

Satchidananda Dehuri
Bhabani Shankar Prasad Mishra
Pradeep Kumar Mallick
Sung-Bae Cho *Editors*



Biologically Inspired Techniques in Many Criteria Decision Making

Proceedings of BITMDM 2021



Smart Innovation, Systems and Technologies

Volume 271

Series Editors

Robert J. Howlett, Bournemouth University and KES International,
Shoreham-by-Sea, UK

Lakhmi C. Jain, KES International, Shoreham-by-Sea, UK

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

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

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

More information about this series at <https://link.springer.com/bookseries/8767>

Satchidananda Dehuri ·
Bhabani Shankar Prasad Mishra ·
Pradeep Kumar Mallick · Sung-Bae Cho
Editors

Biologically Inspired Techniques in Many Criteria Decision Making

Proceedings of BITMDM 2021

Editors

Satchidananda Dehuri
Fakir Mohan University
Balasore, Odisha, India

Pradeep Kumar Mallick
KIIT Deemed to be University
Bhubaneswar, India

Bhabani Shankar Prasad Mishra
KIIT Deemed to be University
Bhubaneswar, Odisha, India

Sung-Bae Cho 
Yonsei University
Seoul, Korea (Republic of)

ISSN 2190-3018

ISSN 2190-3026 (electronic)

Smart Innovation, Systems and Technologies

ISBN 978-981-16-8738-9

ISBN 978-981-16-8739-6 (eBook)

<https://doi.org/10.1007/978-981-16-8739-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

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

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

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

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

Preface

Many criteria decision making (MCDM) is a process of finding a solution in an environment with more than three criteria. Many problems of real life consist of more than three objectives which should be taken into account. Solving problems of such category is a challenging task and requires full attention of deep investigation. Moreover many-objective optimization-the kernel of MCDM brings with it a number of challenges that must be addressed, which highlights the need for new and better algorithms that can efficiently handle the growing number of objectives. In real applications, often simple and easy to understand methods are used and lead to situations in which solutions accepted by decision makers are not optimal solutions. On the other hand, algorithms which do not lead to such situations are very time consuming. The biggest challenge standing in front of the researchers is how to create effective algorithms which will lead to optimal solutions with low time complexity. That is why lots of effort have given to implement biologically inspired algorithms (BIAs), which succeed solving uni-objective problems. Hence to introduce the readers about the state-of-the-art development of biologically inspired techniques and their applications with a great emphasis on MCDM process, this platform has been developed and the ultimate outcome of this platform shall be disseminated through a conference proceedings. This book not only is restricted with the blend of contributions of BIAs and MCDM process but also includes the contributions from nature-inspired algorithms, multi-criteria optimization, and machine learning and soft computing.

Although the scope of the topics of this conference is divided into four different tracks, the contributions received for inclusion in this proceedings go beyond. Let us go through some of the contributions very quickly. De et al. have used locally and globally tuned biogeography-based optimization to identify end vertices for the link recommendations for maximization of social influence. Sahu et al. have used neural networks for liver disease diagnosis. Adhikary et al. have used some of the machine learning algorithms for prediction of used cars. Sahoo et al. have developed a new approach by combining the best features of quasi-opposition-based Roa algorithm (QORA) and artificial neural networks for complexity classification of object-oriented projects. Sahu et al. have contributed a COVID-19 awareness article that is developed to provide latest and correct information regarding the current

pandemic situation. Feature reduction is one of the essential steps for machine learning applications. To select an optimal set of features, Das et al. have proposed a new particle swarm optimization technique in their contribution. Odia language is one of the 30 most spoken languages in the world. It is spoken in the Indian state called Odisha. Odia language lacks online content and resources for natural language processing (NLP) research. There is a great need for a better language model for the low resource Odia language, which can be used for many downstream NLP tasks. Parida et al. have introduced a Bert-based language model, pre-trained on 430,000 Odia sentences. Das et al. have presented a computational approach to assess breast cancer risk in relation to lifestyle factors. Nayak et al. have made an extensive review on convolutional neural network applications for MRI dissection and detection of brain tumor. Das et al. have presented an application of expectation–maximization algorithm to solve lexical divergence in Bangla–Odia machine translation. A study on the performance of machine learning techniques before and after COVID-19 on Indian foreign exchange rate has been made by Pandey et al. Sengupta et al. have proposed a multi-criterion decision making model for online education system perspective. Existing multi-objective evolutionary algorithms can handle multi-objective optimization problems (MOPs) with regular Pareto fronts in which non-dominated solutions are distributed continuously over the objective space. The study by Raju et al., have proposed a clustering-based environmental selection for tackling MOPs with irregular Pareto fronts to address this issue. Das et al. have presented an efficient evolutionary technique for solving nonlinear fixed charge transportation problem in their contribution.

Balasore, India

Bhubaneswar, India

Bhubaneswar, India

Seoul, Korea (Republic of)

December 2021

Satchidananda Dehuri

Bhabani Shankar Prasad Mishra

Pradeep Kumar Mallick

Sung-Bae Cho

Acknowledgements All the papers submitted to this conference were peer-reviewed by the members of this conference reviewer board. We are very grateful to the reviewers for their support and services as without them, this proceedings would not have been possible.

Contents

1	Cloud-Based Smart Grids: Opportunities and Challenges	1
	Atta-ur-Rahman, Nehad M. Ibrahim, Dhiaa Musleh, Mohammed Aftab A. Khan, Sghaier Chabani, and Sujata Dash	
2	A Resource-Aware Load Balancing Strategy for Real-Time, Cross-vertical IoT Applications	15
	Ranjit Kumar Behera, Amrut Patro, and Diptendu Sinha Roy	
3	An Elitist Artificial-Electric-Field-Algorithm-Based Artificial Neural Network for Financial Time Series Forecasting	29
	Sarat Chandra Nayak, Ch. Sanjeev Kumar Dash, Ajit Kumar Behera, and Satchidananda Dehuri	
4	COVID-19 Severity Predictions: An Analysis Using Correlation Measures	39
	Rashmita khilar, T. Subetha, and Mihir Narayan Mohanty	
5	Antenna Array Optimization for Side Lobe Level: A Brief Review	53
	Sarmistha Satrusallya and Mihir Narayan Mohanty	
6	Accuracy Analysis for Predicting Heart Attacks Based on Various Machine Learning Algorithms	61
	Rashmita khilar, T. Subetha, and Mihir Narayan Mohanty	
7	Link Recommendation for Social Influence Maximization	71
	Sagar S. De, Parimal Kumar Giri, and Satchidananda Dehuri	
8	Performance Analysis of State-of-the-Art Classifiers and Stack Ensemble Model for Liver Disease Diagnosis	95
	Barnali Sahu, Supriya Agrawal, Hiranmay Dey, and Chandani Raj	

9	CryptedWe: An End-to-Encryption with Fake News Detection Messaging System	107
	Anukampa Behera, Bibek K. Nayak, Saswat Subhadarshan, and Nilesh Nath	
10	Enabling Data Security in Electronic Voting System Using Blockchain	119
	M. Thangavel, Pratyush Kumar Sinha, Ayusman Mishra, and Bhavesh Kumar Behera	
11	Prediction of Used Car Prices Using Machine Learning	131
	Dibya Ranjan Das Adhikary, Ronit Sahu, and Sthita Pragyna Panda	
12	Complexity Classification of Object-Oriented Projects Based on Class Model Information Using Quasi-Opposition Rao Algorithm-Based Neural Networks	141
	Pulak Sahoo, Ch. Sanjeev Kumar Dash, Satchidananda Dehuri, and J. R. Mohanty	
13	Mood-Based Movie Recommendation System	151
	Soumya S. Acharya, Nandita Nupur, Priyabrat Sahoo, and Paresh Baidya	
14	Covid-19 and Awareness of the Society: A Collection of the Important Facts and Figures Related to the Global Pandemic	159
	Prabhat Kumar Sahu, Parag Bhattacharjee, and Nikunj Agarwal	
15	Implementation of Blockchain-Based Cryptocurrency Prototype Using a PoW Consensus Mechanism	171
	Danish Raza, Pallavi Nanda, and Sudip Mondal	
16	Employing Deep Learning for Early Prediction of Heart Disease	181
	Abdul Aleem, Ayush Raj, Rahul Raj Sahoo, and Amulya Raj	
17	Detection of COVID-19 Cases from Chest Radiography Images	191
	Aniket Kumar, Nishant Niraj, Venkat Narsimam Tenneti, Brijendra Pratap Singh, and Debahuti Mishra	
18	Monitoring the Heart Rate—An Image Processing Approach	203
	Samuka Mohanty, Sumit Pal, Shubhrajit Parida, and Manosmita Swain	
19	Evaluation of Optimal Feature Transformation Using Particle Swarm Optimization	211
	Dibyasundar Das, Suryakant Prusty, Biswajit Swain, and Tushar Sharma	

20 Brain Image Classification Using Optimized Extreme Gradient Boosting Ensemble Classifier	221
Abhishek Das, Saumendra Kumar Mohapatra, and Mihir Narayan Mohanty	
21 Concolic-Based Software Vulnerability Prediction Using Ensemble Learning	231
Swadhin Kumar Barisal, Pushkar Kishore, Gayatri Nayak, Ridhy Pratim Hira, Rohit Kumar, and Ritesh Kumar	
22 Very Short-Term Photovoltaic Power Forecast by Multi-input Long Short-Term Memory Network	243
Sasmita Behera, Debasmita Mohapatra, Aman Kaushal, and Shrabani Sahu	
23 I Hardly Lie: A Multistage Fake News Detection System	253
Suchintan Mishra, Harshit Raj Sinha, Tushar Mitra, and Manadeepa Sahoo	
24 Software Design for Mobile Phone Auto-Silencer	263
Lambodar Jena and Dipti Pratima Minz	
25 Traffic Flow Prediction: An Approach for Traffic Management	273
Utkarsh Jha, Lubesh Kumar Behera, Somnath Mandal, and Pratik Dutta	
26 Detection and Classification of Encephalon Tumor Using Extreme Learning Machine Learning Algorithm Based on Deep Learning Method	285
Premananda Sahu, Prakash Kumar Sarangi, Srikanta Kumar Mohapatra, and Bidush Kumar Sahoo	
27 Comparative Study of Medical Image Segmentation Using Deep Learning Model	297
Pubali Chatterjee, Simran Sahoo, Subrat Kar, and Pritikrishna Biswal	
28 A Review of Challenges and Solution in Peer-to-Peer Energy Trading of Renewable Sources	307
Ritweek Das, Stuti Snata Ray, Gayatri Mohapatra, and Sanjeeb Kumar Kar	
29 Data Transfer Using Light Fidelity (Li-Fi) Technology—A Methodological Comparative Study	315
Lirika Singh, Manish Rout, J. S. Bishal, and Jayashree Ratnam	
30 Study on Surveillance of Crop Field Using Smart Technique	323
Manesh Kumar Behera, Sobhit Panda, Sonali Goel, and Renu Sharma	

31 Smart Student Performance Monitoring System Using Data Mining Techniques	337
Jay Bijay Arjun Das, Saumendra Kumar Mohapatra, and Mihir Narayan Mohanty	
32 BertOdia: BERT Pre-training for Low Resource Odia Language	345
Shantipriya Parida, Satya Prakash Biswal, Biranchi Narayan Nayak, Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Satya Ranjan Dash	
33 A Machine Learning Approach to Analyze Mental Health from Reddit Posts	357
Smriti Nayak, Debolina Mahapatra, Riddhi Chatterjee, Shantipriya Parida, and Satya Ranjan Dash	
34 A Computational Approach to Assess Breast Cancer Risk in Relation with Lifestyle Factors	367
Oindrila Das, Itismita Pradhan, Riddhi Chatterjee, and Satya Ranjan Dash	
35 Digitalization of Education: Rural India's Potential to Adapt to the Digital Transformation as New Normality	377
Ankita Sahu and Swati Samantaray	
36 Methodologies and Tools of Sentiment Analysis: A Review	389
Bijayalaxmi Panda, Chhabi Rani Panigrahi, and Bibudhendu Pati	
37 A Review: Convolutional Neural Network Application for MRI Dissection and Detection of Brain Tumor	403
Dillip Ranjan Nayak, Neelamadhab Padhy, Pradeep Kumar Mallick, and Dilip Kumar Bagal	
38 Neutrosophic Logic and Its Scientific Applications	415
Sitikantha Mallik, Suneeta Mohanty, and Bhabani Shankar Mishra	
39 Application of Expectation–Maximization Algorithm to Solve Lexical Divergence in Bangla–Odia Machine Translation	433
Bishwa Ranjan Das, Hima Bindu Maringanti, and Niladri Sekhar Dash	
40 Impact of Odisha Gramya Bankon Socio-economic Development of Beneficiaries: A Case Study of Balasore and Bhadrak District of Odisha	441
Arutta Bandhu Jena and Debadutta Nayak	
41 Performance of Machine Learning Techniques Before and After COVID-19 on Indian Foreign Exchange Rate	467
Trilok Nath Pandey, Rashmi Ranjan Mahakud, Bichitrana Patra, Parimal Kumar Giri, and Satchidananda Dehuri	

42 Effects of Binning on Logistic Regression-Based Predicted CTR Models	483
Manvik B. Nanda, Bhabani Shankar Prasad Mishra, and Vishal Anand	
43 Proposed Multi-criterion Decision-Making Model—On Online Education System Perspective	495
Ishani Sengupta, Bhabani Shankar Prasad Mishra, and Pradeep Kumar Mallick	
44 Indoor Plant Health Monitoring and Tracking System	507
Nikhil Kumar, Sahil Anjum, Md Iqbal, Asma Mohiuddin, and Subhashree Mishra	
45 Hybrid of Array-Based and Improved Playfair Cipher for Data Security	517
K. R. Harini, D. Vijayaraghavan, S. Sushmidha, Vithya Ganesan, and Pachipala Yellamma	
46 Modeling Internet of Things-Based Solution for Evading Congestion and Blockage in Waterways	527
Madhuri Rao, Narendra Kumar Kamila, Sampa Sahoo, and Kulamala Vinod Kumar	
47 A Multi-objective Evolutionary Algorithm with Clustering-Based Two-Round Selection Strategy	537
M. Sri Srinivasa Raju, Kedar Nath Das, and Saykat Dutta	
48 An Efficient Evolutionary Technique for Solving Non-linear Fixed Charge Transportation Problem	551
Rajeev Das and Kedar Nath Das	
49 Solar PV Application in Aerospace Technologies	561
Saumya Ranjan Lenka, Sonali Goel, and Renu Sharma	
50 Cloud Classification-Based Fine KNN Using Texture Feature and Opponent Color Features	567
Prabira Kumar Sethy and Sidhant Kumar Dash	
51 Machine Learning-Based Diabetes Prediction Using Missing Value Impotency	575
Santi Kumari Behera, Julie Palei, Dayal Kumar Behera, Subhra Swetanisha, and Prabira Kumar Sethy	
52 A Detailed Schematic Study on Feature Extraction Methodologies and Its Applications: A Position Paper	585
Niharika Mohanty, Manaswini Pradhan, and Pradeep Kumar Mallick	

53	Image Colorization Using CNNs	603
	R. I. Minu, S. Vishnuvardhan, Ankit Pasayat, and G. Nagarajan	
54	Music Generation Using Deep Learning	613
	R. I. Minu, G. Nagarajan, Rishabh Bhatia, and Aditya Kunar	
55	Detection of Heart Disease Using Data Mining	627
	G. Kalaiarasi, M. Maheswari, M. Selvi, R. Yogitha, and Prathima Devadas	
56	Smart Agriculture Framework Implemented Using the Internet of Things and Deep Learning	639
	R. Aishwarya, R. Yogitha, L. Lakshmanan, M. Maheshwari, L. Suji Helen, and G. Nagarajan	
57	Effect of COVID-19 on Stock Market Prediction Using Machine Learning	649
	J. Kalaivani, Ronak Singhania, and Shlok Garg	
58	Real-Time Elderly Multiple Biological Parameter Safety Monitoring System Based on mCloud Computing: Telemedicine Era	657
	Sawsan D. Mahmood, Raghda Salam Al Mahdawi, and Shaimaa K. Ahmed	
59	Intelligent Transportation Systems (ITSs) in VANET and MANET	667
	Sami Abduljabbar Rashid, Lukman Audah, and Mustafa Maad Hamdi	
60	Lexicon-Based Argument Extraction from Citizen's Petition in Arabic Language	677
	Sura Sabah Rasheed and Ahmed T. Sadiq	
61	Multi-key Encryption Based on RSA and Block Segmentation	687
	Rana JumaaSarih Al-Janabi and Ali Najam Mahawash Al-Jubouri	
62	Chaotic Pseudo Random Number Generator (cPRNG) Using One-Dimensional Logistic Map	697
	Ayan Mukherjee, Pradeep Kumar Mallick, and Debahuti Mishra	
63	Consumer Buying Behavior and Bottom of Pyramid (BoP): The Diffusion of Marketing Strategy	709
	Artta Bandhu Jena and Lopamudra Pradhan	
	Author Index	733

About the Editors

Satchidananda Dehuri is working as Professor in the Department of Computer Science (erstwhile Information and Communication Technology), Fakir Mohan University, Balasore, Odisha, India, since 2013. He received his M.Tech. and Ph.D. degrees in Computer Science from Utkal University, Vani Vihar, Odisha, in 2001 and 2006, respectively. He visited as BOYSCAST Fellow to the Soft Computing Laboratory, Yonsei University, Seoul, South Korea, under the BOYSCAST Fellowship Program of DST, Government of India, in 2008. In 2010, he received Young Scientist Award in Engineering and Technology for the year 2008 from Odisha Vigyan Academy, Department of Science and Technology, Government of Odisha. His research interests include evolutionary computation, neural networks, pattern recognition, and data mining. He has already published more than 200 research papers in reputed journals and referred conferences and has published five text books for undergraduate and post-graduate students and edited more than ten books of contemporary relevance. Under his direct supervision, 17 Ph.D. scholars have been successfully awarded. His h-index (Google Scholar) is more than 25.

Bhabani Shankar Prasad Mishra born in Talcher, Odisha, India, in 1981. He received the B.Tech. in Computer Science and Engineering from Biju Pattanaik Technical University, Odisha in 2003, M.Tech. degree in Computer Science and Engineering from the KIIT University, in 2005, Ph.D. degree in Computer Science from F. M. University, Balasore, Odisha, India, in 2011, and Post-Doc in 2013 from Soft Computing Laboratory, Yonsei University, South Korea. Currently he is working as Associate Professor at School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India. His research interest includes pattern reorganization, data mining, soft computing, big data, and machine learning. He has published more than 80 research articles in reputed journal and conferences and has edited more than five books of current importance. Under his guidance, 2 Ph.D. scholars are already been awarded. Dr. Mishra was the recipient of the Gold Medal and Silver Medal during his M.Tech. for the best Post-Graduate in the university. He is Member of different technical bodies ISTE, CSI, and IET. His h-index (Google Scholar) is 14.

Dr. Pradeep Kumar Mallick is currently working Senior Associate Professor in the School of Computer Engineering, Kalinga Institute of Industrial technology (KIIT) Deemed to be University, Odisha, India. He has also served as Professor and Head Department of Computer Science and Engineering, Vignana Bharathi Institute of Technology, Hyderabad. He has completed his Post-Doctoral Fellow (PDF) from Kongju National University South Korea, Ph.D. from Siksha ‘O’ Anusandhan University, M.Tech. (CSE) from Biju Patnaik University of Technology (BPUT), and MCA from Fakir Mohan University Balasore, India. Besides academics, he is also involved various administrative activities, Member of Board of Studies to C. V. Raman Global University Bhubaneswar, Member of Doctoral Research Evaluation Committee, Admission Committee, etc. His area of research includes Data Mining, Image Processing, Soft Computing, and Machine Learning. Now he is Editorial Member of Korean Convergence Society for SMB .He has published 13 edited books, 1 text book, 2 international patent, and more than 100 research papers in national and international journals and conference proceedings in his credit.

Sung-Bae Cho received the Ph.D. degree in computer science from KAIST (Korea Advanced Institute of Science and Technology), Taejeon, Korea, in 1993. He was Invited Researcher of Human Information Processing Research Laboratories at Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, from 1993 to 1995, and Visiting Scholar at University of New South Wales, Canberra, Australia in 1998. He was also Visiting Professor at University of British Columbia, Vancouver, Canada from 2005 to 2006, and at King Mongkut’s University of Technology Thonburi, Bangkok, Thailand in 2013. Since 1995, he has been Professor in the Department of Computer Science, Yonsei University, Seoul, Korea. His research interests include hybrid intelligent systems, soft computing, evolutionary computation, neural networks, pattern recognition, intelligent man–machine interfaces, and games. He has published over 300 journal papers and over 750 conference papers.

Chapter 1

Cloud-Based Smart Grids: Opportunities and Challenges



**Atta-ur-Rahman, Nehad M. Ibrahim, Dhiaa Musleh,
Mohammed Aftab A. Khan, Sghaier Chabani, and Sujata Dash**

Abstract Cloud computing offers users a new way to access the computing resources, such as data storage, software, and computing power on demand whenever and wherever they need. It is among the rapidly growing fields in the information and communication technologies (ICT) paradigm that has greatly impacted our daily lives. Recently, smart grids which are a smarter and more enhanced version of traditional electricity system counterpart have been benefited from the integration of cloud computing. This paper introduces cloud computing and smart grids, presents how the electricity grid evolution led to the creation of the smart grid, and showcases the way both fields have been augmented. It further argues how smart grid technologies play a noteworthy role in Saudi Arabia's Vision 2030.

1.1 Introduction

There are several important factors that impact the future of electricity systems around the world. It includes government policies, managing the growing consumer needs, efficiency and quality of service, and the emergence of new intelligent computers and hardware technologies. Governmental policies were created and enforced globally

Atta-ur-Rahman · N. M. Ibrahim · D. Musleh

Department of Computer Science, College of Computer Science and Information Technology,
Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

M. A. A. Khan

Department of Computer Engineering, College of Computer Science and Information Technology,
Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

S. Chabani

Department of Network and Communications, College of Computer Science and Information
Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi
Arabia

S. Dash (✉)

Department of Computer Application, Maharaja Sriram Chandra Bhanja Deo University,
Baripada, Odisha 757003, India

due to environmental concerns, which are driving the entire energy system to efficiency, renewable sources of electricity, and conservation. To solve all these concerns, advanced technologies were used to come up with the idea of smart grid that is more efficient and sustainable. One of the advanced technologies that contributed to the enhancement of the smart grid performance is cloud computing [1].

Cloud computing has been playing a tremendous role in the evolution of many fields of life like health care and so on [2–6]. Moreover, its augmentation to other technologies has been greatly revolutionizing the technological impact over the society. Other than smart grids, clouding computing has been integrated with Internet of things (IoT) and a new term is coined as cloud of things (CoT) [7]. This shows the instrumental nature of the cloud computing technology that enhances the capabilities of the existing technologies upon integration. The main purpose of this integration of cloud computing to other fields is to maximize their potential by adding up the huge computing resources.

Rest of the paper is organized as follows: Sects. 1.2 and 1.3 introduce cloud computing and smart grids, respectively. Section 1.4 covers the integration of cloud computing in smart grids, and Sect. 1.5 highlights the security aspects of this integration. Section 1.6 connects the smart grid with the Saudi Arabia's Vision 2030, while Sect. 1.7 concludes the paper.

1.2 Cloud Computing Overview

Cloud computing is a system with handy, on-demand facilities to access network and computing resources (such as servers and storage options), typically on the Internet via a pay-as-you-need basis. It is a vastly growing field in the IT industry, and nowadays, many providers of cloud computing exist such as Microsoft and Google. Cloud computing offers many services and options from basic large storage facilities, networking operations, processing power through natural languages, artificial intelligence, and standard office applications. There are many more services to mention, but the main concept is that any service that does not require you to be physically near the computer hardware can be provided and delivered through the use of cloud computing technologies [8, 9]. Cloud computing takes advantage of parallel and distributed processing technologies to provide large-scale integrated processing capabilities. These match that of a supercomputer, but with a fraction of the cost, which is incredibly cost-effective [10]. Many companies prefer using cloud computing rather than having their own data center or computing infrastructure to avoid the major costs and complexity of having and regularly maintaining their own IT infrastructure. The companies just need to rent storage to a wide range of smart applications from a cloud service provider to use for a certain tenure [8, 9, 11–13]. The essential characteristics of cloud computing are as follows (Fig. 1.1):

- On-demand and all-time self-service
- Broad network access capabilities

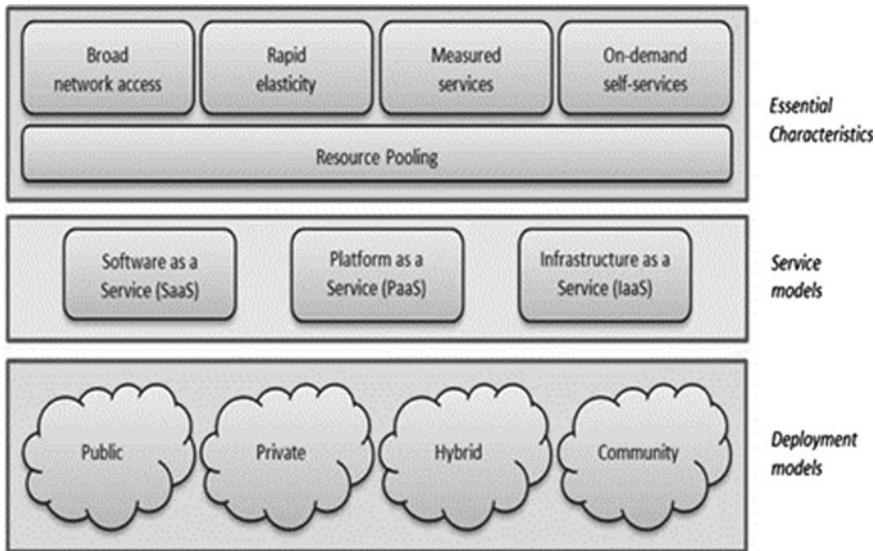


Fig. 1.1 Characteristics of cloud computing

- Resource pooling of computer resources and assets
- Swift elasticity
- Affordability
- Cost optimization services and mechanisms

The most important part in a cloud computing system is its data center that affects the routing and congestion algorithm that a system may experience. It also plays an important role with connecting to the Internet. The inherent task manager component manages the demand efficiently [14–17]. There are three main architectures of cloud computing technologies which are software as a service, platform as a service, and infrastructure as a service, which are shown in Figs. 1.2 and 1.3, respectively. Moreover, the hybridization between these architectures is also in practice [8].

