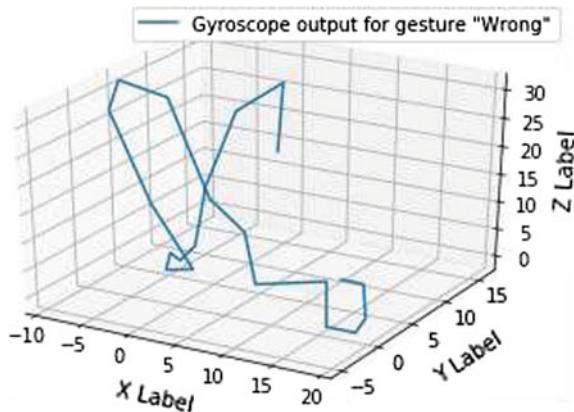
- **CoordinateNormaliser:** This shifts the coordinates to its zero setting which gives better result during warping.
- **AccelerometerNormaliser:** Accelerometer gives the data in  $g$  where  $g \rightarrow$  gravity acceleration, that is,  $9.8 \text{ m/s}^2$ . These data do not add much to the prediction as acceleration might vary for the different person. Instead of using the voltage figure, we would be using the direction vector only. This normaliser takes only the sign and dumps the numeric value.



**Fig. 3** Graphical representation of flex sensor data for gesture ‘X’



**Fig. 4** Graphical representation of accelerometer data for gesture ‘X’



**Fig. 5** Three-dimensional graphical representation of gyroscope data for gesture ‘X’

- **AnalogVoltageScaler:** Flex sensor data are the result of the change in the resistance caused due to finger movement. Magnitude is not of much importance; in this case only the change in data points matters. So we will shift the origin as in CoordinateNormaliser to zero.

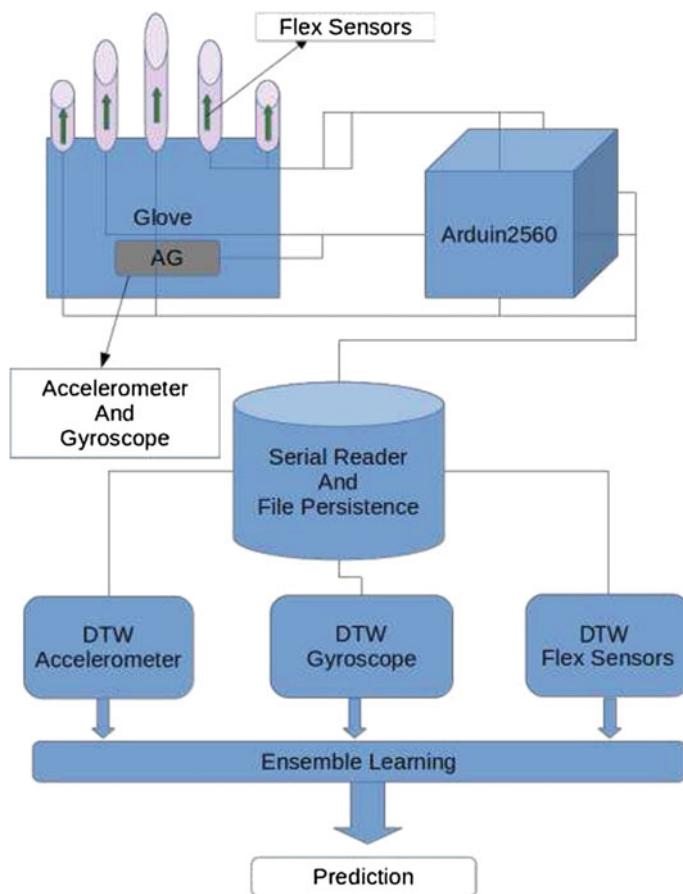
## 2.4 DTW (Dynamic Time Warping)

DTW [1] is an ML-based algorithm used to measure similarities between two temporal sequences or time series which may vary in speed. As in our case, the time

required to complete a particular gesture may change each time. Regarding speed or size, DTW seems to be an appropriate choice. For instance, if we make gesture ‘X’ with our index finger, factors which can vary or subject to change are speed, total time and size of gesture. For instance, ‘X’ done as part of a gesture by person A can vary in size if done by person B, although they both mean the same. We have captured the data using gyroscope, accelerometer and flex sensor in the linear sequence and analyzed the same using DTW.

We have three different parameters as part of data, that is, the accelerometer, gyroscope and flex sensor, to train and predict the gesture. It is reasonable that the DTW classifier is applied separately on the three sensor data independently.

To achieve this, we created three different pipelines and applied DTW [3] to each of those pipelines separately, as shown in Fig. 6.



**Fig. 6** Block diagram for GestTalk

Each pipeline is processed differently based on data points depending on their behaviour; for instance, flex sensor data is less likely to differ from each other, while compared with that of data obtained from accelerometer; hence while estimating gesture from accelerometer data, we are calculating Euclidean distance followed by normalisation.

This combination is obtained using hyper tuning with randomised search cv [2] which helped us in determining what would work best for given data set. Below stated are the results for the same:

```

gyro_attr_names = ["cord_x", "cord_y", "cord_z"]
acc_attr_names = ["tilt_x", "tilt_y", "tilt_z"]
fgr_attr_names = ['thumb', 'index', 'middle', 'ring', 'little']

all_PIPELINES = []
gyro_pipeline = Pipeline([
    ('selector', DataFrameSelector(gyro_attr_names)),
    ('cord_norm', CoordinateNormalizer()),
    ('estimator', DTWClassifier(dist='euclidean', normalize=True)),
])
all_PIPELINES.append(gyro_pipeline)

acc_pipeline = Pipeline([
    ('selector', DataFrameSelector(acc_attr_names)),
    ('acc_norm', AccelerometerNormalizer()),
    ('estimator', DTWClassifier(dist='euclidean', normalize=True)),
])
all_PIPELINES.append(acc_pipeline)

flex_pipeline = Pipeline([
    ('selector', DataFrameSelector(fgr_attr_names)),
    ('std_scaler', AnalogVoltageScaler()),
    ('estimator', DTWClassifier(normalize=True)),
])
all_PIPELINES.append(flex_pipeline)

```

**Code Snippet 1.1.** Pipelines for prediction in Python

The output from the above pipelines returns three best ‘ $k$ ’ matches for the given input along with their distance.

### 3 Results and Discussion

The output of the above pipeline is scaled, and eventually, based on the highest probability, the best match is found. Both steps are elaborated as follows.

**Table 1** Pipeline's result

Pipeline	Probability with gesture $P(G)$
Gyroscope	0.81('X'), 0.92('X'), 0.65('O')
Accelerometer	0.75('X'), 0.55('O'), 0.70('O')
Flex sensor	0.85('X'), 0.78('X'), 0.72('X')

### 3.1 Scaling Output

The classifier provides us with the  $k$ -nearest neighbours comparing performed gesture with the trained data set. We need to scale the distances between the range of 0 and 1. Using MinMaxScaler will do, but we have used a custom method which more or less does the same. Now the distances are mapped to scores or percentage of error.

The scaling differs according to data types. For example, gyroscope provides us with coordinates, whereas in the case of flex sensor, it feed us with fluctuating voltage output. So the normalisation should also be applied for each pipeline rather than on the merged output. Hence, we modified our original classifier results with the above-stated normalising technique.

### 3.2 Percentage Match

Now we have three different outputs from three separate pipelines with predicted gesture and error percentage (rather than distance). More the error, the less is the percent match. So subtracting from 1 (100% match) will give us the probability of the gesture. Again each pipeline provides  $K$  best matches; in our case  $K = 3$ , and we would be having  $K_1 + K_2 + K_3$  number of matches from three pipes. Now finding the sum of probabilities for a given gesture and scaling them will provide us with the net matching percentage. Below is sample output if the user makes the gesture.

'X', this is how a result from the pipeline would look like.

So from Table 1, we can say that prediction for gesture will be 'X' as follows:

Gyroscope and flex sensor pipeline predicts the gesture to be 'X' with the cumulative probability of 0.81 ( $0.81 + 0.92 + 0.85 + 0.78 + 0.72/5$ ) with five matches for 'X' out of six results, while accelerometer predicted it to be 'O' with the cumulative probability of 0.62 ( $0.55 + 0.70/2$ ) with two matches for 'O' out of 3. Hence, in this case, results are in favour of 'X' with five matches from the total of nine outputs, so we predict the gesture to be 'X'.

## 4 Conclusion

With GestTalk glove we were able to achieve an accuracy of ~95% in predicting the performed gesture when it was made by the same person who trained the model.

Results were quite impressive knowing that we used just ten training data series for a particular gesture. While for different users accuracy was around ~60% for all the 15 gestures performed. DTW helps us to recognise gesture dynamically. Owing to home-made sensors, the glove is economical. The accuracy of the glove can be further increased by improving the quality of the sensors and more training data.

## 5 Future Scope

Improvisation in the design of the glove can be achieved by improving the quality of the flex sensor as well as the position of accelerometer and gyroscope, which will give us more reliable data. DTW may not provide accurate results as the data set increases, in such cases recurrent neural network (RNN) would fit appropriately.

More than just converting gesture to speech, this glove can be widely used in some other applications, such as in gaming and VR. This glove can simulate the hand gesture, for example, games like table tennis, driving a vehicle, and so on. Using virtual reality imagine playing the piano virtually through this glove.

**Acknowledgements** This prototype was accomplished as a part of a global hackathon conducted in Persistent Systems. The authors would like to acknowledge the efforts of all the team members of BiT's Please (Samaikya Akarapu, Abhijeet Pal, Nikita Shah, Anirudh Ghosh & Nisha Kumari). Special thanks to Sagar Inamdar for helping us immensely in building the glove.

## References

1. DTW Wikipidea. <https://en.wikipedia.org/wiki/Dynamictimewarping>.
2. Hand gesture interpretation using sensing glove integrated with machinelearning algorithms. *World Academy of Science, Engineering and Technology International Journal of Mechanical and Mechatronics Engineering*, 10(11), 1860–1860 (2016). <https://waset.org/publications/10005809/hand-gesture-interpretation-using-sensingglove-integrated-with-machine-learning-algorithms>.
3. OurGitRepo,<https://github.com/sabyasachi087/aml-gesture-recognition/blob/master/gesturerrecognitionprojectreport.ipynb>.
4. Pypi Library, <https://github.com/pierre-rouanet/dtw>.
5. WHO Report on Hearing Disability. (2016). <http://www.who.int/mediacentre/factsheets/fs300/en/>.

# Deep Learning Algorithms for Accurate Prediction of Image Description for E-commerce Industry



Indrajit Mandal and Ankit Dwivedi

**Abstract** Currently, the demand for automation in information technology industry is increasing at rapid rate. Leading companies in artificial intelligence and computer vision have started investing in the research and making new products that can easily do the redundant work done by humans, ultimately increasing the productivity and reducing the cost of doing work. Industrial artificial intelligence can create new business models, can automate the industrial tasks which are redundant and can be easily done by specifically trained machines. The work is experimented with various feature extraction methods, like Visual Geometry Group (VGG-16) and VGG-19, and a smaller proposed five-layer convolution neural network (CNN-5) to generate captions for an image. The performances of these methods are compared using BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores. Various experiments have been conducted to see the effect of different hyper-parameters, like number of layers, learning rate, dropout, on the model accuracy measured by BLEU score. The effect of the input length (15, 34 and 40) of the sequence of text data on the LSTM network is also being analyzed and reported in the results section. The experiments are conducted on benchmark Flickr-8 K dataset which contains 8000 images and their respective descriptions in a text file. Each image was described by five different people and the text file contains image id and their respective five descriptions for each of the 8000 images.

**Keywords** Artificial intelligence · Automation · Convolution neural network · Long short-term memory · Image caption · Merge architecture

---

I. Mandal (✉) · A. Dwivedi

Research and Innovation Lab, Tata Consultancy Services, Bangalore, India  
e-mail: [indrajit.mandal@tcs.com](mailto:indrajit.mandal@tcs.com)

A. Dwivedi  
e-mail: [ankit.dwivedi2@tcs.com](mailto:ankit.dwivedi2@tcs.com)

## 1 Introduction

Automatically describing a digital image is a challenging problem in the world of artificial intelligence. First, a method to understand the content of the digital image is required. Second, a language model is required from natural language processing to convert the understanding from image to text. Captioning images has a lot of business applications in the today's world. Every day, millions of images are uploaded on the internet. Automatic and cost-effective labeling of those images will be a huge value addition. A robust and cost-effective method to caption those images will also help visually impaired to get the content of the image by converting the text into audio descriptions. Apart from it, e-commerce websites have lots of images of different products to showcase their customers; a good caption on the product's image will help a customer to understand the product better, ultimately increasing the sales of the company.

There are two main approaches to caption images: bottom-up and top-down. Bottom-up approaches [1–3] first identify different objects, humans, and so on, present in an image by applying different object detection methods and then combine those detected items in an image to form a sentence. On the other hand, top-down approaches [4–6] first generate a semantic representation of an image by extracting features from an image and then those features are decoded using recurrent neural network to generate captions for an image.

In the existing work on image captioning using the merge architecture, the analysis of using different neural networks for feature extraction of an image on the accuracy of the model is missing. Therefore, this work compares the effect of using deep pre-trained networks like VGG-16 and VGG-19 with smaller five-layer convolution neural network, in the merge architecture of image captioning. The accuracy of the model is evaluated using the BLEU- $N$  scores, for  $N = 1, 2, 3$  and 4.

## 2 Motivation

Humans have a remarkable ability to look into an image and understand what is there in it. They can easily understand the semantic relationship in the visual scene of an image and can describe it in natural language. Image captioning is a complex problem in computer vision that is far from reaching human-level accuracy. Successful image caption generation model can solve many real-world problems which require human involvement to reach the desired results; it can be a game-changer in many industries and can also help in medical diagnosis and other fields [7–13].

E-commerce, nowadays, is a huge industry which has lots of image data available to showcase their products to their customers. According to the research, 56% of the online customers step down from the buying process due to lack of information provided on the product, which is really a very big loss to e-commerce industries [14]. We can minimize this loss by using image processing and natural language

processing to generate descriptions of the image by using the previous images and their descriptions to train our model so that it can automatically generate descriptions of any new image. It will be a great value addition to the industry if the task of describing the images is done by computers with no or very little human involvement.

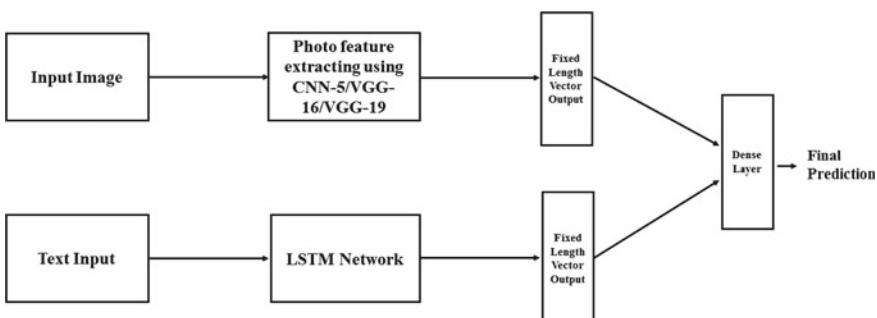
The earliest paintings were the rock paintings of pre-historic times, many of which humans cannot understand even now. Deciphering pre-historic images may open the door to many secrets of the past. Image captioning has lots of potential in analyzing those images and to understand the secrets behind them.

Describing medical images in natural languages will be a great boost in computer-aided diagnosis (CAD) systems, which is one of the leading research subjects in medical imaging and diagnostic radiology [15]. Till now, most of the works in medical imaging have focused on the classification problem, like whether the disease is present or not. The natural language description of the input medical images will help to diagnose the disease more efficiently and in detail.

### 3 Methodology

In this work, a top-down approach is adopted for generating captions of an image. First, a semantic representation of an image was generated by extracting features from an image, which will be a vector of fixed length. Then the LSTM network was used to encode the input linguistic features, which will again be a vector of same fixed length. Finally, both the output feature vectors of the same length will be merged and processed by a dense layer. The output from this dense layer will be finally processed by output dense layer which makes a softmax prediction for next word in the description sequence over the entire vocabulary.

This architecture of image captioning (shown in Fig. 1) is called merge architecture [16], where the feature vectors from images and their respective text description are merged later and processed by a dense layer to train the model. Therefore, RNN is



**Fig. 1** The merge architecture to generate natural language description of an image

only used as an encoder of linguistic features and the vector output from RNN is merged with the image features at a later stage.

In this work, the features of the images were extracted using three different convolution neural networks (CNN-5, VGG16 and VGG-19 were used for experiment purpose). The output from the photo feature extractor will be a vector of length 4096. This vector of 4096 elements will be passed through a dense layer with 256 neurons which will represent the input image as a vector with 256 elements.

A word embedding is used to represent words and text documents by a dense vector representation. The input will be a sequence with a pre-defined length (15, 34 and 40 were used in this work for experiment purpose) which will be first embedded to transform the input text to a dense vector, which will be the input to the long short-term memory (LSTM) recurrent neural network layer with 256 memory units. Therefore, the output from the LSTM network will be a vector of length 256.

Finally, the feature extractor and the text processor will return a fixed-length vector with 256 elements. Both the output vectors are merged together using an addition operation and processed by a dense 256 neuron layer and then by a final output dense layer that makes a softmax prediction over the entire output vocabulary for the next word in the sequence.

## 4 Model

Three different models were experimented in this work to generate natural language descriptions of an image, and the results are compared in the results section.

Model-1: Image features were extracted using VGG-16 network, and the LSTM network was used to encode input text sequence. Finally, both were merged to generate descriptions for an image.

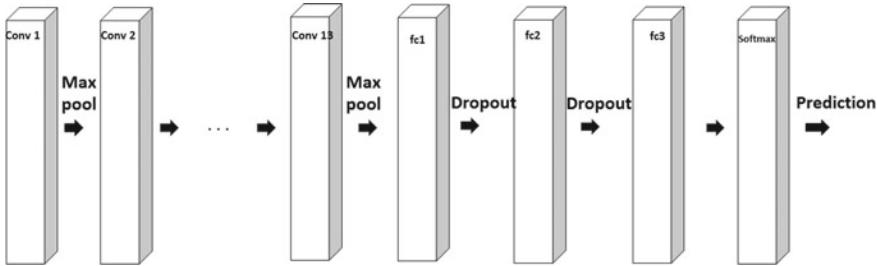
Model-2: Image features were extracted using VGG-19 network, and the LSTM network was used to encode input text sequence. Finally, both were merged to generate descriptions for an image.

Model-3: Image features were extracted using VGG-19 network, and the LSTM network was used to encode input text sequence. Finally, both were merged to generate descriptions for an image.

Three convolution neural networks (VGG-16, VGG-19 and CNN-5) that were used for extracting features from images and the LSTM network used to encode text data are described in the following sections.

### 4.1 VGG-16

Convolution networks (ConvNets) have been very successful in large-scale image and video recognition. ConvNets are also very successful in extracting features from



**Fig. 2** VGG-16 network architecture. Conv1 to Conv13 are 13 convolutional layers which are then followed by three fully connected layers (fc1, fc2 and fc3)

images. In 2015, Simonyan et al. [17] came up with the deep ConvNet which contains a total of 16 layers, 13 convolution layers, and 3 dense layers and achieved the state-of-the-art results in ILSVRC classification challenge, 2014.

Architecture of VGG-16 (Fig. 2).

VGG-16 takes fixed size input image;  $224 \times 224$  RGB image. Then the mean RGB value which was computed on the training set from each pixel is subtracted from the input image. After the pre-processing was done, the pre-processed input image was passed into the stack of convolution layers. In the network, five max-pooling layers ( $2 \times 2$  pixel window with stride 2) were also used to carry out spatial pooling.

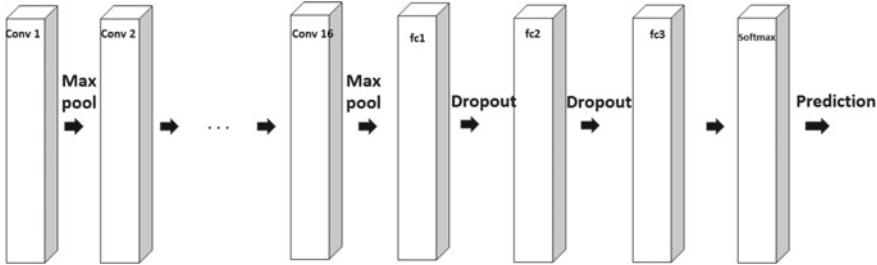
A stack of convolution layers was followed by three fully connected (FC) layers. The first two FC layers have 4096 neurons each and the third FC layer has 1000 neurons which was used to perform 1000-class ILSVRC classification which contains 1000 classes. The final layer was the softmax layer.

The network shows that more complex features can be learned by using multiple small-sized kernels rather than a few large-sized kernels. A  $3 \times 3$  kernel used in the network retains finer properties of an image, that too at lower computational cost.

**VGG-16 as Feature Extractor:** In this work, VGG-16 was used to extract features from images in the dataset. This can be achieved by removing the last layer (the third FC layer used for classification) from the VGG-16 Net and storing the output in features.pkl file. So, each image can be represented as a feature vector of length 4096, since the second last FC layer has length 4096. The extracted features will be merged later with the encoded text to generate captions.

## 4.2 VGG-19

It was denser than the previous architecture VGG-16. VGG-19 [17] was a 19-layer convolution neural network with three more convolution layers added to VGG-16. In both the networks, VGG-16 and VGG-19, a small ( $3 \times 3$ ) kernel was used and the



**Fig. 3** VGG-19 network architecture. Conv1 to Conv16 are 16 convolutional layers which are then followed by three fully connected layers (fc1, fc2 and fc3)

effect of depth on the accuracy was measured. The VGG-19 architecture is shown in Fig. 3.

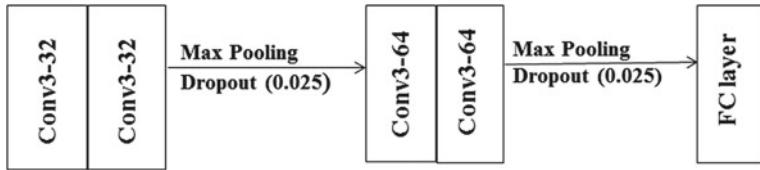
**VGG-19 as Feature Extractor:** VGG-19 was used as another method to extract features of images in the dataset and they were stored in separate features.pkl file. The last fully connected layer with 1000 neurons was removed; therefore features extracted from each image can be seen as a vector of length 4096, which is nothing but size of the second last FC layer.

#### 4.3 *The Proposed Smaller Convolution Neural Network with Five Layers: CNN-5*

To extract the features from images, we have defined a ConvNet, which has similar architecture like VGG-16 but has some noticeable changes. The input to the network is the same as previous:  $224 \times 224$  RGB image. The image was passed through a stack of convolution layers, with  $3 \times 3$  receptive field filters. After the first two convolution layers, max-pooling with pool size  $2 \times 2$  followed by dropout layer is added. The network consists of totally five layers: four convolution layers and one FC layer (max-pooling and dropout is not counted in layers). The first two layers are convolution layers followed by max-pooling and dropout. The last layer is FC layer, which will return features of image as one-dimensional (1-D) vector of length 4096.

Therefore, each  $224 \times 224 \times 3$  input image was converted into a 1-D vector of length 4096, which we are going to use later as an input to our caption-generation network for captioning images (Fig. 4).

The network is very less deep in comparison to VGG-16 and VGG-19 networks. Also, here dropout is used for regularization.



**Fig. 4** Architecture of the proposed CNN-5 network, where Conv3-32 means a convolution layer with a receptive field size of 3 and 32 numbers of channels

#### 4.4 Long Short-Term Memory Network (LSTM)

LSTM is an exceptional sort of recurrent neural networks which are fit for learning long-term dependencies. It was introduced by Hochreiter et al. in 1997 [18]. The major problem with RNN was the inability to model long-term dependencies in text generation. The problem was solved by LSTM.

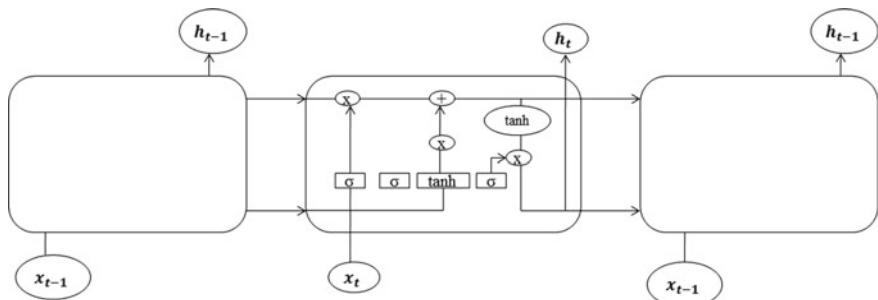
The architecture of LSTM network is shown in Fig. 5. The blocks in the figure are the memory blocks of LSTM called cells. From the previous cell to next cell, two states are being transferred, the cell state and the hidden state. The memory blocks are responsible for storing the memory of the network and this memory is manipulated with the help of three gates, namely forget gate, input gate and the output gate.

**Forget Gate:** It is in charge of expelling the memory from the network, which is never again required. For example, let the following sentence is fed into LSTM:

P-1 is an intelligent person. P-2 is not very intelligent.

When the first full stop will be encountered by LSTM network after person, it will forget the sentence before full stop since this information is not required to predict anything for person P-2. The information that is not required by LSTM is removed by the forget gate using filter multiplications.

It takes two inputs  $h_{t-1}$  and  $x_t$ , where  $h_{t-1}$  is the output from the previous layer and  $x_t$  is the input at time t. The inputs are multiplied with the weights matrices (W) and the bias (b) is added. Then the sigmoid function ( $\sigma_g$ ) is applied to it, which will



**Fig. 5** Architecture of LSTM network

give the output between 0 and 1.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

where  $f_t$  is the forget gate activation function.

**Input Gate:** It will add the information which is important and not redundant to the cell state. The input gate activation vector  $i_t$  can be calculated as:

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

**Output Gate:** This gate is responsible for choosing important and helpful information from the present cell state and showing it out as an output. The output gate activation vector ( $o_t$ ) can be calculated as:

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

The weight matrices and bias vector are the parameters that should to be learned during training.

Mathematically, for the forward pass of LSTM unit with a forget gate, networks compute [16, 19]:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (4)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (5)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (6)$$

$$c_t = f_t * c_{t-1} + i_t * \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (7)$$

$$h_t = o_t * \sigma_h(c_t) \quad (8)$$

where  $*$  denotes Hadamard or entry-wise product, and the starting values of  $h_0$  and  $c_0$  are taken to be zero. The variables involved in the above calculation are described as follows:

$x_t \in R^d$ : Input vector to the LSTM unit

$f_t \in R^h$ : Forget gate's activation vector

$i_t \in R^h$ : Input gate's activation vector

$o_t \in R^h$ : Output gate's activation vector

$h_t \in R^h$ : Output vector of the LSTM unit

$c_t \in R^h$ : Cell state vector

$W \in R^{h \times d}$  and  $U \in R^{h \times h}$  are weight matrices and  $b \in R^h$  is the bias vector; these are the parameters to be learned during training.

**LSTM as encoder of linguistic features:** In this work LSTM with 256 memory units is used to get the encoded vector representation of the input text descriptions of an image. The input text description of pre-defined length (15, 34 and 40 were used in this work for experimental purpose) is first embedded to transform the input text to a dense vector, which will be the input to a long short-term memory (LSTM) recurrent neural network layer with 256 memory units. Finally, the output from the LSTM network will be a vector of length 256.

Therefore, LSTM takes input text (represented by dense vector using embedding) and converts it into a vector of length 256, which is merged by feature vector of image of same length and finally fed into dense layer to train the model.

## 5 Description of Data

The Flickr-8 K [20] is one of the publicly available benchmark dataset used for image captioning. It contains images obtained from the Flickr website. The images do not contain any well-known individual or place with the goal that the whole image can be learnt in light of every single distinctive object introduced in the image. The dataset contains 8000 images in a folder and their descriptions in a separate text file. Each image is described by five different people. The standard dataset division that is used for Flickr-8 K dataset is 6000 images and their respective description for training: 1000 for validation and 1000 for testing. To compare the work with the existing state-of-the-art, same division was done.

## 6 Evaluation Technique

BLEU, or the Bilingual Evaluation Understudy [21], is a score for comparing a machine translation of text or in this case machine description of an image to one or more reference translations or descriptions of an image by humans. The closer a machine translation is to an expert human interpretation, the higher will be the BLEU score. Its output always lies between 0 and 1. This number will reflect how close is the machine-translated text to the reference texts, where values closer to 0 indicate that the machine-generated descriptions are not matching with the reference texts and values closer to 1 indicate the machine-translated text is similar to the reference texts.

Mathematical Details:

It is the geometric mean of modified precision scores calculated on the test corpus and then multiplied by an exponential brevity penalty factor. To compute BLEU score, first compute the modified  $n$ -gram precision  $p_n$ , using n-grams up to length N and positive weights,  $w_n$ , which sums to 1.

Let  $c$  be the length of the machine-generated description and  $r$  be the effective reference corpus length.

$$\text{BLEU} - N = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right) \quad (9)$$

where BP is the brevity penalty given by:  $\text{BP} = 1$  if  $c \geq r$  and  $\text{BP} = \exp(1 - \frac{r}{c})$  if  $c < r$ .

Note that  $N$  is the length of n-gram considered for evaluation, for BLEU- $N$  score,  $w_n = \frac{1}{N}$ , for  $N = 1, 2, 3, 4$ .

In this work BLEU- $N$ , for  $N = 1, 2, 3, 4$ , scores were calculated for individual machine-generated description by comparing them with a set of good quality reference sentences, and finally, the average over test dataset (1000 images) was taken to reach an estimate of the overall quality of the generated descriptions on test set.

## 7 Experimental Results

### 7.1 Comparison of the Proposed Work with Deep Networks like VGG-16 and VGG-19 in the Merge Architecture of Image Captioning

The image captioning model is called skillful if the BLEU scores when evaluated on the test dataset lie in the below range [22]:

BLEU-1: 0.401 to 0.578

BLEU-2: 0.176 to 0.390

BLEU-3: 0.099 to 0.260

BLEU-4: 0.059 to 0.170.

Therefore, the BLEU- $N$ ,  $N = 1, 2, 3$ , and 4, score of the proposed models qualifies as the skillful model for image caption generation.

From Table 1, we can see that Method-2 (extracting features using VGG-16) performed best in terms of BLEU-1, BLEU-2, BLEU-3 scores (shown in bold) and Method-1 (extracting features from CNN-5) performed best in terms of BLEU-4 score.

**Table 1** BLEU- $N$ ,  $N = 1, 2, 3, 4$  scores obtained from the three feature extraction methods implemented in this thesis work

Evaluation Matrix	Method-1	Method-2	Method-3
	CNN-5	VGG-16	VGG-19
	(Proposed method)	(Existing methods)	(Existing methods)
BLEU-1	0.523818	<b>0.525194</b>	0.513013
BLEU-2	0.223779	<b>0.245628</b>	0.231956
BLEU-3	0.137816	<b>0.142223</b>	0.126792
BLEU-4	<b>0.054964</b>	0.048126	0.040078

**Table 2** The effect of learning rate on BLEU- $N$ ,  $N = 1, 2, 3, 4$  scores of the proposed CNN-5

Evaluation matrix	Learning rate = 0.01	Learning rate = 0.001	Learning rate = 0.0001
BLEU-1	0.523818	0.418220	<b>0.529364</b>
BLEU-2	<b>0.223779</b>	0.169478	0.215588
BLEU-3	<b>0.137816</b>	0.107443	0.128624
BLEU-4	<b>0.054964</b>	0.043284	0.045880

## 7.2 Analysis of Number of Layers

The effect of number of layers on convolution neural network used for extracting features from images is analyzed in this section. The existing state-of-the-art models [23–25] used very deep convolution neural networks to extract features from images, but the results shown in Fig. 5 show that very dense networks are not needed to get skillful BLEU scores on the Flickr-8 K dataset. Even a five-layered neural network (CNN-5) gives a comparably good BLEU score. There is no statistically significant change in BLEU- $N$  score, for  $N = 1, 2, 3, 4$ , even if the number of layers is increased significantly. Therefore, the proposed CNN-5, which is computationally less expensive in comparison to deep VGG-16 and VGG-19 layers networks, also gives almost the similar level of accuracy on the test dataset.

## 7.3 Learning Rate Analysis of the Proposed CNN-5 Model

The effect of changing learning rate of the proposed CNN-5 model is analyzed in this section. Experiments were conducted by taking the learning rate 0.01, 0.001 and 0.0001, and the results are shown in Table 2. Learning rate of 0.01 gives better BLEU- $N$ , for  $N = 2, 3, 4$ , scores (shown in bold) while BLEU-1 score was best in the case when learning rate = 0.0001. The BLEU- $N$  score was worst in the case when learning rate was 0.001. Since BLEU-2,3,4 scores are best when learning rate was 0.01, and there is also no significant difference in BLEU-1 score when learning rates were 0.01 and 0.0001; therefore, out of these three, the optimal choice of learning rate will be 0.01.

## 7.4 Impact of Dropout on the Proposed CNN-5

Dropout is a regularization technique used to reduce overfitting and increase the performance of a neural network. In the proposed CNN-5, dropout layer is added after the second and fourth convolution layers (Table 3).

**Table 3** Experimental results showing impact of using dropout layers with  $p = 0.025$  and  $p = 0.50$  on CNN-5

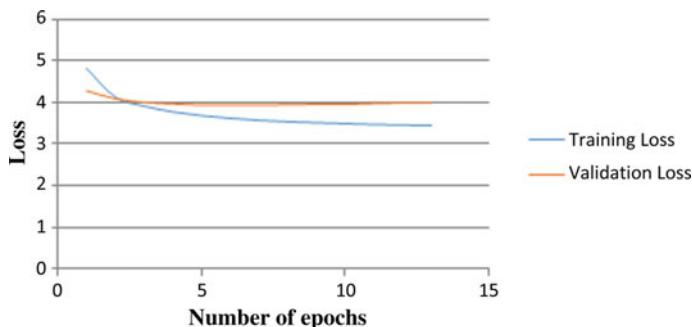
Evaluation matrix	Dropout = 0.025	Dropout = 0.50
BLEU-1	0.452286	<b>0.523818</b>
BLEU-2	0.185308	<b>0.223779</b>
BLEU-3	0.112190	<b>0.137816</b>
BLEU-4	0.040627	<b>0.054964</b>

Dropouts with probabilities,  $p = 0.025$  and  $p = 0.50$ , are applied on CNN-5 network, and the results with dropout probability,  $p = 0.50$  (shown in bold), are far better in terms of BLEU- $N$  score for  $N = 1, 2, 3, 4$ .

## 7.5 Model Convergence: Analysis of Training and Validation Loss

The analysis of training and validation loss with respect to number of epochs is done in this section. While doing experiments, 6000 images and their respective descriptions were taken for training the model and 1000 images with their descriptions were taken as validation dataset. On observing the training and validation loss, it suggests that the model learns fast and rapidly overfits the training dataset. When the skill of the model improves on the validation dataset (less validation loss) at the end of each epoch, model is saved to .h5 file, for later use.

From Fig. 6 we can see that the validation loss did not improve after fifth epoch, which suggests that the model learns very fast and converges after fifth epoch.



**Fig. 6** Training (blue) and validation (orange) loss vs number of epochs

**Table 4** Columns 2, 3 and 4 show BLEU scores after each fold of cross-validation and column 4 shows the row average

Evaluation matrix	FOLD-1	FOLD-2	FOLD-3	Average
BLEU-1	0.541720	0.501843	0.507015	0.516859
BLEU-2	0.256611	0.216968	0.251748	0.241775
BLEU-3	0.160959	0.123462	0.164437	0.149619
BLEU-4	0.066922	0.027334	0.067754	0.054000

## 7.6 Cross-Validation of the Proposed CNN-5 Model

In all the experiments done in this work, while fitting the model, 1000 images and their respective descriptions were left for validation of the model and 1000 images with their respective description were used as a test set to get the BLEU score.

In this section, the stability of the proposed CNN-5 for feature extraction is checked by doing additional three-fold cross-validation over the entire corpus of data.

The entire dataset of 8000 images were divided into three-fold form, of which two folds were taken at a time for training and the third fold is left for validation, and the BLEU- $N$ , for  $N = 1, 2, 3, 4$ , scores over the validation set are reported in Table 4.

The results did not deviate much from the results presented in Table 3 (second column in Table 3 shows the BLEU score obtained from the CNN-5 method). Therefore, the proposed CNN-5 method for feature extraction generalizes well on different validation datasets.

## 7.7 Analysis of Length of Input Sequence to the Text Processor

The text processor (described in Sect. 7.4) takes an input sequence of text with a pre-defined length. In this section, the analysis of the length of the input sequence to the text processor is done and the results show that the input sequence length is an important hyper-parameter that needs to be fixed. For experimental purpose, the input sequence length was taken to be 15, and the resulting BLEU scores are shown in Table 5. Similar analysis is done by taking input sequence length to be 40, and the results are shown in Table 6. By comparing both the tables, we can see that we are getting better BLEU- $N$ , for  $N = 1, 3, 4$ , scores when the length of the input sequence to the text processor is taken to be 15, and the BLEU-2 score is better when the length is 40.

Finally, this hyper-parameter was fixed by calculating the maximum number of words in the longest description, on the entire text corpus of Flickr-8 K dataset containing description of the images. The maximum words in the longest description

**Table 5** BLEU scores when length of the input sequence to the text processor was 15

Evaluation matrix	Method-1	Method-2	Method-3
	CNN-5	VGG-16	VGG-19
BLEU-1	0.464615	0.468372	0.452733
BLEU-2	0.207028	0.217373	0.207424
BLEU-3	0.141287	0.147282	0.138181
BLEU-4	0.056627	0.048272	0.042627

**Table 6** BLEU scores when length of the input sequence to the text processor was 40

Evaluation matrix	Method-1	Method-2	Method-3
	CNN-5	VGG-16	VGG-19
BLEU-1	0.438254	0.435588	0.431614
BLEU-2	0.224079	0.229058	0.232806
BLEU-3	0.123914	0.140545	0.138062
BLEU-4	0.045203	0.056030	0.054052

**Table 7** BLEU scores when length of the input sequence to the text processor was 34

Evaluation matrix	Method-1	Method-2	Method-3
	CNN-5	VGG-16	VGG-19
BLEU-1	0.523818	<b>0.525194</b>	0.513013
BLEU-2	0.223779	<b>0.245628</b>	0.231956
BLEU-3	0.137816	<b>0.142223</b>	0.126792
BLEU-4	<b>0.054964</b>	0.048126	0.040078

were 34. Taking the input sequence length to be 34 has improved the BLEU score to a large extent. In this work, in all the experiments (except Tables 5 and 6) the input sequence length to the text processor is taken to be 34.

The results obtained by taking the input length equal to 34 are shown in Table 7. CNN-5 performed well in terms of BLEU-4 score in comparison to deep models, VGG-16 and VGG-19 () .

## 8 Conclusions

In this section some of the important lessons learnt from the experiments work are discussed. Image captioning is a complex and recent problem in computer vision. This work will be a value addition to the future research, since lot of research is still needed to generate captions for digital images that can reach close to human-level accuracy.

In the available literature of captioning images, comparative analysis of models that uses transfer learning and the models that does not use transfer learning is

missing. This work compares the models that use transfer learning (VGG-16 and VGG-19) and the model that does not use transfer learning (proposed CNN-5). The results (BLEU scores) do not improve much while using VGG-16 and VGG-19 architecture in comparison to CNN-5.

It is a general misconception that increasing the depth of the neural network will make the network to learn better, but this cannot be generalized, as shown in the results section of this work. The best BLEU-1 score was obtained by using a small CNN-5 network. VGG-16 and VGG-19 performed better in terms of BLEU-2, 3, 4 scores, but the accuracy does not increase to that extent compared to the complexity of the model. The reason behind it could be Flickr-8 K dataset, which contains only 8000 images and does not require a network as deep as VGG-16 and VGG-19 to learn the features from images; even a five-layer network is sufficient.

Learning rate plays an important role in finding global optimal using stochastic gradient decent. If the learning rate is very large, then the model may not converge and if the learning rate is too low, then training the model will take a lot of time. Therefore, optimal choice of learning rate is necessary so that the network will take less time to converge. In this work, the optimal learning rate out of tested three was 0.001.

While training any deep learning model, the problem of overfitting is common. Dropout is a regularization technique used to reduce overfitting and increase the performance of a neural network. Using appropriate dropout layers at appropriate places in the network can improve the learning of the model to a large extent.

It is not always enough to check the model accuracy on the left-out fixed test dataset. The model can work well in left-out test dataset, but may not perform well when deployed on other new datasets. Therefore, it is important to check the stability of the model by doing cross-validation. K-fold cross-validation was used in this work and the results show that the model is robust on unseen dataset.

## 9 Discussions and Future Works

### 9.1 Discussions

The accuracy of the caption model depends on many factors, like size of the dataset, vocabulary used for generating descriptions, model used for feature extraction from images and tuning the hyper-parameters involved in the model. Recent research [26] in this field has shown that increasing the size of the dataset will make the model to learn better and ultimately will increase the accuracy of the descriptions generated. Flickr-30 K [27] or MS-COCO [28] datasets contain 30,000 and 2,500,000 images and their respective descriptions can be used to increase the accuracy of the model.

In this work, RNN was used only as an encoder of the text describing images. In this approach of image captioning, the image is merged with the output of RNN after processing the words. This is called merge architecture of image captioning

where image features and the RNN encoded descriptions are merged to train the network. On the other hand, another architecture available in the literature is the inject architecture, where the image is injected into the same RNN that processes the words. In the inject architecture of image captioning, RNN plays an important role because both the image and text descriptions are processed by RNN to predict the next word for describing an image, while in the merge architecture the role of RNN is only as an encoder of the text data [22].

If two different individuals were asked to write a description for a particular image independently, then there are high chances that their descriptions for the same image might differ, in terms of words used to describe an image and also the way different individuals look into the semantic relationship in an image. This difference in perception between the individuals results in different descriptions for the same image. BLEU-1 score of human description is 70 on the Flickr-8 K dataset [26]. Therefore ideally, if the image caption model reaches a human-level accuracy in near future, it will have the BLEU-1 score of 70 on Flickr-8 K dataset.

## 9.2 Future Works

The primary objective of this work is to build an image caption generation model and to compare the effect of various features extraction methods on the accuracy of the generated captions. The experimental outcomes confirm that the objective is reached. However, there is plenty of room still left where this research work can be extended.

The CNN has various hyper-parameters and all of them cannot be tuned using grid search as it requires high computation time and resources. There is a scope to improve the neural network architecture used in this project work.

There are two main approaches to caption images: bottom-up and top-down. Bottom-up approach first identifies different objects, humans, and so on present in an image by applying different methods for object detection, then combines those detected items in an image to form a sentence. On the other hand, top-down approach first generates a semantic representation of an image by extracting features from an image and then these features are decoded using recurrent neural network to generate captions for an image. This work only focuses on top-down approach of image captioning, but as a future work I would like to combine both the techniques and build an ensemble model that can generate captions for images close to human-level accuracy.

In addition, the machine-generated descriptions are sometimes grammatically incorrect. As a future work I would like to extend this work, by implementing an extra pipeline after the machine-generated text, which can take care of the grammatical mistakes in the generated descriptions.

## References

1. Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10* (pp. 15–29). Berlin, Heidelberg: Springer.
2. Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L. , & Choi, Y. (2012). Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*. ACL'12 (Vol. 1, pp. 359–368). Stroudsburg, PA, USA: Association for Computational Linguistics.
3. Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale *n*-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL'11* (pp. 220–228). Stroudsburg, PA, USA: Association for Computational Linguistics.
4. Chen, X., & Zitnick, C. L. (2014). Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654.
5. Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014). Deep captioning with multimodal recurrent neural networks (*m-rnn*). CoRR, abs/1412.6632.
6. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2014). Show and tell: A neural image caption generator. CoRR, abs/1411.4555.
7. Mandal, I. (2015). Developing new machine learning ensembles for quality spine diagnosis. *Knowledge-based systems*, 73, 298–310. <https://doi.org/10.1016/j.knosys.2014.10.012>. ISSN 0950-7051.
8. Mandal, I., & Sairam, N. (2013). Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system. *International Journal of Medical Informatics*, 82(5), 359–377. <https://doi.org/10.1016/j.ijmedinf.2012.10.006>. ISSN 1386-5056.
9. Mandal, Indrajit, & Sairam, N. (2012). New machine-learning algorithms for prediction of Parkinson's disease. *International Journal of Systems Science*, 45(3), 647–666. <https://doi.org/10.1080/00207721.2012.724114>.
10. Mandal, I., & Sairam, N. (2012). Accurate prediction of coronary artery disease using reliable diagnosis system. *Journal of Medical Systems*, 36, 3353. <https://doi.org/10.1007/s10916-012-9828-0>.
11. Mandal, Indrajit. (2014). A novel approach for accurate identification of splice junctions based on hybrid algorithms. *Journal of Biomolecular Structure & Dynamics*, 33(6), 1281–1290. <https://doi.org/10.1080/07391102.2014.944218>.
12. Mandal, Indrajit. (2016). Machine learning algorithms for the creation of clinical health-care enterprise systems. *Enterprise Information Systems*, 11(9), 1374–1400. <https://doi.org/10.1080/17517575.2016.1251617>.
13. Mandal, I. (2015). A novel approach for predicting DNA splice junctions using hybrid machine learning algorithms. *Soft Computing*, 19, 3431. <https://doi.org/10.1007/s00500-014-1550-z>.
14. <https://econsultancy.com/blog/61991-83-of-online-shoppers-need-support-to-complete-a-purchase-stats>.
15. Kisilev, P., Sason, E., Barkan, E., & Hashoul, S. (2011). *Medical image captioning: Learning to describe medical image findings using multi-task-loss CNN*.
16. Tanti, M., Gatt, A., & Camilleri, K. P. (2017). What is the role of recurrent neural networks (RNNs) in an image caption generator? In *INLG*.
17. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.
18. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>. PMID 9377276.
19. Li, X., & Wu, X. (2014). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition, October 15, 2014.

20. Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using Amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.
21. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* (pp. 311–318). CiteSeerX 10.1.1.19.9416 .
22. Marc, T., Albert, G., & Kenneth, C. (2017). *Where to put the image in an image caption generator*.
23. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156–3164).
24. Johnson, J., Karpathy, A., & Fei-Fei, L. (2015). *Densecap: Fully convolutional localization networks for dense captioning*. arXiv preprint [arXiv:1511.07571](https://arxiv.org/abs/1511.07571).
25. Wang, C., Yang, H., Bartz, C., & Meinel, C. (2016). *Image captioning with deep bidirectional LSTMS*. ArXiv preprint [arXiv:1604.00790](https://arxiv.org/abs/1604.00790).
26. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2014). Show and tell: A neural image caption generator. CoRR, abs/1411.4555, 2014.
27. Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2, 67–78.
28. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the ECCV'14* (pp. 740–755).

# Taj-Shanvi Framework for Image Fusion Using Guided Filters



Uma N. Dulhare and Areej Mohammed Khaleed

**Abstract** Multi-focus image fusion aims to produce an all-in-focus image by integrating a series of partially focused images of the same scene. A small defocused (focused) region is usually encompassed by a large focused (defocused) region in the partially focused image; however, many state-of-the-art fusion methods cannot correctly distinguish this small region. To solve this problem, a novel Taj-Shanvi framework, used for multi-focus image fusion algorithm based on multi-scale focus measures and generalized random walk (GRW), is implemented. First, multi-scale decision maps are obtained with multi-scale focus measures. Then, multi-scale guided filters are used to make the decision maps accurately align with the boundaries between focused and defocused regions. Next, GRW is used to combine these decision maps at different scales. After obtaining them, these maps are aligned using the watershed technique, whose edges are further smoothed using the guided filter. Experimental results are obtained by using few quality parameters, namely, entropy, edge structure-based similarity index measure, spatial frequency, mutual information, and so on, to evaluate the quality of the final fused image. Quality parameter assessment demonstrates that the proposed method produces a better quality fused image than conventional image fusion techniques.

**Keywords** Multi-focus · Image fusion · Guided filter · Watershed technique

## 1 Introduction

Image fusion has been commonly used in medical treatment and in diagnostic applications. It is generally used to provide additional information by merging the number of images of a patient [1, 2]. The fused image can be obtained by aggregating informa-

---

U. N. Dulhare (✉) · A. M. Khaleed  
C.S.E. Department, Muffakham Jah College of Engineering and Technology, Hyderabad,  
Telangana, India  
e-mail: [Prof.umadulhare@gmail.com](mailto:Prof.umadulhare@gmail.com)

A. M. Khaleed  
e-mail: [areej.mohammed95@gmail.com](mailto:areej.mohammed95@gmail.com)

tion from various modalities, namely, magnetic resonance image (MRI), computed tomography (CT), and so on. In medical diagnosis, the fused images signify variances in tissue density by CT images and diagnose brain tumors by MRI images. Various formats of multiple images must be combined to predict correct diagnoses and treatment, like cancer.

In this paper, we discuss a novel Taj-Shanvi framework, a multi-focus and multi-modal image fusion method in spatial domain, which is based on multi-scale focus measures, watershed technique, and GRW (generalized random walk) that can effectively segregate the smaller regions and obtain decision map that accurately aligns with the boundaries of defocused and focused regions. Initially, multi-scale decision maps for every input image given to the system as input are obtained using the multi-focus measures. Then, the boundaries between the decision maps are accurately aligned between the defocused and focused regions using the multi-scale guided filters. Next, GRW is used to combine these decision maps at different scales. After obtaining them, these maps are aligned using the watershed technique, whose edges are further smoothed using the guided filter.

## 2 Image Fusion Technique

The image fusion technique proposed here includes various steps that are to be carried out in order for the fusion process to be implemented. The detailed description of every part of the methodology is as follows:

### **Sum of Modified Laplacian**

For clarity of image in multi-focus image fusion, focus measures [3] are used, like EOG (energy of image gradient), variance, SML (sum of the modified Laplacian), SF (spatial frequency), and EOL (energy of Laplacian of the image). Sum of the modified Laplacian has been proved to outperform the other focus measures [4]. Thus, SML is used as a focus measure in this work. In [5], Nary proposed the modified Laplacian (ML).

### **Generalized Random Walks (GRW)**

In image processing and computer vision, random walk (RW) algorithm has been used for segmentation. Initially, random walk starts at one pixel which reaches to any one of the seeds with label. That pixel is represented as the same label along with maximum probability of corresponding seed. According to RW, the generalized random walk (GRW) is proposed to fuse multi-expose images [6].

### **Guided Filter Fusion Algorithm**

For creating an extensively informative fused image by combining various images, an effective image fusion method is used. This method [7] is based on weighted average technique by two-scale decomposition in which a base layer captures large-scale variations in intensity and a detail layer contains small-scale details.

### Watershed Technique for Image Segmentation

Watershed technique is a popular mathematical morphological tool for the image segmentation [8]. The boundaries are determined by the watershed lines that divide an image into various regions [9]. Segmentation by watershed represents the concepts of the three techniques, such as region-based, threshold-based, and edge-based techniques. In the watershed technique, image can be divided based on an image gradient.

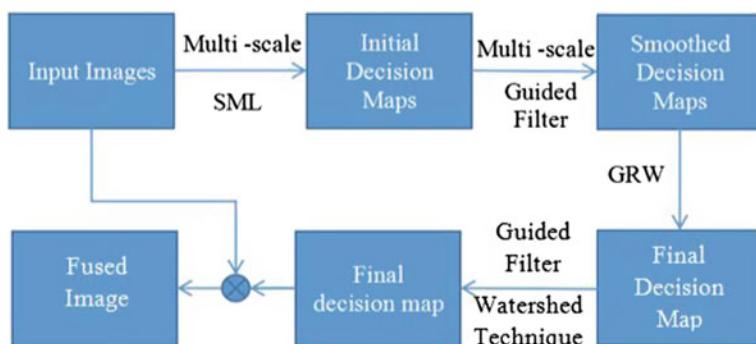
The main advantages of the watershed method are:

- Conventional edge-based methodologies form boundaries that are disconnected and later produce closed regions after post processing, but in watershed technique, the resulting boundaries form connected and closed regions.
- The resulting region consists of boundaries that correspond to contours, which appear in the image as their respective contours of objects.
- Combine all the formed regions to get the entire image region.

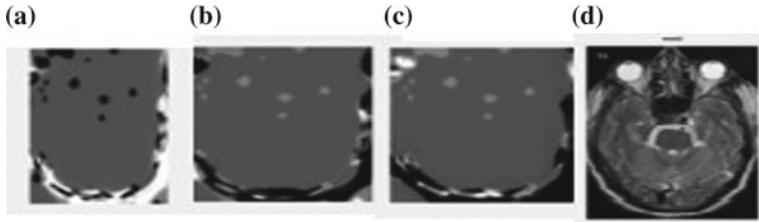
### 3 Schematic Architecture of the System

Block Diagram: Fig. 1 indicates Taj-Shanvi framework of image fusion method.

Figure 1 illustrates various steps involved in the image fusion process. The method begins by calculating the SMLs of the input images of the system at three different scales (1, 2, and 3 here). On comparing the SMLs of the outputs, the boundaries of the obtained decision maps are aligned using the guided filter method. Later, the focused and the defocused regions are identified using a generalized random walk (GRW) algorithm. Using a similarity function, large and small intensities are selected. Finally, the final decision map is constructed using the watershed technique and is smoothed out using guided filters, which leads to the construction of the final fused image.