Infrastructure as a service is known to be a single surface cloud layer where clients share the dedicated resources on a pay-per-use basis. Software as a service is known to be any software that executes a simple task without any necessary interaction and communication with other systems. They can be used as a component in other applications, where they are known to be as mash-ups or plug-ins. Platform as a service is a model that allows users to use infrastructure or acquired applications of a provider without necessarily downloading and installing the software. Cloud computing technologies can be classified according to their deployment model. These are known as [9]:

- Private cloud: A virtual environment that is deployed at an organization where the number of users is restricted to employees of the company only. This model suits

Fig. 1.2 Cloud computing architectures

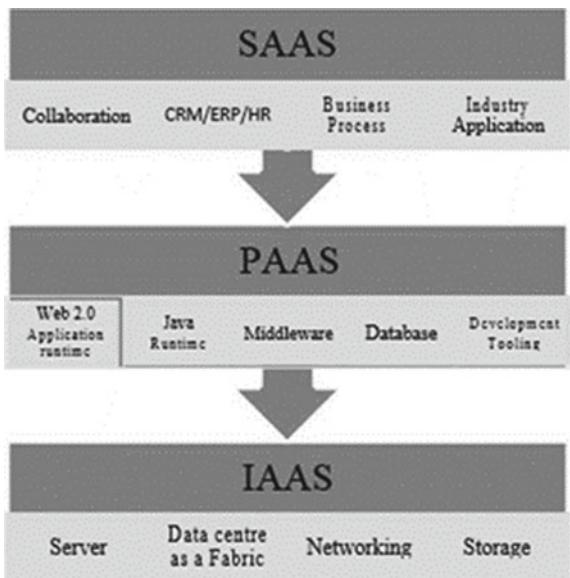
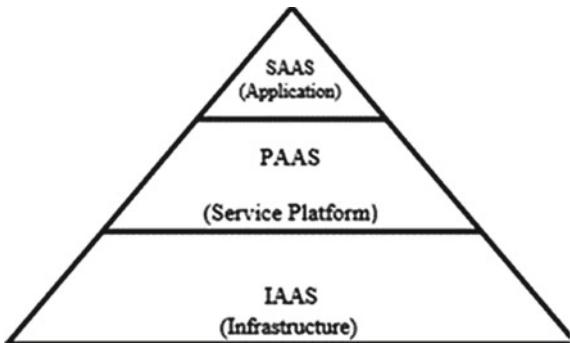


Fig. 1.3 Cloud computing infrastructure



any situation with confidential information that needs to be secured in a proper way.

- Community cloud: It is like the private cloud; however, it can be said that it is a collection of many combined private clouds that share a similar purpose.
- Public cloud: A virtual environment that is available to the public for anyone to use. Consumers can purchase computing resources, services, and storage via a web portal.
- Hybrid cloud: A mixture of both public and private cloud models where specific resources can only be used in the public cloud, and others in the private cloud.

Finally, another concept that is like cloud computing is grid computing. Grid computing optionally uses computing resources, while cloud computing provides on-demand resources, which is a step further to grid computing [8].

1.3 Smart Grid

This section introduces the smart grid technology and highlights recent advancements/evolutions in the field.

1.3.1 *Overview of Smart Grids*

Electrical networks of transmission lines and transformers that deliver electricity from power plants to homes and institutions are known as the grid. With the aid of recent technologies, these grids have become smart in the sense that they allow a two-way communication between the customers and the utility providers, advanced metering infrastructure (AMI) system, and advanced storage techniques. Smart grids hence can be said to be an advanced electric power system that have controls, computers and automation technologies, and equipment [8, 18, 19]. The advanced metering infrastructure (AMI) is basically a combined system of many smart meters, communications networks, and data management systems that offer a two-way communication between the electrical utilities and customers [9]. The smart meters benefit customers and companies because they record real-time readings of many things such as electrical energy consumption, voltage quality, and other useful information like interval and how often the readings are taken can be controlled. Another benefit that these smart meters provide is getting the statuses and performance readings of the electrical devices in a smart grid network, such as transformers, capacitors, circuit breakers, and electrical lines. Fault detection of devices and the replacement of assets can now be performed effectively and efficiently. The analysis data retrieved from these meters can be sent to companies for future use [9]. The general architecture of a smart grid can be found in Fig. 1.4. Electricity and information flow between a customer and the provider in a bidirectional manner. This ensures the delivery of a reliable, cost-effective, and secure service in the whole system [19].

Since smart grids require a great deal of effort to implement, the process of adding features is gradual. It takes many decades to incorporate all the features and achieve an ideal standard of smart grids. They must have the capacity and capability to adapt to any new technologies that might evolve or be invented later. There are some requirements that a smart grid system must follow to ensure this which are [1]:

- Allowing the addition of renewable energy resources



Fig. 1.4 Smart grid architecture

- Allowing customers to be active members and participate in some respects to aid in energy conservation
- Having a secure mechanism of communication between the consumer and the utility facility
- Reducing energy waste as much as possible
- Supporting of data management and processing
- Supporting of self-healing
- Having a real-time operation and control
- Enhancing the utilization of existing assets and resources
- Improving the system safety and operational flexibility and lowering the use of hydrocarbon fuels.

These functionalities allow smart grids to create a new infrastructure that is safe, secure, reliable, resilient, efficient, cost-effective, clean, and sustainable. Although smart grids are very powerful, they do not replace the existing traditional electric systems, but rather are built on top of existing infrastructure. This enables a better utilization of resources and assets [1].

1.3.2 Evolution in the Smart Grids

Smart grids did not come to existence for no reason. They emerged as a solution to modernize the electricity grid, making it more environmentally friendly, and causing the power delivery to be more efficient [20]. To understand where the concept of smart grids came from, first, we need to understand the history of the traditional electrical grid. The first-time electrical grids were introduced in the 1800s. However, they were widely utilized in developed countries only in the 1960s. Around that time, electrical grids had considerable capacities and penetration power. In addition, it also had sufficient reliability and quality, with power sources for it such as nuclear plants, fossil fuel, and hydroelectric technologies. Due to the high technological advances that

occurred in the twenty-first century in the power industry, many technologies were integrated into the power grid, such as information and communication technologies, and smart sensors, which helped make the smart grid vision a reality. Additionally, the power generation advances helped to provide non-environmentally harmful power sources such as wind, tidal, geothermal, and solar energy, not like the previously harmful sources. Furthermore, the readiness and availability of various kinds of power sources helped make the power generation decentralized, which supported the power supply and reduced the costs of power distribution. Lastly, the emergence of new technologies such as communication technologies allowed for the information on the consumption and production of the energy to be available, which contributed to increasing the reliability and efficiency of the grid overall [21].

1.4 Cloud Computing in Smart Grids

Any smart grid system requires generation, transmission, distribution, and usage of power simultaneously along with the storage of large amounts of energy. They need real-time and reliable control mechanisms which can be provided by cloud computing technologies. Cloud computing combined with smart grids can maximize utilization of storage resources, and to improve the overall robustness and load balancing of the system by using control algorithms and the huge amounts of information can be handled within a short span of time. Cloud computing is also needed with smart grids because it can help the system recover in case of a blackout condition, help to monitor and schedule the power systems, and enable reliable evaluation of the power system. This added feature makes the grids robust against such type of vulnerabilities. Smart grids communicate with data centers that utilize cloud computing technologies in a manner that is like Fig. 1.5. The smart grid system sends and receives its data from a data center that is provided by cloud computing servers [8].

When cloud computing is integrated with smart grids, it allows smart scheduling, controlling, marketing, power distribution, protection, and running of all operations of the system [8] (Fig. 1.6).

Fig. 1.5 Smart grid communication with data centers

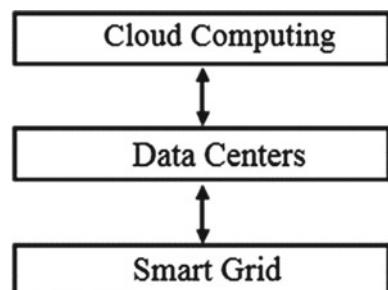
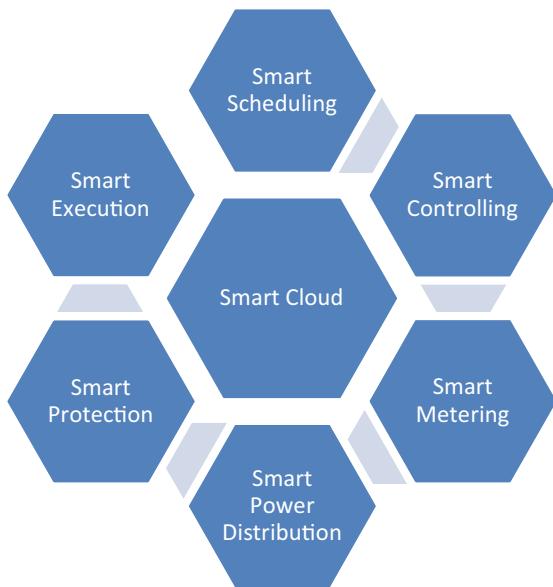


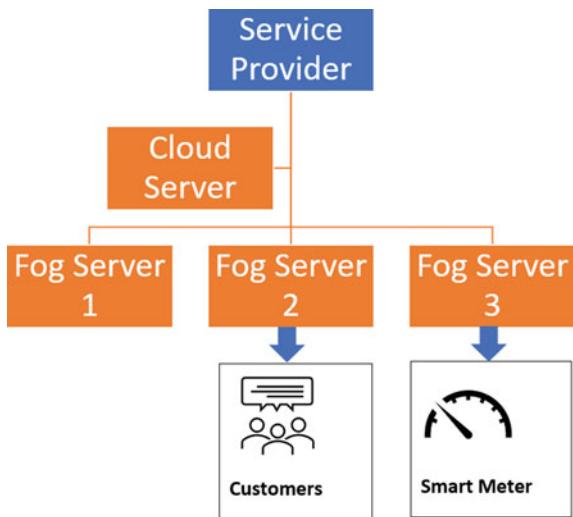
Fig. 1.6 Smart cloud

As mentioned earlier, cloud computing offers a new and improved way to achieve online power system operation analysis and control [8]. Since the integration of cloud computing with smart grids, energy monitoring has become much easier. Energy management systems are cloud-based systems where all the information, programs, and controls are stored on the cloud (over the Internet) and can be accessed from any computer [22]. They can be thought of as an extension or extra service that is provided by smart grids that integrate cloud computing technologies. Introducing cloud computing to smart grids may not be as seamless as it should be, and it may introduce many challenges such as [9, 10]:

- Demand response: which are the changes in electrical usage and consumption by customers from their normal patterns that are usually caused by the changes in electricity service's prices.
- Management: It might be hard to obtain a synchronized and efficient operation of all data centers working together and to manage precisely.

An approach used to overcome these challenges is to use a branch of cloud computing known as fog or edge computing. Fog computing basically connects cloud computing with IoT technologies. It extends the existing cloud computing network at the edges. It allows the information coming from the data centers to be preprocessed, which reduces the load of the smart grid network. Adding to that, fog computing contains a huge number of different devices that interact with each other to perform storage or processing tasks without the interference of another party, hence minimizing the need of many cloud operators. Another reason is that fog computing

Fig. 1.7 Smart grid model with both fog and cloud computing



is used for it to overcome the increasing latency and response time of smart devices that are caused by cloud computing [19, 23].

Looking at it from the service-level point of view, both fog and cloud computing offer very similar services. However, in smart grids, it cannot be said that fog computing can replace cloud computing entirely. A smart grid can be realized by using cloud computing, and all the challenges that come with it can be dealt with fog computing. Figure 1.7 shows how both types of computing work together [19].

Some additional features that fog computing provides to smart grids are [19]:

- Locality: Fog services run outside the current network; hence, it is isolated and not affected by other components that are trying to access its resources. Instead of dealing with big data and its traffic, fog computing splits the data into subgroups of data that are easier to manage.
- Proximity: Fog computing gets closer to end users, allowing them direct access to the source. This simplifies data extraction.
- Privacy: Fog computing splits public and private data found in the cloud, making it much easier to provide end users with data privacy and security.
- Latency: It is reduced since the proximity is reduced. Both the reaction time and congestion of the network are lessened because of this.
- Geo-distribution: All kinds of devices found in a traditional electrical system such as power generators, energy transformers, and so on are now located in the edges of the networks, hence providing services with reduced latency and delay.

Location awareness: Because fog servers are geographically distributed, location of each device that is connected to the network can be determined. This can be useful for individuals and companies.

1.5 Security in Smart Grids

As mentioned previously, security services can be achieved in smart grids by the use of cloud computing. However, cloud computing can also suffer from security issues since many network services are inherently vulnerable to cyber-attacks, which is a problem that threatens the system's safety. Smart grids incorporate cloud computing to reduce security risks to an acceptable level in the system. Furthermore, smart grid has vulnerabilities that threaten the system's safety as well. One example of a vulnerability in smart grids includes intelligent devices that are playing a role in managing both electricity supply and network demand. These intelligent devices may act as an attack of entry points into the network, which might allow the attackers to access the network, violate the system security by breaking the confidentiality and integrity of the transmitted data, and make the service unavailable. Various defense mechanisms were established to protect smart grids from these vulnerabilities. Moreover, cloud security grants spam mail filtering, anti-virus, threat detection and security, and private storage. It improves power system performance, avoids unauthorized access to user data and code, and enhances the network scalability. The firewalls of the cloud utilize the cloud to enhance the security of the smart grid. The application of cloud technologies in the smart grid improves the disaster recovery ability of power systems. Especially, the resilience of the power systems against natural disasters can be enhanced. A disaster recovery plan based on cloud storage provides high reliability, availability, and low downtime. Additionally, the use of a distributed data center scan can provide disaster recovery in a short time [8, 10].

1.6 Smart Grid and Saudi Arabia's 2030 Vision

Saudi Arabia is known for being the largest country in the world in exporting oil that contributes 90% of the government revenue. On April 25, 2016, Saudi Arabia's government announced its 2030 Vision that aims to reinforce economic diversification and reduce the dependence on oil as well as the development of different sectors including education, health, tourism, and infrastructure construction. One of the highest priorities that the vision tries to achieve is to invest in renewable energy resources. A research study focused on designing a smart system for power generation scheduling and optimization tailored for the western power grid of Saudi Arabia by integrating renewable energy sources. The western part of Saudi Arabia is known for its severe weather conditions such as high temperatures and humidity during the summer which results in an increased demand to use air conditioning by people living in this area and eventually creates a high load on the power grid [24]. The study suggests using solar and wind energy as an alternative solution to fuel as they are waste-free and available in a good amount in the region. In this study, four offshore farms and two solar panel platforms were injected into a power transmission grid. The study uses historical weather conditions and electricity load data throughout the

year to develop a model that can predict the grid load conditions compared to the time of the year. The purpose is to turn off some fuel generation and depend on renewable energy using the predicted loading data which helps in planning the operation of the power stations based on the demand. Load prediction models help electric companies to estimate the accurate amount of power to deliver to their customers [25]. Modernizing electricity infrastructure is one of the plans of Saudi 2030 Vision, and one of the most important aspects of modernizing any electricity grid is implementing smart grid technologies. The study discusses the feasibility of implementing smart grid technologies in Saudi Arabia's electricity infrastructure by proposing a cost–benefit analysis to determine the most suitable and profitable smart grid technique for Saudi electricity infrastructure. The insertion of smart grid technologies (SGTs) will help in transforming the current electricity grids to a much smarter and enhanced form. Smart grids offer an economic, sustainable, efficient, and reliable electric system which can overcome the problems of increasing demands and network congestion. However, there are some problems facing SGTs, such as having high initial capital, low interest of some concerned parties, and willingness of consumers to participate. Therefore, the need for a careful analysis to assess the benefits of SGTs arises. The cost–benefit analysis of the study resulted in an estimation of 15.87\$ billions of total benefits compared to 3.36\$ total costs, hence 12.51\$ total profits. The results show that implementing SGTs in Saudi electricity infrastructure is profitable [26]. Consumer awareness also plays a significant role in the process of adopting smart grid technologies. The public attitude toward the adoption of smarter grid in the eastern province of Saudi Arabia was examined. The study shows that the residents of the eastern province are willing to move toward smart grids by using green energy and smart meters. However, they have raised some concerns regarding the issue. Only the consumer awareness is not enough, the government should also provide support by promoting energy efficient and saving behavior [27]. Smart grids are constituent part of smart cities for optimum resource utilization [28]. Neom project is an example of first smart city in the Saudi Arabia that is a fascinating example of future of the smart grids in the kingdom, especially when the kingdom is switching/from oil-based economy to a knowledge-based economy.

1.7 Conclusion

This paper discusses the integration of cloud computing with the smart grids. It approached this topic by first explaining the concept of cloud computing and smart grids proving a general overview and the essential characteristics. It further highlights the significance and emergence of cloud computing as well as fog computing in smart grids, potential benefits, and vulnerabilities or challenges like security issues. Finally, the alignment and significance of cloud-based smart grids with the Saudi Arabia's Vision 2030 have been presented.

References

- Anderson, R.N., Ghafurian, R., Gharavi, H.: Smart grid: the future of the electric energy system (2018)
- Dash, S., Biswas, S., Banerjee, D., Rahman, A.: Edge and fog computing in healthcare—a review. *Scalable Comput.* **20**(2), 191–206 (2019)
- Ahmad, M., Qadir, M.A., Rahman, A., Zagrouba, R., Alhaidari, F., Ali, T., Zahid, F.: Enhanced query processing over semantic cache for cloud based relational databases. *J. Ambient Intell. Human. Comput.* (2020). <https://doi.org/10.1007/s12652-020-01943-x>
- Khan, M.A., Abbas, S., Atta, A., Ditta, A., Alquhayz, H., Khan, M.F., Rahman, A., et al.: Intelligent cloud based heart disease prediction system empowered with supervised machine learning. *Comput. Mater. Continua* **65**(1), 139–151 (2020)
- Rahman, A., Dash, S., Ahmad, M., Iqbal, T.: Mobile cloud computing: a green perspective. In: Udgata, S.K., Sethi, S., Srirama, S.N. (eds.) *Intelligent Systems. Lecture Notes in Networks and Systems*, vol. 185. Springer (2021). https://doi.org/10.1007/978-981-33-6081-5_46
- Rahman, A.: Efficient decision based spectrum mobility scheme for cognitive radio based V2V communication system. *J. Commun.* **13**(9), 498–504 (2018)
- Alhaidari, F., Rahman, A., Zagrouba, R.: Cloud of things: architecture, applications and challenges. *J. Ambient Intell. Human. Comput.* (2020). <https://doi.org/10.1007/s12652-020-02448-3>
- Mishra, N., Kumar, V., Bhardwaj, G.: Role of cloud computing in smart grid. In: 2019 International Conference on Automation, Computational and Technology Management (ICACTM) (2019)
- Naveen, P., Ing, W.K., Danquah, M.K., Sidhu, A.S., Abu-Siada, A.: Cloud computing for energy management in smart grid—an application survey. *IOP Conf. Ser.: Mater. Sci. Eng.* (2016)
- Fang, B., Yin, X., Tan, Y., Li, C., Gao, Y., Cao, Y., Li, J.: The contributions of cloud technologies to smart grid. *Renew. Sustain. Energy Rev.* 1326–1330 (2016)
- Rupabanta Singh, K., Dash, S., Deka, B., Biswas, S.: Mobile technology solutions for COVID-19. In: Al-Turjman, F., et al., (eds.) *Emerging Technologies for Battling COVID-19, Studies in Systems, Decision and Control*, vol. 324. pp. 271–294 (2021)
- Rahman, A., Sultan, K., Naseer, I., Majeed, R., Musleh, D., et al.: Supervised machine learning-based prediction of COVID-19. *Comput. Mater. Continua* **69**(1), 21–34 (2021)
- Dilawari, A., Khan, M.U.G., Al-Otaibi, Y.D., Rehman, Z., Rahman, A., Nam, Y.: Natural language description of videos for smart surveillance. *Appl. Sci.* **11**(9), 3730 (2021)
- Atta-ur-Rahman, Sultan, K., Dash, S., Khan, M.A.: Management of resource usage in mobile cloud computing. *Int. J. Pure Appl. Math.* **119**(16), 255–261 (2018)
- Rahman, A.: GRBF-NN based ambient aware real-time adaptive communication in DVB-S2. *J. Ambient Intell. Human. Comput.* **2020**(12), 1–11 (2020)
- Rahman, A., Alhaidari, F.A., Musleh, D., Mahmud, M., Khan, M.A.: Synchronization of virtual databases: a case of smartphone contacts. *J. Comput. Theor. Nanosci.* **16**(3) (2019)
- Rahman, A., Dash, S., Luhanch, A.K.: Dynamic MODCOD and power allocation in DVB-S2: a hybrid intelligent approach. *Telecommun. Syst.* **76**, 49–61 (2021). <https://doi.org/10.1007/s11235-020-00700-x>
- Kim, S.-C., Ray, P., Reddy, S.S.: Features of smart grid technologies: an overview. *ECTI Trans. Electr. Eng. Electron. Commun.* (2019)
- Okay, F.Y., Ozdemir, S.: A fog computing based smart grid model. In: Int'l Symposium on Networks, Computers and Communications, pp. 1–6 (2016)
- Tuballa, M.L., Abundo, M.L.: A review of the development of smart grid technologies. *Renew. Sustain Energy Rev.* 711–716 (2016)
- Avancini, D.B., Rodrigues, J.J., Martins, S.G., Rabelo, R.A., Al-Muhtadi, J., Solic, P.: Energy meters evolution in smart grids: a review. *J. Cleaner Prod.* (2019)
- Hans, M., Phad, P., Jogi, V., Udayakumar, P.: Energy management of smart grid using cloud computing. In: 2018 International Conference on Information, Communication, Engineering and Technology (2018)

23. Zahoor, S., Javaid, S., Javaid, N., Ashraf, M., et al.: Cloud-fog-based smart grid model for efficient resource management. *Sustainability* (2018)
24. Nurunnabi, M.: Transformation from an oil-based economy to a knowledge-based economy in Saudi Arabia: the direction of Saudi vision 2030. *J. Knowl. Econ.* **8**, 536–564 (2017)
25. Aljahdali, F., Abibod, M.: Design of smart generation by integrating renewable energy into western power grid of Saudi Arabia. **6** (2018)
26. Alqaqel, T., Suryanarayanan, S.: A comprehensive cost-benefit analysis of the penetration of smart grid technologies in the Saudi Arabian electricity infrastructure. **11** (2019)
27. Düşteğör, D., Sultana, N., Felemban, N., Al Qahtani, D.: A smarter electricity grid for the eastern province of Saudi Arabia: perceptions and policy implications. **14** (2017)
28. Rahman, A., Qureshi, I.M., Malik, A.N.: Adaptive resource allocation in OFDM systems using GA and fuzzy rule base system. *World Appl. Sci. J.* **18**(6), 836–844 (2012)

Chapter 2

A Resource-Aware Load Balancing Strategy for Real-Time, Cross-vertical IoT Applications



Ranjit Kumar Behera, Amrut Patro, and Diptendu Sinha Roy

Abstract In this new age, with the maturation of Internet of things (IoT), a multitude of advanced and innovative services are foreseen. Cloud, an accustomed computing technology, being circumscribed by the huge traffic cannot assuage the real-time services demanded by the cross-vertical IoT applications. Over and above this, though the fog computing paradigm, a hopeful alternative providing real-time services, was introduced, still, fog and cloud collaboratively may not be able to bear the tremendous amount of requests that arises from the numerous vertical IoT applications, because of the resource-bounded nature of fog. An indecent resource management and load balancing strategy in the fog infrastructure may further lead to a deterioration in the quality of service (QoS) and failure in providing services in real time. Without disregarding the unignorable issues and challenges in fog, this paper A Resource-Aware Load Balancing Strategy for Real-Time, Cross-vertical IoT Applications has been envisioned. The designed architecture and the proposed model are presented comprehensively with an emphasis on elucidating the resource-aware load balancing strategy. The equations and algorithms for resource management mechanism and load balancing are presented meticulously. Following the end, the efficacy of the proposed methodology is validated using CloudSim and the performance is evaluated in terms of load balance, resource utilization, and power consumption based on employed fog nodes.