**Fig. 1** Schematic Taj-Shanvi framework for fusion method



**Fig. 2** **a–c** The decision maps obtained by three-scale guided filtering. **d** Fused image

Figures 2a–c show the decision maps obtained by three-scale guided filtering. It can be observed that the small-scale map (Fig. 2 (a)) can accurately align the boundaries, and distinguish the small defocused (focused) regions encompassed by the large focused (defocused) regions; however, some small focused (defocused) regions in focused (defocused) regions are incorrectly identified as defocused (focused) regions. On the contrary, the small focused (defocused) regions in focused (defocused) regions can be correctly identified in the large-scale decision maps (Fig. 2b and c), while the boundaries are blurred, and the small defocused (focused) regions encompassed by the large focused (defocused) regions cannot be effectively distinguished. Based on this observation, it is concluded that the brightest and darkest regions in Fig. 2a–c are usually the desired focused and defocused regions, respectively. Hence, the brightest and darkest regions need to be selected in the final decision map, and the other regions can be abandoned. Generalized random walk (GRW) is particularly applicable to this selection problem [5].

## 4 Results

The images used in the multi-focal dataset include 10 set of pair images, where a particular scenario is highlighted. Every set of image describes a scenario via two input images. Building, seascape, books, calendar, ground are set of color images of  $520 \times 520$  resolution displaying less homogeneity. Lab, flower, desk, clock, girl are set of color images of  $640 \times 480$  resolution, displaying high homogeneity and increased structural content, as shown in Table 1. We also use DICOM images which consist of five sets of multi-modal images of  $256 \times 256$  resolution each, where every set of image are inputs from two different medical imaging modalities [10]. Tables 1 and 2 represent the comparison of the image quality parameters of the proposed fusion method with the other conventional methods used for image fusion. Entropy, edge-based structure similarity index measure, spatial frequency, and mutual information are the quality parameters used to analyze the results and the performance of the fusion algorithms.

Tables 1, 2 and 3 depicts a comparison of guided filter (GF), weighted guided image filter (WGIF) and the proposed fusion by Taj-Shanvi framework where multi-

**Table 1** Entropy and edge structure similarity measure Dataset-1 (multi-focus images)

Parameters	Entropy			ESSIM		
	GF	WGIF	Proposed Taj-Shanvi framework	GF	WGIF	Proposed Taj-Shanvi framework
Building	7.8911	7.8174	8.1697	0.5345	0.7722	0.9979
Lab	7.0736	6.9261	7.276	0.8814	0.9441	0.9999
Seascape	7.5275	7.425	7.9203	0.7545	0.9362	1.0000
Flower	7.1805	7.1373	7.5388	0.7784	0.9543	0.9995
Desk	7.2816	7.1997	7.6323	0.7922	0.9394	1.0000
Clock	6.9944	6.955	7.3259	0.8483	0.9699	0.9993
Books	7.4296	7.4885	7.8484	0.946	0.9899	0.9999
Calendar	6.7509	6.6427	6.9415	0.7649	0.9172	0.9997
Girl	7.6536	7.5846	7.9544	0.7088	0.9074	1.0000
Ground	7.5733	7.5695	8.0003	0.8059	0.9209	1.0000

**Table 2** Mutual information and spatial frequency Dataset-1 (multi-focus images)

Parameters	Mutual information			Spatial frequency		
	GF	WGF	Proposed fusion method	GF	WGF	Proposed fusion method
Building	0.5282	0.4509	0.5898	0.5169	0.5501	0.4959
Lab	1.2059	1.0364	1.3838	0.4807	0.4825	0.485
Seascape	0.6645	0.6105	0.677	0.5064	0.513	0.5196
Flower	1.103	0.7628	1.1826	0.4048	0.4126	0.4038
Desk	1.0967	0.8703	1.2919	0.3831	0.3864	0.3879
Clock	1.2139	1.0198	1.3104	0.3065	0.309	0.3091
Books	0.6612	0.642	0.6842	0.3273	0.3306	0.3298
Calendar	0.9352	0.8118	0.9522	0.432	0.4338	0.4344

focal and multi-modal images are taken as the input from Dataset-1 and Dataset-2, respectively.

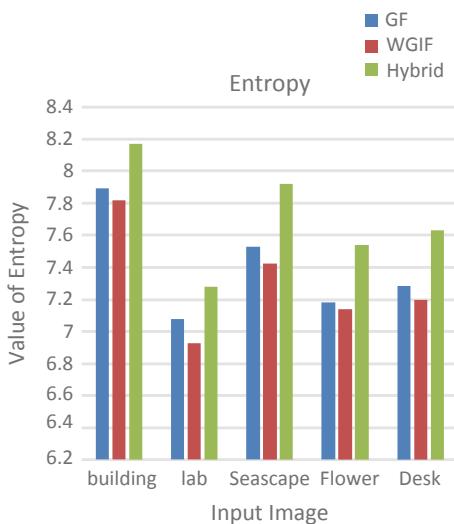
As shown, in Figs. 3, 4, and 5 the proposed hybrid method, Taj-Shanvi framework, is compared with guided filter (GF) fusion method, weighted guided image filter (WGIF) fusion methods.

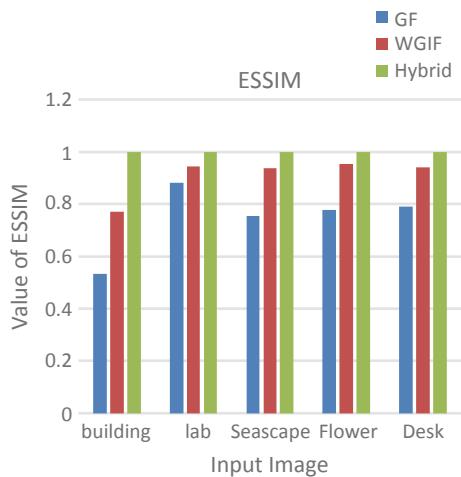
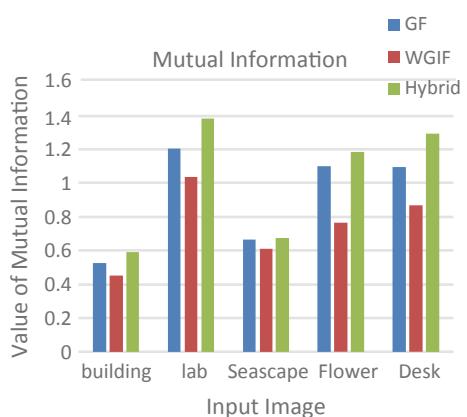
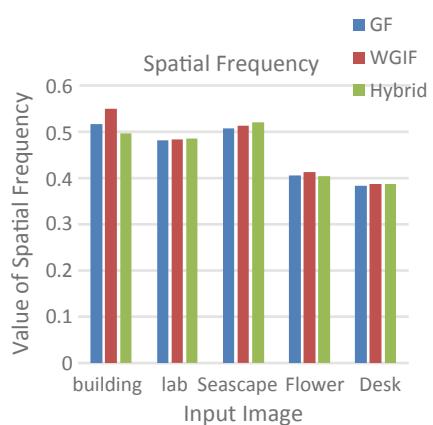
### Dataset-1 (Multi-Focal Images)

Figures 3, 4, 5 and 6 depict a comparison of guided filter (GF), weighted guided image filter (WGIF), and the proposed image fusion by Taj-Shanvi framework in accordance with various quality parameters, where multi-focal images are taken as the input from Dataset-1.

**Table 3** Fused images sample of Dataset1 and Dataset 2 with result

S. no.	Image 1	Image 2	Guided filter fusion	Proposed fusion method
N1				
N2				
N3				
N4				
N5				

**Fig. 3** Entropy comparison

**Fig. 4** ESSIM comparison**Fig. 5** Mutual information evaluation**Fig. 6** Spatial frequency

From the obtained graphs and values, it can be observed that the proposed hybrid approach by Taj-Shanvi framework outperforms guided filter fusion and weighted guided filter fusion methods in almost all the set of images, for both multi-focus and multi-modal image databases.

As weighted guided filter for fusion makes use of a parameter called ‘weight’ and implements an edge-aware technique, it results in higher values of certain parameters that use image sets containing boundary-rich information in comparison to guided filter alone. These observations are clearly cited in the graphical representations depicted in Tables 1 and 2.

The Taj-Shanvi framework used for fusion combines the usage of the watershed technique and guided filtering algorithm, which effectively segregates for boundaries to form the connected and closed regions. The benefit of watershed-based image segmentation is reducing the computational complexity. Quality parameters such as entropy, mutual information, spatial frequency, and edge structure-based similarity measure result in higher values for the proposed method, indicating better fusion results and a high-quality fused image being produced at the end of the fusion technique.

## 5 Conclusion

We have proposed Taj-Shanvi framework based on multi-scale focus measures, generalized random walk (GRW), and the watershed technique. The framework was examined with guided filter (GF), weighted guided image filter (WGIF) on multi-focal and multi-modal dataset and evaluated the fused image quality. The experimental results show that Taj-Shanvi framework outperformed than other methods. This framework can be enhanced further by incorporating more number of images which can assist more in the application of medical imaging involving disease detection and its extent of severity too.

## References

1. [https://en.wikipedia.org/wiki/Image\\_fusion](https://en.wikipedia.org/wiki/Image_fusion).
2. Pajares, G. (2004). A wavelet-based image fusion tutorial. *Pattern Recognition*, 37(9), 1855–1872.
3. Bai, X., Zhang, Y., Zhou, F., & Bindang. (2015). Quadtree-based multi-focus image fusion using a weighted focus-measure. *Information Fusion*, 22, 105–118.
4. Zhang, Y., Bai, X., Wang, T. (2017). Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure. *Information Fusion*.
5. Nayar, S. K., & Nakagawa, Y. (1994). Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8), 824–831.

6. Shen, R., Cheng, I., Shi, J., et al. (2011). Generalized random walks for fusion of multi-exposure images. *IEEE Transactions on Image Processing*, 20(12), 3634–3646.
7. Li, S., Kang, X., & Hu, J. (2013). Image fusion with guided filtering. *IEEE Transactions on Image Processing*, 22(7), 2864–2875.
8. Sarker, M. S. Z., Haw, T. W., & Logeswaran, R.: Morphological based technique for image segmentation. *International Journal of Information Technology*, 14(1).
9. Bhagwat, M., Krishna, R. K., & Pise, V. (2010). Simplified watershed transformation. *International Journal of Computer Science and Communication*, 1(1), 175–177.
10. Nejati, M., Samavi, S., & Shirani, S. (2015). Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25, 72–84.
11. Yang, L., Guo, B. L., Ni, W. (2008). Multimodality medical image fusion based on multi-scale geometric analysis of Contourlet transform. *Euro Computing*, 72, 203211.
12. Singh, S., Gyaourova, A., Bebis, G., & Pavlidis, I. (2004). Infrared and visible image fusion for face recognition. *Proc. SPIE*, 5404, 585596.
13. Kaur, P., & Sharma, E. R. (2015). A study of various multi-focus image fusion techniques. *International Journal of Computer Science and Mobile Computing*, 4(6).
14. Grady, L. (2006). Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11), 1768–1783.
15. Amoda, N., & Kulkarni, R. (2013). Image segmentation and detection using watershed transform and region based image retrieval. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2(2).
16. Sivagami, R., Vaithyanathan, V., Sangeetha, V., Ifjaz, M., Ahmed, K., Sundar, J. A., et al. (2015). Review of image fusion techniques and evaluation metrics for remote sensing applications. *Indian Journal of Science and Technology*, 8(35), <https://doi.org/10.17485/ijst/2015/v8i35/86677>, December 2015.

# **Advances in Network Technologies**

# Effective Classification and Handling of Incoming Data Packets in Mobile Ad Hoc Networks (MANETs) Using Random Forest Ensemble Technique (RF/ET)



Anand Nayyar and Bandana Mahapatra

**Abstract** Mobile ad hoc network is a network type, where wireless network comprises independently moving nodes that cooperatively contribute towards making a network operate successfully. These nodes are tiny electronic devices operating on battery power and have limited operational resources, such as memory, buffer capacity and processing units. These nodes can act as client, server or router, depending upon the network requirement because of the absence of a centralized server control. MANET faces lots of challenges while maintaining a secured and resilient communication due to its limited resources and operational capacity. One of the serious challenges is handling and classifying tons of incoming data at a single point of time for efficient processing. As tons of data enter MANETs network at a single point of time, it is necessary to deploy a new mechanism not only for classification but also for further processing. The heavy and extensive data become unmanageable on the part of mobile nodes while implementing techniques like data analysis to trace anomaly or extracting the required information from the data pool incoming into MANETs. This unmanageable data calls for classification technique so as to segregate data into specific group that can be further utilized for implementing or formulating a specialized mechanism that tends to solve many other research problems. The objective of this paper is to classify the incoming data into MANETs and reduce the data set using the RF/ ET technique. The results showed that the proposed model attained an accuracy level of 86% in handling data packets in MANETs.

**Keywords** Data classification · MANETs · Random forest/ensemble tree · Regression technique · Machine learning

---

A. Nayyar (✉)

Graduate School, Duy Tan University, Da Nang, Vietnam

e-mail: [anandnayyar@duytan.edu.vn](mailto:anandnayyar@duytan.edu.vn)

B. Mahapatra

Computer Science & Engineering Department, Siksha 'O' Anusandhan University, Bhubaneswar, Orissa, India

e-mail: [bandana11@gmail.com](mailto:bandana11@gmail.com)

## 1 Introduction

Mobile ad hoc network (MANET) [1] is considered as the network of today's genre, where the current demand is to get connected anywhere and anytime via any device. This desire is the prime motivation behind the rapid growth of ad hoc networks in the areas of research and development. Many researchers work toward formulating various strategies, algorithms and protocols and perform experiments using new concepts that can solve the inherent problems of MANET due to its ad hoc nature. Ranging from pulling out concepts from tiny creatures belonging to (e.g. ant colony optimization (ACO) [2–4] for MANETs) increases the resource capacity of the electronic devices. For example, researchers are trying their best to attain the goal of a secured, robust and resilient communication medium for the MANETs.

Execution of any new concepts, technology or algorithms into MANETs comes across many challenges, out of which one is the huge unmanageable data pool that streams into MANETs. The data pool incoming into the MANETs not only consists of important information but also tons of unwanted or corrupted information which may not be of any use and makes data handling a tedious task. Extracting relevant information from such an unorganized and chaotic data pool may be time-consuming, as well as may increase the complexity due to the huge search space. Moreover, running a new algorithm on the incoming data packets of MANETs involves lots of computational cost and increases network delay. This calls for the design and development of a new algorithm for MANETs so that data can be segregated into different classes where each class member shares certain common features based on data classification [5].

Data classification is regarded as a process of segregating the data into various categories pertaining to certain characteristics and its metrics involved so that their accessibility becomes easier and more effective in terms of time and complexity involved. The process of classification not only enhances the ease of data location and retrieving but also plays a key role in the fields like risk management, compliance and data security.

The process of classification makes data searching and tracking easy via data tagging and also eliminates the redundancy caused by data duplication, thereby reducing the storage as well as backup costs. The classification is also helpful for meeting the regulatory requirements for business as well as personal activities, including security enforcement that is designed on the basis of types of data that is regularly retrieved, transmitted or copied [6].

Data classification is required to make proper sense of huge quantity of data available at any point of time. This approach provides clarity on stored information, as well as a clear understanding regarding how it can be accessed easily and protected from potential security threats by formulating appropriate defense mechanisms. The data classification can be carried out by using supervised as well as unsupervised methodologies where the supervised or predictive method of classification has sets of possible classes known beforehand, in contrast to unsupervised, descriptive or indirect method that has the sets of possible unknown classes [7].

In this paper, our objective is to apply random forest ensemble technique on incoming packets of MANETs to handle and classify the data in an efficient manner.

### Organization of Paper

Section 2 gives a comprehensive overview of related works performed by other researchers in this area for optimizing data classification and handling in MANETs scenario. Section 3 gives an overview of associated terminologies with regard to the research undertaken. Section 4 highlights the proposed model and gives a detailed presentation of data classification of MANETs using RT/ET technique. Section 5 enlists experimentation and analysis of results. The paper concludes in Sect. 6.

## 2 Related Works

Considering the possible limitations and shortcomings of MANETs in efficient data classification, various researchers have proposed a lot of data classification techniques. Among all these techniques, random forest ensemble technique is highly significant. This section gives a detailed overview of various related techniques proposed by other researchers.

Agarwal and Agarwal [8] highlights all data mining techniques used for anomaly detection proposed by several researchers in different research papers on the basis of decision trees. The authors concluded that ID3, C4.5, GA, SVM are best for anomaly detection.

Chang et al. [9] have emphasized in their article that the concept of randomization pertains to both bagging samples and feature selection, while tree construction in the forest is quite likely to choose the features that are most uninformative for splitting of nodes, thus resulting in low accuracy of the constructed tree while using high-dimensional data. Apart from this, biasness can be observed in the process of feature selection, random toward multi-valued features. Hence, authors have come up with an improvised new  $x$ RF algorithm that can choose appropriate attribute while executing RFs for huge dimensional data. Researchers have also proposed a new means of feature subspace selection toward constructing efficient random forests for classification of high-dimensional data. Further innovative model is proposed for unbiased feature sampling that chooses a set of unbiased features with the aim of node splitting during growth of trees in the forest. The technique also wipes away the redundant attributes (noise) within the data set, resulting in a reduced dimension considering a predefined threshold.

Chen and Ishwaran [10] have contributed a comprehensive review of various applications as well as recent progress in areas of random forest pertaining to genomic data, which includes aspects like prediction, classification, variable selection, pathway analysis, genetic association, pathway, epistasis detection and unsupervised learning.

McClary et al. [11] performed a real-time audio transmission over the MANET. It serves as the initial proof of concept for validation of both the DoEs as a means of

effectively verifying the results of the simulation. The concept basically detects the factor interaction among the network parameters and the decision tree learning as a method of increasing the predictability of end-to-end delay.

Gite and Thakur [12] proposed a machine learning technique for intrusion detection system in MANETs. They made use of the network characteristics and the behavioral differences during the attack. By using both the attack and the normal behavior for training purpose into a machine learning algorithm, the malicious patterns are distinguished according to the new network pattern.

From the literature survey, we can conclude that via the random forest ensemble method of tree construction, all the incoming data packets can be efficiently classified.

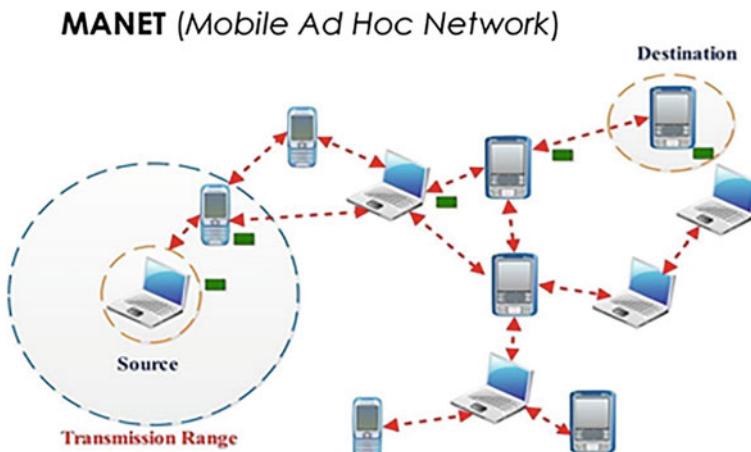
### 3 Related Terminologies

#### 3.1 *Concept of Mobile Ad Hoc Networks*

Mobile ad hoc networks (MANETs) concept was proposed in the year 1972 as the first generation of ad hoc networks. PRNET (packet ratio networks) was proposed as a research outcome for military purpose. It later evolved into SURAN program (Survivable Adaptive Radio Networks) in the year 1980. In the late 1980s, the ad hoc system was further enhanced and implemented as a part of SURAN program. The program was extremely beneficial in improving radio performances by making them smaller, cheaper and robust to counterfeit all sorts of electronic attacks. Finally, in 1990s, the commercial ad hoc network was implemented with notebook computers and other viable communication equipment.

*Mobile ad hoc network*, as the name suggests, is a self-configuring network when infrastructure-less mobile electronic devices are connected together. They are a group of devices that may roam freely in any random direction. Hence, MANETs undergo frequent link changes where every node has to participate in data transmission that is unrelated to itself. Such networks are practically independent in carrying out activities and may be connected to a larger network, making the network highly dynamic with varying topology. MANETs as a self-healing or a self-forming network gained importance in the mid-1990s with the growth of laptops with the 802.11 Wi-Fi standard. Researchers and academicians, since MANETs inception, have contributed by designing efficient protocols in terms of best packet delivery ratio, less overhead, lower end-to-end delay, energy efficiency and high throughput.

MANETs have limited resources in terms of memory capacity, network bandwidth, limited buffer, battery power that compel researchers to design algorithms and protocols to enhance the overall capability of MANET. Figure 1 shows the mobile ad hoc network.



**Fig. 1** Mobile ad hoc network

### 3.2 *Data Classification*

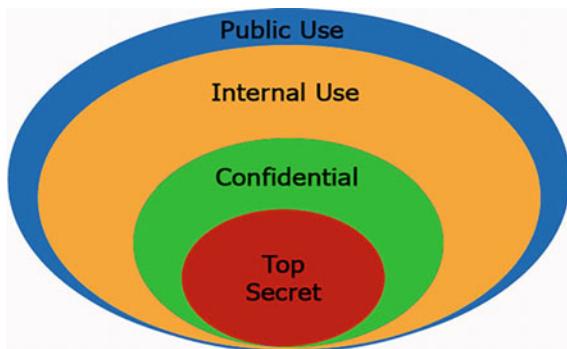
Data classification is regarded as tagging the data with regard to its specification, that is, its type, sensitivity, contribution organization and its repercussions, if manhandled or stolen. Broadly, it can be defined as a method of organizing a considered data into its relevant category which enhances its usage as well as safety. The classification makes the data easy to locate and retrieve. This process is significant when it comes to risk management, compliance and data security. In a nutshell, data classification can be elaborated as a data mining technique that categorizes the unstructured data format into a structured form and groups the data in proper fashion that makes the knowledge extraction as well as discovery easier. Data classification is regarded as a model of intelligent decision-making regarding solving all issues regarding packet classification in real-time scenarios.

The data classification process comprises two phases: learning and execution. The process of learning involves analyzing huge amounts of training data sets for extracting and creating data rules and patterns. Figure 2 demonstrates data classification based on sensitivity.

### 3.3 *Machine Learning*

Machine learning is a sub-discipline, belonging to artificial intelligence, that makes the system capable of learning and improving from the experience rather from programming. It focuses on the development of computer programs that can retrieve the data and utilize it in its learning process [13, 14]. Here the learning process begins with data study, direct experience and instructions to discover patterns that help

**Fig. 2** Data classification based on data sensitivity



them to make better decisions in the future that are built upon the examples that are provided. Machine learning algorithms can be categorized into four categories: supervised machine learning, unsupervised machine learning, semi-supervised machine learning and reinforcement machine learning algorithms.

Supervised method is a learning methodology that requires a labeled training data set for inference purpose, whereas unsupervised method of learning classifies or describes the hidden data structure from an unlabeled data where the classification or categorization is not present in the observation set.

The objective of classification in MANETs is analysis of huge data and attributes to select appropriate features to accurately describe each class using this feature present in the data. In this research paper, we have adopted the method of supervised learning to classify the incoming data attributes and label them into class and records.

The supervised learning can be carried out using the decision tree (DT) method or the support vector machine (SVM) method.

### 3.4 Decision Tree

Decision tree analysis is a general, predictive modeling tool that has a wide range of application spanning over a number of areas. They are constructed through an algorithmic approach that identifies different methods of splitting a data set based upon different conditions. A decision tree can be defined as a multiple variable analysis that allows us to predict, explain, describe or classify an outcome. The attribute of multiple variable analysis of a decision tree provides the option of going beyond a simple one cause, that is, relationship, and discover as well as describe things in the context of multiple influences. The concept of multiple variable analysis has been significant in current problem solving, since all the critical outcomes are based on multiple factors. The decision tree may be described as a filtration method that can handle huge quantity of data for classification using the filtration technique. It can be considered as the basic method or as the classical approach of classification that gives satisfactory results in terms of efficiency and data accuracy. A decision

tree algorithm proves beneficial at tuning between the precision that can be trained very fast and providing sound results of the classified data. Hence, we select the DT method of classification to classify the data incoming into MANETs.

The decision tree ensemble technique also stated as “random forest technique” is primarily utilized in the process of feature selection and classification. The process produces a multiple number of precisely constructed trees set formed with respect to a target feature (Fig. 3) [5].

## 4 Data Classification in MANETs Using RF/ET Technique

Figure 4 highlights the conceptualized model used for performing the task of data processing. The incoming data into the node consists of 42 attributes that occupy the data storage buffer, after which this raw data is sent for further processing by nodes processing unit. Considering this heavy data collection for implementing new algorithm or technique or a new computation, security enforcement or data retrieval proves to be complicated, heavy resource consuming as well as slow. So, this laid the

**Given the training set  $X = x_1, x_2, \dots, x_n$ . With response  $Y = y_1, \dots, y_n$ , bagging repeatedly ( $B$  times) selects a random sample with replacement of the training set and fits trees to these samples.**

**The Random Forest Training Algorithm is given as:**

**For  $b = 1, \dots, B$**

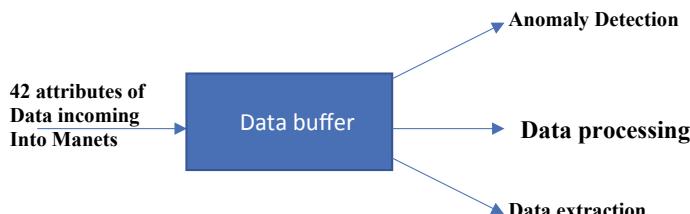
1. Sample, with replacement,  $n$  training examples from  $X, Y$ ; call these  $X_b, Y_b$ .
2. Train a decision or regression tree  $f_b$  on  $X_b, Y_b$ .

After training, prediction for unseen samples  $x^1$ , can be made by averaging the predictions from all the individual regression tree on  $x^1$ .

i.e.:  $\hat{f} = 1/B \sum_{b=1}^B \hat{f}_b(x^1)$  or by taking majority vote in case of decision tree.

Here we aim to classify the incoming data into Manets using Decision Tree/ Ensemble Technique Method

**Fig. 3** Decision tree/ensemble technique algorithm



**Fig. 4** Flow of data being processed in a node

foundation to consider RF/ET technique of data classification that helps in organizing the data into classes, thereby removing the noise factor and making data handling easy. Figure 5 represents the proposed model.

In the proposed model, all the incoming data are stored into the node memory buffer and undergo RF/ET method of classification, giving a classified, organized as well as an arranged form of data as per its sensitivity. The classified data are used for further processing for enforcing security-based algorithms, like data snooping, anomaly detection, identifying attack patterns, etc., or data extraction via effective data mining techniques, and the model is highly fast in performance. Figure 6 highlights the flowchart of the proposed method.

### Algorithm for Proposed Model

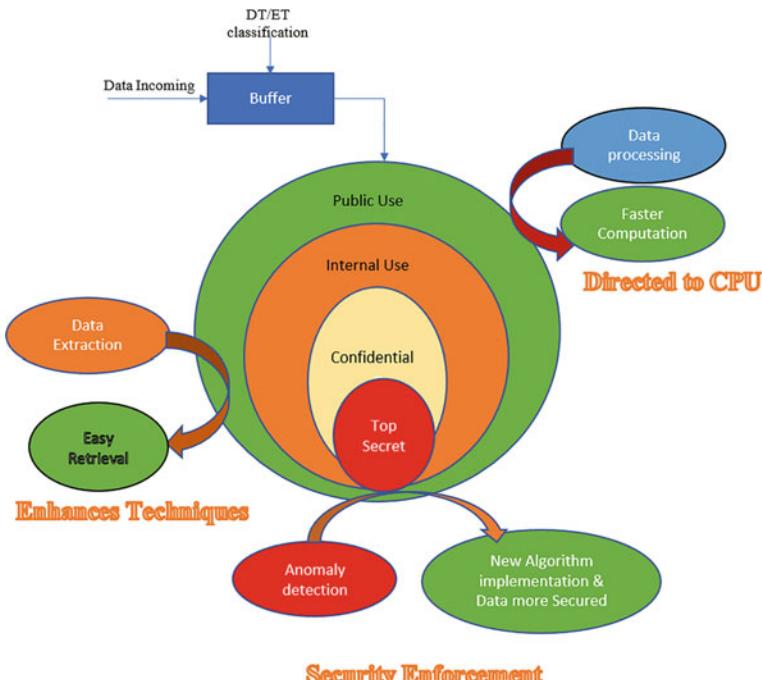
#### Algorithm: RF/ET classification on incoming data

**Begin**

**Step 1.**  $X[] \leftarrow$  Incoming data into MANET buffer

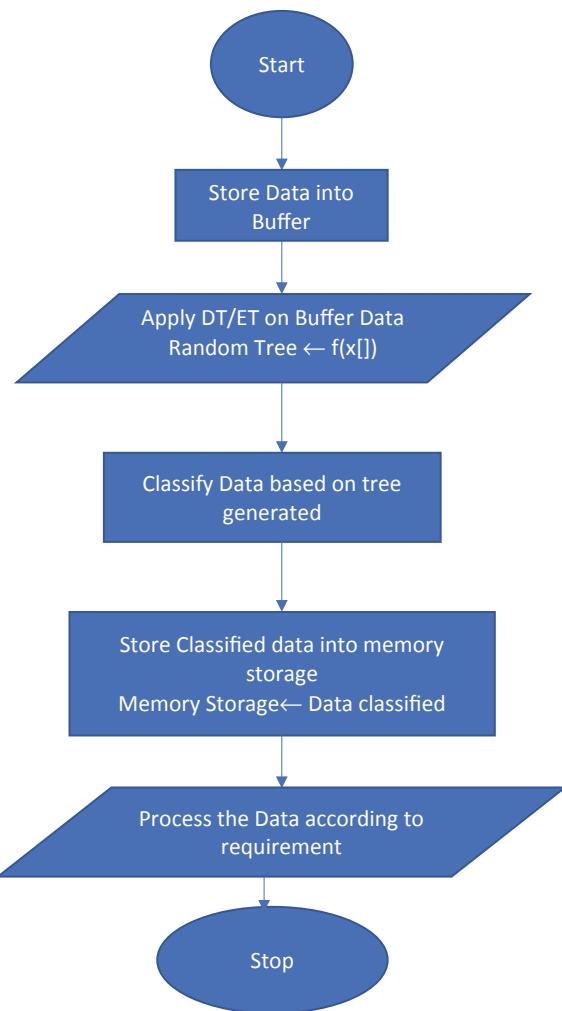
**Step 2.** Apply DT/ET on buffered data, i.e.

Random tree =  $f(x[])$



**Fig. 5** Data classification model using RF/ET technique

**Fig. 6** Flowchart of the proposed model



**Step 3.** Classify data based on generated tree

**Step 4.** Store classified data into memory storage

Memory storage [] ← Classified data

**Step 5.** Consider stored data for further processing

**End.**

## 5 Simulation and Experimental Results

To classify the incoming data into a node within MANET environment, the NSL-KDD-99 data set was simulated using classification tree and regression tree algorithm in MATLAB.

### 5.1 *Simulation Environment and Parameters*

The simulation environment and parameters used for the experiment are shown in Table 1.

### 5.2 *Data Set*

The data set used here is NSL-KDD-99, considered as the primary benchmark for researchers, especially for experimenting in the fields of intrusion detection techniques in MANETs. It is the next version of NSL-KDD Cup-99 data set which solves quite a number of problems. The data set is said to be the standard benchmark for intrusion detection consisting of 42 attributes with 150 tuples. Table 2 shows the detailed classification of the attributes belonging to the NSL-KDD-99 data set.

**Table 1** Simulation parameters used for the experiment

Operating system	Microsoft Windows 8
Simulation Tool	MATLAB, version-16 (b)
Algorithm	Classification tree and regression tree
Attribute set	42 × 150
Type	Classification
Method	Tree
Split criteria	“gdi”
N val to sample	All
Merge leaves	On
Prune	On
N surrogate	0
Max Cat	10
Alg Cat	“auto”
QE Toler	[]
Stream	[]

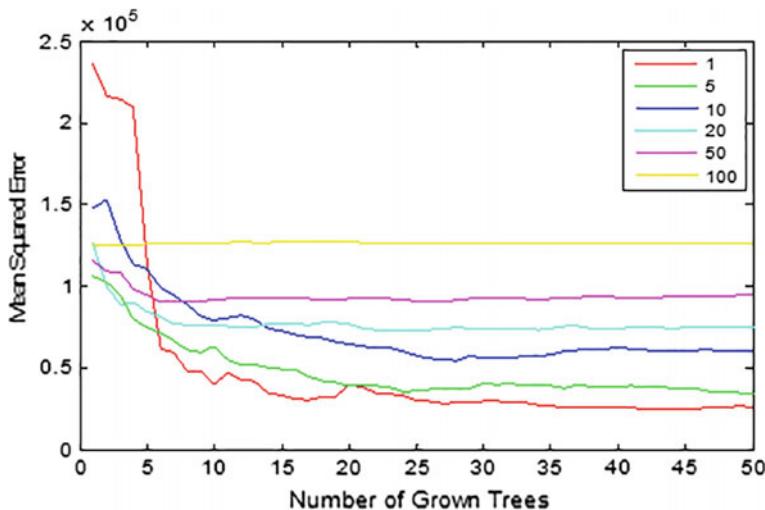
**Table 2** Attributes with class labels in NSL-KDD data set

S. no.	Label	Attribute name	S. no.	Label	Attribute name	S. no.	Label	Attribute name	S. no.	Label	Attribute name
1	B	duration	10	C	hot	23	T	count	32	H	dst_host_count
2	B	protocol_type	11	C	num_failed_logins	24	T	sevorr_rate	33	H	dst_host_srv_count
3	B	service	12	C	logged_in	25	T	renor_rate	34	H	dst_host_same_srv_rate
4	B	src_bytes	13	C	num_compromised	26	T	same_srv_rate	35	H	dst_host_diff_srv_rate
5	B	dst_byes	14	C	root_shell	27	T	diff_srv_rate	36	H	dst_host_same_src_port_rate
6	B	flag	15	C	su_attempted	28	T	srv_count	37	H	dst_host_srv_diff_host_rate
7	B	land	16	C	num_root	29	T	stv_sevorr_rate	38	H	dst_host_sevorr_rate
8	B	wrong_fragment	17	C	num_file_creations	30	T	srv_errror_rate	39	H	dst_host_srv_errror_rate
9	B	urgent	18	C	num_shells	31	T	srv_diff_host_rate	40	H	dst_host_errror_rate
			19	C	num_access_files				41	H	dst_host_srv_errror_rate
			20	C	num_outbound_cmds				42	-	Class
			21	C	is_hot_login						
			22	C	is_guest_login						

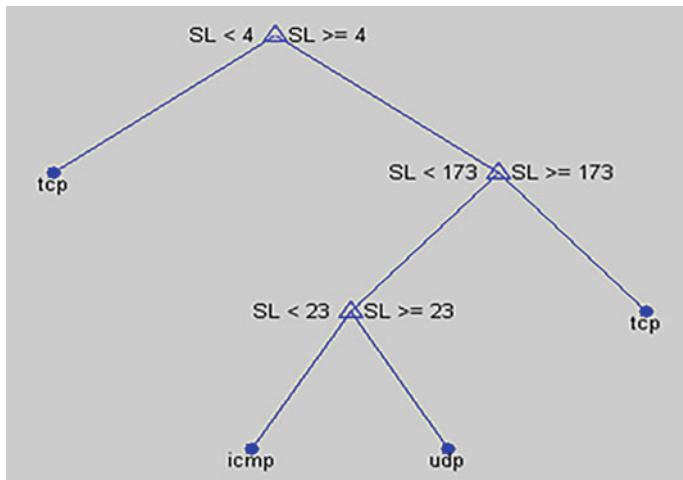
### 5.3 Results

In Fig. 7, the red curve shows the lowest MSE value for the smallest leaf value of the tree taken as 1 consisting of four to five features. As the number of features grows, so does the leaf value, which depicts gradual decrease in the value of MSE.

The path rule extracted from the tree diagram shown in Fig. 8 is given in Table 3.



**Fig. 7** Number of random trees generated using RF/ET technique



**Fig. 8** Random tree generated using RF/ET technique on sample data

**Table 3** Path rule for resultant classification tree

Path	Rule
D1	If SL < 4 then class “tcp”
D2	If SL $\geq$ 4 and SL < 173 and SL < 23 then class “icmp”
D3	If SL $\geq$ 4 and SL < 173 and SL $\geq$ 23 then class “udp”
D4	If SL $\geq$ 4 and SL $\geq$ 173 and hen class “tcp”

## 5.4 Result Analysis

The random forest ensemble technique shows a reduction rate of 86% when applied over the attribute set, giving an accuracy rate of 76%, which confirms with the reduction as well as accuracy rate mentioned in the article [5], “Seven Techniques for Dimensionality Reduction”. The RF/ET method applied over the NSL-KDD-99 data set organizes the informative data into groups or classes, as shown in Fig. 7, which is further elaborated and classified in the form of decision tree with the sub-classification discriminating classes at splitting points, as seen in Fig. 8 into TCP, ICMP and UDP. The node split points and path rule for construction of classification tree are shown in Table 3.

## 6 Conclusion

The research paper discusses the importance of data preprocessing and how classification and proper organization of data improves the data quality as well as enhances the complexity of the algorithms implemented over the data, in terms of both time and space. Various research contributions have been discussed in the state-of-the-art. The paper also discusses in detail the various terminologies and their concepts, which include unsupervised machine learning and its subclass random forest ensemble technique with its benefits in regards to data preprocessing and classification. Finally, the experimental evaluation is conducted where preprocessing and classification using the random forest ensemble technique is done on NSL-KDD-99 data set. The resultant data are classified into four classes, namely TCP, UDP and ICMP, giving 76% as the rate of accuracy and 86% as the rate of reduction, which complies with the comparative results obtained in the state-of-the-art, where it also shows better accuracy as well as reduction rate in comparison to other classical data reduction approaches. The future scope may aim at using the advanced classification algorithms in order to categorize as well as organize the data, improving efficiency in the fields of data handling and management, as well as study the performance gain subject to comparative analysis and review performed over the classical methods of classification techniques.

## References

1. Kaur, M., & Nayyar, A. (2013). A comprehensive review of mobile ad hoc networks (MANETS). *International journal of emerging trends & technology in computer science (IJETTCS)*, 2(6).
2. Nayyar, A., & Singh, R. (2014). A comprehensive review of ant colony optimization (ACO) based energy-efficient routing protocols for wireless sensor networks. *International Journal of Wireless Networks and Broadband Technologies (IJWNBT)*, 3(3), 33–55.
3. Nayyar, A., & Singh, R. (2017). Ant colony optimization (ACO) based routing protocols for wireless sensor networks (WSN): A survey. *International Journal of Advanced Computer Science and Applications*, 8, 148–155.
4. Nayyar, A., & Singh, R. (2016, March). Ant colony optimization—Computational swarm intelligence technique. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1493–1499). IEEE.
5. Silipo, R., Adae, I., Hart, A., & Berthold, M. (2014). Seven techniques for dimensionality reduction. *White Paper by KNIME. com AG*, 1–21.
6. Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5), 2297–2307.
7. Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. G., & Andina, D. (2010, November). Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In *IECON 2010-36th annual conference on IEEE industrial electronics society* (pp. 2845–2850). IEEE.
8. Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60, 708–713.
9. Chang, Y. H., Gray, J. W., & Tomlin, C. J. (2014). Exact reconstruction of gene regulatory networks using compressive sensing. *BMC Bioinformatics*, 15(1), 400.
10. Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6), 323–329.
11. McClary, D. W., Syrotiuk, V. R., & Lecuire, V. (2008). Adaptive audio streaming in mobile ad hoc networks using neural networks. *Ad Hoc Networks*, 6(4), 524–538.
12. Gite, P., & Thakur, S. (2015). An effective intrusion detection system for routing attacks in manet using machine learning technique. *International Journal of Computer Applications*, 113(9).
13. Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, Berlin.
14. Dietterich, T. G. (2002, August). Machine learning for sequential data: A review. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (pp. 15–30). Springer, Berlin, Heidelberg.

# Community Structure Identification in Social Networks Inspired by Parliamentary Political Competitions



**Harish Kumar Shakya, Nazeer Shaik, Kuldeep Singh, G. R. Sinha  
and Bhaskar Biswas**

**Abstract** Revealing the concealed community structure, that is vital to understanding the options of networks, is a vital drawback in network and graph analysis. Throughout the last decade, several approaches are projected to resolve this difficult drawback in various ways in which there are totally different measures or knowledge structures. The social network is very popular in the current scenario. Owing to various usefulness in daily routine life, it is a very hot area of research in the current environment. Community structure identification is one of the well-known research works in the social networks. In this research problems are solved by many researchers through different techniques, that is, data mining techniques, soft computing, evolutionary and swarm algorithms. In this experiment, we introduced a novel algorithm called political competition algorithm (PCA) inspired by the Indian parliamentary election procedure. We validate the outcome of the political competition algorithm with the help of various well-known dataset, that is, American Football Club, Books on US Politics, Karate Club and Strike datasets.

**Keywords** Community detection · Genetic algorithm · Modularity · Political competition algorithm · Social networks

---

H. K. Shakya (✉) · N. Shaik

Department of Computer Science & Engineering, Bapatla Engineering College,  
Guntur, India

e-mail: [hkshakya.rs.cse@iitbhu.ac.in](mailto:hkshakya.rs.cse@iitbhu.ac.in)

N. Shaik

e-mail: [nazeer.shaik@becbapatla.ac.in](mailto:nazeer.shaik@becbapatla.ac.in)

K. Singh · B. Biswas

Department of Computer Science & Engineering, Indian Institute of Technology (BHU),  
Varanasi, India

e-mail: [kuldeep.rs.cse13@iitbhu.ac.in](mailto:kuldeep.rs.cse13@iitbhu.ac.in)

B. Biswas

e-mail: [bhaskar.cse@iitbhu.ac.in](mailto:bhaskar.cse@iitbhu.ac.in)

G. R. Sinha

International Institute of Information Technology (IIIT), Bangalore, India

e-mail: [drgrsinha@ieee.org](mailto:drgrsinha@ieee.org)

## 1 Introduction

Social networks are mapping of associations and flows between individuals, groups, organizations and any other entities. The vertices in the networks denote persons or entities while the edges denote connections or flow among the nodes. We analyze such networks in order to gain a visual and mathematical idea of relationships among humans and organizations.

A subset of vertexes in a graph is called a community [1] if all the nodes in the subset satisfy a property of relative cohesiveness. The partition of the set of nodes into such communities is called a community structure. In social networks, the process of generating community structures based on certain properties and degrees of cohesiveness among the nodes is called community detection [2].

Community detection may also result in partitions of the graph which are not necessarily strict. This implies that overlapping communities exist where graph nodes belong to more than one community of partition. The presence of overlapping communities suggests that a node has a membership value or degree associated with each community. Each node has a degree of membership that ranges from 0 to 1 for each community. These degrees are used to determine the communities that overlap and also the overlapping of nodes [3].

In this paper, we have used the political competitions algorithm (PCA) for community detection in social networks. During this experiment, we are trying to find the best proposed algorithm for disjoint community detection through an algorithm. In this experiment, PCA is a political competition algorithm based on the Indian parliamentary election procedure. Everybody knows that democratic election process is the best way to find the quality of a person among the crowds. First, we study the election process of the Indian democracy and then convert into an algorithm format.

The first step in the planned methodology is initialization of the population. In this step we follow two internal procedures: assigning communities to individuals and another one is distributing population into parties, which is the complete initialization process of our proposed algorithm. Therefore, it is a straightforward and efficient algorithm for disjoint community structure identification in social networks. The second step of the proposed algorithm is internal party elections. In this process we find the best individuals for the party management and prepare the candidate for the inter-party election. After the second step we find the party leaders and followers. Third and last step is called the inter-party election and main election. At the end of the above three steps of the proposed algorithm, the best leader among all party leaders is selected as solution for the particular iteration. In addition, the present population is sent for party elections for next iterations. A detailed explanation of the political competition algorithm is given in the Proposed Method section.

## 2 Literature Review

A number of algorithms have been developed to detect the community structure in social networks, such as classical algorithms, FN [4] and GN [5], as well as existing high-accuracy algorithm, TGA [6]. Many modified variety of genetic algorithms are available for overlapping and disjoint community detection such as CONGA [7], CPM [8], GA-Net+ [9, 10], GaoCD [11, 12], and so on. GaoCD is a method based on genetic algorithm for identifying the overlapping community with link clustering. The equation starts and finds the connection in networks by upgrading the objective value and segment thickness (D) [13] that point delineate connection networks to hub networks upheld a special genotype representation technique. The quantity of networks recognized by GaoCD is frequently mechanically decided, with no past information. GA-Net+ [9, 10], arranged by Pizzuti, first embraces the hereditary run to locate covering networks. It proposes a procedure to exchange hub diagram to line chart, among which hubs approve edges of the hub diagram, through edges approve contiguous connections of edges of hub chart. The line diagram is then utilized on the ground that the contribution of the hereditary administers, and in each age, the line chart is exchanged to the hub chart to measure the wellness. When picked, the chart is exchanged yet again for back-to-back emphasis of GA. The change between line chart and hub diagram costs bottomless calculation and lessens the adequacy. CPM is the most remarkable and broadly utilized equation. In any case, CPM incorporates a strict network definition and is not sufficiently flexible for genuine system. Once the system is simply excessively thick, CPM discovers substantial inward circle networks; in any case, once the system is excessively inadequate, it finds no inner circles, even the slightest bit. Also, along these lines, the inclusion of CPM, generally, relies upon the element of the system, giving no world viewpoint to the total system. There are diverse calculations for covering network identification, determining the SCP of Kumpula [14] Lancichinetti's equation [15], and so on. Each one of them might need past data, or have an inclusion issue, or suffer from effectiveness. Differential evolution algorithm is also good for the community structure identification in social networks. It does not require prior information but faces some other problems related to the real-world datasets [16].

## 3 Proposed Method

The proposed algorithm is inspired from the basic parliamentary election methods adopted by most of the democratic countries and is modified to identify the communities in various social network graphs. For better results, node similarity is used to improve the fitness of the population. Since the proposed algorithm is inspired by parliamentary elections, various steps in the algorithm are similar to day-to-day democratic elections [17].

In these elections and also in our algorithm, the total population is divided into various groups called parties, and party elections are contested within the parties. Among the population of a party, top individuals are selected and are called party leaders. These party leaders try to impact the rest of the population of their parties. Rest of the population in the party, other than leaders, is often termed as followers. Influencing the followers not only improves their fitness but also improves their chances to get selected as a leader in the upcoming iterations. After the party elections, overall fitness of every party is evaluated. This party fitness is a weighted average of fitness of leaders and followers [18].

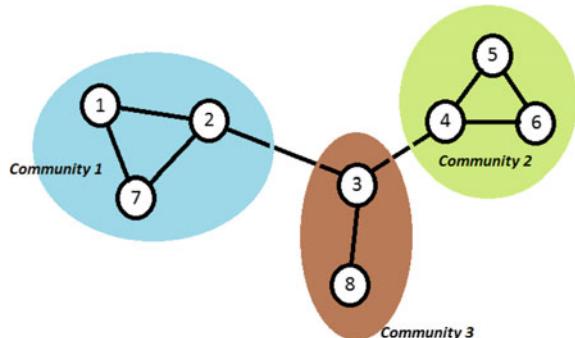
In any democratic elections, weaker parties merge themselves to form a bigger and powerful party or alliance. Merging helps them in competing with other powerful groups. Likewise, in the proposed algorithm, parties with less party fitness tries to merge with a merging probability  $P_m$ . Similarly, the party with worst party fitness is dissolved with a probability  $P_d$ . Since dissolving any party is not a regular practice, probability of dissolving ( $P_d$ ) is quite low [19].

**Representation of individuals:** For implementing this method for community structure identification in social graphs, we tried to initialize each individual as a column vector of dimension  $N$ , where  $N$  is the number of vertices in our social graph (Fig. 1). This vector is filled with the community numbers. For example, for the given graph of eight nodes, the individual will be shown as Fig. 2.

Therefore, any individual is a potential solution for community detection in the given graph with its elements representing their corresponding communities. Fitness value of the individual is calculated with the help of Q-function value.

Hence, any random permutation of numbers in an individual represents a type of community distribution in the graph. And its Q-function value is used to determine the fitness of the distribution.

**Fig. 1** A graph of 8 nodes with 3 communities



1	1	3	2	2	2	1	3
Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 8

**Fig. 2** Array of nodes with community numbers filled in the array

### 3.1 Parliamentary Competition Algorithm

Parliamentary competition algorithm is an iterative algorithm which tends to find a better solution as the iteration proceeds. This algorithm is classified into three major parts:

- Population initialization
- Internal party elections
- Inter-party election

Population initialization is performed once and this population is modified iteratively in the complete algorithm. Internal party elections and main elections are performed iteratively and these operations bring changes in the populations to improve the fitness values [3].

- Population Initialization

Parliamentary competition algorithm starts with the initialization of population of size  $P_n$  with a random distribution of communities. For population initialization, a two-dimensional matrix is formed with size  $P_{nx}$ , where  $n$  is number of nodes in the graph. In this two-dimensional matrix, every column represents an individual.

- Assigning communities to individuals: For every individual, a random number  $R$ , is selected between 2 and  $n - 1$ . This randomly generated number represents the number of communities in the corresponding individual. The range of this number is selected between 2 and  $n - 1$  so as to ensure that there are at least 2 and maximum  $n - 1$  communities. This is done to rule out the possibility of worst cases: all the nodes are in the same community (only 1 community) or all the nodes are in different communities ( $n$  communities).

For every individual, the column matrix is filled randomly with numbers between 1 and  $R$ , where  $R$  is a randomly generated number for that individual in the previous step. This random distribution of numbers in column matrix represents the random allocation of communities to the nodes.

Hence, after the completion of this step, every individual (column matrix) in the population of  $P_n$  is a potential solution to the optimization problem.

- Distributing population into parties: The entire population of  $P_n$  is divided equally into  $M$  political parties. Hence, every population gets  $L$  individuals.

$$L = P_n / M \quad (1)$$

- Internal Party Elections

For party elections, the following steps are followed:

1. Among the members of the party, top members are selected as leaders and rest of the population is called followers. To find leaders, Q-function value of every member is calculated. Top one-third members are termed as leaders.
2. The selected party leaders try to influence the followers. The followers tend to follow the leaders and try to move toward the leaders. In an N-dimensional

vector space, leaders represent the points with better fitness value. As the followers tend to move toward the leaders, their fitness value improves. This is done using the following formula:

$$P'_0 = P_0 + \eta \frac{\left\{ \sum_{i=1}^{\theta} Q(P_i)(P_i - P_0) \right\}}{\sum_{i=1}^{\theta} Q(P_i)} \quad (2)$$

Here,

$P_0$  is the original N-dimensional vector of the follower

$P'_0$  is the updated N-dimensional vector of the follower

$P_i$  (where  $i \in [1, \theta]$ ) are the  $\theta$  leaders in any political party.

3. Once the followers are influenced, we perform the following operations to validate the newly generated vector of each follower.

- After the computation of above formula, the values are not necessarily integers. Hence these values are rounded off to their nearest integers.
- We check if there is any community with no nodes. Such communities are deleted and adjustments are made in community number of other communities. For example,

If the N-dimensional looks like-

1	3	3	1	4	4	3	4	1
---	---	---	---	---	---	---	---	---

In the above example, community 2 does not have any node. Hence this problem is validated as:

1	2	2	1	3	3	2	3	1
---	---	---	---	---	---	---	---	---

After party elections, the fitness value of followers changes and is improved.

- Inter Party Election

Overall elections in the population are carried out in the following steps:

1. Evaluating power of existing parties. This is calculated using weighted average of fitness of leaders and followers of parties. Power of  $i$ th party is given by:

$$\text{Power}_i = \frac{\{m \times \text{avg(Leader)} + n \times \text{avg(Follower)}\}}{m + n} \quad (3)$$

2. With a probability of  $P_m$ , the worst two parties are selected and merged to form a single party of bigger population.  $P_m \in (0,1)$

3. With a probability of  $P_d$ , worst existing party is dissolved by deleting the party population of the particular party,  $P_d \in (0,1)$ , although dissolving any party is not a regular practice. Therefore, the value of  $P_d$  is generally taken as <5%.
4. All the leaders of the parties try to improve themselves. This is done using node similarity principle. For every leader from all parties, random nodes are chosen with a probability  $P_f$  and all the nodes connected to this chosen node are assigned the same value as selected node. Physically, we are assigning same community to connected nodes.

At the end of the above four steps of main election, the best leader among all party leaders is selected as solution for the particular iteration. And the present population is sent for party elections for next iterations.

## 4 Experimental Result and Discussion

### 4.1 Experimental Description

The proposed work was kept running on a Microsoft Windows 10 ( $\times 64$ ) working framework utilizing MATLAB 11 programming platform; Intel® Core™ i5-3230 M CPU @2.60 MHz processor, 3.00 GB memory and 1 TB hard disk. The constraint values of the experiment fix is shown earlier. These could be altered as per the situation, depending on the size of the network, variation in convergence time and extent of overlap required among communities. We have employed the well-known dataset for this experiment. Table 1 gives the following details:

- Strike: Michael considered workers' strike designs in a wood-handling office later than new administration proposed changes to the worker's compensation packages. The investigation depended on age and racial gathering. Set of 24 workers were assembled into three such gatherings relying upon 34 relationships between workers amid various timetables of strikes [20].
- Karate: This dataset is about the examination of a karate club association by Zachary. The system comprises 34 people from a karate club as centre points and 78 relationships among people speaking to kinships in the club which was seen

**Table 1** Details of datasets

Datasets	Nodes	Edges
Strike	24	38
Karate Club	34	78
Dolphin	62	159
Books on US Politics	105	441
American College Football	115	613

over a period of 2 years. This dataset is a social network of kinships between 34 individuals from a karate club at a US university in the 1970s [21].