2.1 Introduction

For a while now, the Internet of things (IoT) has emerged as an agglomeration of various communication and information technologies, creating a topical noise in both academia and as industries [1]. IoT being a gigantic network with a profusion of interconnected devices and objects has been empowering diverse intelligent and

R. K. Behera (✉) · A. Patro

National Institute of Science and Technology, Berhampur, India

D. S. Roy

National Institute of Technology, Shillong, Meghalaya, India

innovative applications like smart healthcare systems, smart city and smart homes, as well as vehicle to vehicle communications-based modern transportation systems, and so forth. Nevertheless, yet to be achieved goal in the IoT scenario is the phenomenal mutation of such smart IoT applications to cross-vertical services and applications [2]. Cross-vertical services which can also be referred as cross-domain applications are the consolidation of individual smart applications or verticals. By way of illustration, collaboration of smart parking, traffic management system, and smart vehicles can give rise to an enhanced and an advanced automated transportation model. Such collaborations of different verticals would not only save time but would also prove to be considerably worthwhile for the mankind. Hence, it can be said that the development of aforesaid cross-vertical or cross-domain applications is an indispensable to actualize such intelligent IoT applications [3].

Cloud being the most pervasive computing technology since its origination has been predominantly used by the IoT applications of various verticals for the purpose of storage and computation [4]. Notwithstanding the long-term and the impregnable support which was provided by cloud, gradually several unignorable issues and drawbacks started to be noticed in cloud [5]. At this juncture, noticing the challenges in cloud, fog computing emerged as an evolutionary technology in the field of Internet and computation [6].

Howbeit, the terms like fog and cross-vertical applications seem to be unsophisticated but it is the antithesis. The requests that arrive from various collaborated verticals also has some criticality based on delay sensitivity. Requests from cross-verticals related to VR or gaming systems, irrigation systems, smart home droids, etc., are tolerable to delays and hence can be considered as soft tasks. On the other hand, requests emerging from cross-verticals related to healthcare or disaster management systems are highly sensitive and can merely tolerate delays. Even a minuscule retardation in computing such real-time or critical tasks would exacerbate the situation and may lead to great disasters. The management of requested resources in which the workloads or the tasks are distributed among the fog nodes with an aim of avoiding bottlenecks is known as load balancing.

In the recent times, a manifold of research efforts have been made exploring the concepts like load balancing and cross-vertical applications in fog environment [7, 8]. Load balancing acts as a key factor in identifying the competency of resource management and allocation strategies. The authors in [9] have investigated on the trade-off between transmission delay and power consumption in the fog-cloud system and have tried resolving the workload allocation problems formulating toward minimum power consumption and service procrastination, while the authors in [10] have proposed a load balancing and optimization strategy using dynamic resource allocation method based on genetic algorithm and reinforcement learning. Their proposed model seamlessly monitors the network traffic, collects the information about the load in each server, handles and distributes the requests among the servers equally using the dynamic resource allocation methodology. In [11], researchers have proposed a centralized load balancing policy in which the workload is being uniformly distributed among the fog nodes reducing the response time. At every location, the unreliable fog resources form a cluster under a corresponding

controller node that maintains metadata about idle and busy nodes, forwarding the requests to a suitable fog node based on the resource availability. Fuzzy logic's accuracy and pace in determining the priority are impregnable, while probabilistic neural network proves to highly accurate and predicts the target much faster than the multilayer perceptron networks. Utilizing such invincible algorithms, an effective load balancing strategy using probabilistic neural networks is introduced in [12], which proves to be a competent load balancing model in fog, especially for health-care applications. The authors in [13] have presented a dynamic resource allocation methodology (DRAM), trying to achieve maximum utilization of resources in the fog nodes. Their proposed model consists of two primary methods, static resource allocation, and the dynamic service migration. Firstly, using the static resource allocation method, the tasks are allocated on fog nodes based on task resource requirements in descending order, and using dynamic service migration, the tasks are again migrated to a suitable fog node trying to achieve a better load balance variance. However, the proposed strategy proves to be successful in achieving high resource utilization in fog nodes, yet some unignorable drawbacks have been noticed in their proposal such as allocating tasks statically and again migrating them to fog nodes would result in an extra cost (migration cost) which is not considered in the paper. Migration of partially executed tasks to other fog nodes would not prove to be a good idea as it causes unnecessary pause and delay in execution.

An improper resource management mechanism would undoubtedly lead to starvation of resources among the requests and a failure in providing the demanded resources to the tasks in the fog layer. Eventually, the starving tasks are forwarded to cloud overwhelmed by the high traffic, which further boosts the chances of tasks being failed to be provided with needed resources in real time. Hence, without overlooking the severe issues and challenges seen in fog, this paper A Resource-Aware, Load Balancing Strategy for Real-Time, Cross-vertical IoT Applications has been proposed with a goal of attaining maximum load balancing and resource utilization with minimized power consumption, being mindful of providing momentary real-time services to the needy IoT verticals.

The remaining sections of the paper are organized as follows. Section 2.2 comprehensively presents the overall clustered network architecture, while Sect. 2.3 elaborates the proposed resource-aware load balancing strategy along with a discussion on the mathematical concepts and the algorithms used in implementing the proposal. Section 2.4 focuses on the experimental setup followed by the performance evaluation presented in Sect. 2.5. At the end, a conclusion to the proposed model and its efficacy is drawn in Sect. 2.6, also directing toward the future work.

2.2 Overall Clustered Network Architecture for Cross-vertical IoT Applications

In this section, the overall clustered network architecture considered in our model has been presented in a comprehensive manner. As it can be seen from Fig. 2.1, the network architecture consists of three different layers alike the traditional fog–cloud architecture. The first layer which can also be deemed as the data acquisition layer is comprised of the IoT users, and the smart systems and devices which form the crux of the IoT world. Not only the smart devices are able to connect with the users but also are capable of communicating with other similar IoT objects (D2D) for interoperability and with the fog layer (D2F) for the purpose of storage or computation.

Above the data acquisition layer, there lies the fog infrastructure which consists of the geographically distributed network of fog nodes capable of providing resources to the needy cross-vertical applications. It is presumed that the requisites like various files, previously stored contexts or data, metadata about the applications, etc., for processing are present in the fog layer itself. It has also been considered that there are 64 fog nodes which are clustered using k-means ++ clustering algorithm into fractals of 8 fog nodes, with each cluster handling, scheduling, and processing the requests arising from its respective geographical extent. Moreover, it is made sure that the controller nodes are dexterous of communicating with the nodes of neighboring clusters (FC-2-FC) in case of overhead in their own cluster. Each controller node consists of several components like dynamic node distributor, task classifier, soft task manager, and real-time task manager whose functionalities are comprehensively discussed along with a schematic diagram and algorithms in Sect. 2.3. The

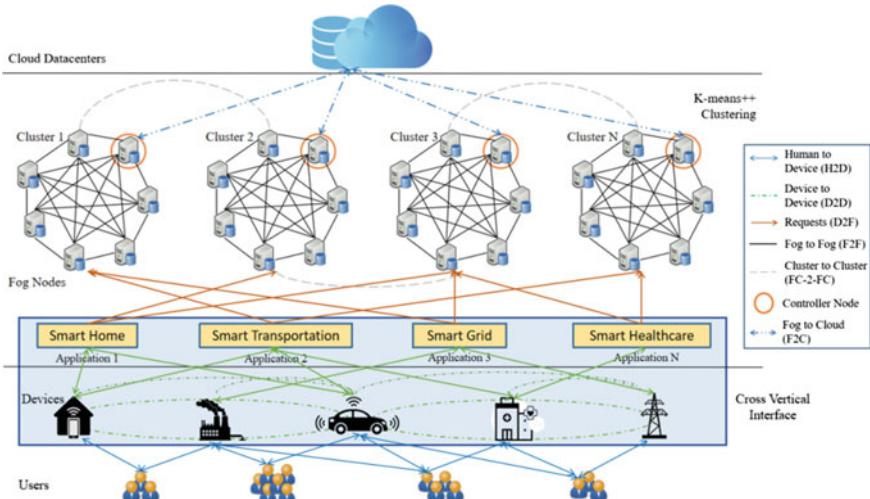


Fig. 2.1 Cluster-based network architecture

cloud data centers form the crowning layer of the entire network architecture delivering ubiquitous support for storage and computation to the applications in case of unbearable traffic and resource scarcity in the entire fog layer.

2.3 Proposed Resource-Aware Load Balancing Strategy for CVIA

In this section, a discussion on the overall working of the proposed model is conducted meticulously along with presenting the mathematical equations and algorithms for better perspicuity. Figure 2.2 depicts the overall flow of activities in scheduling the requests and balancing the workload among the fog nodes.

Collectively, the requests that arrive from the cross-vertical interface, directly or indirectly connected to the users alias clients, are received by the controller node of the respective geographic cluster. The controller node consists of a task classifier (TC) which classifies the task into soft task or real-time task, which is further forwarded to the respective manager. The controller node with the help of dynamic node distribution (DND) technique reserves certain number of nodes in the cluster for real-time tasks and for soft tasks based on the ratio between the arrival rates of both types. The task forwarded to RT task manager schedules it to the nearest real-time node for execution, and in case the required resource units are not available in the scheduled node, then the request is shifted to the neighboring real-time nodes. In case

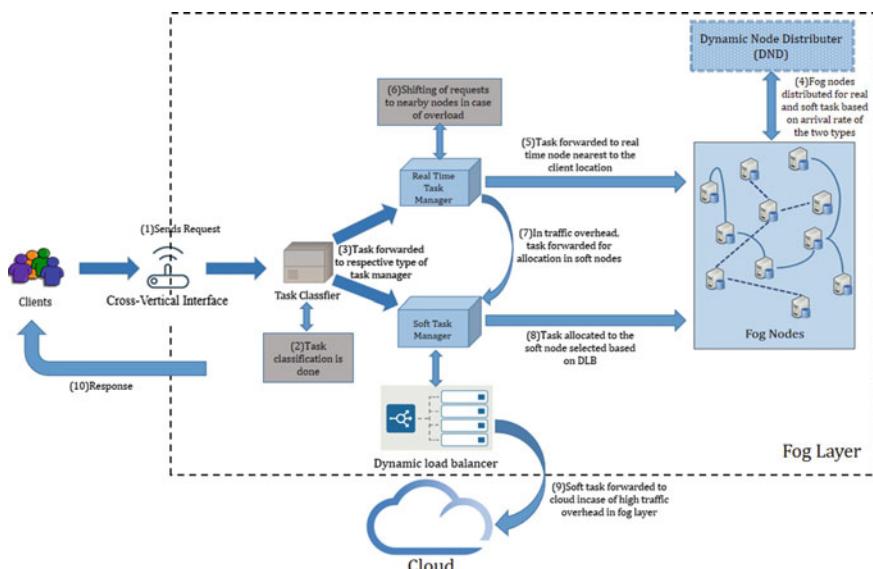


Fig. 2.2 Workflow of the proposed resource-aware LB strategy

of resource scarcity, the task is forwarded to the soft task manager for scheduling in soft task nodes without any delays. The soft task manager at first calculates the resource utilization in the active fog nodes excluding the empty fog nodes which are kept shut down till the resources are available in active nodes. The node with minimum load balance variance is culled for allocation of resources. In situations when there are no active nodes with enough resource units available for allocation, then at that instant of time an empty node is turned on and resource allocation is done on that node. Moreover, the forwarded requests from the real-time nodes are scheduled to the active soft task node with available space without any delay. If there is no soft task node with available space, then the nearest soft task node is selected and the real-time task is executed without any retardation preempting the executing soft task nodes till the required units are not satisfied.

2.3.1 K-Means ++ Based Fog Nodes Clustering

The presupposed fog infrastructure would consist of 64 fog nodes with a series of 8 fog nodes. The clustering of the 64 nodes is done using the k-means ++ clustering into fractals of 8. Each fractal contains 8 fog nodes. Each cluster contains a seed node or the controller node which acts as a manager for handling the requests from several cross-verticals within the specified geographical extent. The clustering scale can be configured effortlessly, and such a clustered architecture not only helps organizing the fog clusters but also makes fog nodes communicate with adjacent fog nodes with lower energy consumption and high bandwidth. The distance between the nodes is computed with the help of Euclidean distance formula. The distance between two clusters $C(A_i = 1, 2, \dots, N)$ and $C(B_i = 1, 2, \dots, N)$ can be calculated as follows:

$$\begin{aligned} \|C(B) - C(A)\| &= \sqrt{\left[\sum_{i=1}^N C(A_i) - C(B_i) \right]^2} \\ &= \sqrt{C(A_1) - C(B_1)^2 + \dots + C(A_N) - C(B_N)^2} \end{aligned} \quad (2.1)$$

The K-means ++ clustering algorithm is an enhancement of K-means clustering and can be presented as:

- (a) Select a cluster center (C_1) evenly and arbitrarily from O .
- (b) Choose another cluster center C_i , where $q \in O$ with probability $\frac{Q*Q}{\sum_{q \in O} Q*Q}$.
- (c) Repeat until there are K centers, i.e., $Cr = \{Cr_1, Cr_2, \dots, Cr_k\}$.
- (d) From the centers, start implementing the accustomed K-means clustering.

2.3.2 Variance-Based Dynamic Resource Allocation Strategy

Load balance in fog environment refers to the act of distributing workloads across a group of fog computing nodes. In this subsection, the mathematical modeling and algorithm for scheduling soft tasks to the reserved fog nodes have been presented comprehensively.

Let Cap_m be the capacity of a fog node f_m . The capacity of the fog nodes refers to the number of resource units it contains, while the resource requirement can be denoted as $\text{Req}_n = \{\text{CPU}_{\text{req}}, \text{RAM}_{\text{req}}, \text{Storage}_{\text{req}}, \text{start}_t, \text{duration}_t\}$. Let $a_n^m(t)$ be a variable to judge whether App_n ($1 \leq n \leq N$) is assigned to a fog node f_m ($1 \leq m \leq M$) at time instant t ,

$$a_n^m(t) = \begin{cases} 1, & \text{if } \text{app}_n \text{ is assigned to } f_m \\ 0, & \text{otherwise} \end{cases} \quad (2.2)$$

$$\text{ResUt}_m(t) = \frac{1}{\text{Cap}_m} \sum_{n=1}^N (a_n^m(t) \cdot \text{req}_n) \quad (2.3)$$

$$\text{EmpS}_m(t) = \sum_{n=1}^N a_n^m(t) \quad (2.4)$$

$$\text{EmpFN}(t) = \sum_{m=1}^M E_m(t) \quad (2.5)$$

where

$$E_m(t) = \begin{cases} 1, & \text{if } \text{EmpS}_m(t) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2.6)$$

Algorithm 1 Variance-based Dynamic Resource Allocation

```

Input:  $STaskList$ ,  $ResSoftFList$ ,  $EmpSoftFList$ ,  $RealReq_{forwarded}$ 
Output: Allocation of tasks and returns records the allocation records
(1) if  $RealReq_{forwarded}$  then
(2)   for  $f_m$  in  $ResSoftFList$ ,  $EmpSoftFList$  do
(3)     if  $spaceAvailableFor(f_m) \geq RealReq_{forwarded}$  then
(4)       Allocate the request and generate  $ar_i$ .
(5)       break
(6)     end if
(7)   end for
(8)   if  $f_m$  not available with enough space then
(9)     Preempt soft tasks from nearest node for  $RealReq_{forwarded}$ 
(10)    Append preempted soft tasks in  $WaitList$ 
(11)    Allocate the request and generate  $ar_i$ .
(12)  end if
(13) end if
(14) for  $Req_i$  in  $WaitList$ ,  $STaskList$  do  $Req_i = selectTaskFrom()$ 
(15) if  $length(EmpSoftFList) > 0$  then
(16)   for  $f_m$  in  $EmpSoftFList$  do
(17)      $availableSpace = spaceAvailableFor(f_m)$ 
(18)     if  $availableSpace \geq Req_i$  then
(19)        $varAftAlloc = Calculate variance for allocating Req_i on f_m$ 
(20)     end if
(21)   end for
(22)   Allocate to  $f_m$  with minimum  $varAftAlloc$  and generate  $ar_i$ 
(23) end if
(24) if  $f_m$  not available with enough space then
(25)    $forwardToCloud(Req_i)$ 
(26)   Generate  $ar_i$ .
(27) end if
(28) end for
(29) return  $AR$ 

```

And, $ResUt_m(t)$, $EmpS_m(t)$, and $EmpFN(t)$ represent the resource utilization of f_m , number of tasks allocated to f_m , and employed or the active number of fog nodes at time instant t , respectively, whereas the mean resource utilization and the load balance variance can be derived and calculated using the following equations.

$$\text{MeanRU}(t) = \frac{1}{\text{EmpFN}(t)} \sum_{m=1}^M (\text{ResUt}_m(t) \cdot E_m(t)) \quad (2.7)$$

$$\text{Varience}_m(t) = (\text{ResUt}_m(t) - \text{MeanRU}(t))^2 \quad (2.8)$$

$$\text{AvgVarience}(t) = \frac{1}{\text{EmpFN}(t)} \sum_{m=1}^M (\text{Varience}_m(t) \cdot E_m(t)) \quad (2.9)$$

Algorithm 1 presented provides a pseudo-code representation of the load balance variance-based dynamic resource allocation for soft tasks, while Algorithm 2 provides the pseudo-code representation of real-time task allocation.

Algorithm 2 Real-Time Task Allocation

```

Input:  $RT_{List}$ ,  $ResReal_{FNLst}$ ,  $EmpReal_{FLst}$ 
Output: Allocation of real time tasks and returns records
(1) for  $i = 1$  to  $i \leq length(RT_{List})$  do
(2)    $Req_i = TaskSelection(RT_{List})$ 
(3)   Select real time node nearest to client location
(4)   if  $spaceAvailableFor(nearestRNode) \geq Req_i$  then
(5)     Allocate on  $nearestRNode$  and generate  $ar_i$ 
(6)   else
(7)     for  $fm$  in  $EmpReal_{FLst}$ ,  $ResReal_{FNLst}$  do
(8)       Search for nearby  $fm$ 
(9)       if  $spaceAvailableFor(fm) \geq Req_i$  then
(10)         Allocate on  $nearestRNode$  and generate  $ar_i$ 
(11)         If  $fm$  in  $ResReal_{FNLst}$  do
(12)           Move  $fm$  to  $EmpReal_{FLst}$ 
(13)         end if
(14)         break
(15)       end if
(16)     end for
(17)   end if
(18) Call Algorithm(1) to allocate in soft task nodes
(19)end for
(20)return  $AR$ 

```

2.4 Experimental Setup

For implementing the model and evaluating the efficacy of our model, the CloudSim simulation toolkit was used, which is a framework for modeling and simulating large-scale cloud and fog computing environments. Two object data centers were created at first, namely CloudDatacenter representing the entire cloud layer and FogDatacenter representing the fog infrastructure. The allocation policy used for virtual machines was SpaceSharedAllocation policy, and various other helper methods for calculating the variance and for shutting down hosts were defined along with customization of the submitCloudlets() method in the DatacenterBroker class. Keeping the number of nodes in the fog infrastructure fixed as 64, the number of requests arriving was increased from 50 to 500 with a gradation of 50 tasks each time, so as to validate the impact on the efficiency of the model with low as well as high incoming tasks.

2.5 Performance Evaluation

The punctilious comparative study on the performance of our model based on certain parameters like average load balance variance, average resource utilization, and the number of active nodes which was conducted during the experiments has been

compared with the DRAM model proposed in [13] to measure the competency of our proposed model.

2.5.1 Evaluation on Active Nodes

The number of active nodes indirectly indicates the resource utilization and the usage of power consumption. Figure 2.3 depicts the comparison analysis of the number of active nodes between DRAM and the proposed model.

From the graph, it can be seen that initially when the number of requests arriving from the cross-vertical applications was low the proposed model seems to be much superior than the DRAM showing nearly about 20% less number of active fog nodes, indicating 20% less power consumption than DRAM. However, it can also be deciphered from the graph that, gradually with the increase in the number of requests, the efficiency of the proposed model seems to be matching that of DRAM. This decrease in the efficiency may be a cause of the reservation of nodes for real tasks, curtailing the resources for soft tasks. Nevertheless, if we consider the overall performance, it can be said that the proposed model has shown superior results than DRAM.

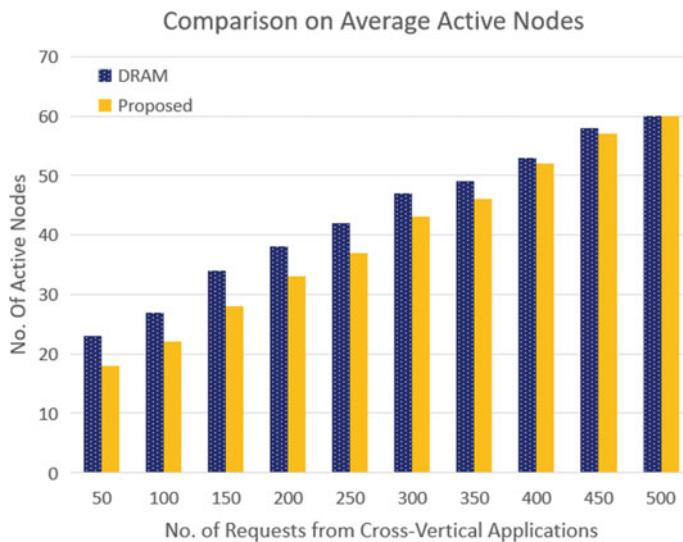


Fig. 2.3 Comparison on average active fog nodes

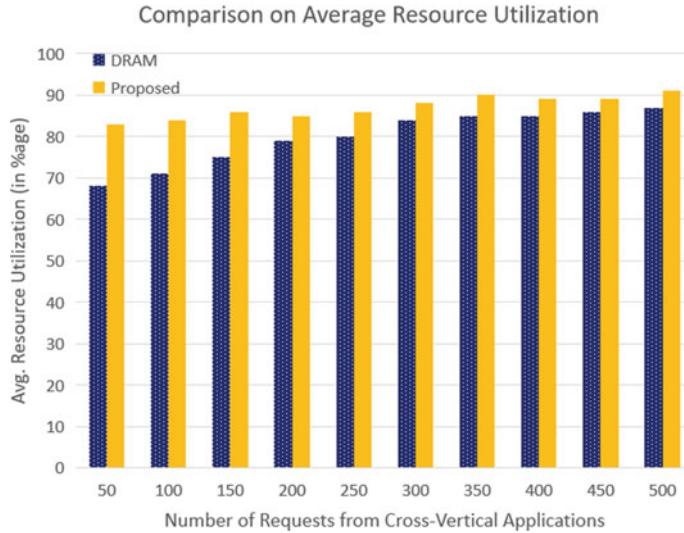


Fig. 2.4 Comparison on average resource utilization

2.5.2 *Evaluation on Resource Utilization*

Resource utilization plays a major role in deciding the load balance variance; therefore, we have considered this parameter for evaluating the performance of the model. Resource utilization generally refers to the number of resource units used to process the requests from applications on a fog node. Figure 2.4 presents the comparison on the percentage of resource utilization achieved by the proposed model versus the percentage of resource utilization by DRAM. Putting a glance on the graph, it can be considered that the increase in the number of requests does not impact the resource utilization achieved by our model. It can be said that the performance achieved by DRAM has been quite lower than expectations in the initial stages. Nonetheless, with gradations in the number of requests the resource utilization by DRAM increases gradually. However, without unheeding the fact that the proposed model has shown promising results in terms of resource utilization it can be said that the model is highly efficient and superior.

2.5.3 *Evaluation on Load Balance Variance*

Figure 2.5 presents a comparative study between the proposed model and the DRAM based on load balance variance. Skimming through the graph, it is clearly understandable that the proposed model outcomes the performance of DRAM in terms of load balance variance. It can be seen that through the entire experiment from the initial

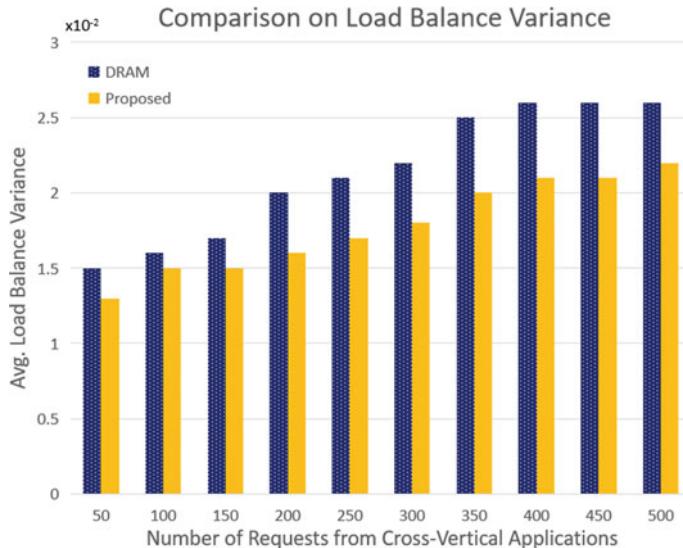


Fig. 2.5 Comparison on load balance variance

stage with lower number of incoming tasks till the later stages with highly incoming tasks, the proposed model has shown nearly about 20% better efficiency than DRAM.

Hence, scrupulously observing the entire empirical study it can be said that the proposed Resource-Aware, Load Balancing Strategy for Real-Time, Cross-Vertical IoT Applications proves to be an efficient load balancing methodology in terms of achieving high resource utilization, minimum load balance variance, and less power consumption.