- Dolphin: This is a coordinated social network of bottlenose dolphins. The vertices are the bottlenose dolphins of a bottlenose dolphin network living off Doubtful Sound, a fjord in New Zealand. An edge demonstrates a continuous affiliation. The dolphins were seen somewhere in the range of 1994 and 2001. The network was isolated into two clusters relying upon the affiliation examples of dolphins [22].
- American College Football: This system is a recreation arrangement in view of an American college football standard season plan in the fall of 2000. The hubs speak to groups, and an edge demonstrates that matches are played between the two associated groups. In the general season, 115 groups go to 12 gatherings of various sizes. The lion's share of matches is played between groups inside a similar meeting; hence the 12 gatherings establish the system's 12 genuine networks [23].
- Books about US Politics: A network of books about US political issues distributed around the season of the 2004 presidential election decision and sold by the online book shop Amazon.com. Edges between books represent frequent co-purchasing of books by the same buyers. The network system was compiled by V. Krebs and is unpublished, however, can be found on Krebs' website [23].

## 4.2 Experimental Analysis

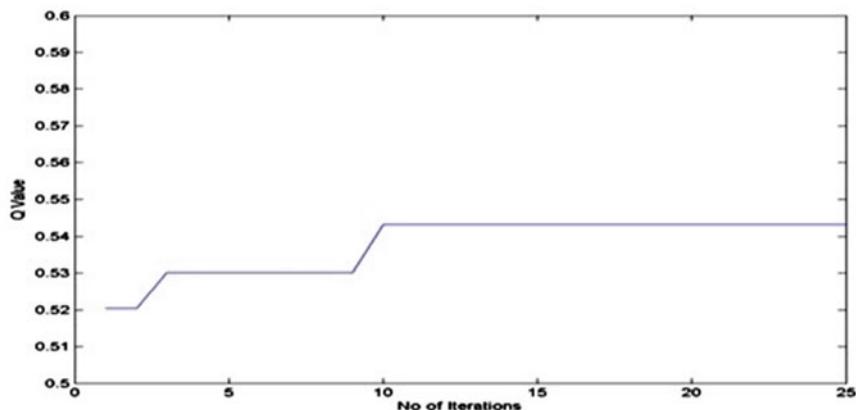
In political competition algorithm, we perform the three basic internal operations: population initialization, internal party election and, lastly, inter-party election (main election). In the first part for population generation, we generate a random population in two-dimensional matrix, assign the communities and distribute the population into the parties. These two tasks have completed in this particular step. The second part is the internal party election; in this step we update the population according to the fitness value. In this algorithm we used the modularity as a fitness function. Modularity is the well-known fitness function in the world of social networks. The range of value is between 0 and 1. We have divided the total population into two parts: first is the leaders and second is the followers based on the modularity fitness function. One-third of the total population are the leaders and remaining are the followers. After every internal party election, update the value of fitness function and vary the leaders and followers according to the fitness value. The third and the last step is evaluating the power of existing parties according to the average fitness value of the followers and leaders. All leaders from all the parties try to improve the power with the help of node similarity concept. After the end of the above steps we found the best leaders for all parties and remaining population again go for the internal party election for next iteration. Till the end of the complete iteration, we found the best leaders and suitable community.

In Table 2, we have shown that the PCA algorithm performs better compared with other traditional techniques. In the given table, value represents the fitness value and we have used the modularity as a fitness function. In Strike dataset, the PCA algorithm has performed the best value of 0.584, remaining FN, GN and TGA found the modularity fitness value as 0.501, 0.534 and 0.565, respectively. Traditional algorithms FN, GN, TGA are the best algorithms in the evolutionary algorithm category. Similarly in Books on US Politics dataset, PCA found the best modularity value of 0.525 compared to other algorithms; FN modularity value is 0.502 and GN fitness value is 0.516. TGA got the nearest fitness value, 0.524 of the PCA. In the same case for the Karate Club dataset, PCA has performed the best and found the highest value of 0.442 compared to other algorithms. In the American Football dataset, PCA got the highest modularity value in the given table 0.599 compared to other values. It means that the PCA algorithm performed the best for the large dataset because in Table 1, American Football is the largest dataset and second largest is the Books on US Politics. Both have performed the best and found the highest modularity value. According to Table 1, we know that the Strike and Karate datasets are the smallest in the given table. PCA performed the best for the smallest and largest dataset compared to evolutionary and the traditional algorithms. In the Dolphin dataset, PCA has got the modularity value of 0.497 and TGA got the 0.524, remaining are the GN 0.470 and FN 0.371. We know that in this case PCA is not the best value but got the second position and compared to GN and FN performance. In Table 2, we used totally five datasets and PCA performed the best in four datasets and got the second position in remaining one dataset. The overall performance of PCA is good compared to other algorithms. According to experiments, political competition algorithm (PCA) is a good option for the next future in community detection in social networks.

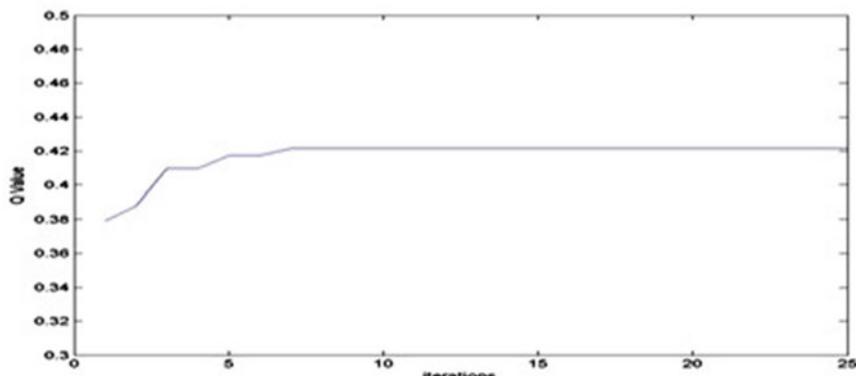
In this experiment, PCA performed the best in four datasets; the graphical representation of the output value for the different datasets is given. Figure 3 shows the smallest Strike dataset and graph represents the modularity value according to the number of iterations. In this experiment, we run all the given algorithms on 25th iteration. Figure 4 represents the modularity value of Karate Club dataset. Similarly Fig. 5 shows the graphical representation of the modularity value for Dolphin sociality dataset. In this graph the value is too much fluctuated. If we increase the number of iterations then we will get good value of modularity for the PCA algorithm. Figures 6 and 7 show PCA performed the best compared to other algorithms, and graphical

**Table 2** Modularity values on various datasets using PCA

Datasets/algorithms	FN	GN	TGA	PCA
Strike	0.501	0.534	0.565	<b>0.584</b>
Karate Club	0.252	0.401	0.403	<b>0.442</b>
Dolphin sociality	0.371	0.470	0.524	0.497
Books on US Politics	0.502	0.516	0.524	<b>0.525</b>
American College Football	0.454	0.599	0.593	<b>0.599</b>



**Fig. 3** Modularity versus no. of iterations for strike dataset



**Fig. 4** Modularity versus no. of iterations for Karate Club dataset

representation is also drawn in it. According to Table 2, PCA has performed the best compared to other given algorithms.

## 5 Conclusion and Future Work

In this experiment, a unique community detection method, parliamentary competition algorithm, that tries to optimize network modularity exploitation with fitness has been proposed. To the simplest of our information, it is the first time PCA has been applied to community detection in social network issues. Though PCA is first projected and not any modifications or additions are performed to the algorithm, the effective experimental results obtained from the real-world datasets are promising.

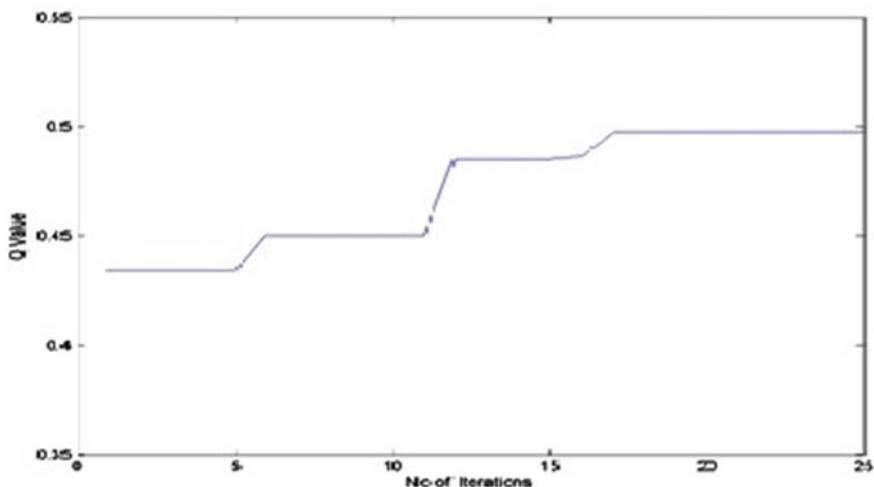


Fig. 5 Modularity versus no. of iterations for Dolphin Sociality dataset

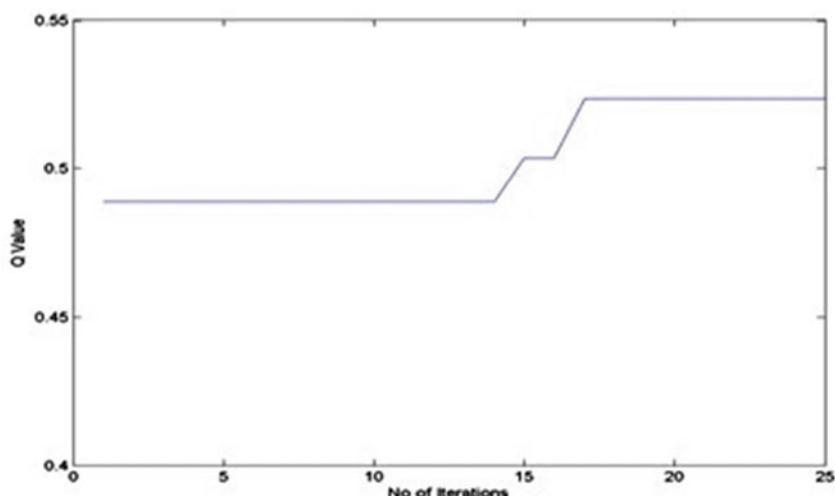
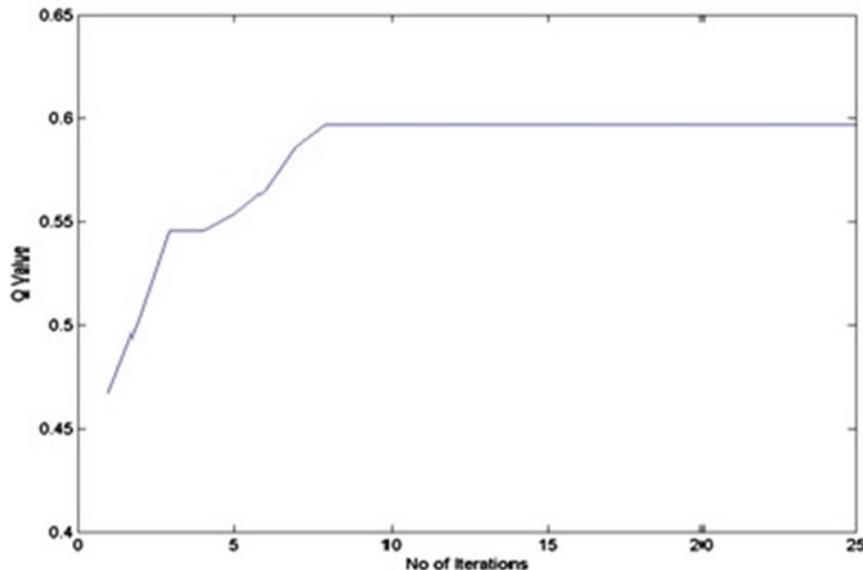


Fig. 6 Modularity versus no. of iterations for Books on US Politics dataset



**Fig. 7** Modularity versus no. of iterations for American College Football dataset

The designed PCA will facilitate to investigate the community structure and observe communities. The limitation of this work is that solely modularity measure has been used because the fitness functions seek out the communities of networks. In our further work, PCA is going to be generalized for multi-objective functions in large networks. We will apply the other fitness functions, like NMI, F-measure, accuracy and many more, with PCA. We will use the PCA for the overlapping community detection with artificial and real-world datasets.

## References

1. Brutz, M., & Meyer, F. G. (2015). A modular multiscale approach to overlapping community detection. arXiv preprint arXiv: 1501.05623 (2015). <https://doi.org/2015arXiv150105623B>.
2. Sah, P., Singh, L. O., Clauset, A., & Bansal, S. (2014). Exploring community structure in biological networks with random graphs. *BMC Bioinformatics*, 15(1), 220. <https://doi.org/10.1186/1471-2105-15-220>.
3. Shen, H., Cheng, X., Cai, K., & Hu, M. B. (2009). Detect overlapping and hierarchical community structure in networks. *Physica A: Statistical Mechanics and its Applications*, 388(8), 1706–1712. <https://doi.org/10.1016/j.physa.2008.12.021>.
4. Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), 066–133. <https://doi.org/10.1103/PhysRevE.69.066133>.
5. Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821–7826. <https://doi.org/10.1073/pnas.122653799>.

6. Gog, A., Dumitrescu, D., & Hirsbrunner, B. (2007). Community detection in complex networks using collaborative evolutionary algorithms. In *European Conference on Artificial Life* (pp. 886–894). Berlin, Heidelberg: Springer. [https://doi.org/10.1007/978-3-540-74913-4\\_89](https://doi.org/10.1007/978-3-540-74913-4_89).
7. Gregory, S., An algorithm to find overlapping community structure in networks. In *European Conference on Principles of Data Mining and Knowledge Discovery* ([pp. 91–102]. Berlin, Heidelberg: Springer. [https://doi.org/10.1007/978-3-540-74976-9\\_12](https://doi.org/10.1007/978-3-540-74976-9_12).
8. Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814–818. <https://doi.org/10.1038/nature03607>.
9. Shakya, H. K., Singh, K., & Biswas, B. (2018). Community detection in social network with regenerative genetic algorithm. *International Journal of Pure and Applied Mathematics*, 118(5), 397–411.
10. Pizzuti, C. (2009). Overlapped community detection in complex networks. In Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation (pp. 859–866) (2009). <http://doi.org/10.1145/1569901.1570019>.
11. Cai, Y., Shi, C., Dong, Y., Ke, Q., & Wu, B. (2011). A novel genetic algorithm for overlapping community detection. In *International Conference on Advanced Data Mining and Applications* (pp. 97–108). Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-25853-4\\_8](https://doi.org/10.1007/978-3-642-25853-4_8).
12. Shakya, H. K., Singh, K., & Biswas, B. (2017). An efficient genetic algorithm for fuzzy community detection in social network (Vol. 712, pp. 63–72). Singapore: Springer (2017). [https://doi.org/10.1007/978-981-10-5780-9\\_6](https://doi.org/10.1007/978-981-10-5780-9_6).
13. Ahn, Y. Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multi-scale complexity in networks. *Nature*, 466(7307), 761–764 (2010). <https://doi.org/10.1038/nature09182>.
14. Kumpula, J. M., Kivelä, M., Kaski, K., & Saramäki, J. (2008). Sequential algorithm for fast clique percolation. *Physical Review E*, 78(2), 026109. <https://doi.org/10.1103/PhysRevE.78.026109>.
15. Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3), 033015. <https://doi.org/10.1088/1367-2630/11/3/033015>.
16. Shakya, H. K., Singh, K., & Biswas, B. (2014). Community detection using differential evolution algorithm with multiple objective function. *International Journal of Urban Design for Ubiquitous Computing*, 2(1), 7–14 (2014). <http://dx.doi.org/10.21742/ijuduc.2014.2.1.02>.
17. Borji, A. (2007). A new global optimization algorithm inspired by parliamentary political competitions In *Mexican International Conference on Artificial Intelligence* (pp. 61–71). Berlin, Heidelberg: Springer. [https://doi.org/10.1007/978-3-540-76631-5\\_7](https://doi.org/10.1007/978-3-540-76631-5_7).
18. Borji, A. (2008). Heuristic function optimization inspired by social competitive behaviors. *Journal of Applied Sciences*, 8(11), 2105–2111. <https://doi.org/10.3923/jas.2008.2105.2111>.
19. Borji, A., & Hamidi, M. (2009). A new approach to global optimization motivated by parliamentary political competitions. *International Journal of Innovative Computing, Information and Control*, 5(6), 1643–1653.
20. Michael, J. H. (1997). Labor dispute reconciliation in a forest products manufacturing facility. *Forest products journal*, 47(11/12), 41.
21. Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4), 452–473 (1977). <https://doi.org/10.1086/jar.33.4.3629752>.
22. Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., & Dawson, S. M.: The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, 54(4), 396–405 (2003). <http://www.jstor.org/stable/25063281>.
23. NEWMAN M E J. Network data [DB/OL], 2011-06-14. <http://www-personal.umich.edu/~mejn/netdata/>.

# An Integrated Technique to Ensure Confidentiality and Integrity in Data Transmission Through the Strongest and Authentic Hotspot Selection Mechanism



**Shiladitya Bhattacharjee, Divya Midhun Chakkaravarthy,  
Midhun Chakkaravarthy and Lukman Bin Ab. Rahim**

**Abstract** Wireless technology has consistently been a popular resolution for handling the growing demand of data receiving in mobile linkages as the availability of high-speed internet is increasing rapidly. Although, the legacy of such technology shortages of seamless inter-communication amid the wireless link as well as mobile cellular linkages on one individual pointer as well as among the wireless hotspots conversely. Subsequently, data confidentiality and integrity are also primary concerns in data transmission through any network. There can be various weak and strong hotspots, since it is also possible to have fake hotspot in any network. A number of researches have been conducted to ensure data security by selecting the strong and authentic hotspot in wireless network infrastructure. However, the present literature fails to offer any security mechanism that can select the strongest, authentic hotspot and secure data confidentiality as well as integrity at the same time. Hence, this research aims to build an integrated technique that can offer adequate level of data confidentiality and integrity along with the selection of a strongest and most authentic hotspot. The proposed integrated technique comprises a unique hotspot selection and a SDES-based authentication mechanism to establish a secure communication link by picking the strongest hotspot. It encrypts the input data with SDES to assure confidentiality and uses an idle error checking technique with a backup scheme for fortuitous information loss to protect data integrity during transmission. The experi-

---

S. Bhattacharjee · L. B. Ab. Rahim

High Performance Cloud Computing Center (HPC3), Universiti Teknologi PETRONAS,  
Seri Iskandar, Perak Darul Ridzuan, Malaysia  
e-mail: [shiladitya.b@utp.edu.my](mailto:shiladitya.b@utp.edu.my)

L. B. Ab. Rahim  
e-mail: [lukmanrahim@utp.edu.my](mailto:lukmanrahim@utp.edu.my)

D. Midhun Chakkaravarthy (✉) · Midhun Chakkaravarthy  
Faculty of Computer Science and Multimedia, Wisma Lincoln, Petaling Jaya,  
Selangor Darul Ehsan, Malaysia  
e-mail: [divya.phd.research@gmail.com](mailto:divya.phd.research@gmail.com)

Midhun Chakkaravarthy  
e-mail: [midhun.research@gmail.com](mailto:midhun.research@gmail.com)

mental results show its superiority over other existing techniques by improving QoE demands and fairness matrices. It offers higher avalanche effect and entropy values which justify its capacity to offer adequate data confidentiality. The capacity to offer lower percentage of information loss proves its efficiency in offering data integrity.

**Keywords** Avalanche effect and entropy values · Backup system for accidental data loss · Data confidentiality and integrity · Strongest and authentic hotspot · SDES-based authentication · SNR and percentage of information loss · QoE demands and fairness matrices

## 1 Introduction

Technically, hotspots can combine single or multiple wireless access points (APs) planted inside any particular location. According to [1, 2], a wide percentage of users are trying to interconnect with a minor subsection of access point (AP) within the particular wireless local area network. Accordingly, the wireless LAN administrators are often facing the issues of fulfilling user clogging at some specific prevalent places (hotspots) in any linkage. These specific type of handler attentions to produce an unhinged consignment on the network impede the capability formation issue and convert it to rigid to lodge huge as well as intense weight in dissimilar portions of the network deprived of compelling, expensive and atop-manufacturing [3, 4]. As per [5], another common issue in wireless network is the short range of available hotspot. Owing to the short range, a large number of hotspots are needed to cover a certain network area. The selection of hotspots according to signal strength is one of the most important issues. Another most important issue in any wireless system is selection of authentic hotspots [6, 7]. Sometimes, some fake hotspots with very high signal strength are generated by any illicit users, and if one user connects any device using these hotspots, he/she can lose private information. So to fight with these types of obstacle, we have designed a hotspot selection protocol that helps to establish secure and safe connection by analyzing its signal strength and the authenticity [8].

According to [9, 10], among the various recently used techniques, use pre-consumer verification agreement to check valid users through entry keys and switch of data admittance by means of pre-packet confirmation. Consumer information safety is attained using different information encryption. The authorization of user is done with the assistance of numerous encryption contrivances verifying the ranks of safety. Distinct systems that deal with verification and safety at the medium access control (MAC) and network coats are being used in modern wireless LANs. Wireless LAN levels, for instance, IEEE 802.11 and Home RF comprise a voting ability for verification, and confidentiality is built on collective keys. It is recognized as wired equivalent privacy (WEP) task [11, 12]. In this system, a mutual key is constructed into the admittance sockets and its wireless customers. The selection of network architecture requires a software bid at the network level to achieve pre-packet substantiation where admittance keys are supplied and verified by means of a federal authorizer-verifier unit. 802.1X executes entry regulator at the admittance points and CHOICE does authentication at the admittance router in the admittance subnet. How-

ever, these techniques are not efficient to provide the optimal solution for selection of hotspots in a wireless infrastructure [13–15].

A number of standard public or private key cryptographies have been used to resolve different confidentiality issues. Yet, they comprise several limitations in terms of high time and space complexities by involving plenty of iterations, as well as complex initialization vector and key sizes, during the encryption process [16]. Alternatively, the applications of DNA structure-based pattern matching and hash function are unable to stop data notching during transmission [17, 18]. Thus, efforts to resolve confidentiality issues by using a number of stenographic techniques in data transmission have failed due to low hiding capacity and complex hiding procedure (Anandaprova [19–21]). Consequently, the existing error control techniques for resolving data integrity issues have missed the mark when the length of error bits is more than eight. In general, cryptography or steganography technique offers high confidentiality, however, they require a large number of iteration and initialization vector to ensure data security [22]. As a consequence, time and space complexities rise considerably. Transmission errors hamper data integrity as they cause data loss and liquefy data confidentiality by licking information. The transmission errors increase due to data size and high channel congestion. The existing error control techniques have failed to solve data error of more than 8 bits at a time, whether they are discrete or continuous.

As yet, discussions show that data transmission by establishing communication link through a hotspot in any wireless network infrastructure suffers due to various causes. Hence, this research primarily aims to build an integrated technique to resolve these problems in a connectional means. The additional goals of this research are:

- Establishing a secure and authentic communication link for data transmission through the strongest hotspot selection in any wireless network infrastructure
- Ensuring adequate confidentiality during the data transmission through the hotspot-linked wireless network channel
- Securing desired level of data integrity by detecting and correcting communication errors as well as reducing information loss percentage in data transmission over wireless channel with the hotspot selection mechanism.

The other parts of this article is organized as follows: Sect. 2 comprises the contextual revision to analyze the contemporary research breach and to plan the projected integrated system; Sect. 3 depicts the constriction of the projected combinatorial technique to achieve the desired level of data confidentiality and integrity with the selection of an authentic and strongest hotspot in any wireless network infrastructure; Sect. 4 describes the required experimental setup and few important parameters for describing the achievements of the planned system in various features; Sect. 5 shows the performance of the planned technique in terms of result analysis; and finally, Sect. 6 concludes the strengths and weaknesses and predicts a future work to enhance the performance of the intended combined technique.

## 2 Background Study

As per the so far discussions, the background study is separated into four subsections. The first subsection comprises the strengths and drawbacks of various hotspot selection mechanisms. The second part depicts the strengths and shortcomings of few authentication techniques. The third and final parts discuss about the strengths and limitations of various accessible errors control and encryption technique to enhance data confidentiality as well as data integrity.

The article [23] proposed a three-dimensional position exhibiting tactic for prime Wi-Fi access point interior design using a signal forfeiture prototype depending on development outline. Here an exertion has been prepared in submission perspective; for example, Wi-Fi assignment in inner places. It also offers the important instruction or guideline to wireless planners about equal exposure and dimensions grounded placement of distinct access point. However, this research is limited to mandate cosmos using limited quantity of request nodes and this research is not perfectly suitable for complete volumetric coverage of network. Besides these, it is incapable to resolve diverse problems correlated to data communication, for instance, network intervention, overlying of network exposure necessity and access point parting.

According to [24], Wi-Fi access point locations and their propagation parameters enhance the time requirements during any connection establishment. Therefore, the authors designed a technique which is used for automatic access point localization and propagation factors estimation with the help of using an internal navigation solution. The work provides a precise Wi-Fi standing solution through no price to figure and preserve a Wi-Fi databank. This technique is easy to implement and is user-friendly and robust to change the indoor environments. However, this technique is inefficient to control the multiple network selection, and it is also unable to control the data loss during the transmission. After establishment of connection to any network, authentication of connected network is also an important aspect. Owing to unauthentic or illegal interference, the integrity and confidentiality of connected data may suffer.

The authors of [25] applied a method to reduce the iteration of execution with 128-bit key size and they also looked up the table based on S-box. This technique is an embedded system to cover crucial data and involves low cost for implementation. However, it works very slowly and extra care was required to protect the symmetric key. Mohammad Ahmed Alomari et al. (2011) introduced an outline for loading encryption in mobile appliances by means of applying XTS-AES encryption process. This encryption technique offers safety during the storage of massive data and can process a large amount of data in parallel. As the bit size of the input block is very small, hence the proposed encryption technique is unable to prevent nonlinear attacks, such as rectangular attack and square attack. The article [26] proposed a multipath-based routing protocol for enhancement of privacy with cryptography in ad hoc networks. This protocol is very simple and could be easily implemented in different ad hoc devices. But, it produces redundant data along with the multipath routing and increased channel overhead. Article [27] proposed an hash key-based

encryption of video transmission to raise confidentiality. It prevents some special attacks, such as regrouping and erasure attacks, but frame loss may occur at the time of decryption, as well as at the time of decompression.