2.6 Conclusion

In this era of modern technology, with the maturation of IoT, a multitude of advanced services are being foreseen. Cloud, an accustomed computing technology, cannot assuage the real-time services demanded by the cross-vertical IoT applications due to the high traffic and latency. Over and above this, though the fog computing paradigm, a hopeful alternative providing real-time services, was introduced, it failed to take the huge burden of the tremendous amount of data generated. An indecent resource management strategy in the fog infrastructure may further lead to a deterioration in the QoS and failure in providing services to the real-time tasks arriving from several cross-vertical applications. Without disregarding the unignorable issues and challenges in fog, this paper A Resource-Aware Load Balancing Strategy for Real-Time, Cross-Vertical IoT Applications has been presented. Through rigorous simulation carried out using CloudSim, and meticulous comparative study on the performance

of the model, it can be said that the model has achieved promising results and has proved to be an efficient load balancing strategy achieving high resource utilization and low power consumption. In the future, we shall be investigating on a detailed model exploring the task classification and node distribution methodologies.

References

1. Chiang, M., Zhang, T.: Fog and IoT: an overview of research opportunities. *IEEE Internet Things J.* **3**(6), 854–864 (2016)
2. Roy, D.S., Behera, R.K., Hemant Kumar Reddy, K., Buyya, R.: A context-aware fog enabled scheme for real-time cross-vertical IoT applications. *IEEE Internet Things J.* **6**(2), 2400–2412 (2018)
3. Behera, R.K., Hemant Kumar Reddy, K., Roy, D.S.: A novel context migration model for fog-enabled cross-vertical IoT applications. In: International Conference on Innovative Computing and Communications, pp. 287–295. Springer, Singapore
4. Stergiou, C., et al.: Secure integration of IoT and cloud computing. *Future Gener. Comput. Syst.* **78**, 964–975 (2018)
5. Biswas, A.R., Giaffreda, R.: IoT and cloud convergence: opportunities and challenges. In: 2014 IEEE World Forum on Internet of Things (WF-IoT), IEEE (2014)
6. Bonomi, F., et al.: Fog computing and its role in the internet of things. In: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing (2012)
7. Reddy, K.H.K., Behera, R.K., Chakrabarty, A., Roy, D.S.: A service delay minimization scheme for QoS-constrained, context-aware unified IoT applications. *IEEE Internet Things J.* **7**(10), 10527–10534 (2020)
8. Puthal, D., et al.: Secure and sustainable load balancing of edge data centers in fog computing. *IEEE Commun. Mag.* **56**(5), 60–65 (2018)
9. Deng, R., Lu, R., Lai, C., Luan, T.H., Liang, H.: Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption. *IEEE Internet Things J.* **3**(6), 1171–1181 (2016). <https://doi.org/10.1109/JIOT.2016.2565516>
10. Talaat, F.M., et al.: A load balancing and optimization strategy (LBOS) using reinforcement learning in fog computing environment. *J. Ambient. Intell. Humanized Comput.* 1–16 (2020)
11. Manju, A.B., Sumathy, S.: Efficient load balancing algorithm for task preprocessing in fog computing environment. In: Satapathy, S.C., Bhateja, V., Das, S. (eds.) Smart Intelligent Computing and Applications, pp. 291–298. Springer Singapore, Singapore (2019)
12. Talaat, F.M., et al.: Effective load balancing strategy (ELBS) for real-time fog computing environment using fuzzy and probabilistic neural networks. *J. Netw. Syst. Manage.* **27**(4), 883–929 (2019)
13. Xu, X., et al.: Dynamic resource allocation for load balancing in fog environment. *Wireless Commun. Mobile Comput.* **2018** (2018)

Chapter 3

An Elitist



Artificial-Electric-Field-Algorithm-Based Artificial Neural Network for Financial Time Series Forecasting

Sarat Chandra Nayak, Ch. Sanjeev Kumar Dash, Ajit Kumar Behera, and Satchidananda Dehuri

Abstract This article introduces the concept of elitism in the recently developed artificial electric field algorithm (AEFA) and termed it as eAEFA. The elitism method helps AEFA to preserve the best individuals from iteration to iteration through directly placing best fit particles into the population for the next generation and thus, strengthening the optimization capacity of AEFA. The proposed eAEFA is then used to find the optimal parameters of an artificial neural network (ANN) to form a hybrid model eAEFA + ANN. Financial time series (FTS) data behaves arbitrarily, highly volatile in nature, and possess nonlinearity, hence difficult to predict. We evaluated AEFA + ANN on predicting four FTS such as NASDAQ index, IND/USD exchange rate, WTI crude oil prices, and Bitcoin closing prices through four error statistics. Analysis of experimental outcomes are in favour of proposed model when compared with other.

3.1 Introduction

ANN is the most popular approximation algorithm used in the domain of data mining. The generalization ability of ANN mainly depends on its training, during which the optimal model parameters such as initial weights and biases are decided in an iterative manner. Several optimization methods are available in literature for ANN training, among which gradient descent-based techniques are widely used. Besides strong mathematical justifications behind these methods, few issues like slow convergence, poor accuracy, and sticking at local minima are associated with them. Last two

S. C. Nayak (✉)

Department of Artificial Intelligence and Machine Learning, CMR College of Engineering & Technology, Hyderabad 501401, India

Ch. Sanjeev Kumar Dash · A. K. Behera

Department of Computer Science, Silicon Institute of Technology, Bhubaneswar, India

S. Dehuri

Department of Information and Communication Technology, Fakir Mohan University, Vyasa Vihar, Balasore 756019, India

decades have seen ample number of non-derivative optimization methods inspired from natural phenomena [1, 2]. These nature and bio-inspired algorithms are used to solve real-world problems with varied complexities [3]. The usability of these methods is bottlenecked by number of parameters to be tuned, convergence rate, accuracy, execution time etc. A technique that needs minimal parameters without compromising the accuracy can be a better choice.

Evolutionary learning methods such as GA, PSO, and DE are more proficient methods in searching the optimal parameters of ANN [3]. Since no single technique is found suitable in solving all sort of problems, continuous improvements in existing methods have been carried out by researchers through enhancement in an algorithm [4, 5] or hybridization of them [6–9, 19]. Recently, AEFA has been anticipated as an optimization method inspired by the principle of electrostatic force [10]. AEFA is based on strong theoretical concept of charged particles, electric field, and force of attraction/repulsion between two charged particles in an electric field. The learning capacity, convergence rate, and acceleration updates of AEFA have been established in [10] through solving some benchmark optimization problems. AEFA starts with random solutions, fitness evaluation, reproduction, and updating the velocity and position of the particles in the search space. The updated solution is then compared with the previous and better-fit one is retained. However, there might be few good solutions in the previous step known as elite solutions. In elitism mechanism, these elite solutions are directly carried forward for the next generation without modification. The worst solutions are replaced by elite solutions. In any generation, the worst solutions are substituted by the elite individuals identified in the previous generation. The elitism mechanism of replacing worst solutions with elite one through each generation is applied with many swarms and evolutionary algorithms [11, 12].

FTS data such as stock closing price series, crude oil prices, exchange rates, and cryptocurrency prices are highly volatile in nature as they are influenced by many socio-economic and political factors. It is hard to predict their movements. However, recent advancement in computational intelligence technologies and availability of huge datasets intensified the process of FTS modelling and forecasting [6–9, 13–15]. Though ample number of nature-inspired learning algorithms for ANN training are observed in the literature in predicting FTS, an algorithm that needs fewer controlling parameters and possess good accuracy is highly desired.

This study has two objectives. First, we tried to introduce the elitism concept in AEFA (i.e. eAEFA) to improve its generalization ability. Secondly, we designed a hybrid model i.e. eAEFA+ANN. The model is used to predict the closing prices of four FTS such as NASDAQ index, IND/USD exchange rate, WTI crude oil prices, and Bitcoin closing prices.

We discuss the preliminaries in Sect. 3.2, demonstrate the model building in Sect. 3.3, forecast process in Sect. 3.4 and simulative outcomes in Sect. 3.5 followed by concluding notes.

3.2 Preliminaries

The background methods such as the basic AEFA and ANN are concisely presented in this section. For details, prospective readers are suggested to read the base articles.

3.2.1 AEFA

AEFA is designed on the principle of Coulomb's law of electrostatic force [10]. It simulates the charged particles as agents and measures their strength in terms of their charges. The particles are moveable in the search domain through electrostatic force of attraction/repulsion among them. The charges possessed by the particles are used for interaction, and positions of the charges are considered as the potential solutions for the problem. The overall AEFA steps are shown in Fig. 3.1. According to AEFA, the particle having highest charge is measured as the best individual; it attracts other particles having inferior charge and moves in the search domain. The mathematical justification of AEFA is illustrated in [10]. Here, we simulate a potential solution of ANN as a charge particle, and its fitness function as the quantity of charge associated with that element. The velocity and position of a particle at time instant t are updated as per Equations 3.1 and 3.2, respectively.

$$V_i^d(t+1) = \text{rand}_i * V_i^d(t) + \text{acceleration}_i^d(t) \quad (3.1)$$

$$X_i^d(t+1) = X_i^d(t) + V_i^d(t+1) \quad (3.2)$$

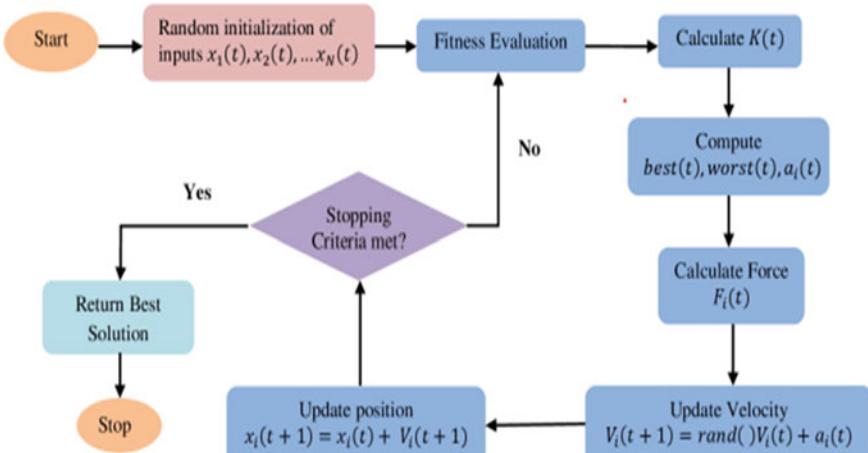


Fig. 3.1 AEFA process

3.2.2 ANN

A typical ANN architecture is depicted in Fig. 3.2. The first layer links input variables of the given problem. The subsequent layer is to capture non-linear associations among variables.

At each neuron j , in the hidden layer, the weighted output y_j is calculated using Eq. 3.3.

$$y_j = f \left(b_j + \sum_{i=1}^n w_{ij} * x_i \right) \quad (3.3)$$

where x_i is the i th input component, w_{ij} is weight value between i th input neuron and j th hidden neuron, b_j is the bias, f is a nonlinear activation function. Suppose, there are m numbers of nodes in this hidden layer, then for the next hidden layer, these m outputs become the input. Then, for each neuron j of the next hidden layer, input is as in Eq. 3.4.

$$y_j = f \left(b_j + \sum_{i=1}^m w_{ij} * y_i \right) \quad (3.4)$$

This signal flows in the forward direction through each hidden layer until it reaches the output layer. The output y_{est} is calculated using Eq. 3.5.

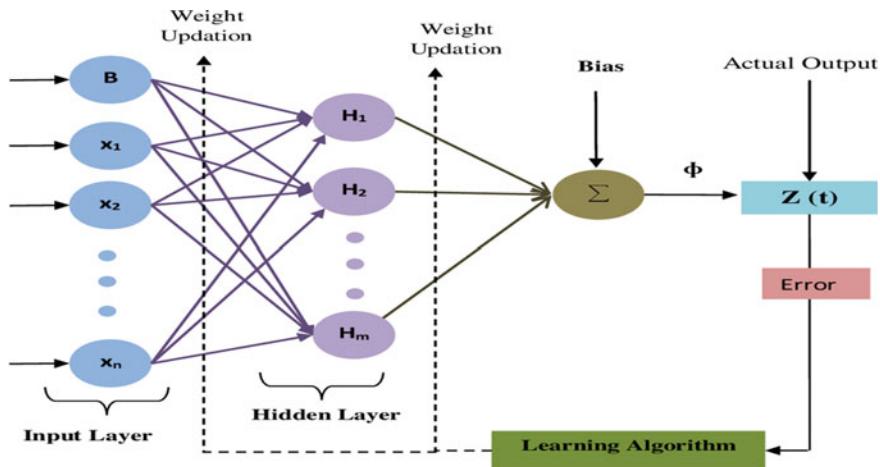


Fig. 3.2 Single hidden layer neural network

$$y_{\text{est}} = f \left(b_o + \sum_{j=1}^m v_j * y_j \right), \quad (3.5)$$

where, v_j is the weight between j th hidden to output neuron, y_j is the weighted sum calculated as in Eq. 3.1, and b_o is the output bias. The error is calculated by using Eq. 3.6.

$$\text{Error}_i = y_i - y_{\text{est}} \quad (3.6)$$

3.3 Proposed Approach

In this section, we first discuss the eAEFA+ANN approach and then the FTS forecasting using the proposed approach.

3.3.1 eAEFA + ANN

This section first describes about design of eAEFA+ANN model and then the FTS forecasting process. As discussed earlier, elitism is a mechanism to preserve the best individuals from generation to generation. By this way, the system never loses the best individuals found during the optimization process. Elitism can be done by placing one or more of the best individuals directly into the population for the next generation. The overall eAEFA+ANN process is depicted in Fig. 3.3. The process starts with a random initial population of solutions. An individual of the population represents a potential initial weight and bias set for the ANN. This population and the input samples are fed to the ANN model and the fitness is evaluated. Based on the fitness, a set of elite solutions are selected. The remaining of the population is undergone with the regular operators of AEFA. At the end of the current generation, the updated and original solutions are compared, and the better one is carried over. Here, the worst solutions are replaced by the elite solutions, and the process entered to the next generation. In this way, the elite solutions are carried forward through successive generations. Finally, the best solution is preserved and used for testing.

3.3.2 eAEFA + ANN-Based Forecasting

An ANN with single hidden layer as depicted in Fig. 3.2 is used in this study as the base model. The size of input layer is equal to the size of input vector. The hidden

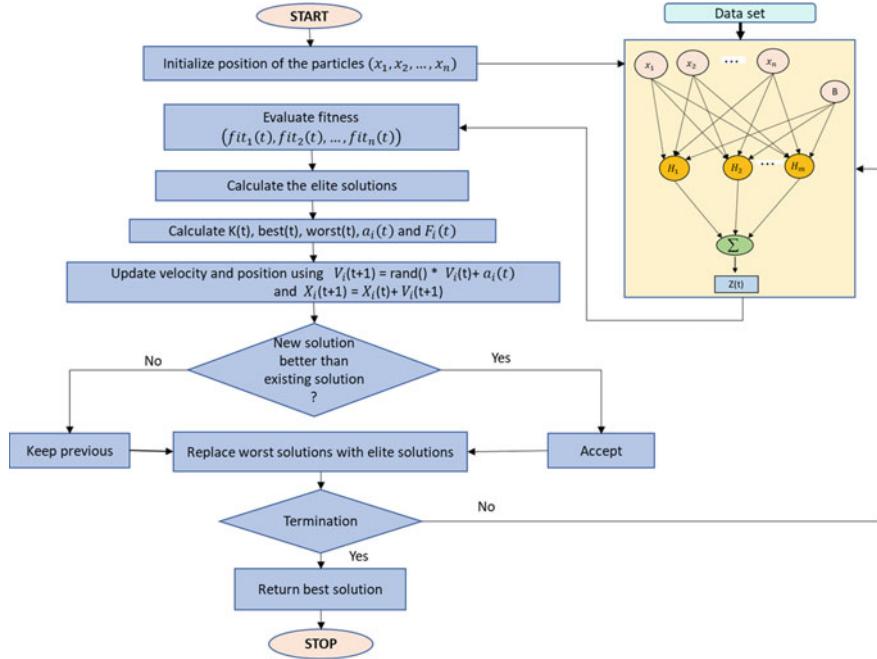


Fig. 3.3 The eAEFA + ANN process

layer size is decided experimentally. Since the number of dependent variable is one (the model is predicting one value), the output layer is fixed with one neuron only. The size of input vector is decided using a rolling window method [16, 18]. The inputs are normalized with sigmoid method and fed to the model [17]. Both hidden and output neurons used a sigmoid activation on the weighted sum of inputs. The absolute difference of true and estimated output at the output layer is considered as the fitness. The model is then trained with eAEFA as discussed in 3.1.

3.4 Experimental Result and Analysis

The proposed approach is then used to forecast the next data points on the four FTS. The real closing prices for the FTS are collected from the free source in [16] from 15 June 2020 to 15 June 2021. The information about these is listed in Table 3.1.

In order to establish the better performance of the proposed forecasting approach, four hybrid models such as ANN trained with genetic algorithm (GA+ANN), differential evolution (DE+ANN), and gradient descent method (GD+ANN) as well as two popular forecasts such as support vector machine (SVM) and auto-regressive integrated moving average (ARIMA) are designed in his study. All the seven forecasts are evaluated in a similar way. The mean absolute percentage of error (MAPE)

Table 3.1 Statistic summary of four FTS

FTS name	Short name	Minimum	Mean	Median	Variance	Maximum	Std. dev	Correlation coefficient
NASDAQ daily closing prices	NASDAQ	9.7260e + 03	1.2306e + 04	1.2520e + 04	1.6534e + 06	1.4174e + 04	1.2858e + 03	-0.1122
EURO/USD exchange rate prices	EUR/USD	1.1178	1.1898	1.1908	7.1739e-04	1.2341	0.0268	-0.0730
Daily Crude oil prices	Cnude Oil	35.7900	50.3249	46.7800	113.3724	71.7600	10.6476	-0.1084
Bitcoin/USD exchange rate	BIT/USD	9.0454e + 03	2.8595e + 04	1.9536e + 04	3.3475e + 08	6.3503e + 04	1.8296e + 04	-0.0701

Table 3.2 MAPE statistics from all models

FTS	eAEFA + ANN	AEFA + ANN	GA + ANN	DE + ANN	GD + ANN	SVM	ARIMA
NASDAQ	0.5093	0.5763	0.7205	0.7533	1.3217	0.9452	1.7544
EUR/USD	0.4765	0.5377	0.5540	0.6038	0.9287	0.7920	1.2045
Crude Oil	0.3823	0.5008	0.5327	0.5255	1.0326	0.9347	1.3025
BIT/USD	0.2460	0.4652	0.4855	0.4821	1.0273	0.7035	1.2147

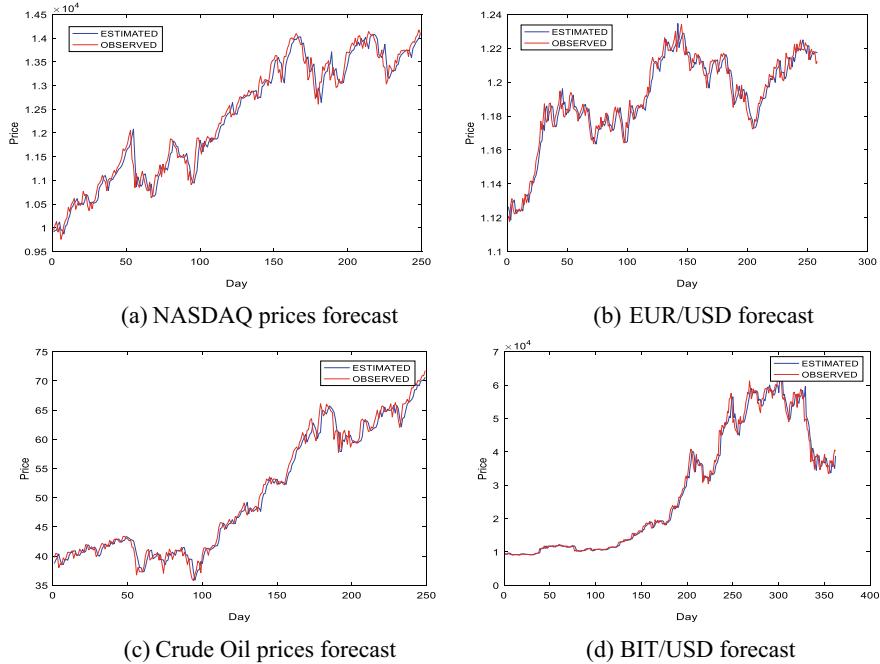
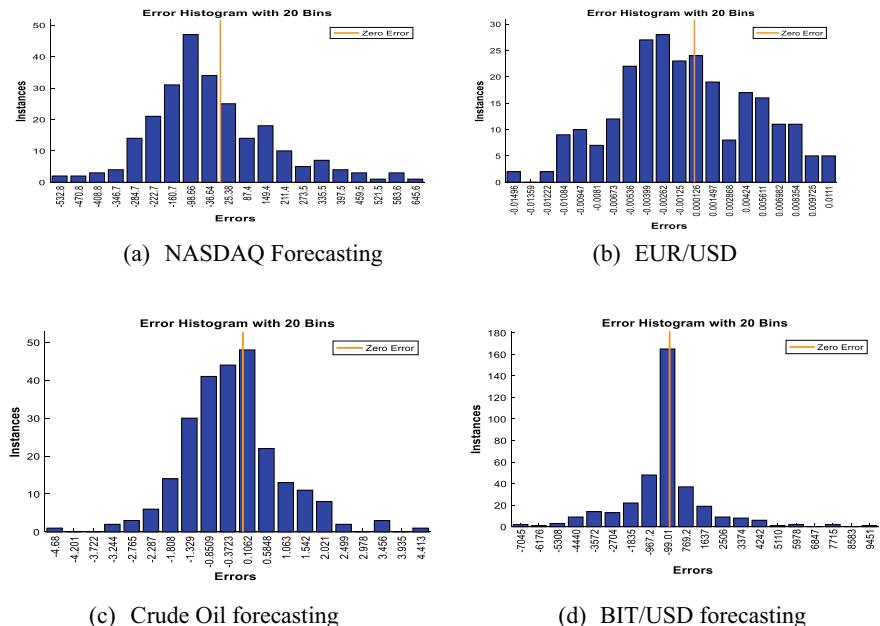
indicator as shown in Eq. 3.7 is used to evaluate the models and the simulated values are recorded in Table 3.2. The total number of samples is represented by N .

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \frac{|\text{Observed}_i - \text{Estimated}_i|}{\text{Observed}_i} \times 100\% \quad (3.7)$$

The observations from Table 3.2 are as follows. For all FTS, the MAPE generated by the eAEFA+ANN is lower than others. The MAPE by the proposed model from NASDAQ and EUR/USD are bit higher (i.e. 0.5093 and 0.4750, respectively) than that from crude oil and BIT/USD (i.e. 0.3823 and 0.2460, respectively). The use of elitism helps in improvising in accuracy. The performance of GA+ANN and GD+ANN are found nearer to each other. The SVM, GD+ANN, and ARIMA models shown inferior performance compared to hybrid methods. The forecast plots and residual histograms from eAEFA+ANN-based forecasting is depicted in Figures 3.4 and 3.5. This evidence is in support of the goodness of the proposed approach.

3.5 Conclusion

The article presented a hybrid forecast called eAEFA+ANN, where the optimal neuron parameters are decided by eAEFA. The concept of elitism has been incorporated to the recently developed AEFA in order to strengthen its optimization capacity. The elitism method helps AEFA to preserve the best individuals from iteration to iteration through directly placing best-fit particles into the population for the next generation. The proposed approach is applied to forecast the future value of four highly nonlinear FTS (NASDAQ, EURO/USD, crude oil, and BIT/USD prices) and found efficient in modelling and forecasting such volatile data. Comparative study in terms of MAPE with three hybrid methods and three conventional models established the superiority of eAEFA+ANN. The usability of the proposed method may be tested with other data mining problems.