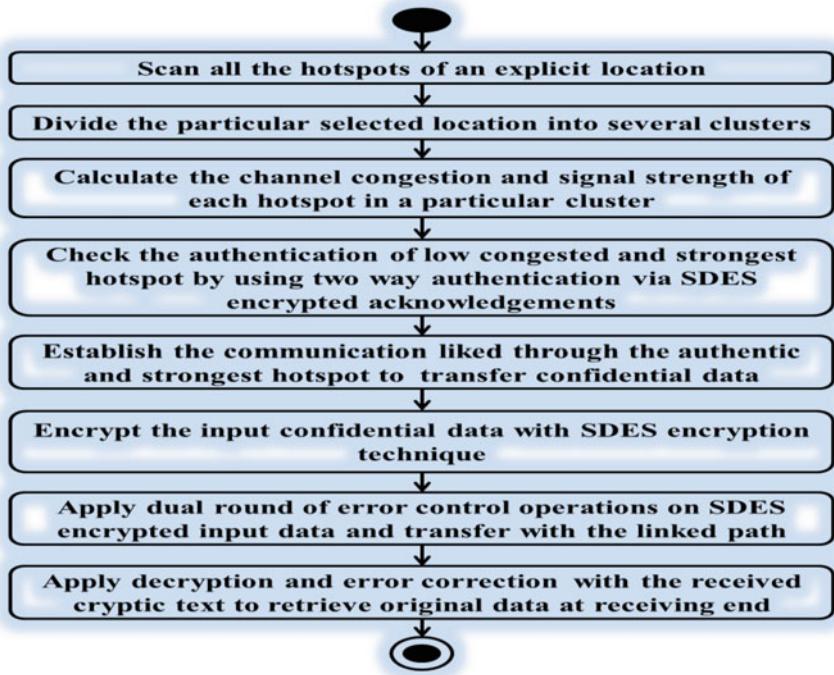
Article [28] has applied dissimilar set ciphers along with the storage utility for detecting the errors during storage. This effort is efficient to recognize the uncorrected faults tad with precise competence and it enlarges the uncovering abilities. Nonetheless, the situation fails to perceive numerous bit errors all at once as it needs higher execution time. An approximation for errors detection using even codes has been proposed by [29]. Noise protection and Hoffman code have been integrated in this technique to offer robustness. On the other hand, the erroneous prefix codes cause data losses during the data retrieval process. Meanwhile, the authors in article [30] have proposed an error detection test of effective post-silicon validation. It has solved the long error detection latency issue. But, it fails to solve the logical bugs, whereas electrical bugs may crash the system.

Therefore, the discussions in Sects. 1 and 2 show that various issues exist in selection of strongest and authentic hotspot to transfer information with adequate level of integrity and confidentiality. Among the various existing techniques, there is no passable method to bid robust verification arrangement for posing obligatory information acquiring in the course of founding of one linking. Furthermore, few security techniques among the existing are efficient to offer data confidentiality, however, they require high amount of time as well as space for execution. The high time and space complexity decreases data integrity by causing information forfeiture because of unnecessary channel outflows. Largely, the present study is unable to refer any combined method which takes care of these issues altogether in a combinatorial way. Hence there are some research gaps that exist to build an integrated technique which can select the strongest hotspot by checking its authenticity. At the same time, it will protect the data confidentiality and integrity by protecting various security attacks and transmission errors, whether large or small, discrete or continuous, with minimal time and space complexities. Furthermore, there should be a mechanism to back up for accidental data loss by regenerating the information at the receiving end to enhance the additional data integrity.

### 3 Proposed Technique

From the background study in Sect. 2, we have seen that no single integrated technique exists that can take care of selecting the strongest and most authentic hotspot network to transfer data with adequate level of data security and integrity. Hence, this research proposed an integrated technique which can take care of all these security aspects during the data transfer with any hotspot-linked network path. The executions flow of the projected integrated technique is displayed in Fig. 1.

Figure 1 shows the distinct steps of proposed integrated technique for execution. Initially, it finds all the hotspots in a particular area and divides them into few hotspot clusters. The signal strength and congestion level of each of the hotspot for every cluster are then calculated to find the strongest hotspot with less congested



**Fig. 1** Work flow of the projected integrated technique

communication channel. After finding the strongest hotspot, two-way authentication technique has been applied to check the authenticity of the designated hotspot with the SDES-encrypted acknowledgments [31]. The communication link has been established through strongest and authentic hotspot. In the meantime, the confidential data is encrypted with SDES and a dual round of error control technique has been applied to generate and incorporate the error control bit on the encrypted string to transfer the confidential data. As a final point, the cryptic confidential data needs to be transferred with the linked communication channel. Every individual part of the projected combined technique is further discussed in the following subsection more precisely.

### 3.1 Selection of Strongest and Authentic Hotspot

The selection process of the robust hotspot has been described here with the help of research article, written by [32]. Let  $N$  be the whole amount of extremely claimed clusters, that is, where the traffic burden is greater in comparison to usual and  $P$  be the overall aspirant hotspots which need to be linked at least through a request cluster. Furthermore, suppose  $Stren_{xy}$  be the signal power of hotspot  $y$  in some request cluster  $x$ ,  $loc_x$  be the location of demand cluster  $x$ ,  $TL_x$  be the mediocre circulation burden

of one request cluster  $x$ . Let  $Bin_{xy}$  is a binary variable (while the worth of it is one, then the request cluster  $x$  is allocated to some hotspot  $y$ ),  $BW_y$  is the bandwidth of any hotspot  $y$  and  $Con_y$  is the congestion level of any hotspot  $y$ . After declaration of the variable, the selection of the strongest hotspot can be done by the following.

1. Scan entire hotspots in one specific network area.
2. Split the zone into few amount of cluster depending on the hotspot obtainability so that one cluster should comprise at least one single hotspot to be associated.
3. Compute the jamming density of each hotspot by the subsequent formulation.

$$Con_y = \frac{1}{BW_y} \sum_{x=1}^{x=N} TL_x \times Bin_{xy} \quad \text{where, } \sum_{x=1}^{x=N} Bin_{xy} \leq 1 \text{ and } 1 \leq y \leq P \quad (1)$$

Here  $Bin_{xy}$  becomes 1 when the signal strength  $Stren_{ij}$  touches to default threshold value.

4. After calculating the congestion level of each individual hotspot using Step 3, select the hotspot with minimum congestion level.

### ***3.2 Checking the Authenticity of Selected Hotspot***

The authentication mechanism will validate the user to the server and the authenticity of source, destination and communication channel. Initially, the user authentication information is encrypted with SDES and sent along with the secret key to the server for validating. The construction of SDES encryption can be read from the research article, published by [31]. After the validation of user, the server also sends the SDES-encrypted acknowledgment along with the secret key to confess the user that the used communication link and selected hotspot are authentic and the user can establish the communication link. By decrypting the cryptic acknowledgment, the user device gets confirmation that the selected link is highly secured and gets connected with it and uses it for confidential data transmission. This two-way validation process reduces the complexity of WPA and WPA2 encryption-based authentication system. As the time and space complexities of SDES encryption are very low, the execution of two-way validation process offer high throughput by reducing latency time and transmission delay.

### ***3.3 Encryption and Dual Rounds of Error Control Techniques***

The SDES encryption technique has been applied on the confidential input data for transmission. Construction of SDES encryption technique is defined by [31] in the

research article. The initialization vector of SDES encryption is an 8-bit string and the initial secret key size is 10 bits. So, the input trusted data needs to be converted into a number of 8-bit strings. Let suppose M number of 8-bit threads are produced from the confidential input threads. After applying the SDES encryption technique, the M number of 8 bits encrypted string will be generated and stored in the string array  $\langle Enc_M \rangle$  and all the 10-bit primary keys are stored in the string array  $\langle Key_M \rangle$ . Both the strings are concatenated and stored into the cryptic string array  $\langle Crypt_M \rangle$  by the following Eq. (2).

$$Crypt_M = (Enc_M \times 10^{l(Key_M)}) + Key_M \quad \text{where, } l(Key_M) = \left( \left\lfloor \log_{10}^{Key_M} \right\rfloor + 1 \right) \quad (2)$$

After the concatenation operation the dual step of error regulator bit creation and combination by taking  $\langle Crypt_M \rangle$  by way of inputs are described with the help of Algorithm 1 and [33].

### Algorithm 1: Creation and Combination of Error Regulator Bits

```

(a1) Declare , XorStr, ErrCon', ErrCon'', Str as string arrays and num as integer variable and
      initialize num = 0;
(a2)   for (int m = 0; m < ( $\frac{M}{2}$ ) ; m = m + 2)
(a3)     for (int j = 0; j ≤ 18; j++)
           XorStr[m][j] = Crypt[m][j]⊕Crypt[m + 1][j];
      End for
// Primary step of error regulatory thread creation
(a4)   for (int k = 0; k < 18; k++)
           ErrCon'[m][k] = Crypt[m][k];
           ErrCon'[m][(m + 1) * 18 + k] = Crypt[m + 1][k];
           ErrCon'[m][(m + 2) * 18 + k] = XorStr[m][k];
      End for
// Subsequent steps of error control thread creation
      ErrCon''[m][0] = ErrCon'[m][0];
      ErrCon''[m][1] = ErrCon'[m][1];
(a5)   for (int j = 0; j ≤ 54;)
           Str[m][num] = ErrCon'[m][num]⊕ErrCon'[m][num + 1];
           ErrCon''[m][j + 2] = Str[m][i];
           ErrCon[m][j + 3] = Str[m][j];
           j = j + 2;
           num = num + 1;
      End for
End For

```

In Algorithm 1, Step (a1) states the string arrays as well as an integer variable for execution of the proposed dual round of error control operation. Step (a2) and Step (a3) perform the XOR exploitation amid a piece of the dual unceasing cryptic

threads of *Crypt* as well as collect into array *XorStr*. Phase (a4) concatenates the two continuous strings to form the string array *Crypt* and the corresponding XOR thread from the array *XorStr* to produce error control array *ErrCon'* after performing the first round of error control string generation and incorporation operation. Finally, Stage (a5) describes the generation of error control string *ErrCon''* by performing the second round of fault regulator bit creation as well as combination. Here, the final step, the XOR operations have been performed between each dual unremitting bits of *ErrCon'* and each resultant XOR tad has been concatenated afterward with each pair of input bits. After performing the double steps of error regulator processes, the final string *ErrCon''* has been transferred through the secure communication channel, linked through the selected strongest and authentic hotspot.

## 4 Assessment Platform

This section describes few significant parameters that will be used in measuring the performance of the proposed integrated technique in distinct aspects at the result analysis unit. Furthermore, this section includes an experimental setup to build this integrated technique.

### 4.1 Experimental Setup

The performance evaluation of our suggested unified technique has been accomplished in the Linux atmosphere. The observation and preparation about data link level and the network component, for example, adapter of mobile armies are measured through wpa\_supplicant utensils (offered by Linux) and DHCP consumer element. Interruption of a single hotspot is observed by collaborating through Linux kernel by means of a datagram worried Netlink Plug. Hotspots display scheme that supports skimming the linkage network and also forms an organized index of probable applicant hotspot intended for communicating connector and with the linkage, organized along with DHCP. The performance of our projected integrated technique has been measured with the state of video conferencing. Few important disclaimers for performing this test are displayed in Table 1.

### 4.2 Definitions of Some Important Parameters

The enactments related to the suggested technique can be measured by means of measuring the execution rate and capability to protect the data integrity and confidentiality of the proposed integrated technique. So, a small number of factors are demarcated in this subpart for depicting the clear idea about them.

**Table 1** Investigational stipulations

Compulsory constraints	Disclaimers
Ethernet inter-frame load	20 Bytes
Voice payload	20 Bytes
UDP header dimension	8 Bytes
Load per call	39,200 bps
Max wireless network volume	11 Mbps
Run-off traffic burden	1040 kbps
Least bandwidth involved	2–5 Mbps

**(A) Jitter**

Jitter is the aberration or dislodgment of signal pulsations for any tall frequency digital motion. The nonconformity can be substantiated by means of amplitude, phase control and width of the signal pulse. The Jitter caused by various signals (electromagnetic) travels surrounding the communication channel. The Jitter is an important parameter to measure the strength of any communication link. Conservatively, while the Jitter is huge, network connection is reflected as frail. While the timer frequency of any motion is  $f_c$  and the segment noise of the motion is  $\theta(t)$  aimed at a period of  $t$ , the Jitter for each of the sequence ( $J_p$ ) can be exemplified as:

$$J_p = \frac{\theta(t)}{2\pi f_c} \quad (3)$$

**(B) Traffic Load**

Traffic load in cellular network is the calculation of entire calls in a specific period of time. Here, traffic load signifies whole amount of nodes linked through a specific network. In general, traffic load is assessed by calculating Erlang. The Erlang entity is essentially required to calculate in the telecommunication circulation and is portrayed by mark  $E_r$ . Erlang is designed as the proportion of entire amount of calls per unit period to the regular allotment period. The circulation load for information traffic can be measured using Erlang as:

$$A = \lambda \times T \quad (4)$$

Here,  $A$  is a regular traffic consignment in the form of Erlang,  $\lambda$  is the whole quantity of data transmission for each unit period and  $T$  is normal packet transfer period.

**(C) Avalanche Effect**

In cryptography grounded verification scheme, the avalanche effect, computes the result of bit altering over the entire encrypted conveyed information. The bit altering happens due to numerous safety attacks or restraint of communication scheme. Avalanche effect controls the power of an encryption method and determines how robust is the verification method. The avalanche effect can be formulated as:

$$AE = \left( \frac{\text{Flipped bits of cipher text}}{\text{Total bits of cipher text}} \right) \times 100 \quad (5)$$

In general, a cryptography founded verification scheme is considered as robust, if it harvests advanced avalanche effect.

#### (D) *Entropy*

In coding theory, entropy is needed to quantify the hesitation related with an arbitrary variable. In cryptography, entropy necessity be delivered by the encryption for presence into the plaintext of a communication with the aim of it can resolve the quantity of constructing dangerous plaintext information. Entropy obtained by a cryptography is described as:

$$H(S) = \sum_{i=0}^{2N-1} P(S_i) \log_2 \frac{1}{P(S_i)} \quad (6)$$

In Eq. (6),  $P(S_i)$  indicates the possibility of occurring the symbol ( $S_i$ ). The idyllic *Entropy* worth for an encoded message must be 8. So, a cryptography founded verification system is supposed to be robust if it crops the entropy standards nearer to 8 or vice versa.

#### (E) *Percentage of Information Loss*

Information loss (IL) is the damage of conveyed information in the form of packets throughout the communication. Packet dropping or exploitation happens because of linkage disappointment, large transmission delay, network overcrowding and so on. Data loss could be an obvious issue in any data transmission. It can be formulated as:

$$IL = \frac{(\text{Transmitted Data} - \text{Received data})}{\text{Transmitted Data}} \times 100 \quad (7)$$

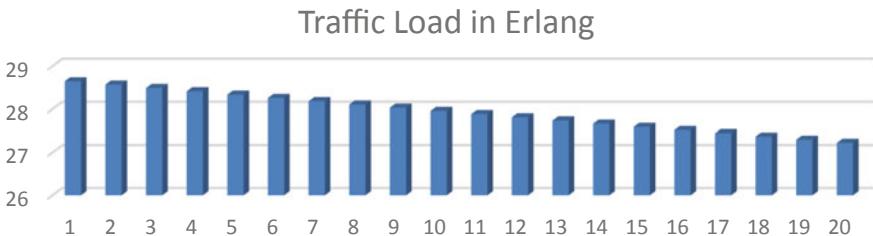
## 5 Result Analysis

In this section, the achieved result has been analyzed according to several aspects, such as the strength of the proposed integrated technique to select the lowest congested hotspot and the capacity to protect data confidentiality as well as the data integrity. Hence, the analyses of results in different aspects have been separated into two parts. The first part discusses about the selection of strongest and lowest congested hotspot in various demand clusters and the latter part analyzes the capacity of proposed integrated technique to ensure data security in terms of improving data confidentiality and integrity.

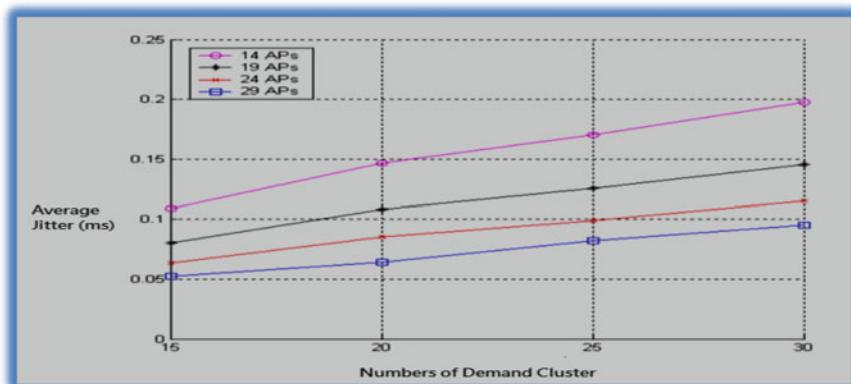
### 5.1 Determination of Lowest Congested and Strongest Hotspot

Initially, this section analyzes the capability of the said technique to select the lowest congested and strongest hotspot. Figure 1 shows that the proposed integrated technique analyzes the traffic load and channel congestions to select the lowest congested traffic in terms of calculating the Erlang values. As per the definition in Assessment platform section, if the traffic load of any hotspot is higher then the corresponding Erlang value will also be high, or vice versa. The traffic load of various hotspots in a specific demand cluster has been calculated in terms of Erlang values and plotted in Fig. 2.

The proposed integrated technique calculates the congestion level of any hotspot by calculating the traffic load in terms of calculating the Erlang value using Eq. (4). Figure 2 shows that hotspot 20 offers the lowest traffic load in the network. The Assessment platform section shows that the signal strength can be calculated by calculating the average Jitter value of any demand hotspot cluster. The Jitter can be calculated by using the Eq. (3) and plotted in Fig. 3.



**Fig. 2** Traffic load of various hotspots



**Fig. 3** Jitter offered by the various demand clusters

Figure 3 shows that the demand cluster 15 with 29 access points offers lowest average Jitter. As per the definition, if the Jitter is low then the signal strength is high. Hence, the demand cluster 15 with 29 APs offers highest signal strength rather than other demand clusters. Thus, Fig. 2 justifies the first objective of this research work.

## 5.2 Capacity to Offer Data Confidentiality and Integrity

As per the Assessment platform section, the data confidentiality can be measured in terms of scheming avalanche effects and entropy values, offered by any security technique. Conferring to the definitions, a security technique can offer higher data confidentiality, if this one offers advanced avalanche effect as well as entropy values. Avalanche effect and entropy values obtained by the planned integrated technique and additional corresponding prevailing security techniques have been computed with the help of Eqs. (5) and (6) and plotted in Table 2.

Table 2 displays that the anticipated combined process offers highest avalanche effect percentage and entropy values among other important security methods available in the current literature and mentioned in the list. Hence, as per the definitions, the anticipated system produces advanced data confidentiality among the rest. Furthermore, the data integrity offered by a safety method could be confirmed by calculating percentage of information loss. As per the definition, if the percentage of information loss is large, then the worn method is not well-organized to offer the desired information integrity. The percentage of information loss produced by the planned combined technique with the various congestion levels has been calculated by using Eq. (7) and strategized in Table 3.

Table 3 displays that the planned combined method offered very low percentage of information loss, though the congestion level of the linked channel through a specific hotspot is high. Hence, the anticipated combined technique produces desired level of integrity by producing minimum percentage of information loss. Thus, Tables 2 and 3 display that the planned method is effective in offering higher confidentiality as well as integrity, which justify our second and third objectives.

**Table 2** Avalanche effect and entropy values offered by distinct security techniques

Security techniques	Avalanche effect (in %)	Entropy values
Proposed combined technique	69.3	7.84
AES	65.2	7.73
Triple DES	64.3	7.53
3-DES	62.2	7.67
Blowfish	59.2	7.52
RSA	61.6	7.51

**Table 3** Percentage of information loss with different congestion levels

Congestion level (kB/s)	Percentage of information loss
100	0.003
300	0.007
500	0.011
800	0.018
1000	0.021

## 6 Conclusion and Future Work

Hotspot selection for wireless communication is a very challenging task. There are many factors involved during the selection of a perfect hotspot. Among them, analysis of network congestion, traffic load and signal strength are the most important factors. Apart from that, the security of data is another important aspect of any communication system. So, the validation of a hotspot needs to be checked before establishing any communication link. This process secures the data transmission. However, to address all these aspects, no technique is available. Therefore, we resolve these issues by our proposed technique which can address all these issues integrated. In the Result analysis section, we can see that the planned joined technique is effective in analyzing the signal strength as well as traffic load of any hotspot during the selection process. It minimizes the percentage of information loss and enhances avalanche effect and entropy values for ensuring the desired level of data integrity and confidentiality. Our proposed integrated technique is gifted to resolve numerous problems of hotspot selection procedure in a united way. Nevertheless, it is not capable to diminish the data loss completely. So to reduce the data loss further precisely, the projected cohesive technique requires enhancement to improve data integrity level more specifically.

## References

1. Louta, M., Zournatzis, P., Kraounakis, S., Sarigiannidis, P., & Demetropoulos, I. (2011). Towards realization of the ABC vision: A comparative survey of access network selection. In *2011 IEEE Symposium on Computers and Communications (ISCC)* (pp. 472–477), June 28–July 1, 2011.
2. Zhihong, L., Yaping, L., Zhenghu, G., & Lin, C. (2011). A Muti-rate access point selection policy in IEEE 802.11 WLANs. In *2011 International Conference on Multimedia Technology (ICMT)* (pp. 63–67), July 26–28, 2011.
3. Bennai, M., Sydor, J., & Rahman, M. (2010). Automatic channel selection for cognitive radio systems. In *2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)* (pp. 1831–1835), September 26–30, 2010.
4. Tuysuz, M. F., & Mantar, H. A. (2010). Access point selection and reducing the handoff latency to support VoIP traffic. In *2010 International Conference on Computer Engineering and Systems (ICCES)* (pp. 58–63), November 30–December 2, 2010.
5. Gharsellaoui, A., Chahine, M. K., & Mazzini, G. (2011). Optimizing access point selection in wireless local area networks. In *2011 International Conference on Communications and*

- Information Technology (ICCIT)* (pp. 47–52), March 29–31, 2011.
- 6. Mingyi, H., Garcia, A., & Barrera, J. (2011). Joint distributed access point selection and power allocation in cognitive radio networks. In *INFOCOM, 2011 Proceedings* (pp. 2516–2524). IEEE, April 10–15, 2011.
  - 7. Balachandran, K., Kang, J. H., Karakayali, K., & Rege, K. (2011). Cell selection with downlink resource partitioning in heterogeneous networks. In *2011 IEEE International Conference on Communications Workshops (ICC)* (pp. 1–6), June 5–9, 2011.
  - 8. Haoxuan, M., Zhanbin, W., Jingcheng, W., Langwen, Z., & Liu, Z. (2014). A novel access point selection strategy for indoor location with Wi-Fi. In *The 26th Chinese Control and Decision Conference (2014 CCDC)* (pp. 5260–5265), May 31–June 2, 2014.
  - 9. Yaqing, Z., & Sampalli, S. (2010). Client-based intrusion prevention system for 802.11 wireless LANs. In *2010 IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)* (pp. 100–107), October 11–13, 2010.
  - 10. Choi, J., Chang, S.-Y., Diko, K., & Hu, Y.-C. (2011). Secure MAC-layer protocol for captive portals in wireless hotspots. In *2011 IEEE International Conference on Communications (ICC)* (pp. 1–5), June 5–9, 2011.
  - 11. Chew, C. C., Funabiki, N., & Fujiita, S. (2014). Extensions of active access-point selection algorithm for wireless mesh networks using IEEE802.11ac Protocol. In *2014 Second International Symposium on Computing and Networking (CANDAR)* (pp. 310–314), December 10–12, 2014.
  - 12. Ruikai, M., Nguyen, D. H. N., & Tho, L.-N. (2015). Joint access point selection and linear precoding game for MIMO multiple-access channels. In *Wireless Communications and Networking Conference (WCNC)* (pp. 753–758). IEEE, March 9–12, 2015.
  - 13. Chauhan, N., & Yadav, R. K. (2012). Security analysis of identity based cryptography and certificate based in wimax network using Omnet++ simulator. In *2012 Second International Conference on Advanced Computing & Communication Technologies (ACCT)* (pp. 509–512), January 7–8, 2012.
  - 14. Soon, K. H., Kim, E., & Hwangnam, K. (2012). QoE-driven Wi-Fi selection mechanism for next generation smartphones. In *2012 First IEEE Workshop on Enabling Technologies for Smartphone and Internet of Things (ETSIoT)* (pp. 13–18), June 18, 2012.
  - 15. Corena, J. C., & Ohtsuki, T. (2012). A Multiple-MAC-based protocol to identify misbehaving nodes in network coding. In *Vehicular Technology Conference (VTC Fall)* (pp. 1–5), September 3–6, 2012.
  - 16. Yi, X., Chen, N., Jia, Z., & Chen, X. (2010). Trusted communication system based on RSA authentication. In *2010 Second International Workshop on Education Technology and Computer Science (IEEE)* (pp. 329–332).
  - 17. Weixin, B., Luo, Y., Xu, D., & Yu, Q. (2014). Fingerprint ridge orientation field reconstruction using the best quadratic approximation by orthogonal polynomials in two discrete variables. *Pattern Recognition*, 47(10), 3304–3313.
  - 18. Abushariah, A. A. M., Gunawan, T. S., Khalifa, O. O., & Abushariah, M. A. M. (2010). English digits speech recognition system based on Hidden Markov Models'. In *2010 International Conference on Computer and Communication Engineering (ICCCE)* (pp. 1–5).
  - 19. Majumder, A., & Changder, S. (2013). A novel approach for text steganography: Generating text summary using reflection symmetry. *Procedia Technology*, 10, 112–120.
  - 20. Sghaier, G., & Nidal, N. (2012). An audio/video crypto—Adaptive optical steganography technique. In *2012 8th International (IEEE) Wireless Communications and Mobile Computing Conference (IWCMC)* (pp. 1057–1062).
  - 21. Natarajan, M., & Lopamudra, N. (2010). A review of the audio and video steganalysis algorithms. In *Proceedings of the 48th Annual Southeast Regional Conference (ACM)* (p. 81).
  - 22. Sumit, B., & Radu, S. (2014). Trusted DB: A trusted hardware-based database with privacy and data confidentiality. *IEEE Transactions on Knowledge and Data Engineering*, 26(3), 752–765.
  - 23. Lee, G. (2015). 3D coverage location modeling of Wi-Fi access point placement in indoor environment. *Computers, Environment and Urban Systems*.
  - 24. Zhuang, Y., Syed; Z., Georgy, J., & El-Sheimy, N. (2015). Autonomous smartphone-based WiFi positioning system by using access points localization and crowdsourcing. *Pervasive and Mobile Computing*, 18, 118–136.