**Fig. 3.4** Forecast plot of eAEFA + ANN-based model**Fig. 3.5** Residual histograms from different financial time series data

References

1. Fister Jr, I., Yang, X.S., Fister, I., Brest, J., Fister, D.: A brief review of nature-inspired algorithms for optimization. arXiv preprint [arXiv:1307.4186](https://arxiv.org/abs/1307.4186) (2013)
2. Yang, X.S.: Nature-inspired optimization algorithms: challenges and open problems. *J. Comput. Sci.* **46**, 101104 (2020)
3. Darwish, A.: Bio-inspired computing: algorithms review, deep analysis, and the scope of applications. *Future Comput. Inform. J.* **3**(2), 231–246 (2018)
4. Opara, K., Arabas, J.: Comparison of mutation strategies in differential evolution—a probabilistic perspective. *Swarm Evol. Comput.* **39**, 53–69 (2018)
5. Jiang, S., Wang, Y., Ji, Z.: Convergence analysis and performance of an improved gravitational search algorithm. *Appl. Soft Comput.* **24**, 363–384 (2014)
6. Nayak, S.C., Misra, B.B.: A chemical-reaction-optimization-based neuro-fuzzy hybrid network for stock closing price prediction. *Financial Innov.* **5**(1), 1–34 (2019)
7. Nayak, S., Ansari, M.: COA-HONN: cooperative optimization algorithm based higher order neural networks for stock forecasting. *Recent Adv. Comput. Sci. Commun.* **13**(1) (2020)
8. Chirotra, H., Abdulkareem, S., Herawan, T.: Evolutionary neural network model for West Texas intermediate crude oil price prediction. *Appl. Energy* **142**, 266–273 (2015)
9. Nayak, S.C., Das, S., Ansari, M.D.: TLBO-FLN: teaching-learning based optimization of functional link neural networks for stock closing price prediction. *Int. J. Sens. Wireless Commun. Control* **10**(4), 522–532 (2020)
10. Yadav, A.: AEFA: artificial electric field algorithm for global optimization. *Swarm Evol. Comput.* **48**, 93–108 (2019)
11. Rao, R., Patel, V.: An elitist teaching-learning-based optimization algorithm for solving complex constrained optimization problems. *Int. J. Ind. Eng. Comput.* **3**(4), 535–560 (2012)
12. Rajasekhar, A., Rani, R., Ramya, K., Abraham, A.: Elitist teaching learning opposition based algorithm for global optimization. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1124–1129. IEEE (2012)
13. Hamdi, M., Aloui, C.: Forecasting crude oil price using artificial neural networks: a literature survey. *Econ. Bull.* **3**(2), 1339–1359 (2015)
14. Nayak, S.C.: A fireworks algorithm based Pi-Sigma neural network (FWA-PSNN) for modelling and forecasting chaotic crude oil price time series. *EAJ Endorsed Trans. Energy Web* **7**(28) (2020)
15. Nayak, S.C.: Bitcoin closing price movement prediction with optimal functional link neural networks. *Evol. Intell.* 1–15 (2021)
16. Nayak, S.C., Misra, B.B., Behera, H.S.: ACFLN: artificial chemical functional link network for prediction of stock market index. *Evol. Syst.* **10**(4), 567–592 (2019)
17. Nayak, S.C., Misra, B.B., Behera, H.S.: Impact of data normalization on stock index forecasting. *Int. J. Comput. Inf. Syst. Ind. Manage. Appl.* **6**(2014), 257–269 (2014)
18. Dash, C.S.K., Behera, A.K., Nayak, S.C., Dehuri, S.: QORA-ANN: quasi opposition based Rao algorithm and artificial neural network for cryptocurrency prediction. In: 2021 6th International Conference for Convergence in Technology (I2CT), pp. 1–5. IEEE (2021)
19. Behera, A.K., Panda, M., Dehuri, S.: Software reliability prediction by recurrent artificial chemical link network. *Int. J. Syst. Assur. Eng. Manage.* 1–14 (2021)

Chapter 4

COVID-19 Severity Predictions: An Analysis Using Correlation Measures



Rashmita khilar, T. Subetha, and Mihir Narayan Mohanty

Abstract The outbreak of coronavirus worldwide infected people much and affected their lives and economy very badly. There are several anticipation techniques such as maintaining distance, health hygiene, and refrain congregation. The situation also led various researchers to discover remedies to overcome these situations using machine learning algorithms. This paper provides an early and necessarily selective review, discussing the contribution made by machine learning to fight against deadly disease, COVID-19, and it is grouped into drug discovery, trend analysis, medical image analysis, and other machine learning techniques. Correlation analysis between attributes like gender, age, and death rate is also performed using popular correlation tests such as Pearson, Spearman rank, and Kendall rank correlation test to reveal the relationships between the attributes. This paper also focuses on available amenities for providing the treatment for the patients such as the number of primary health centers, community health centers, sub-district hospitals, district hospitals, beds available in those centers. The results show that age is positively correlated with the deceased rate irrespective of the algorithm. This paper also focuses on the impending research directions in COVID-19.

4.1 Introduction

From past one year, the entire world is getting affected by coronavirus disease (COVID-19) [1, 2]. All the countries and territories in the world acquire infections from this virus. The effective incidents till August 2021 are around 21.3 crores, and the fatality rate is around 44.5 lakhs. Though the world's cases are so high, the

R. khilar (✉)

Department of IT, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India

T. Subetha

Department of IT, BVRIT HYDERABAD College of Engineering for Women, Hyderabad, India

M. N. Mohanty

ITER, Siksha 'O' Anusandhan (Deemed To Be University), Bhubaneswar, Odisha, India

death rate is comparatively low than SARS and MERS. However, this coronavirus's transmissibility and infectivity are to a high degree, and the deterrent measures of COVID-19 encompass social distancing, sanitizing hands frequently, and refraining touching nose and mouth. We cannot destroy COVID-19 [3] virus because we have this coronavirus for a long back. The people must adapt themselves to COVID-19 and the ways to protect themselves from the virus. As we live in the digitized world, many researchers, from the young to the senior scientists, take this situation as a challenge, and as a result, hundreds of researchers are forming groups and collecting data from various countries, genders, and locality too and solutions for the same. The research started not only among the doctors and drug specialists to find out the vaccine but also among the machine learning researchers to bring out various disease trends. The novel coronavirus outbreak worldwide has challenged various researchers, mathematicians, pharmacists, etc., to find this virus's actual patterns and behavior. Many researchers started working on machine learning concepts and algorithms to find a great place among various scientists. Computer engineers with ML and data science use various prediction algorithms to predict and detect the severity of the virus pertaining to various age groups, gender, residing locality, traveling history, etc.

In this paper, COVID-19 Indian data obtained from various sources have been investigated. This paper helps the doctors and researchers predict the virus's verity on the patient's body based on age factor, gender, locality, and travel places. We have found the correlation analysis with three popular correlation tests to dig out the relationship between attributes like gender, age, and death rate. The administrators can use the discovered analysis to make judgments about the infected people and make arrangements accordingly, thereby reducing the mortality rate. We have also analyzed various parameters collected from trustable sources in India till June 30, and the experimental analysis shows that the age and mortality rate are positively correlated irrespective of all the three algorithms.

The paper is systematized as follows. Section 4.2 depicts the works consummated in COVID-19. Section 4.3 discusses the analysis carried out in India's COVID-19 data, followed by a conclusion in Sect. 4.4.

4.2 Literature Survey

COVID-19 research can be categorized into drug-based research, medical image-based research, trend analysis-based research, and other machine learning analysis-based research. In this section, we will discuss the works carried out in each area.

4.2.1 Drug-Based Research

Alimadadi et al. [4] have classified the COVID-19 virus based on genomes and predicted the severity based on the genome types. This paper has laid faster access for significant data mining concepts by bioinformaticians and scientists by integrating global COVID-19 patient data for diagnosis and predictive purposes. Randhawa et al. [5] propose digital signal processing (MLDSP) with the help of machine learning for genome analyses. They used a decision tree classifier to validate their result.

With the help of the above approach, they have analyzed and classified the genomes for the cause of the disease, and the machine learning approach is applied to obtain a reliable real-time taxonomic classification. The authors found two different viruses, such as the Sarbeco virus and Betacoronavirus, through this approach. The authors stated that their results were promising, and a comparative study on the viruses is obtained, which will be useful in this pandemic period.

4.2.2 Medical Image-Based Research

Khalifa et al. [6] disclose the inflames from X-ray images of the chest. Generative adversarial network (GAN) is used to detect the type of X-ray image. The authors concluded that the ResNet 18 obtains an appropriate testing accuracy measurement model using the deep transfer method. They have achieved 99% accuracy along with other performance metrics like F -score, precision, and recall. One major drawback in this paper is that the datasets are not original, the originality of their dataset is only 10%, and the dataset is limited to only X-ray images. An AI-empowered image acquisition method [7] is developed to help the laboratory technicians and radiologists develop an automated scanning procedure. They have concluded that the inputs from fusing information from X-ray and CT scan work efficiently compared to other systems. An experimental analysis is performed to apply the machine learning method search and discriminate the chest X-rays of pneumonia and COVID-19 [8]. Since the research industry is flooded with many machine learning algorithms, prediction techniques are applied for scrutinizing the new confirmed cases. Many machine learning models are also applied positively to the growing body of knowledge. Machine learning algorithms are applied to distinguish patients affected by COVID from chest X-ray (CXR) or CT scan. Liu et al. [9] and many others also stated that a better algorithm is required to identify the severity of the disease.

Chung et al. [9–11] use a concept called COVIDX-Net in the deep learning method. The COVIDX-Net method applies a deep learning method to study the X-ray and CT scan reports. Their method is used to assist the radiologist in automatically diagnosing the patient. A deep convolutional neural network is adapted [12] to disclose the status of classifying the patient status, either negative or positive COVID-19 case. The above method provides promising results with X-ray image input. However, the author stated that only 20% of X-ray images are modeled, trained,

and tested. Classification of disease yields a good result. The above methods need to be tested clinically in real time by the authors. Many other researchers also adopted a deep convolutional neural network technique to identify and analyze the disease's severity.

4.2.3 Trend Analysis-Based Research

The researchers have analyzed the outbreak of disease based on the lockdown system [13]. The authors have analyzed and predicted the progression of the disease. They also predicted that if the lockdown works, India will have less than 66,224 cases by May 1, 2020. The risk factors in a country are analyzed using LSTM [14]. The authors have also combined the weather information for predicting the outbreak of COVID-19. Their results are tested among 170 countries, and the authors have developed a tool for categorizing the risk factor based upon the country and climatic nature.

The researchers have also stated that this work can be expanded and explored disparate data modalities like flight, tourists, etc., and can also predict such a pandemic outbreak's economic effect. Marcello Ienca et al. and many other researchers stated that every day, every minute, a large number of datasets are produced. Therefore, an advanced computational model and machine learning techniques are applied for tracing the source of this deadly disease. The authors also predicated on how to reduce the spread of disease. The author has integrated the patient's database with their travel history for the identification of cases. A QR code scanning technology is for online reporting and also for containment purposes. This information is also applied to the people who are under quarantine. Sharma et al. [15] predicted the dire subject with persistent COVID-19 infection using a prognostic model to identify the virus's three clinical features. They built an XGBoost prediction model to check the severity of the virus in patients, and also the death rate is also calculated based on the severity. With a mean age of 58 years, the above method is clinically useful as it identifies the severity and death rate because, based on this, the doctor can proceed with further treatment.

Shi et al. [16] categorize the prediction into two types: data science/machine learning techniques and stochastic theory mathematical models. The authors assemble data from varied platforms to predict the severity.

4.2.4 Other Machine Learning Analysis-Based Research

Ardabili et al. [17] provided analogy analysis of COVID-19 pandemic outbreak using ML and soft computing models. They investigated two models, named MLP and ANFIS. The system also provides a practical approach to calculate the number of people infected and the number of people who died. This calculation is virtually required to find out ICU beds become free from time to time, and it is utmost required

for modeling the mortality rate. This concept is mostly required with other new facilities by every nation. Bonacini et al. [18] performed a lockdown analysis to control the outbreak of the disease, and they have also developed a feedback control system for developing suggestions to mitigate the policies to decrease risk failure. Liu et al. [19] enable geo-6 spatial synchronizations of COVID-19 activities. They have also used an augmentation technique to deal with historical diseases and the outbreak of pandemic disease. Their model is normalized by using z -score values during training and prediction. They have fixed predictors based on the number of standard deviations. The above method outperforms a collection of baseline models. Onder et al. [20] monitor the subjects and analyze the accurate reporting of patient aspects and testing policies. They have clustered the patients based on their severity level. COVID-19 is detected using built-in smartphone sensors [21], a low-cost solution, due to the availability and rapid usage of smartphones by most people. The AI-empowered framework takes the input of smartphone sensors to forecast the rate of the asperity of COVID-19, and also, the above framework predicts the result of the disease. Wang et al. [22] stated that the virus is pandemic and sensitive to the public. Early recognition is the only solution for this issue. Therefore, the government has to take serious steps in delivering all sorts of information regarding health issues to the public in an accurate time. The above paper discussed Taiwan's issues and how they alerted people about all types of awareness to the public at the correct time. To make this possible, a data analytics concept is applied, and now we can see fewer cases in Taiwan.

4.3 Correlation Analyses on COVID-19 India Data

In this section, various analyses are performed on the India statistical data obtained from various sources such as COVID-19 India (<https://www.covid19.india.org/>), ICMR, etc., till June 15, 2020, to predict the growing trends of cases in India [23]. Though many predicted algorithms are developed to forecast the COVID-19 cases, very few papers analyze the causality relationship between the variables. Analyzing the similarities between variables can make us understand how the variables are closely related and how the variables are dependent on others. This inference will make us understand the situation and help the authority take the necessary precaution measures resulting in the mortality rate reduction. A sample of 26,669 is taken from COVID-19 India, and the correlation measures are calculated using the R statistics tool. The popular correlation techniques such as Pearson, Spearman, and Kendall rank correlation are utilized to find the variables' causal relationships. Table 4.1 lists the attributes taken for consideration.

Table 4.1 Attributes taken for correlation analysis

Attribute name	Attribute description
Age	Ranges from 0 months to 100
Gender	Male and female
Current status	Hospitalized/recovered/deceased

4.3.1 Pearson Correlation

It is adopted to find the association between variables, and it is widely used as it adopts the covariance method for finding the associations. Pearson's correlation coefficient formula is given in Eq. 4.1 [24].

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.1)$$

4.3.2 Spearman Rank Correlation

It determines the relationship between two variables and depicts it with a monotonic function. The formula for calculating the Spearman rank correlation is given in Eq. 4.2 [25].

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (4.2)$$

4.3.3 Kendall Rank Correlation

Kendall's Tau [26] finds the relationships between columns of ranked data, and it returns either 0 or 1 depending upon the similarity. The formula for calculating the Kendall rank correlation is given in Eq. 4.3.

$$\rho = \frac{A - B}{A + B} \quad (4.3)$$

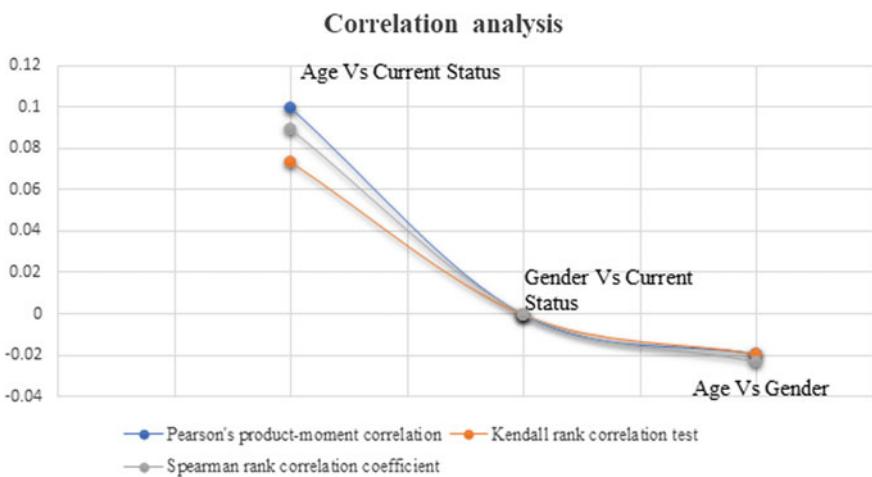
where A represents the accordant pairs and B represents the different pairs. The analysis results obtained on these correlation measures are given in Table 4.2.

Table 4.2 Results obtained in correlation analysis

	Age versus current status	Gender versus current status	Age versus gender
Pearson's product-moment correlation	0.09912459	-0.000356454	-0.01930927
Kendall rank correlation test	0.07328732	-0.000360185	-0.01931225
Spearman rank correlation coefficient	0.08907711	-0.00036078	-0.02344122

A correlation graph is drawn between Pearson's, Kendall's, and Spearman's rank on age versus gender and the affected people's current status. The graph is shown in Fig. 4.1.

From the results obtained, it is clear that the death rate is positively correlated with age. There is a high similarity between these two attributes, and the government should monitor older adults to decrease the deceased rate. It is evident from Table 4.2 that though gender and the current status are negatively correlated, the obtained similarity value is significantly less, and it is near to 0. So, we can infer that a little relationship exists between gender and current status. However, it is evident that no causality exists between age and gender. The overall inference from the table is that there is a high causality between deceased rate and age. This is due to their age-related health ailments, stress, immunity, etc. The analysis is also done on India's total and daily cases until June 2020, it is visualized in Figs. 4.2 and 4.3, and we can see a surge in daily cases in India. As we can see in the figure, though the cases are

**Fig. 4.1** Correlation graph analysis

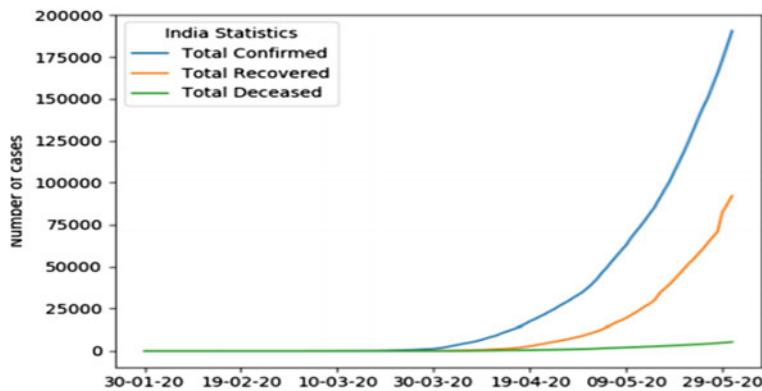


Fig. 4.2 Total number of cases in India till may 2020

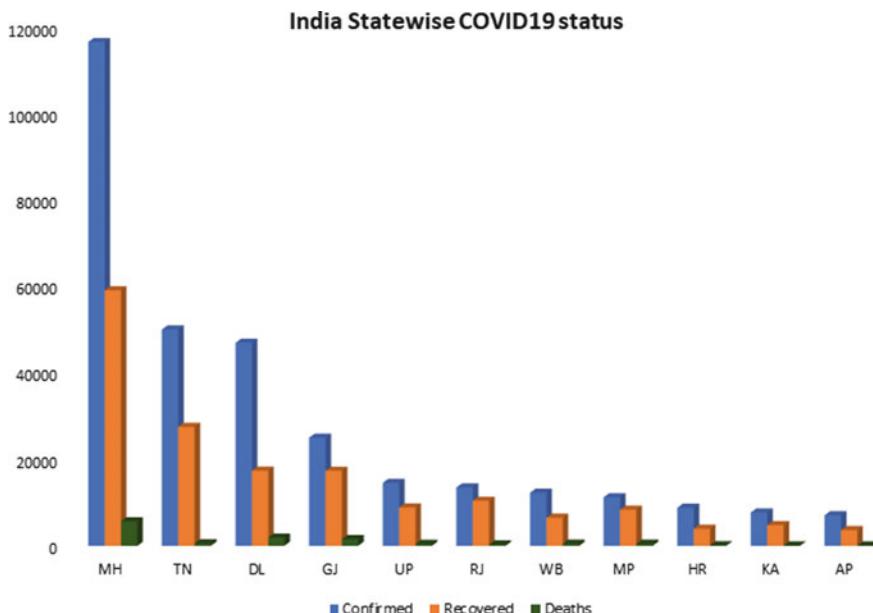


Fig. 4.3 Daily surging cases in individual states of India till June 2020

rising in India, the death rate is relatively low. The rise is due to our demographic conditions, immunity level, etc.

The frequency of affected persons by age and gender is also calculated, and it is given in Figs. 4.3 and 4.4. Here, 1 represents male, and 2 represents female. We can confirm from the figure that men are affected more compared to women. Experts also say that men are dying twice the rate of women in COVID-19. This is because women have higher immunity than men, and the two X chromosomes in women

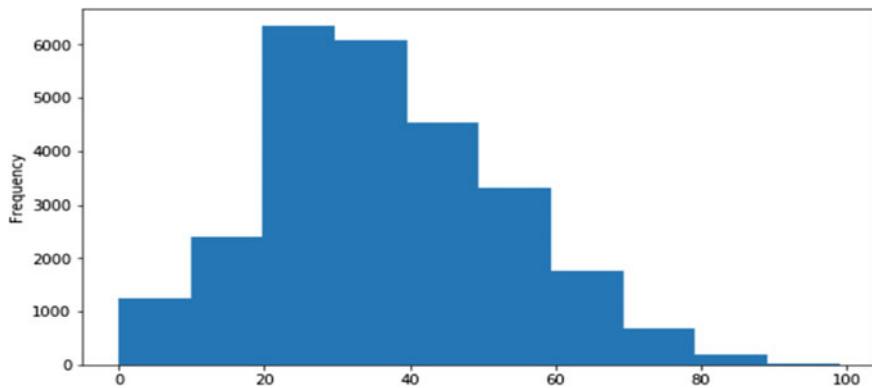
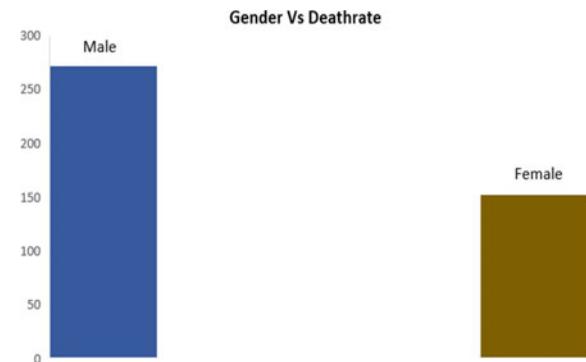


Fig. 4.4 The number of people gets affected based on age

make them stronger than men. Behaviors such as smoking, alcohol consumption, and stress can also be the cause of more men victims. They also add that men tend to ignore physical distancing and do not take the symptoms seriously.

Though India ranks amid the top ten countries in the confirmed COVID-19 cases, the disease infections affected are much lower than its population. Nevertheless, India's testing rates are significantly lower, reigning out any purpose for relief. The analysis done on the number of testing performed in India and the individual states given by the ICMR testing report is visualized in Figs. 4.5 and 4.6. This figure shows that Tamil Nadu is taking more tests per day and understands the importance of random tests compared to other states. Even though the number of infections in India is less than in many countries in March and April, we can find cases are surging every day to turn the country into the top 10 countries affected with the most number of cases. India's case trajectory is rising as all other countries' graph started to slow down, which is a very alarming situation for India. Among all the states in India, Tamil Nadu's testing number is relatively better, and it is highlighted in the graph that compares tests per case over time. In other states such as Maharashtra, Madhya

Fig. 4.5 The number of people gets affected based on gender



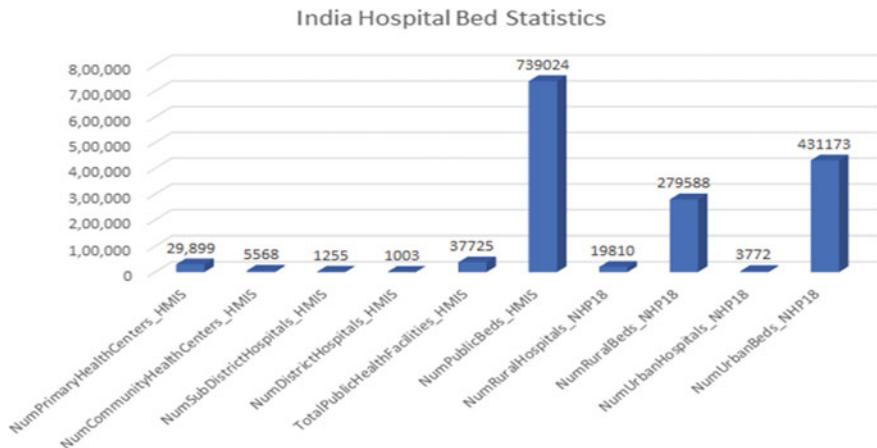


Fig. 4.6 Analysis of hospital beds

Pradesh, Delhi, and Gujarat, testing strategies are inferior, and their metric shows they are affected with relatively higher cases and lower tests per confirmed case.

Hospital bed count in India is also analyzed, and it is shown in Fig. 4.5. The figure highlights the number of primary health centers, number of community health centers, number of sub-district hospitals, number of district hospitals, beds available in those centers. From the figure, it is evident that if cases are surging in India daily, bed scarcity will occur, and the people's treatment will be delayed. Since this is the total available beds in India, it will be shared among all the patients with all health ailments. So the government can allocate only some amount of the beds for COVID-19, which will increase the mortality rate in India in the upcoming days as the recovered rate depends only on the early treatment of COVID-19. Thus, the government must take some adequate measures in increasing the bed's availability.

4.3.4 Chatbot—Information Checker

We have also developed a simple contextual AI chatbot for COVID-19 information check using RASA. The required data is collected from the WHO FAQ and Coronavirus Disease 2019 (COVID-19) FAQ. RASA [27] is a contextual AI structure for the creation of a chatbot. Customized intents are created in nlu.md and their responses in domain.yml. The sample interactions are created in stories.md. A basic version of COVID-19 information checker is created with 25 intents and responses. In the future, custom actions such as COVID data tracker and COVID area tracker will be implemented and deployed in real time. The sample screenshots for the developed COVID-19 chatbot are given in Fig. 4.7.

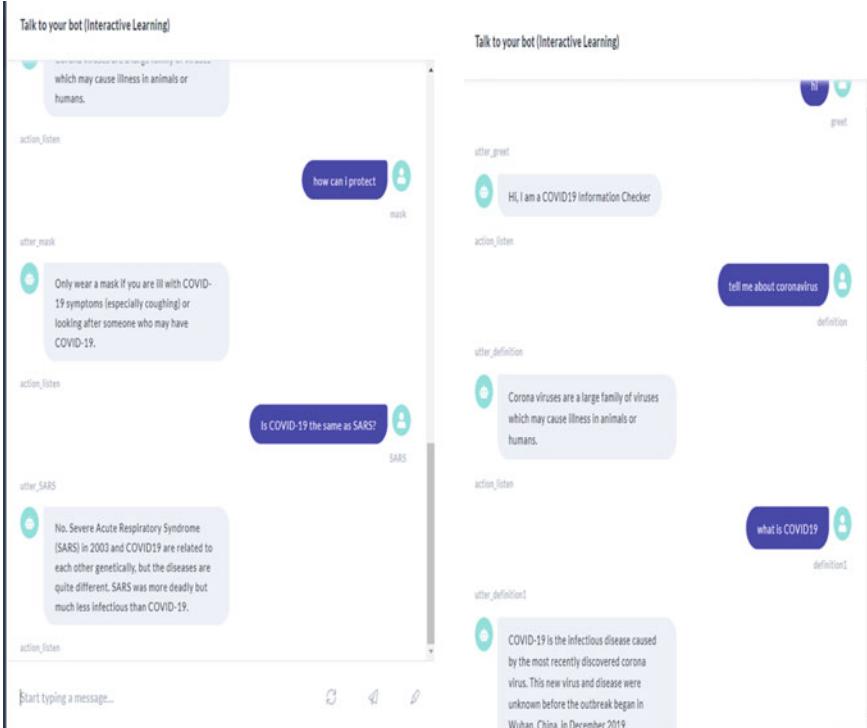


Fig. 4.7 Sample screenshots of COVID-19 information checker

4.4 Future Research Directions in COVID-19

There are still many open issues, and they deserve further research in COVID-19. Some of them are discussed in this section—drug discovery of COVID-19 using AI and ML techniques—trend analysis of COVID-19—medical CT scan image analysis for COVID-19 symptoms—use of virtual reality/augmented reality in COVID-19—applications of IOT such as smart sensors in COVID-19—big data solutions for COVID-19—AI-enabled chatbots for COVID-19 discussions and identification—telemedicine solutions during COVID-19 period—solutions to ambient-assisted living during COVID-19—smart electronic medical report maintenance system during COVID-19—crowd sourcing data collection during COVID-19—fake news prediction during corona crisis.