25. Cai-hong, L., Jin-shui, J., & Zi-long, L. (2013). Implementation of DES encryption arithmetic based on FPGA. *AASRI Procedia*, 5, 209–213.
26. Ahmed, A. M., & Khairulmizam, S. (2011). A framework for GPU-accelerated AES-XTS encryption in mobile devices. In *IEEE Region 10 Conference (IEEE) TENCON 2011–2011* (pp. 144–148).
27. Wang, X., David, H. M., & Mark, L. S. (2013). Stop-and-wait automatic repeat request schemes for molecular communications. In *2013 First International Black Sea Conference on (IEEE) Communications and Networking (BlackSeaCom)* (pp. 84–88).
28. Reviriego, P., Flanagan, F. M., Liu, S.-F., & Maestro, J. A. (2012). Error-detection enhanced decoding of difference set codes for memory applications. *IEEE Transactions on Device and Materials Reliability*, 12(2), 335–340.
29. Paulo, P. E., Fábio, P., & Jaymel, S. L. (2012). Exact and approximation algorithms for error-detecting even codes. *Theoretical Computer Science*, 440, 60–72.
30. Ted, H., Li, Y., Park, S.-B., Mui, D., & Lin, D. (2010). QED: Quick error detection tests for effective post-silicon validation. In *2010 IEEE International (IEEE) Test Conference (ITC)* (pp. 1–10).
31. Puangpronpitag, S., Kasabai, P., & Pansa, D. (2012). An enhancement of the SDP security description (SDES) for key protection. In *2012 9th International Conference on Computer, Telecommunications and Information Technology* (pp. 1–4). IEEE.
32. Bhattacharjee, S., Rahim, L. B. A., Zakaria, M. N., & Aziz, I. B. A. (2018). A protocol for selecting the strongest and authentic hotspot for transferring big data in wireless infrastructure. In *2018 International Conference on Computer and Information Sciences*. IEEE.
33. Bhattacharjee, S., Rahim, L. B. A., & Aziz, I. B. A. (2014). A multibit burst error detection and correction mechanism for application layer. *International Conference on Computer and Information Sciences (ICCOINS)*. IEEE.

# Model to Improve Quality of Service in Wireless Sensor Network



Vivek Deshpande and Vladimir Poulikov

**Abstract** Communication channel is a transmitting medium by which two nodes can communicate with each other in a network. A channel can send the information also called as data in the form of packets to its receivers. If a channel is communicating in an open environment, there can be some environmental conditions that can affect the performance of communication. There can be parameters such as SNR, BNR or channels throughput for every channel which can be affected by the external environment. To improve the quality of service, in this research model, the proposed algorithms can recommend the best and healthy parameters for every channel and the healthy channel itself to have better communication between nodes. The proposed algorithms can be applied to the channel to reduce the dimensionality of data or feature selection from the available set of features automatically. This approach will automatically take care of suitable and useful features for every channel. The technique is further applied to recommend a single communication channel among all available channels, based on the most suitable features that are selected for every channel. The proposed model is mainly focused on wireless low throughput network with ISM band. Using machine learning techniques the model will be trained by itself so that it can return the one of the best-recommended channel with the best parameters to have the best communication in a network. One of the machine learning techniques is reinforcement learning, which will help the machines to learn by itself and give accurate results based on various channel selection parameters.

**Keywords** Nodes · Packets · SNR · BER · Channel throughput · QoS · Recommendation · Reinforcement learning technique

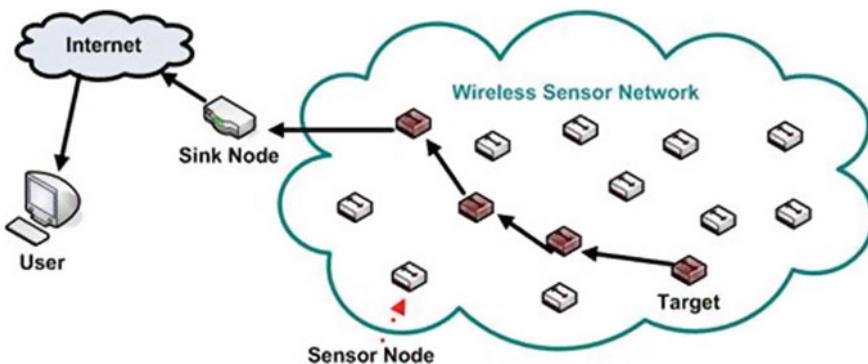
---

V. Deshpande (✉) · V. Poulikov  
Technical University of Sofia, Sofia, Bulgaria  
e-mail: [vsd.deshpande@gmail.com](mailto:vsd.deshpande@gmail.com)

V. Deshpande  
Vishwakarma Institute of Technology, Pune, India

## 1 Introduction

The channel recommendation system is a combination of communication networking and machine learning techniques. It is used to recommend a healthy channel in a wireless low throughput network with ISM band for hassle-free communication between sender and receiver. Design and implementation of the model to improve quality of service for the selection of a good channel in wireless low throughput network (ISM band) using machine learning techniques is the system which can be used by the ISPs or network providers to select the best channel in a network. The recommended channel will help to transmit the data with better quality and speed [1, 2] (Fig. 1).



**Fig. 1** Wireless sensor networks

Wireless sensor network is a special type of network that requires special attention for maintaining quality of service. The entire nodes are spatially distributed and act as autonomous devices individually. The nodes in the WSN use sensors to monitor the physical or environmental parameters, like temperature, humidity, vibrations, pollution, motion, and so on. The nodes in the wireless sensor networks comprise sensing device, computing machinery and communication device to communicate to other node or to the data collector, generally called as sink.

In this paper, the quality of service will be the main concern which will be focused on. The model states that in a specific bandwidth, there may be n number of channels present which are transmitting some data. How to find out based on parameters, which channel's transmission rate is higher? So by means of reinforcement learning, system can train the network to select one of the best available channels.

## 2 Related works

Wireless sensor network helps community to sense physical or environmental parameters. The sensed parameters may or may not be in digital form. After digitizing the input data, the generation of packets will be done. All the generated packets are then sent to the destination via ZigBee communication network.

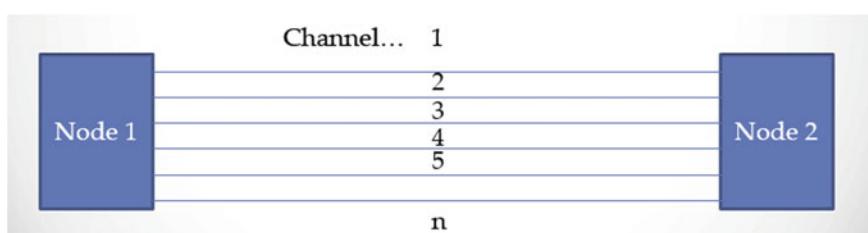
The quality of service (QoS) which will be provided to the wireless sensor network is totally different than it provides to other networks. Quality of service (QoS) terminology relates with performance of a network, such as a bandwidth, delay, throughput, and so on, experienced by the customers of the network. It can be measured quantitatively by packet delivery ratio, packet loss ratio, end-to-end delay in milliseconds, and so on [3]. QoS is particularly important for the traffic with particular requirements. Recent networks have introduced voice over IP to allow computer networks to adopt as useful as telephony for audio and video conversations, and support the new applications with strict network performance requirements as well.

The parameter, SNR, can be defined as ratio of power of signal to power of noise. At the same time, bit error rate is defined as ratio of number of errors to total number of bits sent on the communication channel.

The communication between the nodes at the MAC layer may be with multiple channels. Refer Fig. 2 for multiple channel communication in WSN.

Machine-to-machine is often known as M-M communications. The internet of things (IoT) incorporates collaboration of potentially multi-billions of objects and communicates with one another with low throughput connectivity. The expected advantages of having particular IoT by wireless sensor networks (LTN) include lower costs, the creation of new applications and reduction in EMI/RFI effect. LTN is supposed to be critical to the success of standardization and implementation, and to cater the basis for many new and innovative applications.

In this model the dimensionality reduction is done on the dataset so as to reduce the number of features from dataset. This dimensionality reduction can be done only on numeric data, which again has to be normalized to certain level so as to get the expected output. Feature reduction never harms the data and gives the same results or accuracy as the original data. Up till now, most of the previous research is performed at the specific level such as specific networks or typical machine learning



**Fig. 2** Communication channel between nodes

**Table 1** Literature review for exiting work in machine learning for networking

Title	Author	Year	Methodology
QoS-driven channel selection algorithm for cognitive radio network	Navikkumar Modi, Philippe Mary, Christophe Moy	2015	Multi-user problem is addressed by proposing distributed RQoS-UCB algorithm
Optimization of Channel allocation in wireless BANs by means of reinforcement learning	Tauseef Ahmed, Faisal Ahmed, Yannick Le Moullec	2016	Traffic load conditions parameter is checked for channel assignment. RL – CAA algorithm is proposed by the author
Implementation of channel selection for LTE in Unlicensed bands using Q-learning and game theory	A. Castañé, J. Pérez-Romero, O. Sallent	2016	Two different approaches, one based on Q-learning and game-theory as well
QoS-driven channel selection algorithm for opportunistic spectrum access	Navikkumar Modi, Philippe Mary, Christophe Moy	2016	An OSA algorithm, QoSUCB and based on channel quality information and availability is achieved
Channel selection in multi-hop cognitive radio network using reinforcement learning	A. R. Syed, K. L. A. Yau, H. Mohamad, N. Ramli, W. Hashim	2017	Author proposes channel selection technique using reinforce learning. Working on cognitive radio multi-hop network in order to recommend good operating channel

algorithms. Refer to Table 1 for existing work done where machine learning is used to recommend the channel selection.

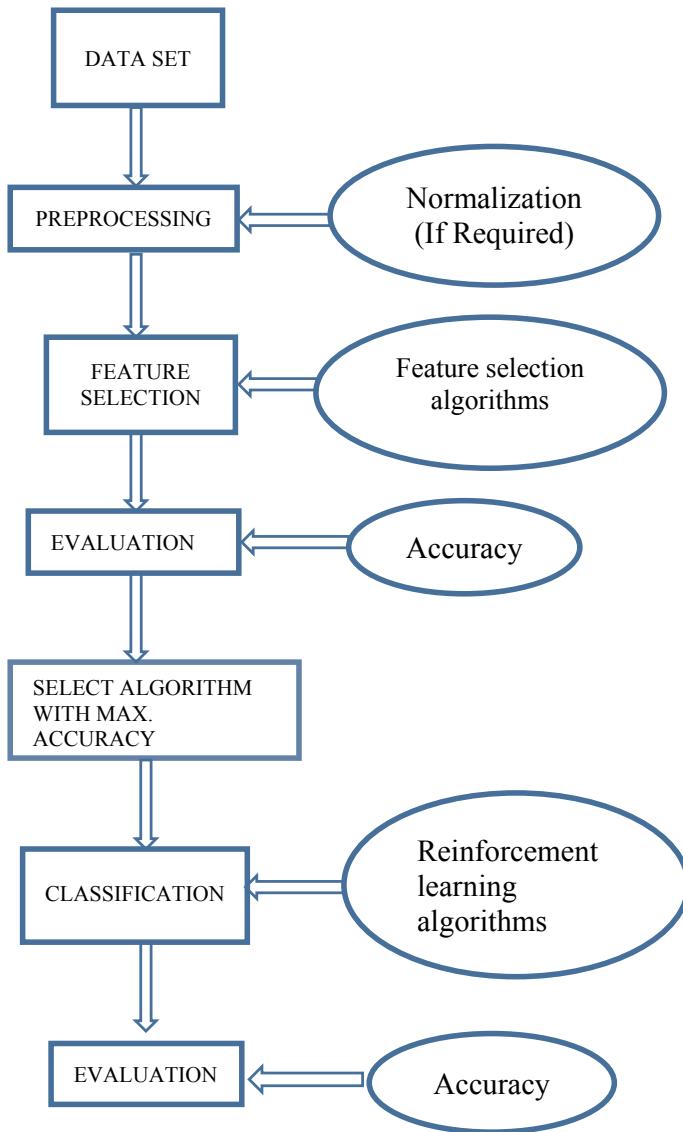
In this paper, best wireless communication channel is recommended for hassle-free communication between a sender and a receiver with the help of reinforcement learning. The machine learning techniques are best suitable for feature extraction from a given networking dataset.

### 3 Methodology

Refer to Fig. 3 for the methodology used in the proposed research.

*Dataset:*

For experiments, real data have been used in terms of real values of every parameter of about 50 channels. This dataset is collected from actual environment setup which has numeric values for each defined parameter of the channel for more than 50 channels, including correct as well as incorrect values [4–6].



**Fig. 3** Architecture of channel recommendation system

*Preprocessing:*

The very first step is data preprocessing. Preprocessing is done to remove inconsistent data and noisy data from the dataset, so that unnecessary computation is avoided. This process is required only if algorithms like PCA is used. This process is not to be followed for each and every algorithm that is applied on the dataset. Preprocessing step is done for normalization of data to normalize it to a certain range. Normalization does not update or modify or alter the data but the representation of the data changes.

*Feature Selection:*

Feature selection technique is used for selecting the most useful or vital features from the available set of features. A total of ten features for every channel on which feature reduction is applied are selected so that it will return the most useful features from a set of features. Every feature is discussed in detail in next section [7–9].

*Evaluation:*

Evaluation has a meaning of finding out the accuracy of each feature selection algorithm and compare them to choose the one best out of it. Evaluation is also done after classification algorithms applied on the dataset to find out which algorithm gives the more accuracy.

*Classification:*

Classification here means to apply the reinforcement learning algorithm to the features which are selected after applying the feature selection algorithms. Every channel feature will get the award or penalty according to the action taken by the agent. Finally, every channel will have some weight in the form of award or penalty and then that channel is recommended to the user which is having more rewards [10].

## 4 Experiments and results

For experiments, the real-time data are used from a simulated environment set up in an open environment. The data contain all numeric values for every channel along with the ten features per channel. Here, the channel can be used again for communication; hence the data are repeated for every channel approximately for ten times. So, for each channel there are repeated records with the same number of parameters. In this research all the parameters are considered for the experiment as features.

Out of available parameters, some of the channel selection parameters are discussed in detail as follows [11]. The various channel selection parameters are:

- (a) Signal-to-noise ratio
- (b) Bit error rate
- (c) Channel capacity
- (d) Channel occupancy

- (e) Channel utilization
- (f) Interference
- (g) Co-existence
- (h) Channel throughput
- (i) Delay/data latency
- (j) Jitter.

These parameters are the inputs to the algorithm and tested accordingly.

As Table 2 shows, all the figures related to the model are implemented in this paper. The model that is compared in this paper is based on python and excel. In python the .csv file is read and some calculations are done on that file. From the coding perspective there is some output and also from excel, there are calculations done for every channel to calculate the fitness of every channel. After that both the outputs, that is, of python and excel are compared and according to that error calculation is done.

## 5 Conclusion

In this paper, a reinforcement learning (RL)-based channel selection mechanism is implemented on a wireless low throughput network (WLTN) in order to select one of the best possible operating channels. In this proposed algorithm principal component analysis is used for the selection of best parameter which gives more accuracy to get one of the best channels to recommend for transmission. Based on the parameters which will be best suitable for selecting parameters, the reinforcement learning technique will be applied for best channel selection.

Designing and implementing the model with the quality of service for selecting one of the best channels in a network by considering only specific network, that is, wireless low throughput network only, is the main task. This increases the scope to implement the model for other types of networks. Also for comparative studies, support vector machine can be used over PCA for better results. Reinforcement learning technique, the very first advantageous thing in this work model, will be that algorithm learns itself and select the best channel according to the requirements based on various parameters. Furthermore, the usage of GPUs can also accelerate the working speed of the model.

**Table 2** Comparison of implemented models

Existing model		Proposed model		Errors
Channel	Fitness	Channel	Fitness	
33	74.95558	33	75.08088	0.125297
13	73.25099	48	73.251087	0.000097
14		41		
18		21		
19		23		
21		24		
22		34		
23		19		
24		43		
31		32		
32		36		
34		18		
36		45		
39		31		
41		22		
43		49		
45		39		
48		13		
49		14		
18	73.11506	24	73.115059	-0.00001
24		18		
31		45		
45		31		
22	73.6763	43	73.676295	-0.000005
43		22		
48		48		
43	74.01387	43	74.01387	0.000003
16	72.76573	16	72.76573	0.000003
19	73.51142	32	73.51142	0.000001
32		19		
36	73.60936	36	73.60936	0.000003
23	72.61691	23	72.61691	0
34	72.50393	34	72.50393	-0.000003
49				
13	73.51162	13	73.51162	0.000002
39		39		
14	73.12928	14	73.12927	-0.000005
21		41		
41		21		

## References

1. Modi, N. K., Mary, P., & Moy, C. (2015). QoS driven channel selection algorithm for cognitive radio network: Multi-user multi-armed bandit approach. *IEEE Trans. Cogn. Commun. Network*.
2. Thorat, M., & Deshpande, V. (2016). *Assessment of fairness against quality of service parameters in wireless sensor networks*. Accepted for oral presentation in IEEE Thirteenth International Conference on Wireless and Optical Communications Networks (WOCN 16), Hyderabad.
3. Ahmed, T., Ahmed, F., & Le Moullec, Y. (2016). Optimization of channel allocation in wireless body area networks by means of reinforcement learning. In *The 2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*.
4. Castañé, A., Pérez-Romero, J., & Sallent, O. (2016). On the implementation of channel selection for LTE in unlicensed bands using Q-learning and game theory algorithms. *IEEE J.*
5. Syed, A. R., Yau, K. L. A., Mohamad, H., Ramli, N., & Hashim, W. (2017). Channel selection in multi-hop cognitive radio network using reinforcement learning: An experimental study. Malaysian Ministry of Science, Technology and Innovation (MOSTI). *IEEE J.*
6. Akyildiz, F., Lee, W., Vuran, M. C., & Mohanty, S. (2006). Cognitive radio wireless networks. *The International Journal of Computer and Telecommunications Networking*, 50(13), 2127–2159.
7. Jouini, W., Ernst, D., Moy, C., & Palicot, J. (2010). Upper confidence bound based decision making strategies and dynamic spectrum access. In *International Conference on Communications (ICC'10)*, May 2010.
8. 3GPP workshop on LTE in unlicensed spectrum, Sophia Antipolis, France, June 13, 2014. [http://www.3gpp.org/ftp/workshop/2014-06-13\\_LTE-U](http://www.3gpp.org/ftp/workshop/2014-06-13_LTE-U).
9. Hämäläinen, M., et al. (2015). ETSI TC SmartBAN: Overview of the wireless body area network standard. In *2015 9th International Symposium on Medical Information and Communication Technology (ISMICT)*, Kamakura, pp. 1–5.
10. Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. New York, NY, USA: Cambridge University Press.
11. Nanavati, A., & Deshpande, V. S. (2015). Analysis of QOS parameters of sensor network to improve reliability. In *Fourth Post Graduate Conference*, Pune, pp 15–19.

# Performance Analysis of the Mobile WSN for Efficient Localization



Kailas Tambe and G. Krishna Mohan

**Abstract** Localization is a most important thing in wireless sensor network and when node is mobile, then it's a challenging task for the user to maintain the information of each node. From last some years many user try to maintain the position of each node by using the different method and maintain it in table so we can able to get the data in less time with less energy consummation. But when all nodes are moving in some random direction in a particular area, then it is little difficult to maintain its information, and it can be achieved by using different methods to store information in one place by gathering the information from the neighboring network. In the current situation by using algorithm where node position is tracked by some constant time interval “ $t$ ”, which we maintain in the table that contains its current position at a particular time interval “ $t$ ”, as well as they try to predict the future position at a particular time interval “ $2t$ ”. In the proposed algorithm to keep the localization error minimum, we have selected two neighboring nodes for each node and every node updates its current position and predicted the future position after every fixed time interval. The minimum distance can be calculated by performing trilateration among two neighboring nodes with unknown position node. RSS is mainly used in range-based localization. These coordinate differences between current and predicted positions for time  $t$  and  $2t$  time slot give us a localization error. With the presented algorithm, we have found the efficient time period where average localization error will be minimum with minimum energy consumption. In this paper with quality of service other parameter we try to calculate like Packet delivery ratio (PDR), throughput and Delay in the network with energy consumed.

**Keywords** E2E delay · Energy consummation · Localization · PDR

---

K. Tambe (✉) · G. Krishna Mohan

Computer Science and Engineering, Koneru Laxmaiah Deemed to be University,  
Green Fields, Vaddeswaram, Guntur District 522502, Andhra Pradesh, India  
e-mail: [kailashtambe@gmail.com](mailto:kailashtambe@gmail.com)

## 1 Introduction

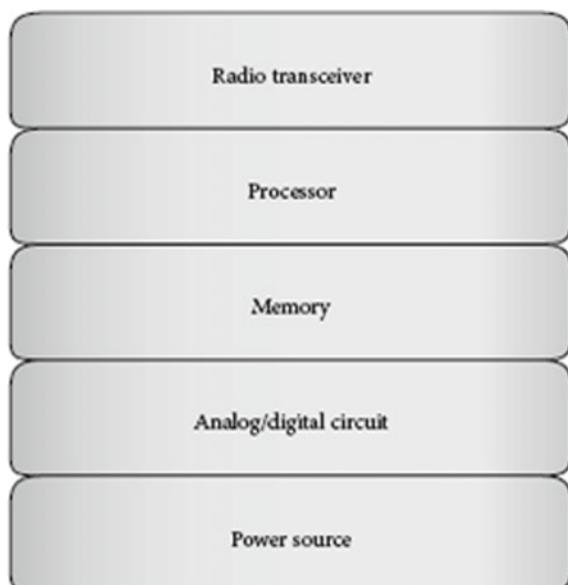
A Wireless Sensor Network (WSN) comprises an enormous low-cost tiny devices, or small sensor nodes, equipped with wireless communication and which supports several applications such as in health care, home automation, surveillance security, automobile system, and industrial application. Sensor node localization and its position is an important and significant concern of the wireless sensor network (Fig. 1).

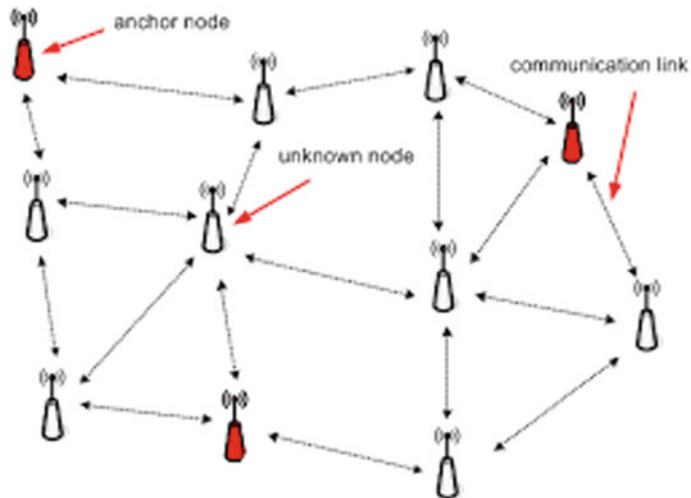
Finding the exact position of each node in the network is localization. In such applications, data is meaningful only when it carries its own location information. If each node has their position, which plays a crucial role in areas such as military application, indoor localization, objects tracking, animal habitat monitoring. When the density of nodes is high, then the manual deployment is quite unrealistic and expensive. Subsequently, through location estimation and gathering techniques, sensor nodes are able to obtain their locations. Global Positioning System (GPS) is an utmost method used for localization [1, 2].

In wireless sensor network, we used two types of nodes, i.e., anchor nodes who are aware about their location and carry all information related to own position as well as current network and other nodes those who are not aware about their locations and network information they need to know their location (Fig. 2).

The main constraint is costs of nodes are the prime important for the range based localization. Node localization for static scenario and static node is not a big task, where there are several ways to know its location. Find its exact position but for dynamic scenario when all nodes are moving continuously and positions are changing then it is a great challenge to find the exact location of all nodes. Localization can be

**Fig. 1** Architecture of sensor node





**Fig. 2** Localization network with different sensor nodes

done in such a scenario using GPS, by attaching GPS unit to all nodes in the network but this solution is not a cost-effective rather its required more cost. Apart from that GPS will not give results in some cases like the atmospheric or topographical obstacles, i.e., dark forest, large rows of highlands in such scenario Line of Sight (LoS) is not clear for GPS [3]. The size of node for deployment in many applications is need in small size but after adding GPS module and its apparatus size of the node will be bulky and as we add this unit power consumption will be high (Fig. 3).

The GPS module and constraint for limited battery life for the sensor node the life time of the network is reduced. GPS is always suitable for the resources where no constraints are for hardware and climatic conditions that time we can use the GPS. But where the resource constraint are consider and bad environmental condition may occurs that time we are not able to GPS system. Due to such things which will work in this punitive environment which is dynamic in nature and must be easily deployable. The localization algorithm can be a good solution to such problems where localization accuracy will be increased with exact position and average localization error will be minimized with less energy consumption and is able to get the maximum throughput and can support dynamic environment where all nodes are moving continuously and changing the network. To address this issue, we have presented here an algorithm for localization, which works on cooperative and sharing approach among all sensor nodes.

The localization techniques are classified into range-based and range-free localization.

**Range-based localization:** Its has limited range so not able to use for dynamic environment as its possible to increase the network size in dynamic condition but its give the maximum throughput in less time because of its fixed size nature.



**Fig. 3** Working of GPS system

**Range-free localization:** It has no limitation on range so able to use freely in dynamic environment. Just it required more time as compare to short-range network.

The proposed algorithm falls in range-based localization. Before it proceeds for implementation of the efficient localization algorithm, we need to understand all range-based techniques such as RSSI, TOA, AOA, TDOA, trilateration, and triangulation, which will work under scenarios where every node is moving [3–5]. For an implementation of point of view, we have chosen RSS techniques where after every fixed period of time interval, the location of the all nodes will update the current location as well as predicted location for the next interval period. It means we will get the current and predicted locations concurrently. The change between the expected position of nodes at the previous interval of period and the actual location of nodes at the next interval period is known as a localization error.

In this work, we tried to minimize the localization error. In range-based technique, till now, various tactics have been presented with explicit hardware to improve accuracy by using distance and angle between two nodes. Till now, different approaches have been projected with hardware module fitted with nodes explicitly in range-based technique to receive a fine accurateness by measuring distance or angle between two nodes where at low-cost range-free techniques provide coarse accuracy. In mobile wireless sensor network, mobile anchor node is aware of their position using GPS continuously moving across the sensor area and broadcast their locations after a fixed time of interval. All unknown sensor nodes in the proximity of the anchor node receive the broadcasted message and based on the same, they are able to compute their current location.

To do localization for the stationary nodes, many algorithms were available but while nodes are mobile is quite challenging. In this algorithm, every node maintains a list of neighbors in its proximity and by making a pairing with two nearest neighboring nodes. This presented localization algorithm solves a set of trigonometric equations for the coordinates of the unknown node's location.

To avoid communication overhead by using antennas to assess the position of nodes and giving the routing data and its comparative location to neighbors in the network. Unknown nodes are those who do not know their location and can compute their location from anchor nodes along with the predicted location for the moving node's direction to achieve a better accuracy as outcomes with less time and less energy. By keeping utilization of minimum energy along with efficient location estimation of the sensor nodes is the main aim. The entire network lifespan can increase by saving the energy of all sensor nodes with a selective approach algorithm [6, 7].

## 2 Proposed Work

The position of each moving node needs to be traced regularly over the period of time after a fixed interval period. The predicted position of a node that is moving with constant speed  $v$  at instance of time at  $t$  seconds will be on the circle where center of the circle is node's current location and product of the speed and time of the node. Here, the challenge is to discover the actual location of every node at fixed time period "t".

### 2.1 Localization Algorithm

The localization algorithm is presented hereby where the starting location of all nodes will be from a random point and moves continuously with some fixed speed. All nodes are able to send and receive the data about their location of neighbors [8–10].