4.5 Conclusions

COVID-19 is threatening the world with its huge transmissibility and infectivity. Many researchers and scientists are in the process of discovering patterns, drugs, and solutions to various issues affecting the world due to this pandemic situation. We have reviewed various researches carried out in the field of drug discovery, trend analysis, medical image analysis, data mining techniques to extract various solutions for COVID-19. Various analyses are also performed by considering the essential attributes like age, gender, recovered rate, mortality rate, number of tests performed, etc. Three correlation tests have also been performed on these attributes resulting in a positive correlation with age and death rate. Some of the challenges and future directions in COVID-19 are also discussed in this paper. In summary, the review on COVID-19 depicts major progresses in various aspects. However, the research works carried on COVID-19 still have not addressed various challenges of perfect medical image analysis, smart sensors to detect COVID-19 patients, drug discovery to its full extent.

References

1. Dietrich, A.M., Kuester, K., Muller, G.J., Schoenle, R.S.: News and uncertainty about COVID-19: survey evidence and short-run economic impact. Becker Friedman Institute for Economic White Paper (2020)
2. Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B., Xu, Y.: A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. *Gen. Psychiatry* **33**(2), e100213 (2020). <https://doi.org/10.1136/gpsych-2020-100213>
3. Srinivasa Rao, A., Vazquez, J.A.: Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine. *Infect. Control Hosp. Epidemiol.* 1–5 (2020). <https://doi.org/10.1017/ice.2020.61>
4. Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P.B., Joe, B., Cheng, X.: Artificial intelligence and machine learning to fight COVID-19. *Physiol. Genomics* **52**(4), 200–202 (2020). <https://doi.org/10.1152/physiolgenomics.00029.2020>
5. Randhawa, G.S., Soltsysak, M.P., El Roz, H., de Souza, C.P., Hill, K.A., Kari, L.: Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *Plos One* **15** (2020). <https://doi.org/10.1371/journal.pone.0232391>
6. Khalifa, N.E.M., Taha, M.H.N., Hassanien, A.E., Elghamrawy, S.: Detection of coronavirus (COVID-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest X-ray dataset. *arXiv preprint arXiv:2004.01184* (2020)
7. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D.: Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE Rev. Biomed. Eng.* (2020). <https://doi.org/10.1109/RBME.2020.2987975>
8. VanBerlo, B., Ross, M.: Investigation of explainable predictions of COVID-19 infection from chest X-rays with machine learning. *Artificial Intelligence Lab* (2020)
9. Liu, H., Liu, F., Li, J., Zhang, T., Wang, D., Lan, W.: Clinical and CT imaging features of the COVID-19 pneumonia: focus on pregnant women and children. *J. Infect.* **80**(5), pp e7-e13 (2020). <https://doi.org/10.1016/j.jinf.2020.03.007>
10. Chung, M., Bernheim, A., Mei, X., Zhang, N., Huang, M., Zeng, X., Jacobi, A.: CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology* **295**(1), 202–207 (2020)

11. Kroft, L.J., van der Velden, L., Girón, I.H., Roelofs, J.J., de Roos, A., Geleijns, J.: Added value of ultra-low-dose computed tomography, dose equivalent to chest X-ray radiography, for diagnosing chest pathology. *J. Thorac. Imaging* **34**(3), 179 (2019)
12. Hemdan, E.E.D., Shouman, M.A., Karar, M.E.: Covidx-net: a framework of deep learning classifiers to diagnose COVID-19 in X-ray images. arXiv preprint [arXiv:2003.11055](https://arxiv.org/abs/2003.11055) (2020)
13. Das, S.: Prediction of covid-19 disease progression in india: under the effect of national lockdown. arXiv preprint [arXiv:2004.03147](https://arxiv.org/abs/2004.03147) (2020)
14. Pal, R., Sekh, A.A., Kar, S., Prasad, D.K.: Neural network based country wise risk prediction of COVID-19. arXiv preprint [arXiv:2004.00959](https://arxiv.org/abs/2004.00959) (2020)
15. Sharma, S.K., Gupta, A., Biswas, A., Sharma, A., Malhotra, A., Prasad, K.T., Broor, S.: Aetiology, outcomes and predictors of mortality in acute respiratory distress syndrome from a tertiary care centre in north India. *Indian J. Med. Res.* **143**(6), 782 (2016)
16. Shi, S., Qin, M., Shen, B., Cai, Y., Liu, T., Yang, F., Gong, W., Liu, X., Liang, J., Zhao, Q., Huang, H.: Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. *JAMA cardiol.* (2020). <https://doi.org/10.1001/jamacardio.2020.0950>
17. Ardabili, S.F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A.R., Reuter, U., Rabczuk, T., Atkinson, P.M.: COVID-19 outbreak prediction with machine learning. Available at SSRN (2020): <https://ssrn.com/abstract=3580188> or <https://doi.org/10.2139/ssrn.3580188>
18. Bonacini, L., Gallo, G., Patriarca, F.: Drawing policy suggestions to fight covid—from hardly reliable data. A machine-learning contribution on lockdowns analysis. Tech. Rep. GLO Discussion Paper (2020)
19. Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J.T., Vespignani, A., Santillana, M.: A machine learning methodology for real-time forecasting of the 2019–2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. arXiv preprint [arXiv:2004.04019](https://arxiv.org/abs/2004.04019) (2020)
20. Onder, G., Rezza, G., Brusaferro, S.: Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA* **323**(18), 1775–1776 (2020). <https://doi.org/10.1001/jama.2020.4683>
21. Maghdid, H.S., Ghafoor, K.Z., Sadiq, A.S., Curran, K., Rabie, K.: A novel ai-enabled framework to diagnose coronavirus COVID-19 using smartphone embedded sensors: design study. arXiv preprint [arXiv:2003.07434](https://arxiv.org/abs/2003.07434) (2020)
22. Wang, C.J., Ng, C.Y., Brook, R.H.: Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. *JAMA* **323**(14), 1341–1342 (2020)
23. Raghavendran, C.V., Satish, G.N., Krishna, V., Basha, S.M.: Predicting rise and spread of COVID-19 epidemic using time series forecasting models in machine learning. *Int. J. Emerg. Technol.* **11**(4), 56–61 (2020)
24. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: *Noise Reduction in Speech Processing*, pp. 1–4. Springer, Berlin, Heidelberg (2009)
25. de Winter, J.C., Gosling, S.D., Potter, J.: Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. *Psychol. Methods* **21**(3), 273 (2016)
26. Abdi, H.: The Kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics*, pp. 508–510. Sage, Thousand Oaks, CA (2007)
27. Deepika, K., Tilekya, V., Mamatha, J., Subetha, T.: Jollity chatbot—a contextual AI assistant. In: *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1196–1200. IEEE (2020)
28. Cucinotta, D., Vanelli, M.: WHO declares COVID-19 a pandemic. *Acta bio-medica: Atenei Parmensis* **91**(1), 157–160 (2020). <https://doi.org/10.23750/abm.v91i1.9397>

Chapter 5

Antenna Array Optimization for Side Lobe Level: A Brief Review



Sarmistha Satrusallya and Mihir Narayan Mohanty

Abstract In recent year, requirement of antenna is necessary for long-distance communication with minimized interference for wireless communication. Antenna array becomes an essential part of communication to achieve directivity, gain, and bandwidth. Array antenna optimization is one of the suitable methods to enhance the performance of the same. In this work, a short review of earlier optimization method for different parameters of the array is discussed. This review describes the methods used for reduction of side lobes in the array. The nature-inspired optimization methods provide better performance for optimal reduction of side lobes.

5.1 Introduction

Wireless communication requires point-to-point communication for efficient transmission of signal. For this reason, a highly directive beam of radiation is required. An antenna array comprising of several radiating elements in electrical or geometrical configuration fulfils the demand. The radiation pattern of the array depends on the geometry of the array and the pattern of the individual elements. Optimization of array is defined as a procedure to maximize or minimize certain measures of the radiation pattern. The performance of the array depends on the side lobe level, null control, and beam width. Different optimization methods were used to study the effect of the parameters for several applications in different frequency range.

5.2 Side Lobes in Antenna Array

The radiation pattern of the antenna is represented with local maxima of the far field radiation pattern known as lobes. The main lobe and the side lobes are the patterns of antenna at various angles. For directional antenna, the main lobe is designed

S. Satrusallya (✉) · M. N. Mohanty

Department of Electronics and Communication Engineering, S'O' A Deemed to be University, Bhubaneswar, Odisha 751030, India

to have maximum signal strength in the specified direction. The other lobes are called side lobes and represent unwanted radiation in undesired direction. Side lobes present behind the main lobe is back lobe. Excessive side lobes are responsible for interference to other components. In receiving part, side lobes are responsible for increasing the noise level in the signal. The power density of side lobes is much more less as compared to main lobe. But, it is desirable to minimize the side lobe level (SLL) which is measured in decibel.

5.3 Optimization in Antenna Array

Optimization algorithm is a procedure which is executed considering number of iteration of various solutions to obtain an optimum or satisfactory solution. Antenna optimization aims at creating efficient performance, serviceability, and cost effectiveness. This process involves selection of appropriate objective functions, design variables, parameters, and constraints.

In this paper, a comparative study of different optimization methods is discussed for minimization of side lobe level (SLL). Section 1 represents the introduction, Sect. 2 describes the effect of side lobes in antenna array. Need of optimization in array is explained in Sect. 3, and 4 analyzes the methods required to minimize the SLL. Conclusion of the study is represented in Sect. 5.

5.4 Optimization Methods

The array used for communication may be linear and circular. The arrangement of array elements are uniform or nonuniform. Nonuniform circular arrays were considered for optimum side lobe reduction using two evolutionary optimization methods. Biogeography-based optimization (BBO) and self-adaptive differential evolution were used to determine the optimum set of weights and position with side lobe reduction for a fixed major lobe beam width. BBO is based on the concept of nature's way of species distribution. The best solution and weak solution depend on island with a high habitat stability index (HSI) and low HSI. The BBO algorithm consists of three steps:

- (i) Creation of a set of solution to the problem
- (ii) Random selection
- (iii) Application of migration and mutation steps to reach optimal solution.

The BBO algorithm provides a better reduction in SLL as compared to genetic algorithm [1].

Real-coded genetic algorithm (RGA) is another method to determine the optimal set of current excitation weights of the antenna elements and the optimum inter element testing. The method is used to introduce deeper null in the interference

direction and to suppress SLL with respect to main beam width for a symmetric linear array. The RGA algorithm performs the following steps [2]:

- (i) Randomly generates an initial population within the variable constraint range
- (ii) Computes and saves the fitness for each individual in the current in the current population
- (iii) Defines the selection probability for each individual so that it is proportional to its fitness
- (iv) Generates the next population by probabilistically selecting the individual from the previous current population to produce off spring via genetic operators.
- (v) Repeat step 2 until a satisfactory solution is obtained.

This work described the design of nonuniformly excited symmetric linear antenna array with optimized nonuniform spacing between elements. The optimization method is suitable deeper nulls in the interference direction and reduced SLL.

Genetic algorithm was used for widening of beam width in the array. An eight-element linear array had a beam broadening factor of 2.44 using the GA [3]. The same method was used for thinned array for SLL reduction. Thinning of array require removal of selective elements to obtain the desired value. An array of 200 elements were considered for thinning using GA. The lowest possible peak side lobe of -31.02 db was achieved by turning off 14 elements [4]. Genetic algorithm and particle swarm optimization were compared for rectangular planner array antenna of different number elements. The array exhibit better performance for PSO with a reduced SLL of -30.39 db [5]. Differential evolution (DE) was analysed for micro strip array of 20 elements. The DE algorithm had a reduced SLL of -32.58 db [6]. Authors discussed fire fly algorithm (FA) to enhance the performance of the array [7]. FA is a method belongs to swarm brainpower family. The flashing pattern of each fire fly is considered for the method. Three numbers of idealized rules have been used for simplification of the method.

- (i) All flies are unisex
- (ii) Attractiveness is proportional to brightness of the flies
- (iii) The brightness of a fireflies is determined by the cost function of the problem.

The FA was used for minimizing the SLL of array by optimizing the element positions, current excitation, and phase of elements. The fitness function of the array is represented as [7].

$$\text{Fitness} = \min(\max\{20 \log|AF(\varphi)|\}) \quad (5.1)$$

The firefly algorithm converged fast in less iteration considering re-optimization of the amplitude excitation. The increase in number of elements in the array reduced the SLL of the antenna. Gray wolf optimization (GWO) was also discussed for minimum side lobe level along with null replacement in the specified direction in [8]. GWO is an algorithm which mimics the social hierarchy and hunting mechanism

of gray wolfs. The algorithms consider both male and females as leaders known as alphas. The beta wolves followed by delta wolves and the lowest ranking wolves are the omegas. The main phases of GWO are as follows [8]:

- (i) Tracking, chasing, and approaching the prey
- (ii) Pursuing, encircling, and harassing the prey until it stops moving
- (iii) Attack toward the prey.

The GWO algorithm was compared with nature-inspired evolutionary algorithm to have minimum side lobe level. Flower pollination optimization (FPA) is a nature-inspired algorithm used to solve multi-objective optimization problem. Authors in [9, 10] used FPA for linear array to obtain optimized antenna position in order to obtain minimum SLL and placement of deep null in the desired direction. Pollination in flower is the process of reproduce. The process in flower is the way to produce best-fit flowers. The rules of the algorithm is as follows [10]:

- (i) Movement of pollen carrying pollinators obeys Le'vy flights, and biotic and cross-pollination are considered as global pollination
- (ii) A biotic and self-pollination are used for local pollination.
- (iii) Flower constancy which is proportional to the similarity of two flowers can be considered as the reproduction probability
- (iv) Switch probability controls the local and global pollination. Physical proximity and factors, such as wind, bias the pollination activities toward local pollination.

Equally spaced and unequal spaced linear array were considered for FPA. The array obtain specified placement of null and minimum SLL as compared to PSO and CSO. Hierarchal particle swarm optimization (HPSO) algorithm was also applied to optimize SLL for nonuniform linear array considering the element spacing, excitation of amplitude, and phase variation of elements [11]. Particle swarm optimization used a group of particles to find the global optimum. The velocity of the particle depends on its relative position to own best location and swarm's best location. For HPSO, the velocity of the particle is updated to optimize the array. The linear array of 12 and 128 elements were considered for the application. The algorithm had a SLL of -26 db for $N = 12$. A modified form of PSO was used for optimization of linear array. PSO suffers premature convergence and trapping in the local minimum. A fuzzy logic controller was used to adjust the control parameters to adapt the PSO parameter to a new situation. The fuzzy system is capable of adjusting inertia weight and acceleration coefficient. The algorithm fuzzy particle swarm optimization (FPSO) was compared with GA and PSO for linear antenna array. Though the method exhibit better performance, it easily suffers from low convergence, weak local search ability, and partial optimism [12]. Authors in [13] suggested two algorithm ant lion optimization (ALO) and grasshopper optimization algorithm (GOA) for linear array. These methods are useful for minimizing the SLL by optimizing the excitation current amplitude of the elements. The ALO follow the behavior of the ant lion and GOA follow the behavior of the grasshopper swarm in nature. A linear array of 10 elements performs equally for both the algorithms. Sparse antenna array

has low mutual coupling, directive beam, and cost effective. The array is of high side lobe level due to different inter-element spacing. PSO was used for optimizing the element spacing and excitation amplitude to suppress the peak side lobe level. The change in spacing reduced the level to -22.07db for a 16 element linear array [14]. An improved fruit fly optimization algorithm (FOA) was proposed for synthesis of antenna array by adding a new search mechanism to enhance the efficiency of the algorithm for high dimensional problem. The average engine linear generation mechanism of candidate solution of FOA (AE-LMGS-FOA) had the choice of adding an alternative search engine and choosing controller parameters. The initial weight and search coefficient are considered as the constant parameter, and the weight coefficient is the tunable parameter. The LMGS-FOA was applied to linear and planar array for efficient performance [15]. In [16], social group optimization algorithm (SGOA) was used for synthesis of circular antenna array. The synthesis process involves both nonuniform amplitude and nonuniform spacing between the elements. SGOA mimics the social behavior of the human beings. Group solving capability is an effective way of solving problems as it combines the capability of each individual in the group. Every individual in the society is capable solving a problem and may be treated as a possible solution. A highly knowledgeable person in the group will have more impact and influence in the strategy. The circular array synthesis for 30 elements using the method had a SLL of -15.83 db. Improved chicken swarm optimization (ICSO) was analyzed for linear, circular, and random array for optimization of maximum side lobe level. The method introduced local search factor, weighting factor, and global search method into the update method of chicken swarm optimization (CSO). The convergence rate and stability of ICSO were best for each optimization case. The only drawback of the method is the number of parameters as compared to CSO [17].

The works described are summarized in Table 5.1.

5.5 Conclusion

This paper presents different optimization methods for reduction of side lobes in antenna array. Optimization methods are applicable to all forms of array to enhance the performance. The number of elements in the array, and the type of method determine the reduction of side lobe. The lower value of side lobe decreases the interference in antenna both in transmitter and receiver.

References

- Dib, N., Sharaqa, A.: On the optimal design of non-uniform circular antenna arrays. *J. Appl. Electromagn.* **14**(1) (2012)
- Goswami, B., Mandal, D.: A genetic algorithm for the level control of nulls and side lobes in linear antenna arrays. *J. King Saud Univ.-Comput. Inf. Sci.* **25**(2), 117–126 (2013)
- Beenamole, K.S., Sona, O.K., Ghanta, H., Devaraj, V., Meena, A., Singh, A.K.: Antenna array beam width widening with phase only optimization based on genetic algorithm
- Devi, G.G., Raju, G.S.N., Sridevi, P.V.: Application of genetic algorithm for reduction of sidelobes from thinned arrays. *Adv. Model. Anal. B* **58**(1), 35–52 (2015)
- Chenchuratnam, R., Rao, N.V.: Analysis of rectangular planar array with different distributions and optimization of sidelobe level using GA and PSO
- Kumari, K.K., Sridevi, D.P.: Pattern synthesis of non uniform amplitude equally spaced microstrip array antenna using GA, PSO and DE algorithms. *Int. J. Adv. Res. Eng. Technol.* **7**, 132–147 (2016)
- Kaur, K., Banga, V.K.: Synthesis of linear antenna array using firefly algorithm. *Int. J. Sci. Eng. Res.* **4**(8), 601–606 (2013)
- Saxena, P., Kothari, A.: Optimal pattern synthesis of linear antenna array using grey wolf optimization algorithm. *Int. J. Antennas Propag.* (2016)
- Saxena, P., Kothari, A.: Linear antenna array optimization using flower pollination algorithm. *Springerplus* **5**(1), 1–15 (2016)

Table 5.1 Different optimization methods

References	Optimization method	Antenna array	No of elements	SLL(dB)
[1]	BBO	Circular	12	−13.84
[2]	RGA	Linear	20	−15.5
[3]	GA	Phased		
[7]	FA	Linear	10	−26.72
[4]	GA	Thinned	200	−31.02
[5]	PSO	Planner	10X10	−59.11
[6]	DE	Nonuniform	20	−32.58
[8]	GWO	Linear		−23.42
[9]	FPA	Linear	20	−23.45
[11]	HPSO	Linear	12	−26
[12]	FPSO	Linear	16	−36.43
[13]	ALO GOA	Linear	10	−39.42 −30.09

(continued)

Table 5.1 (continued)

References	Optimization method	Antenna array	No of elements	SLL(dB)
[14]	PSO	Linear	16	−22.07
[15]	AE-LMGS-FOA	Linear		−23.1
[17]	ICSO	Linear		−30.15

10. Singh, U., Salgotra, R.: Synthesis of linear antenna array using flower pollination algorithm. *Neural Comput. Appl.* **29**(2), 435–445 (2018)
11. Ghosh, J., Poonia, A.K., Varma, R.: Multi-objective hierarchical particle swarm optimization of linear antenna array with low side lobe and beam-width. *Int. J. Appl. Eng. Res.* **12**(8), 1628–1632 (2017)
12. Mohamed, B., Boufeldja, K.: Optimal complex weights antenna array with efficient fuzzy particle swarm optimization algorithm. *J. Theor. Appl. Inf. Technol.* **95**(19) (2017)
13. Amaireh, A.A., Alzoubi, A., Dib, N.I.: Design of linear antenna arrays using antlion and grasshopper optimization algorithms. In: 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) (pp. 1–6). IEEE (2017)
14. Ullah, N., Huiling, Z., Rahim, T., ur Rahman, S., MuhammadKamal, M.: Reduced side lobe level of sparse linear antenna array by optimized spacing and excitation amplitude using particle swarm optimization. In: 2017 7th IEEE International Symposium on Microwave, Antenna, Propagation, and EMC Technologies (MAPE) (pp. 96–99). IEEE (2017)
15. Darvish, A., Ebrahimzadeh, A.: Improved fruit-fly optimization algorithm and its applications in antenna arrays synthesis. *IEEE Trans. Antennas Propag.* **66**(4), 1756–1766 (2018)
16. Chakravarthy, V.S., Chowdary, P.S.R., Satpathy, S.C., Terlapu, S.K., Anguera, J.: Antenna array synthesis using social group optimization. In: Microelectronics, Electromagnetics and Telecommunications (pp. 895–905). Springer, Singapore (2018)
17. Liang, S., Fang, Z., Sun, G., Liu, Y., Qu, G., Zhang, Y.: Sidelobe reductions of antenna arrays via an improved chicken swarm optimization approach. *IEEE Access* **8**, 37664–37683 (2020)

Chapter 6

Accuracy Analysis for Predicting Heart Attacks Based on Various Machine Learning Algorithms



Rashmita khilar, T. Subetha, and Mihir Narayan Mohanty

Abstract Heart disease is a major concern in today's world. It is increasing day by day. Diagnosis of this disease is a difficult task and it has to be predicted earlier in beforehand it causes any damage to other organs. This paper finds out a best classification algorithm that can effectively predict the occurrence of heart disease. In this research paper, a comparative study is made with various machine learning algorithms like K-means, SVM, XGBoost, K-nearest, descision tree, random forest, etc., to find out the method which is most effective and accurate to predict the occurrence of heart attack. The dataset is split into two groups such as group 1 consists of 80% training sets and group 2 contains 20% of testing data. Various classifiers are also compared and analyzed, and the results are denoted in a graph. The system is evaluated with various models by computing r² squared error and negative mean squared error tests. From the above said algorithms, GaussianNB produces a better accuracy results in testing than others for predicting heart disease.

6.1 Introduction

Machine learning finds its applications in various fields. One such case is health care industry. ML finds its most applications in health care system thereby helps doctors to solve out many critical conditions of a patient. Nowadays, heart disease is increasing day by day throughout the world. Sometimes, this deadly disease takes away the life of people without giving any signs [1]. Machine learning concepts are implemented to predict the deadly disease. Various attributes are taken from various patients and by reading those values doctors predict the occurrence of heart attack. The objective

R. khilar (✉)

Department of IT, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India

T. Subetha

Department of IT, BVRIT Hyderabad College of Engineering for Women, Hyderabad, India

M. N. Mohanty

ITER, Siksha 'O' Anusandhan(Deemed to be University), Bhubaneswar, Odisha, India

is to find out whether the patient is mostly likely to be diagnosed with any of the deadly cardiovascular heart diseases based on their various medical attributes like age, sex, type of chest pain, sugar leve both in before and after food, colesterol [2], etc.

The above attributes from various patients are taken and given as input to predict the percentage of occurrence of the disease. Various machine learning algorithms, like linear regression, K-nearest, random forest, SVM, descision tree, KNN classifiers, etc., are utilized to classify and predict heart attack [3]. Various researchers have implemented different algorithms to predict the the heart attack. Each of them have found their efficiency rate for prediction [4]. Though various researchers have implemented various algorithms and analyze various parameters and attributes, still accurate prediction is lacking in this field. This system uses 14 different attributes and predicts the maximum likelihood for the occurrence of heart disease using various classifiers [5]. It enables significant knowledge between the medical parameters for the prediction. The datasets from UCI repository are taken and fed to the above said classifiers. Those 14 attributes are taken and classified with the above said machine learning algorithms help to forecast to get heart attack or not [6].

Most of the algorithms are efficient and produce more than 80% of accuracy. In this research article, the efficiency and effectiveness of the above said algorithms are compared to attain the top algorithm for the prediction of heart attack [7, 22]. After various iterations and classification, it is obtained that GaussianNB algorithm provides a better prediction results.