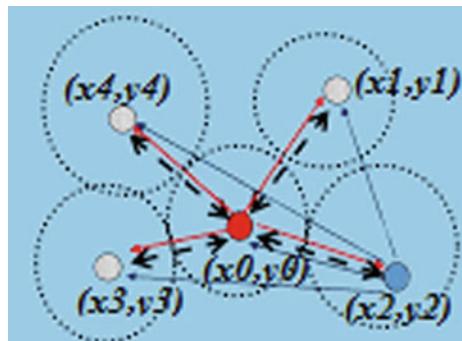
This algorithm needs some prerequisite and assumptions. In this scenario, node density is maintained in the network to have at least two neighbors in their proximity. As Fengqui Y. et al. presented the starting locations placement is done randomly for all nodes. As we know that every node must maintain a list of neighboring nodes, so that using a cooperative approach where the distance between this two nodes can be calculated using Time of Arrival (ToA), Time Difference of Arrival (TDoA), and Received Signal Strength Indicator (RSSI) (Fig. 4).

In Fig. 4, nodes  $A, B, C, D$ , and  $E$  are each other's neighbor and  $(x_0, y_0)$ ,  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $(x_3, y_3)$ , and  $(x_4, y_4)$  are location of the nodes, respectively.

From Fig. 3, we find the trigonometric equations as follows:

$$(x_0 - x_i)^2 + (y_0 - y_i)^2 = R \quad (1)$$

**Fig. 4** Neighboring nodes estimating position



$$(x_0 - x_0)^2 + (y_0 - y_0)^2 = R \quad (2)$$

$$(x_0 - x_1)^2 + (y_0 - y_1)^2 = R \quad (3)$$

$$(x_0 - x_2)^2 + (y_0 - y_2)^2 = R \quad (4)$$

$$(x_0 - x_3)^2 + (y_0 - y_3)^2 = R \quad (5)$$

$$(x_0 - x_4)^2 + (y_0 - y_4)^2 = R \quad (6)$$

Here,  $i = 00, 01, 02, 03, 04$ .

Whereas, coordinates of the predicted positions are denoted as  $(x_0, y_0)$ ,  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $(x_3, y_3)$ , and  $(x_4, y_4)$  for circle  $a$ ,  $b$ ,  $c$ ,  $d$ , and  $e$ , respectively. By solving such trigonometric equation, we will get ten unknown coordinates values (Fig. 5).

### The Proposed Algorithm Description

Step 1: After every fixed interval period of  $t$  seconds, all nodes send a sample message which consists of the node ID and their current position.

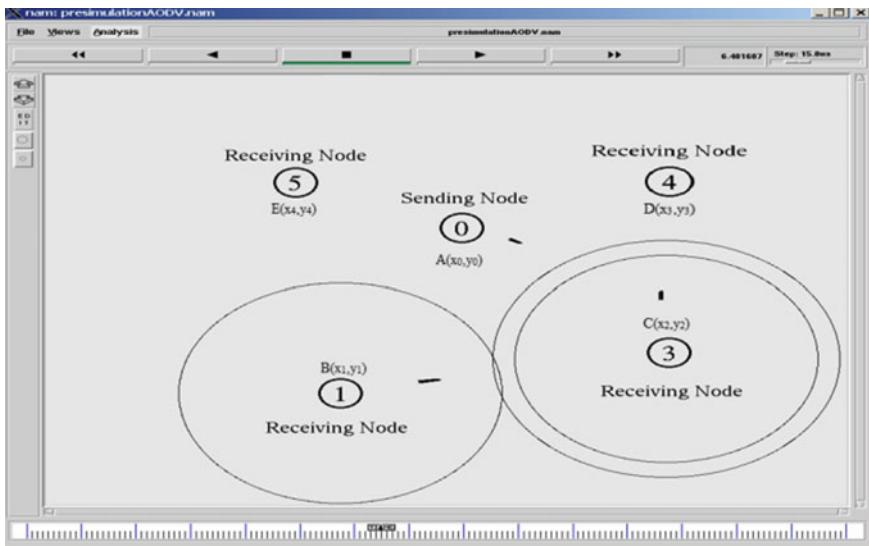
Step 2: On reception of sample message from neighboring nodes, every node prepare a list of neighbors in their proximity and maintain a list of neighbors.

Step 3: Ultimately, all nodes also have the information of list of neighbors of their immediate neighbors.

Step 4: To apply trilateration, RSS methods need to select any two nodes from their neighboring list to find the exact position.

Step 5: By solving given computation for  $(x_i, y_i)$ , where  $i = 1-5$ . If the results are the unique, the location information is correct. If there is any difference in results, we need to repeat Step 3–Step 5.

$$d = \sqrt{(x - x_i)^2 + (y - y_i)^2} \quad (5)$$



**Fig. 5** Sensor node communication by finding the location

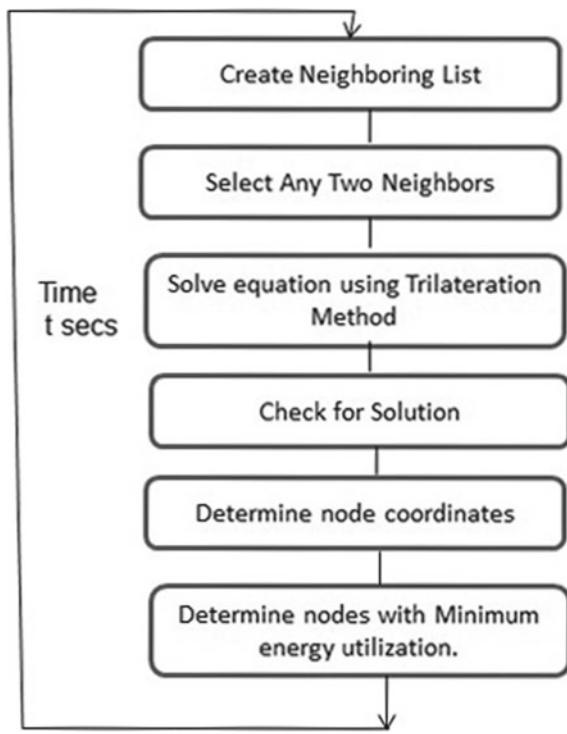
In the present algorithm, a set of nodes are moving with constant speed, which estimates the location data of neighbors of neighbor. Where all nodes start randomly with fixed constant speed  $v$ . To obtain the distance amid nodes, we used trilateration technique to compute the location of unknown node (Fig. 6).

As shown in Fig. 3, the distance as a length AB, AC, AD, and AE are computed and from this, the location of node will be obtained. Once predefined fixed interval time period of interval “ $t$ ” seconds the all steps will be repeated. To ten each nodes position they choose a neighbors from their neighbor list as shown in Fig. 2 and get it ten the position. Same time while fixing current location of node also predicts its predicted position using nodes speed  $v$  and the time periodinterval “ $t$ ”, here we get the predicted position. For next time interval of  $2t$  result value will be computed as a actual output position of node. Here, we will get the localization error as a variance between the actual location as well as predicted position of the node [10].

Here, while setting up an experiment, the scenario we have taken as localization of mobile nodes where the positions of nodes are moving in arbitrary direction continuously [11, 12]. We considered the size of network as  $100 * 100$  in simulation scenario where node density is chosen as six nodes. The range of communication is chosen for communication between two nodes over the maximum distance is 6.5 m.

Here, in the scenario, all nodes are moving with a fixed speed, where the location of each node is calculated after every fixed interval of time “ $t$ ”. At every interval of time “ $t$ ”, the algorithm gives us one predicted location and another is actual location. We considered as estimated position, i.e.,  $\epsilon_i$  and real position, i.e.,  $f_i$ . Here, localization error is denoted as  $\tau_i$  and can be determined using the difference between the actual and predicted position, i.e.,  $\tau_i = |f_i - \epsilon_i|$ . Localization error will be calculated for

**Fig. 6** Proposed algorithm steps



each node and then by averaging, the error of localization for all nodes is called as ALE (Fig. 7) [12–15].

ALE can be calculated by taking root mean square error of all nodes divided by number of nodes present in the network. Here by taking the difference between the coordinates of predicted location and actual location for that interval value of each node [16].

#### Node Configuration

S. No.	Parameter	Details
1	llType	LL
2	macType	Mac/802_11
3	ifqType	Queue/DropTail/PriQueue
4	ifqLen	50
5	antType	Antenna/OmniAntenna
6	propType	Propagation/TwoRayGround
7	phyType	Phy/WirelessPhy
8	channelType	Channel/WirelessChannel

(continued)

(continued)

S. No.	Parameter	Details
9	topoInstance	\$topo
10	energyModel	\$opt(energymodel)
11	rxPower	0.6,0.8
12	txPower	0.9,1.2
13	initialEnergy	\$opt(initialenergy)
14	agentTrace	ON
15	routerTrace	ON
16	macTrace	OFF

### 3 Results

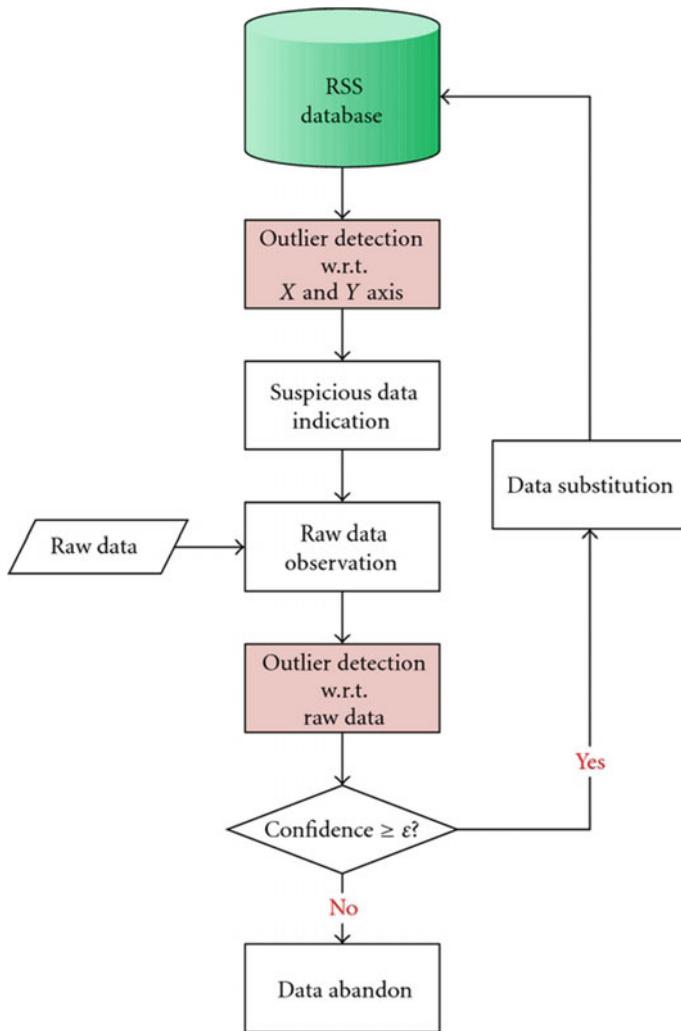
In this result and analysis, we got result for each node on different time intervals. To do this analysis, we used an output file, which is generated from the system. This analysis can be done in two ways, i.e., node-based approach and time-based approach. We will check it by doing the analysis on node-based approach.

As shown in Fig. 8 graph for node based analysis how much data transfer, localization error for all nodes taken into consideration and after taking average of all nodes on different interval time period. On node 4, the localization error is minimum. It means that at different time intervals after taking the average of localization error for all nodes, we found node 4 is getting minimum localization error over all time intervals prescribed in the scenario.

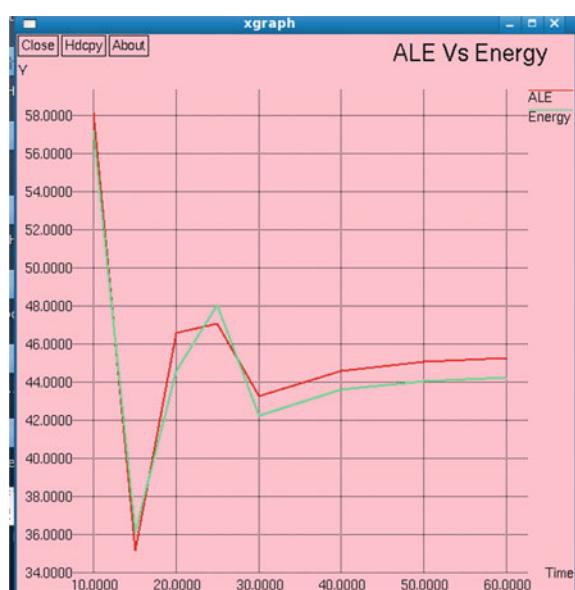
In this result and analysis, we got result on different time intervals for all nodes. To do this analysis, we used an output file, which is generated from the system. This analysis can be done by considering time-based approach. We will check it by doing analysis on different time intervals used in the scenario. As shown in the graph of Fig. 9 for time-based analysis approach, at which time interval nodes gives minimum localization error out of all time interval periods. As shown in the graph, localization error is gradually decreasing and its lowest value for time interval 10 where nodes average localization error is minimum. If we use this time interval efficiently for localization, we will get the minimum value for localization error for all nodes.

How energy can be calculated:

Energy can be calculated by using different parameters how much data send & received and how much time it required to deliver to it by considering the bandwidth of the system [16, 17]. All details of energy calculation is as follows:



**Fig. 7** How RSS technique works to get location

**Fig. 8** Node-based analysis**Fig. 9** Time-based analysis

```

# Variable Energy configuration for Nodes
set upper_limit1 99
set upper_limit2 0.9
for { set i 0 } { $i< $val(nn) } { incr i } {
set energy [expr rand()*500]
set file($i) [open energyfile($i).tr w]
puts $file($i) "$energy"
$ns_node-config -initialEnergy $energy
set node_($i) [$ns_node]
set p_id $i
set data_file($i) [open data_file($i).tr w]
set rand1 [expr double(rand()+$upper_limit2)]
set rand2 [expr double(rand()+$upper_limit1)]
set rand3 [expr double(rand()+$upper_limit2)]
puts $data_file($i) "$energy"
puts $data_file($i) "$rand1"
puts $data_file($i) "$rand2"
puts $data_file($i) "$rand3"
close $file($i)
close $data_file($i)
}

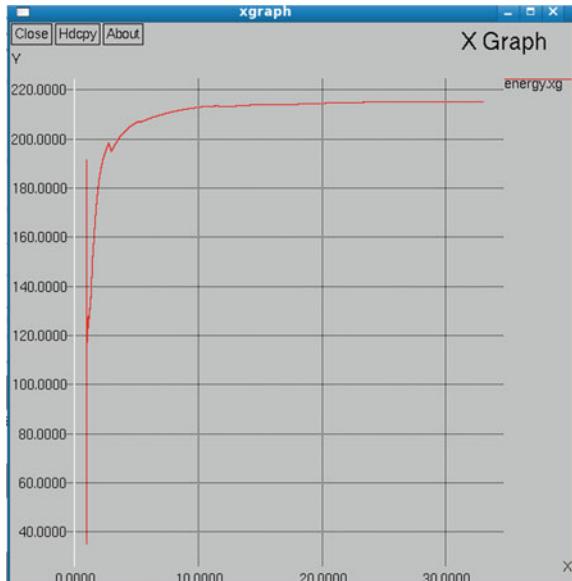
```

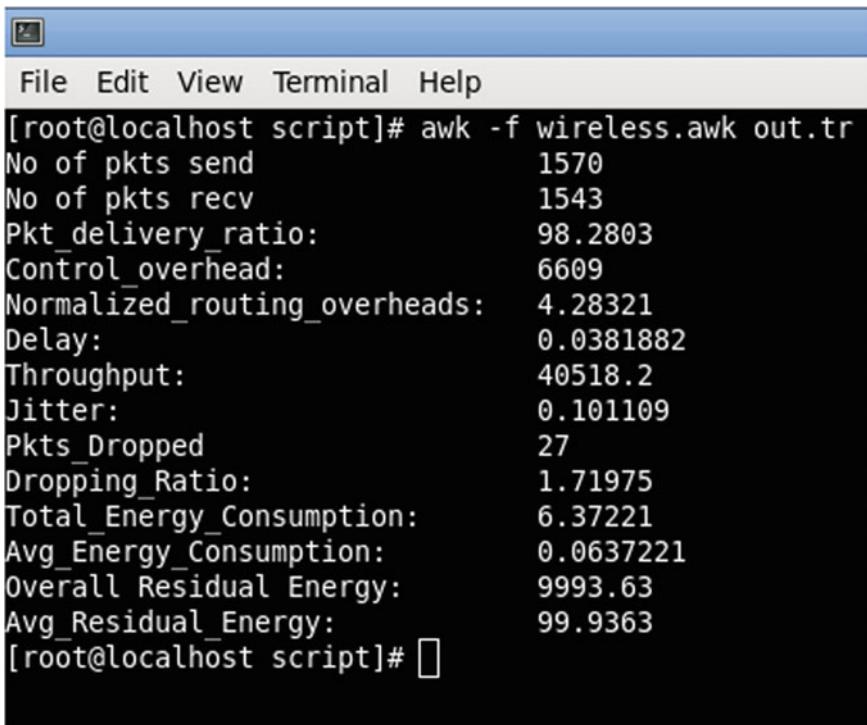
By using the above method, we got the following result from where we can decide how much energy is required.

As far as energy consumption calculation is concerned, our objective is very clear that we want to identify at what time interval which node gives minimum localization error with minimum energy consumption [18].

We have used the energy model with 50 Jules for each node at the start and after completion, the remaining energy with each node for different time periods. As shown in Fig. 10, time interval 10–15 ms is giving minimum consumption with minimum Average Localization Error (ALE) (Fig. 11).

**Fig. 10** Reduces the energy consumption by using maximum resources utilization





The screenshot shows a terminal window with a blue header bar containing icons for file operations. Below the header is a menu bar with "File", "Edit", "View", "Terminal", and "Help". The main area of the terminal displays the output of an awk script on a file named wireless.awk. The output lists various network performance metrics:

Metric	Value
No of pkts send	1570
No of pkts recv	1543
Pkt_delivery_ratio:	98.2803
Control_overhead:	6609
Normalized_routing_overheads:	4.28321
Delay:	0.0381882
Throughput:	40518.2
Jitter:	0.101109
Pkts_Dropped	27
Dropping_Ratio:	1.71975
Total_Energy_Consumption:	6.37221
Avg_Energy_Consumption:	0.0637221
Overall Residual Energy:	9993.63
Avg_Residual_Energy:	99.9363

[root@localhost script]#

**Fig. 11** Minimum consumption of energy with minimum ALE

#### 4 Conclusion

We presented an algorithm for the mobile WSN, where the available localization algorithms are not suitable for the given scenario. Hereby making pair with four neighboring nodes in a given network with range-free localization using it will give more scope to go with accuracy, which will help to reduce the localization error. From the above method, we are able to calculate the PDF and throughput with less energy consummation and minimum localization error, hence, we are improving the QoS of the network. In future scope, we can compare this with existing techniques to improve the QoS of the network to minimize average localization error.

## References

1. Rehman, M. N., Hanuranto, A. T., & Mayasari, R. (2017). Trilateration and iterative multilateration algorithm for localization schemes on wireless sensor network. In *IEEE International Conference on Control, Electronics, Renewable Energy and Communications*.
2. Leila, C., Faouzi, S., & Louiza, B. (2017). Localization protocols for mobile wireless sensor networks: A survey. *Computers and Electrical Engineering*.
3. Paul, A. K., & Sato, T. (2017). Localization in wireless sensor networks: A survey on algorithms, measurement techniques, applications and challenges. *Journal of Sensor and Actuator Networks*, 6, 24. <https://doi.org/10.3390/jsan6040024>.
4. Han, G., Jhang, C., & Jiang, J. (2017). Mobile anchor nodes path planning algorithms using network-density-based clustering in wireless sensor networks. *Journal of Network and Computer Applications*.
5. Tambe, K., Mohan, G. K., et.al. (2016). A novel approach of efficient localization scheme for wireless sensor network. In *IJST*, December 2016.
6. Livinsa, Z., & Jaya Shri, S. (2016). Monitoring moving target and energy saving localization algorithm in wireless sensor networks. In *IJST*, Janauary 2016.
7. Santar, P. S., et.al. (2015). Range free localization techniques in wireless sensor networks: A review. *Procedia Computer Science* 7–16.
8. Han, G., Chao, J., Zhang, C., & Shu, L. (2014). The impacts of mobility models on DV-hop based localization in mobile wireless sensor networks. *Journal of Network and Computer Applications*, 2014.
9. Quai, W., Han, J., & Jun, L. (2013). A linearization reference node selection strategy for accurate multilateration localization in wreless sensor networks. In *IEEE*, February 2013.
10. Kathole, A. B., & Pande, Y. (2013). Survey of topology based reactive routing protocols in vanet. *International Journal of Scientific & Engineering Research*, 4(6). ISSN 2229–5518.
11. Sundaram, B., & Kavitha, R. (2012). Minimizing the localization error in wireless sensor network. *Procedia Engineering*.
12. Hatware, I. V., Kathole, A. B., & Bompilwar, M. D. (2012). *Detection of misbehaving nodes in ad hoc routing*. *International Journal of Emerging Technology and Advanced Engineering*, 2(2). Website: [www.ijetae.com](http://www.ijetae.com). ISSN 2250-2459.
13. Amitangshu, P. (2010). Localization algorithms in Wireless Sensor Net-works: Current approaches and future challenges. *Network Protocols and Algorithms*.
14. Mao, G., & Fidan, B. (2009). *Localization algorithms and strategies for wireless sensor networks*. Hershey, PA, USA: Imprint of IGI Publishing.
15. Ssu, K. F., & Ou, C. H. (2007). Localization with mobile anchor points in wireless sensor networks. *IEEE Transaction Vehicular Technology*.
16. Yu, G., & Yu, F. (2007). A localization algorithm for mobile wireless sensor networks. In *IEEE International Conference on Integration Technology*, April 2007.
17. Nissanka, P., Hari, B., et.al. (2005). Mobile assisted localization in wireless sensor networks. In *Proceedings of IEEE INFOCOM*, Miami, FL, March 2005.
18. Hu, L., & David, E. (2004). Localization for mobile sensor networks. In *MobiCom*.
19. Optimization of vehicular adhoc network using cloud computing.

# Author Index

## A

- Ade, Joseph, 19  
Adewumi, Adewole, 19, 31  
Ahuja, Ravin, 19, 31  
Alkatheeri, Yazeed, 231  
Alrajawy, Ibrahim, 249  
Al-Shbami, Ahmed, 249  
Ameen, Ali, 231, 249  
Amudha, S., 281  
Arunkumar, K., 205  
Assibong, Patrick A., 295  
Ayeni, Foluso, 31  
Azeez, Nureni A., 19

## B

- Babu, Yeddu Vijaya, 43  
Bakar, Zainab Binti Abu, 217  
Bandara, Akila, 71  
Barhate, Rahul, 343  
Bhattacharjee, Shiladitya, 459  
Bhattacharyya, Rishab, 103  
Biswas, Bhaskar, 445

## C

- Chakrabarti, Amlan, 55, 325  
Chakraborty, Basabi, 325  
Chandrakar, Preeti, 167  
Chandrasekaran, K., 377  
Chavan, Sumit, 343  
Chenna Keshava, B.S., 377  
Choudhury, Sabyasachi Roy, 391

## D

- Damasevičius, Robertas, 31, 295  
Das, Aditya, 103  
Deshpande, Vivek, 475, 477  
Devendran, A., 205  
Devi, Bali, 179  
Dhondse, Amol, 343  
Dinakar, B.R., 267  
Diwan, Prabhat, 95  
Dulhare, Uma N., 419  
Duraiasamy, Balagamesh, 231  
Dwivedi, Ankit, 401

## E

- Elizabeth Shanthi, I., 281

## F

- Fahad, S.K.Ahammad, 217  
Fegade, TanujaK., 311

## G

- Ganguli, Bhaswati, 55  
Ghosh, Pramit, 103  
Goel, Varun, 153  
Goswami, Saptarsi, 325  
Govil, Himanshu, 95

## H

- Haddad, Adel, 249  
Hettiarachchi, Kusal, 71  
Hettiarachchi, Yashodha, 71

**I**

Isaac, Osama, 231, 249

**J**

Jain, Harsh, 391  
 Jangir, Vishal, 153  
 Jyoti, Gautam, 131

**K**

Kane, Indrajeet, 343  
 Kanungo, P., 191  
 Kar, Madhuchanda, 55  
 Kar, T., 191  
 Khadilkar, Kunal, 343  
 Khaleed, Areej Mohammed, 419  
 Khalifa, Gamal S.A., 231  
 Khatri, Sapna, 365  
 Komal, Malsa, 131  
 Krishna Mohan, G., 485, 487  
 Kulkarni, Siddhivinayak, 343

**M**

Mahapatra, Bandana, 431  
 Majumdar, Atanu, 103  
 Malakar, Sourav, 325  
 Mandal, Indrajit, 401  
 Martinovič, Jan, 115  
 Maskeliunas, Rytis, 31  
 Midhun Chakkaravarthy, 459  
 Midhun Chakkaravarthy, Divya, 249, 459  
 Misra, Sanjay, 19, 31, 295  
 Mithon, Ahmed M., 217  
 Monika, 95  
 Mukherjee, Jhilam, 55  
 Munasinghe, Sidath, 71

**N**

Nayyar, Anand, 431  
 Nitima, Malsa, 131  
 Nusari, Mohammed, 231

**P**

Pandey, Devendra Kumar, 365  
 Patil, Malini, 267  
 Pawar, B.V., 311  
 Penkar, Daniel, 365  
 Poddar, Triparna, 55

Poulkov, Vladimir, 475, 477

Prabhune, Ashish, 391  
 Ptošek, Vít, 115

**R**

Rahim, Lukman Bin Ab., 459  
 Ramani, Jaiprakash, 365  
 Rapant, Lukáš, 115

**S**

Senthil, D., 3  
 Shaik, Nazeer, 445  
 Shakya, Harish Kumar, 445  
 Shanbhag, Varun, 391  
 Shankar, Venkatesh Gauri, 153, 167, 179  
 Sharma, Neha, 267, 295  
 Sholarin, Muyiwa Adeniyi, 295  
 Singh, Kuldeep, 445  
 Sinha, G.R., 445  
 Sisodia, Dilip Singh, 167  
 Srivastava, Devesh K., 179  
 Srivastava, Sumit, 179  
 Sumukha, P.K., 377  
 Suseendran, G., 3, 141

**T**

Taiwo, Adebola, 31  
 Tambe, Kailas, 485, 487  
 Thayasilvam, Uthayasanker, 71  
 Thiagaraj, M., 141  
 Tripathi, Mahesh Kumar, 95

**U**

Usha, D., 377

**V**

Vikas, Singhal, 131  
 Vyver, Charles Van der, 19

**W**

Wijesinghe, Ishara, 71  
 Wogu, Ikedinachi Ayodele Power, 295

**Y**

Yafooz, Wael M.S., 217