6.2 Literature Survey

Many works related to heart disease prediction is carried out. An efficient cardiovascular heart disease prediction is carried out, and the efficiency of algorithms like linear regression, KNN, and SVM classifiers is calculated and the results are compared [8, 9].

However, it becomes difficult to identify the patients with high risk rate because of the multi-factorial nature of various dangerous risk factors like diabetes, high blood pressure, high cholesterol, etc. [10]. In this situation, machine learning and data mining concepts come to the rescuepoint. The development of this new technology helps the doctors to predict heart attack accurately [11]. Doctors and various researchers have turned ML techniques as a tools. ML along with the concepts like pattern recognition and classification approaches help researchers for accurate prediction [8, 12]. Genetic algorithms outperform total search in intricate datasets and provide a better accurate method for predictiong heart attack.

6.3 Dataset Description

The heart attack disease analysis and prediction dataset are taken from Kaggle [13] is a UCI Cleveland dataset which is a.csv file with 304 records of data. The dataset contains 14 columns, out of which the last column is the class label. The remaining 13 columns contain various features which are essential in determining the risk of a heart attack. The dataset has 165 records of heart disease patients and 138 records without heart disease. The dataset does not have any null values, and so it is a clean database and the description is given in Table 6.1.

Table 6.1 Dataset description

S. No	Attribute name	Range levels
1	Age	In years
2	Sex	Male = 1, Female = 2
3	Chest pain type (cp)	0-Typical angina 1-Atypical angina 2-Non-anginal pain 3-Asymptomatic
4	Resting BloodPressure(trestbps)	Measured in mm-Hg 120 < -Normal > 130–140-Abnormal
5	Cholestrol(chol)	Measured in mg/dl 200
6	Fasting blood sugar(fbs)	> 120 mg/dl 1 = true, 0 = false
7	Restecg—resting electrocardiographic results	0: Nothing to note 1: ST-T Wave abnormality can range from mild symptoms to severe problems 2: Possible or definite left ventricular hypertrophy
8	Thalach	140,173
9	Exang—exercise-induced angina	1 = yes, 0 = no
10	Oldpeak	Numeric Value
11	Slope	0: better heart rate 1: Flat sloping 2: Down sloping
12	Ca—colored vessel	Numeric Value
13	Thal—thallium stress result	1–3—Normal 6-defect
14	Target	1-Yes, 0-No

6.4 Methodology

6.4.1 Data Preprocessing

The real-world data is riotous and has to be preprocessed before constructing the machine learning model for accurate results. Data preprocessing is performed to pull out the riotous and soiled data. The UCI Cleveland dataset does not have any records with null values. So it is a clean dataset. Data transformation is then accomplished to translate the data from one pattern to another for uncomplicated processing.

6.4.2 Feature Extraction

The UCI Cleveland dataset has a total of 14 attributes with one class label. The features are made understandable by plotting a histogram with the features. The sample histogram taken for the first three features is shown in Fig. 6.1. From the exploratory data analysis, it is observed that the patients whose chest pains lie in the category of 1, 2, 3 are most likely possess heart disease. The range attributes like patients who have the value 1 in resting electrocardiographic, value 0 in exercise-induced angina, value 2 in peak slope, value 0 in the total of significant vessels dyed by fluoroscopy, and value 2 in thallium stress results have a significant impact on developing heart disease. The remaining value attributes like value above 140 in blood pressure, value above 200 in cholesterol, and value above 140 in the maximal heart rate attained to significantly impact developing heart disease.

The correlation matrix is built for all the features compared to the target label to identify the prominent features. The developed correlation matrix and the bar graph to plot the correlation matrix compared to the target class label are given in Figs. 6.2 and 6.3.

6.4.3 Machine Learning Algorithms

The heart disease forecasting for a patient is found by building a proper classification algorithm. Various machine learning supervised algorithms are used and associated to find the top suitable algorithm. The dataset is divided into 0.8 training and 0.2 testing. The algorithms chosen for comparison are logistic regression, KNN classifier, decision tree, XGBoost classifier, and random forest. The accuracy of the system for various classifiers is also compared and analyzed. The hyperparameters are also tuned for better performances.

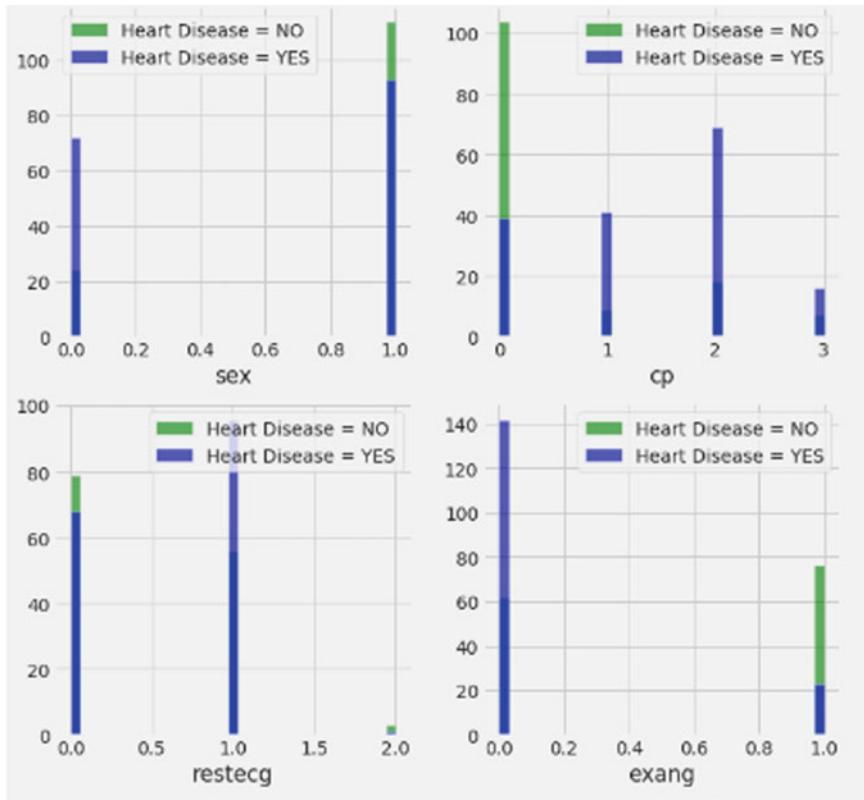


Fig. 6.1 Histogram plotted for sex, cp, restecg, and exang features

6.4.3.1 Logistic Regression

Logistic regression [14] algorithm inspects the amalgamation that exists between one independent variable and diploid dependent variables. The difference between the logistic regression and linear regression is the continuous dependent variable. The testing and training accuracy obtained for heart disease prediction using the logistic regression algorithm are 86.79% and 83.81%.

6.4.3.2 K-Nearest Neighbor

KNN classifier [15] performs classification by establishing the closest neighbors and utilizes this distance calculation to predict a new class label. The system's main drawback is the computational complexity that current computational-powered systems can overcome. The testing and training accuracy obtained for heart disease prediction using KNN classifier algorithm are also 86.79% and 83.81%.

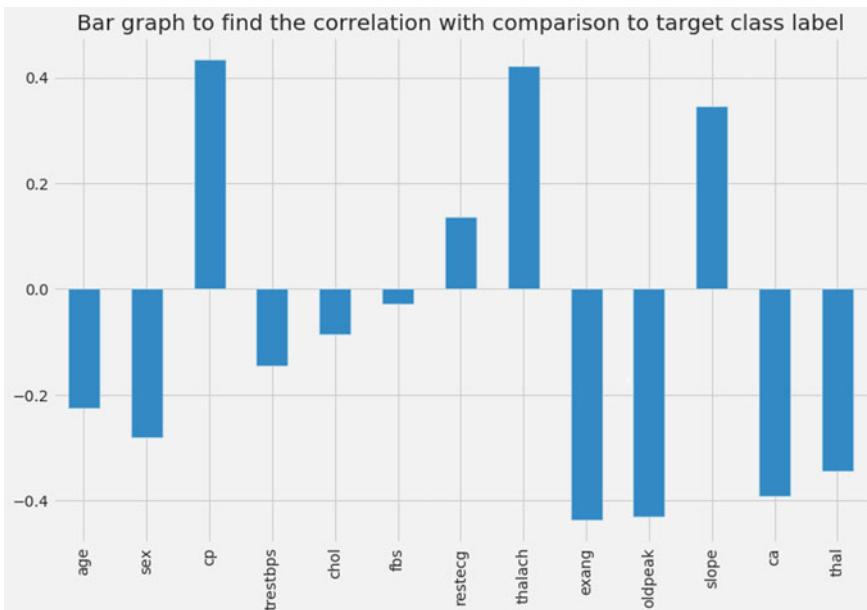


Fig. 6.2 Correlation graph

6.4.3.3 Support Vector Machine

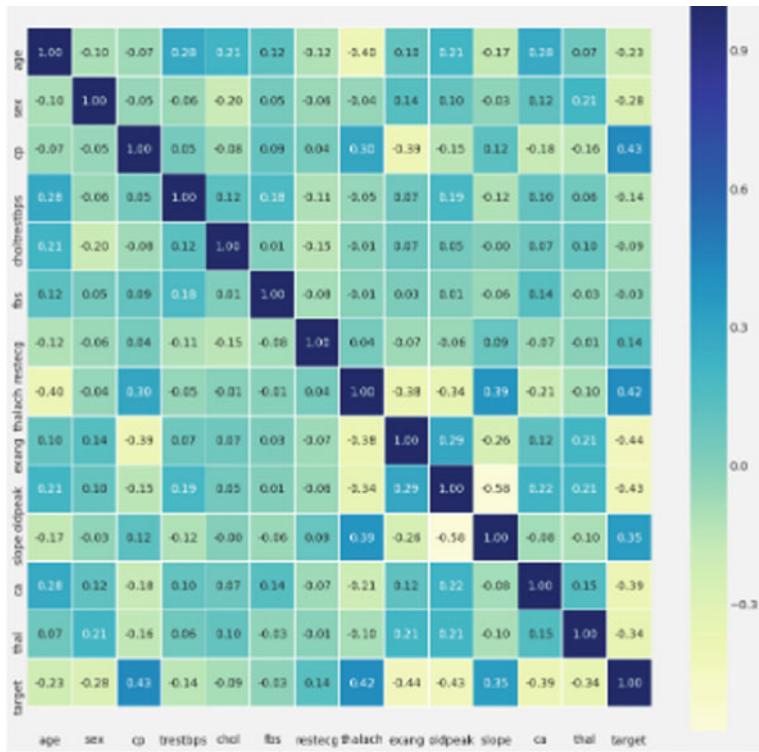
Support vector machine [16] classifies by constructing a hyperplane. If there are more than two classes, then it will perform multi-class SVM. Since our problem is two classes, a single SVM is enough to classify the data points into either patients having heart disease or not. The testing and training accuracy obtained are 93.40% and 83.91%.

6.4.3.4 Decision Tree

It builds a tree-like classifier to traverse the tree to reach the final leaf outcomes. The class label will be present in the leaf outcomes. The testing and training accuracy obtained for heart disease prediction using the decision tree algorithm are 100% and 78.02%.

6.4.3.5 Random Forest

A random forest classifier [17] is an ensemble classifier that constructs a weak classifier's combination to produce final classification results. The random forest

**Fig. 6.3** Correlation matrix

algorithm's testing and training accuracy for heart disease prediction are 100% and 82.42%.

6.4.3.6 XGBoost Classifier

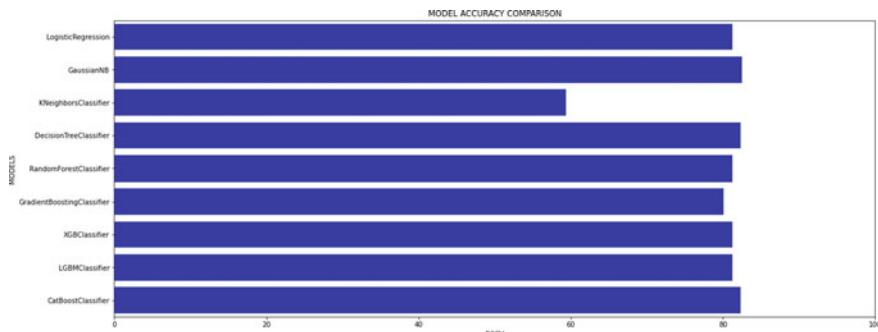
XGBoost classifier [18] is also an ensemble classifier that votes all the classifier decisions and finally predicts the output. The advantage of using the XGBoost classifier is the advantage of utilizing the results of one or more classifiers. The XGBoost classifier algorithm's testing and training accuracy for heart disease prediction are 98.58% and 83.52%.

6.4.4 Comparison Chart

In addition to the above classifier, nine other classifier models are also constructed and evaluated using a cross-validation score to find the optimal classifier. The system

Table 6.2 Comparison table

Name of the classifier/regression model	R ² error	Negative mean square error
Linear regression	0.327	0.377
PLS regression	0.322	0.376
Ridge	0.336	0.375
Lasso	0.045	0.440
Elastic net	0.075	0.435
KNN	-0.245	0.499
Decision tree	-0.047	0.481
Bagging	0.241	0.399
Random forest	0.277	0.392
Gradient boosting	0.244	0.400
XGBoost	0.106	0.438
LGBM	0.359	0.369
Cat boost	0.3100	0.381

**Fig. 6.4** Accuracy comparison chart

is evaluated with various models by computing r² squared error and negative mean squared error tests. The table constructed with values is shown in Table 6.2. The comparison chart for the constructed models is shown in Fig. 6.4.

The comparative table shows that GaussianNB produces better accuracy results in testing than others in predicting heart disease.

6.5 Conclusion

This paper compares the efficiency effectiveness of machine learning algorithms for heart disease prediction. The system is evaluated with various models by computing

r² squared error and negative mean squared error tests. The UCI Cleveland dataset is used to test the results by taking 14 attributes. By taking various iterations and from the above graph, it analyzed that the GaussianNB algorithm predicts a better way for predicting heart disease than any other machine learning algorithms. Though a variety of efficient algorithms are there to predict heart disease, the department of health care system requires the most efficient algorithms to predict the heart disease by taking as many possible attributes.

References

1. Jabbar, M.A., Chandraand, P., Deekshatulu, B.L.: Heart disease prediction system using associative classification and genetic algorithm. In: International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-(ICECIT)
2. Jabbar, M.A., Chandra, P., Deekshatulu, B.L.: Prediction of risk score for heart disease using associative classification and hybrid feature subset selection. In: 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 628–634. Kochi, India (2012). <https://doi.org/10.1109/ISDA.2012.6416610>
3. Normawati, D., Winarti, S.: Feature selection with combination classifier use rules-based data mining for diagnosis of coronary heart disease. In: 2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA), pp. 1–6. Yogyakarta, Indonesia (2018). <https://doi.org/10.1109/TSSA.2018.8708849>
4. Cheng, C., Chen, J.: Reinforced rough set theory based on modified MEPA in classifying cardiovascular disease. In: Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), pp. 178–182. Haikou, China (2007). <https://doi.org/10.1109/FSKD.2007.467>
5. Mohapatra, S.K., Palo, H.K., Mohanty, M.N.: Detection of arrhythmia using neural network. In: ICITKM, pp. 97–100. (2017)
6. Jabbar, M.A., Samreen, S.: Heart disease prediction system based on hidden naïve bayes classifier. In: 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), pp. 1–5. Bangalore, India (2016). <https://doi.org/10.1109/CIMCA.2016.8053261>
7. Tao, R., et al.: Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods. IEEE Trans. Biomed. Eng. **66**(6), 1658–1667 (2019). <https://doi.org/10.1109/TBME.2018.2877649>
8. Zulvia, F.E., Kuo, R.J., Roflin, E.: An initial screening method for tuberculosis diseases using a multi-objective gradient evolution-based support vector machine and C5.0 decision tree. In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), pp. 204–209. Turin, Italy, (2017). <https://doi.org/10.1109/COMPSAC.2017.57>
9. Mehta, D.B., Varnagar, N.C.: Newfangled approach for early detection and prevention of ischemic heart disease using data mining. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1158–1162. Tirunelveli, India (2019). <https://doi.org/10.1109/ICOEI.2019.8862544>
10. Gavhane, G.K., Pandya, I., Devadkar, K.: Prediction of heart disease using machine learning. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1275–1278. Coimbatore, India (2018). <https://doi.org/10.1109/ICECA.2018.8474922>
11. Singh, A., Kumar, R.: Heart disease prediction using machine learning algorithms. In: 2020 International Conference on Electrical and Electronics Engineering (ICE3), pp. 452–457. Gorakhpur, India (2020). <https://doi.org/10.1109/ICE348803.2020.9122958>
12. Motarwar, P., Duraphe, A., Suganya, G., Premalatha, M.: Cognitive approach for heart disease prediction using machine learning. In: 2020 International Conference on Emerging Trends in

- Information Technology and Engineering (ic-ETITE), pp. 1–5. Vellore, India (2020). <https://doi.org/10.1109/ic-ETITE47903.2020.9242>
- 13. Mohapatra, S.K., Mohanty, M.N.: ECG analysis: A brief review. *Recent Adv. Comput. Sci. Commun. (Formerly: Recent Patents Comput. Sci.)* **14**(2), 344–359
 - 14. Cunningham, P., Delany, S.J.: k-nearest neighbour classifiers. arXiv preprint [arXiv:2004.04523](https://arxiv.org/abs/2004.04523) (2020)
 - 15. Noble, W.S.: What is a support vector machine? *Nat. Biotechnol.* **24**(12), 1565–1567 (2006)
 - 16. Song, Y.Y., Ying, L.U.: Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**(2), 130 (2015)
 - 17. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H.: Xgboost: extreme gradient boosting. R Package Version 0.4–2, 1(4) (2015)
 - 18. Mohapatra, S.K., Behera, S., Mohanty, M.N.: A comparative analysis of cardiac data classification using support vector machine with various kernels. In: 2020 International Conference on Communication and Signal Processing (ICCSP), pp. 515–519. IEEE (2020)
 - 19. Belgiu, M., Drăguț, L.: Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote. Sens.* **114**, 24–31 (2016)
 - 20. Lei, W.K., Li, B.N., Dong, M.C., Fu, B.B.: An application of morphological feature extraction and support vector machines in computerized ECG interpretation. In: 2007 Sixth Mexican International Conference on Artificial Intelligence, Special Session (MICAI), pp. 82–90. Aquascalientes, Mexico (2007). <https://doi.org/10.1109/MICAI.2007.32>

Chapter 7

Link Recommendation for Social Influence Maximization



Sagar S. De, Parimal Kumar Giri, and Satchidananda Dehuri

Abstract In social network analysis, the influence maximization problem recognizes a small set of seed nodes that effectively maximizes the aggregated influence under a cascading propagation model. The approach has vast implications in viral marketing, government decision making, epidemic control, and many more. In the last decades, researchers developed many strategies that effectively identify the seed nodes. Although, the seed identification process is an NP-hard problem. Sometimes, due to the network structure, the best-known seeds are incapable of propagating influence throughout the network. We observed that a tiny structural modification by adding a few links sometimes increases the aggregated influence notably. From the literature, we have observed that no prior work exists in this regard. In this paper, first, we have applied multi-objective optimization for identifying initial seeds considering aggregated influence and timestep to achieve the coverage as objectives. Then for suitable non-dominated seeds, the proposed approach computed the minimum number of required missing links against every uninfluenced node to sustain the propagation. Finally, a set of statistical techniques and locally globally tuned biogeography-based optimization are used to identify end vertices for the links recommendation. The recommended links connect non-influenced components to the influenced component and allow further influence propagation.

S. S. De (✉)

S.N. Bose National Centre for Basic Sciences, Block-JD, Sector-III, Salt Lake City, Kolkata, West Bengal 700106, India

P. K. Giri

Department of Computer Science, College of Engineering Bhubaneswar, Patia, Bhubaneswar, Odisha 751024, India

S. Dehuri

Department of Information and Communication Technology, Fakir Mohan University, Vyasa Vihar, Balasore, Odisha 756020, India

7.1 Introduction

Online social networks (OSNs) have grown in popularity over the previous few decades. Interactions between people within the OSNs grow dramatically. OSNs have now established themselves as the de facto means of communication and expression. In a social network system, information propagation follows cascading phenomena aka word-of-mouth analogy [1]. The cascading phenomena imply a sequence of changes of behavior that function as a chain reaction. The cascading aspect of propagation has proven to be extremely successful in spreading messages, views, influence, and ideas rapidly among people. At the same time, the process is also very cost effective. Online social network sites such as Instagram, Twitter, and Facebook have emerged as natural choices chosen by enterprises for product promotion, marketing, camping, and opinion surveys. The diffusion phenomenon in social networks and its vast implementation in real-world applications [2–4] has attracted researchers from different disciplines such as sociology, computer engineering, business administration, physics, and many more. In the last decades, social influence analysis and network modeling have become tremendously popular among the research and business communities.

In general, one of the most difficult problems in the study of complex network analysis is *social influence analysis*. Social influence analysis is crucial for comprehending the spread of behavior. The influence analysis study considers mainly two aspects—(1) How the propagation occurs within a network? (2) Which set of nodes (known as seeds) should be rewarded so that the cascading propagation can be effective? Therefore, an associated problem is how to choose seed nodes that maximize aggregated influence within the network. The *influence maximization (IM)* problem is the name of the problem. The influence maximization problem recognizes a small subset (usually a predetermined size of k) of seed nodes that effectively maximizes the aggregated influence under a cascading propagation model. Several propagation models are available in the IM literature. The *independent cascade model (IC model)* and the *linear threshold propagation model (LT model)* are the most employed models. Network propagation can be progressive or non-progressive. In the progressive propagation model, a previously activated node will not change its state in future timesteps. Most real-life IM problems come with constraints such as a given budget and time to propagate influence. However, for the simplicity of the presentation, in this article, we have considered the progressive linear threshold propagation model and timestep to maximize propagated influence.

Although, this proposed strategy is adequate for any set of constraints and any propagation model. A constrained IM problem is a multi-objective optimization problem. In [5], De and Dehury proposed *non-dominated sorting biogeography-based optimization (NSBBO)* as a multi-objective optimizer. The NSBBO successfully solved the multi-objective influence maximization-cost minimization (IM-CM) problem. In this article, the optimization objectives are (1) maximize aggregated influence and (2) minimize propagation time. *Non-dominated sorting genetic algorithm II (NSGA-II)* [6] is applied for the identification of a non-dominated set of seed nodes.

Moreover, the strategy identifies several non-dominated seed sets. Without any preference of objective, all non-dominated seeds are equally optimum. Nevertheless, sometimes due to the network structure and employed propagation model $\sigma_{G,M}$, the best-known seeds are incapable of propagating influence throughout the network. The propagation remains restricted within a fraction. We observed that a tiny structural modification by adding a few links sometimes increases the aggregated influence notably. To our knowledge, no work was carried out toward influence maximization by structural modification. After obtaining the potential seed sets, we have distinguished the uninfluenced portion of the network for a chosen seed set and then computed the minimum number of required missing links against every uninfluenced node to sustain the propagation. We have established Eq. 7.5 and used it to determine the number of additional links required for every non-influenced node to make them active.

Using above information, non-influenced nodes are ranked. *Locally globally tuned biogeography-based optimization (LGBBO)* [7, 8] is used to represent the minimum number of nodes that are efficient to propagate influence within the non-influenced portion of the network. The identified nodes are the one end of the missing links. Lastly, with the help of few statistical calculations, our approach recognizes another end node of these links. We establish Eq. 7.5 to examine the feasibility of linking a non-influenced node to an already activated node. When the required number of missing links is less than or equal to the number of seeds, there is no need to execute the last step. Preferably, the link ends can be directly connected to the seed nodes. The recommended links connect non-influenced components with the influenced portion of the network and allow further influence propagation. The goodness of the two-stage algorithm is that rather than arbitrarily adding missing links throughout the network, we have gradually identified the non-influenced portion of the network (a small segment of the network) and tried to activate them. Thus, the two-stage strategy significantly improved computation time. Extensive empirical investigations on a variety of benchmarked social networks reveal that the proposed method is effective at discovering seed nodes and recommending missing linkages in order to enhance influence propagation.

The arrangement of the paper is follows as. Sections 7.2 and 7.3 present the background and literature of the study, respectively. Section 7.3.1 highlights the current research gap. Sections 7.4 and 7.5 confer the proposed approach toward link recommendation for influence maximization and respective empirical studies. Section 7.6 concludes the study.

7.2 Backgrounds

The ideas of social influence, influence propagation, and the problem of influence maximization are briefly reviewed in this section, followed by a discussion of multi-objective optimization.

7.2.1 Influence Propagation, Social Influence, Influence Maximization Problem

Social influence, according to Rashotte, is the change in an individual's thoughts, feelings, attitudes, or behaviors as a result of interaction with other people or groups [9]. Social influence occurs when people intentionally or unintentionally impact one's thoughts, feelings, or actions. Social influence can be viewed throughout OSNs in various forms. The social influence is evaluated in a variety of applications when it comes to data mining and big data analysis, e.g., political campaigning, diffusion modeling, viral marketing, and recommender systems. Ryan–Gross's [10] studied the delayed adaptation after the first exposer. In Fig. 7.1, certain years, they have shown the percentage of operators who have had their first hearing and the percentage of operators who accept hybrid seed.

In 2003, at first Kempe et al. [11] proposed the influence maximization problem in the way of algorithm. They have studied a social network which represents a graph $G = (V, E)$, where V is the set of nodes in the graph G (i.e., users), E is the set of (directed/undirected) edges in G (i.e., social links between users), and W represents a matrix of edge weight. Each members $w_{i,j}$ in W denotes the weight of the edge between the node i and j , which represents the probability of node i activates node j . The objective of the IM problem is to determine the initial set of users of size k that have maximum influence in graph G under a given propagation model M .

The influence of any seed set is determined by the process of information dissemination among users. Viral marketing is an example of information diffusion, in which a corporation may want to spread the acceptance of a new product from a few

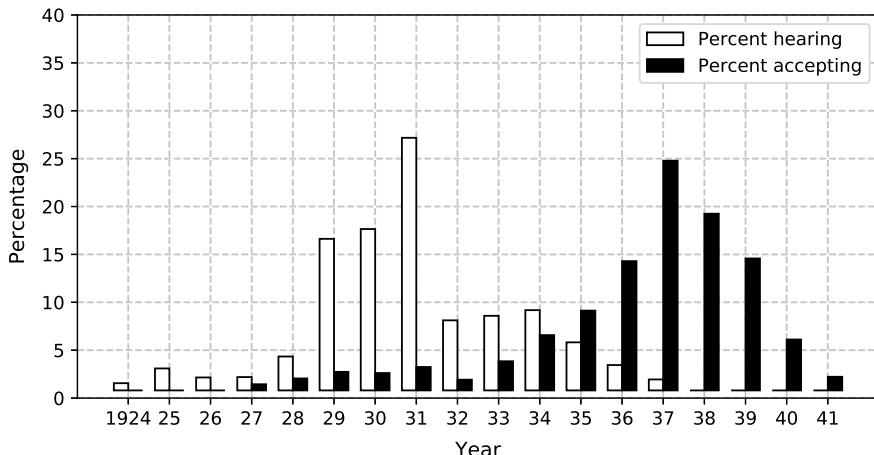


Fig. 7.1 Ryan–Gross study of hybrid seed corn delayed adaptation (the figure is reproduced from the Ryan–Gross study [10])

early adopters through social linkages among consumers. We formally construct the diffusion model and the influence spread under the model to quantify information diffusion.

Definition 1 (*Diffusion Model and Influence Spread*) Consider a graph $G = (V, E, W)$ be a social graph, a user set $S \subseteq V$, and a diffusion model M represent a stochastic process S for spreading information on G . The influence spread of S , viz. $\sigma_{G,M}(S)$, is the expected number of users influenced by S (e.g., users who adopt the new product in viral marketing), provided $\sigma_{G,M}(\cdot)$ be a set of non-negative function such as $\sigma_{G,M} : 2V \rightarrow R \geq 0$ on any subset of users.

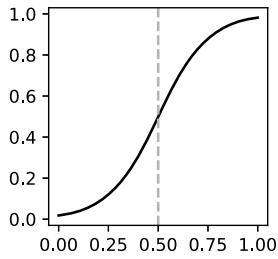
Diffusion models have been developed in recent years to formulate the diffusion process and compute the influence spread [12–15]. We focus on progressive diffusion models in this paper, which means that activated nodes cannot be deactivated in the following steps. Currently, the economics and sociology communities have developed a range of diffusion models. Below listing named as few diffusion models:

- Independent cascade model (IC model)
- Linear threshold model (LT model)
- Shortest path model (SP model)
- Majority threshold model (MT model)
- Constant threshold model (CT model)
- Unanimous threshold model (UT model).

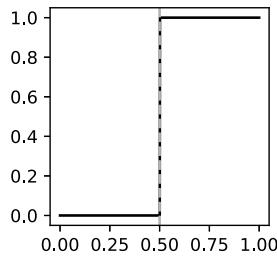
Although various diffusion models have been proposed in the literature, the independent cascade (IC) and linear threshold (LT) models are the most commonly utilized in the research of social influence problems. The primary feature of the independent cascade model is the independence of the influence dispersed across each edge of the entire graph; in other words, one person is enough to activate another.

This model is appropriate for modeling the diffusion of ideas, information, viruses, etc. Every edge $(u, v) \in E$ has an associated influence probability $p_{uv} \in [0, 1]$, that is, the probability that node u succeeds in its attempt to activate its neighbor node v . The threshold model can be used to describe the adoption of a new, untested product. People are more inclined to adopt a new product if the total of their friends' recommendations reaches a certain threshold value. In general, the case is when an industry adopts a new standard. While in the IC model, every edge (u, v) was associated with an influence probability p_{uv} ; the same edges are this time assigned with influence weight $w_{uv} \in [0, 1]$ in the linear threshold model. This represents the degree to which a user u can influence another user v . Therefore, the weights are normalized so that the sum of all incoming edges of each node is lower or equal to 1, i.e., $u \in \mathcal{N}_{in}(v) \quad W_{uv} \leq 1$. Figure 7.2 presents the concept of influence response for the independent cascade and linear threshold models.

Figure 7.3 presents the cascading propagation analogy of the linear threshold model. With a threshold value of 0.40, the figure shows the propagated influence in every timestep. Double-lined nodes (10, 11) are the seeds of the propagation. The blue-colored nodes are the newly influenced nodes at the respective timestep.

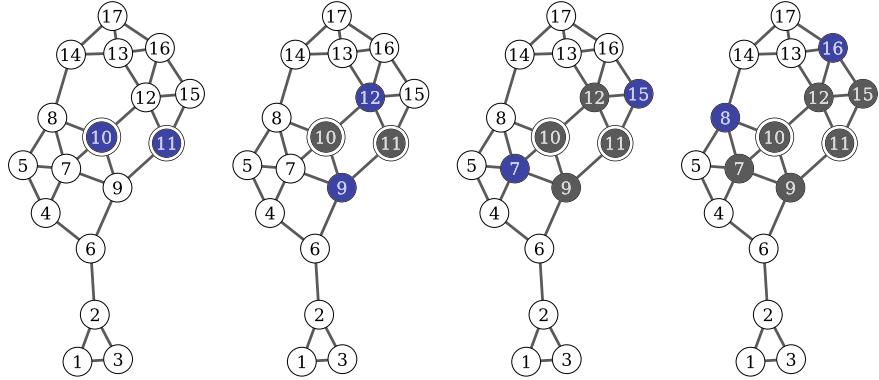


(a) Response of an independent cascade model.



(b) Linear threshold model with a threshold value of 0.50.

Fig. 7.2 Influence propagation response

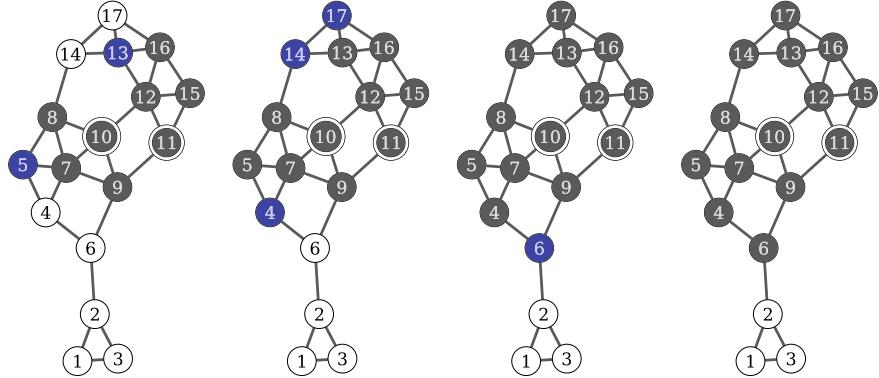


(a) Timestep-0

(b) Timestep-1

(c) Timestep-2

(d) Timestep-3



(e) Timestep-4

(f) Timestep-5

(g) Timestep-6

(h) Final

Fig. 7.3 Linear threshold propagation

Other propagation models, such as time aware, context aware, and propagation with negative feedback, are also present in the literature in addition to those two well-known models [15, 16]. The influence maximization problem is described as follows, based on the formalization of the influence spread:

Definition 2 (*Influence Maximization (IM)*) Consider a graph G be social graph, a diffusion model M with positive integer k selects a subset S of k users as IM from V to maximize the influence spread $\sigma_{G,M}(S)$ based on seed set under the diffusion model M .

The influence function $\sigma_{G,M}(\cdot)$ appears to be strongly dependent on the diffusion process.

7.2.2 Overview of Multi-objective Optimization

A multi-objective optimization (MOO) considers multiple objectives to be optimized simultaneously. As a result, the objective function $f(x)$ is a vector of optimization functions with k dimensions. Mathematically, the problem can be expressed as follows:

$$\begin{aligned} \text{Optimize}_{x} \quad & \mathbf{f}(x) = [f_1(x), f_2(x), \dots, f_k(x)]^T \\ \text{Subject to} \quad & g_p(x) \leq 0, \quad \forall p = 1, 2, \dots, n \\ & h_q(x) = 0, \quad \forall q = 1, 2, \dots, m \end{aligned} \tag{7.1}$$

where $g(x)$ and $h(x)$ are n and m number of inequality and equality constraints. There is no one solution that concurrently optimizes all objectives in a nontrivial multi-objective optimization problem (MOP). It is crucial to separate the non-dominated or Pareto optimum solutions from the competing options. In MOO, a solution x_1 is said to dominate another solution x_2 , if both conditions 1 and 2 are true:

1. In all objectives, the solution x_1 is no worse than x_2 , or $f_i(x_1) \clubsuit f_i(x_2)$ for all $i = 1, 2, \dots, k$.
2. Similarly, the solution x_1 is better than x_2 in at least one objective, or $f_i(x_1) \spadesuit f_i(x_2)$ for all $i = 1, 2, \dots, k$.

The operators \clubsuit and \spadesuit are defined as follows:

$$\clubsuit = \begin{cases} \leq, & \text{Minimization objective} \\ \geq, & \text{Maximization objective} \end{cases}$$

$$\spadesuit = \begin{cases} <, & \text{Minimization objective} \\ >, & \text{Maximization objective} \end{cases}$$

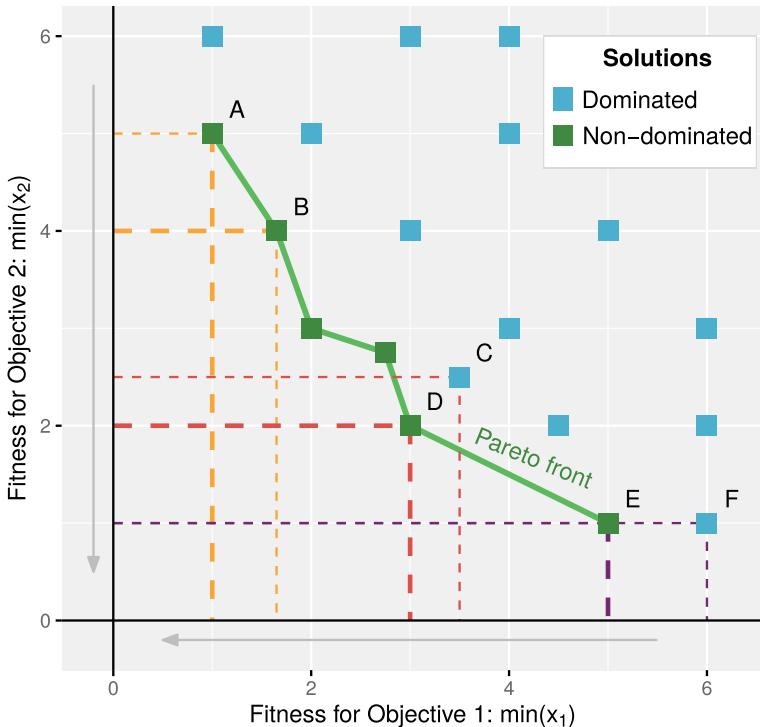


Fig. 7.4 Non-dominated solutions for a bi-objective minimization problem

All Pareto optimum solutions are considered equally desirable without any additional subjective preference. The concept of ‘optimum’ in MOPs can be understood as a ‘trade-offs’ of non-dominated solutions because they have more than one objective function.

Figure 7.4 shows the basic idea of Pareto front and non-dominated solutions. Here, the search space $\Omega \in \mathbb{R}^2$; $\mathbf{x} = [x_1, x_2]^T$ such that $x_1, x_2 \in [0, 6]$. The two-objective functions are $f_1(\mathbf{x}) = \min(x_1)$ and $f_2(\mathbf{x}) = \min(x_2)$. A possible solution is represented by a square. The non-dominated or Pareto optimal solutions are represented by the green squares. The Pareto front is the green line that connects Pareto optimal solutions. The cost for each objective function is indicated by dotted lines. The superiority of the objective is indicated by a bold-dotted line. They are non-dominated with regard to each other since solution A dominates solution B in f_1 and solution B dominates solution A in f_2 . In both objectives, solution D outperforms solution C. Because both solutions are identical in f_2 , but solution E dominates solution F in f_1 , and solution E dominates solution F.

Many multi-objective optimizers have been proposed in the last decade, including the multi-objective genetic algorithm (MOGA) [17], the strength Pareto evolutionary algorithm 2 (SPEA2) [18], the non-dominated sorting genetic algorithm II (NSGA-II) [19],

II) [19], and the multi-objective evolutionary algorithm based on decomposition (MOEA/D) [20]. Coello Coello [21] and Zitzler et al. [22] provide in-depth studies on multi-objective optimization. NSGA-II is one of the most widely used multi-objective optimizers among them.

7.3 Related Works and Research Gap

Domingos and Richardson [23, 24] were the first authors to introduce influence maximization as an algorithmic problem. Their methods were probabilistic. But, Kempe et al. [11] have developed first to formulate the problem as discrete optimization. They have used hill-climbing greedy approximation algorithm which guarantees the influence spread within $(1 - 1/e)$ of the optimal influence. In their approach, they have also run Monte Carlo simulations model significantly in large quantities to obtain an accurate estimate of the influence spread. Moreover, the algorithm obtained an excellent performance. Kimura and Saito [14] have proposed shortest path-based influence cascade models and provide efficient algorithms of computer influence spread. Leskovec et al. [25] presented ‘cost-effective lazy forward’ (CELF) model. They have used the greedy optimization strategy and submodularity property of the influence maximization to reduce the number of evolution. Gong et al. [26] have extended the work of Leskovec et al. and improved the CELF approach. Chen et al. [27] proposed NewGreedy and MixedGreedy algorithms for the IC model with uniform probabilities. However, their performance is non-steady and sometimes even worse than CELF. Zhou et al. in [28, 29] have enhanced the CELF model using upper bound-based lazy forward (UBLF).

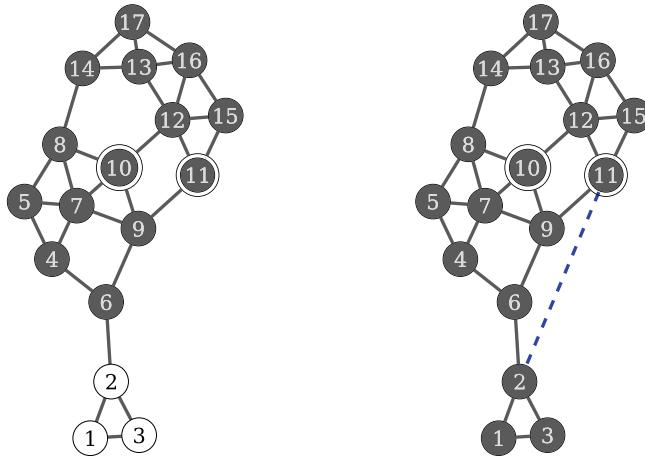
In this method, the Monte Carlo calls in the first round are drastically compared with the CELF. In general, the maximum influence paths (MIP) between every pair of nodes in the network via a Dijkstra shortest path algorithm ignore MIPs with probability smaller than an influence threshold θ , effectively restricting influence to a local region. Wang et al. [30] have proposed a novel community-based greedy algorithm for mining top- k influential nodes. Barbieri et al. [31] studied social influence for modeling perspective. Guo et al. [32] investigated the influence maximization problem for item-based data. Goyal et al. [33] have proposed an alternative approach to influence maximization instead of assuming influence probabilities as input, which directly uses the past available data. In [34], authors discussed the integral influence maximization problem in case of repeated activations which are involved. The complementary problem of learning influences probabilities from the available data is studied in the works [35, 36].

Finally, Yang and Liu [37] have used multi-objective discrete particle swarm optimization algorithm to identify influential seeds with minimum cost worked on the IM-CM problem. De and Dehuri [5] proposed non-dominated sorting biogeography-based optimization (NSBBO) algorithm to solve the IM-CM problem. The authors also proposed a multi-objective ranking strategy for efficiently utilizing preexisting knowledge for faster convergence.

7.3.1 Current Research Gap

Modeling social networks for influence maximization is a non-deterministic polynomial (NP-hard) problem [38]. Previous research successfully applied approximation and evolutionary strategies to recognize seed nodes.

However, sometimes the set of discovered seeds covers only a portion of the network due to the network structure and propagation model. Figure 7.5a depicts such an example. In this figure, we have tried to identify two seeds under the linear threshold propagation model. Under this propagation model, a node can be activated with at least a threshold value of 0.4 (i.e., at least 40% of the neighboring nodes need to be activated to become a node active). In the best-case scenario, it is found that the propagation can cover utmost 14 nodes out of 17 nodes. In this figure, the dark nodes are influenced by the beforehand described propagation model, while white-colored nodes are not influenced by the described propagation. Figures 7.14a and 7.15a also shows similar situations for the two well-researched networks named Zachary's karate club and Dolphins social network, respectively. Thus, there is a scope of further improved propagated influence by adding a few missing links. In Fig. 7.5a, the two networked components are connected through a bridge between nodes 2 and 6. In this figure, it is clear that the non-influenced component will not get influenced as the bridge-connected vertex could have a 0.33 threshold value at most. As 0.33 is less than 0.4, the propagation will not happen under this consideration. However, the addition of a missing link such as 2–11 could allow it to propagate influence further. Figures fig:zacharyspsltsp propagationb and 7.15b depict how the influence can be further propagated after adding few missing links. However, to



(a) Without recommended link. (b) With recommended link.

Fig. 7.5 Influence coverage considering structural modification

the best of our knowledge, no attempt was made to restructure the network that maximizes influence. In the below section, we have proposed a strategy to recommend missing links and identify seed nodes that further maximize influence.

7.4 Our Proposed Approach

Our proposed approach is a multistage strategy. Here first we have recognized several seeds with the desired timestep to propagate influence. Next, compute the number of missing links required to propagate influence to the non-influenced portion of the network and lastly recommended the minimal set of missing links. In the below subsections, we have discussed them in detail.

7.4.1 Identification of Initial Seeds

An efficient influence maximization problem has at least two optimization objectives, namely (1) maximize aggregated influence and (2) minimize the time of influence propagation. Therefore, the simplest form of the problem is a bi-objective optimization problem, where both the objectives have to be optimized simultaneously. Consider a social graph G and a diffusion model M , and a influence maximization problem identifies a set S of k nodes from seed V , where k is a positive integer, as the set that maximizes the influence spread $\sigma_{G,M}(S)$, as well as minimizes the propagation timestep $\mathcal{T}(S)$ for influence propagation, i.e., $\mathcal{T}(S) = \arg \max_S \mathcal{T}(v) \forall v \in \sigma_{G,M}(S)$. The influence maximization problem is mathematically represented using Eq. 7.2.

$$\begin{aligned} & \underset{S}{\text{Maximize}} && [\sigma_{G,M}(S), -\mathcal{T}(S)]^T \\ & \text{Subject to} && |S| = k, \\ & && S \in V \end{aligned} \tag{7.2}$$

In Eq. 7.2, we have used $-\mathcal{T}(S)$ as in optimization $\underset{S}{\text{maximize}} \mathcal{T}(S)$ is equivalent to minimize $-\mathcal{T}(S)$.

Usually, social networks do not offer at least one solution that simultaneously optimizes both objectives. A trade-off, therefore, becomes vital for deciding on such optimization. However, in this particular problem, maximizing the aggregated influence objective has a greater priority. As a bi-objective optimization problem, the influence maximization problem can be solved using any multi-objective solvers. In this study, we have used NSGA-II to recognize seeds. This approach identifies several non-dominated seed sets and corresponding node coverage.

7.4.2 Solution Encoding

Encoding of the solutions is still one of the basic requirements of NSGA-II. The two basic types of encoding systems used in evolutionary algorithms are *binary encoding* and *real-coded encoding*. The binary encoding of a network solution employs a n bit string, where n is the network's size. The string's index i denotes the network's i th node. A 1 bit indicates that the node is a seed node, while a 0 indicates that it is not. The binary encoding for the solutions (seeds are 10, 11) reported in Fig. 7.3 is shown in Fig. 7.6.

The size of the binary string in this encoding is the same as the network's size. As a result, the method is ineffective when dealing with vast social networks. The depiction itself may take up more memory than is available.

The genuine-coded encoding skim, on the other hand, simply preserves the collection of crucial actors in their current form. As a result, they are capable of providing model solutions regardless of network size. The method necessitates a less quantity of memory. In Fig. 7.7, the real-coded encoding for the solutions presented in Fig. 7.3 has been shown.

Although real-coded encoding is better suited for big network optimization issues, evolutionary procedures bring a few additional challenges. Below is a list of them.

1. The crossover and mutation operations may produce duplicate solutions. Figures 7.8 and 7.9 depict the situations for the crossover and for the mutation, respectively.
2. Set size may be shrunk (i.e., $|S| < k$) due to more than one time occurrences of a node within a solution. An ordered set can reduce the likelihood of duplicate generation. However, the ordering does not guarantee that the duplicate actor will be completely removed. Figure 7.10a illustrates how a crossover can reduce the size of the solution, and Fig. 7.10b shows the same set reduces the duplicate generation on an ordered set. Figure 7.11 illustrates the shrink of set size due to mutation.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0

Fig. 7.6 Binary encoding of the seeds (10, 11) presented in Fig. 7.3

1	2
10	11

Fig. 7.7 Real-coded encoding of the seeds (10, 11) presented in Fig. 7.3

Parent 1	$v(3) v(2) v(1)$	$v(5) v(6)$	Parent 3	$v(3) v(2) v(8)$	$v(5) v(7)$
Parent 2	$v(2) v(4) v(1)$	$v(7) v(5)$	Parent 4	$v(2) v(3) v(1)$	$v(6) v(4)$
↓					
Offspring 1	$v(3) v(2) v(1) v(7) v(5)$		Offspring 3	$v(3) v(2) v(8) v(6) v(4)$	
Offspring 2	$v(2) v(4) v(1) v(5) v(6)$		Offspring 4	$v(2) v(3) v(1) v(5) v(7)$	
↓					
Offspring 1	$v(3) v(2) v(1) v(7) v(5)$		Offspring 3	$v(3) v(2) v(8) v(6) v(4)$	
Offspring 2	$v(2) v(4) v(1) v(5) v(6)$		Offspring 4	$v(2) v(3) v(1) v(5) v(7)$	

Fig. 7.8 Duplicate solutions generated after crossover. In the illustration, although all four parents are unique, after crossover they produce duplicate solutions, i.e., offspring 1 and offspring 4

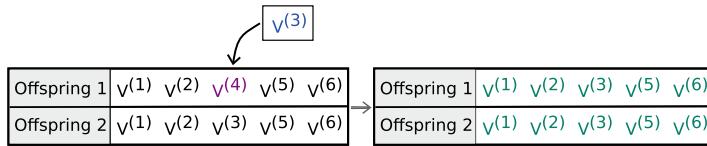


Fig. 7.9 Situation where duplicate solutions are generated after mutation. In offspring 1, mutation happens to $v^{(4)}$. After the mutation, $v^{(4)}$ is replaced with $v^{(3)}$. Therefore, both the solutions (offspring 1 and offspring 2) become same after the mutation

Parent 1	$v(3) v(2) v(1)$	$v(5) v(6)$	Parent 1	$v(1) v(2) v(3)$	$v(5) v(6)$
Parent 2	$v(5) v(4) v(7)$	$v(1) v(2)$	Parent 2	$v(1) v(2) v(4)$	$v(5) v(7)$
↓					
Offspring 1	$v(3) v(2) v(1) v(1) v(2)$		Offspring 1	$v(1) v(2) v(3) v(5) v(7)$	
Offspring 2	$v(5) v(4) v(7) v(5) v(6)$		Offspring 2	$v(1) v(2) v(4) v(5) v(6)$	
↓					
Offspring 1	$v(3) v(2) v(1)$		Offspring 1	$v(1) v(2) v(3) v(5) v(7)$	
Offspring 2	$v(5) v(4) v(7) v(6)$		Offspring 2	$v(1) v(2) v(4) v(5) v(6)$	

(a) Unordered

(b) Ordered

Fig. 7.10 Unordered set shrinks due to crossing. Using an ordered set, the problem was eliminated

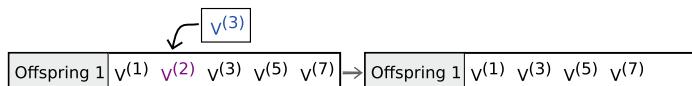


Fig. 7.11 Shrink due to mutation. In offspring 1, mutation happens to $v^{(2)}$. $v^{(2)}$ is replaced with $v^{(3)}$. As no duplicate elements can be present in a solution, the solution size become less after the mutation

7.4.3 Link Recommendation

For the sake of the link recommendation toward influence maximization, the first and most important thing is to make a few non-influenced nodes active so that the newly activated nodes can further propagate influence. Thus, identifying these sets of nodes is essential. To recommend a minimum number of links, it is also vital to determine the number of new links required to make a non-influenced node to active. A node can be activated by adding a few additional links to few already activated nodes. Thus, for each node present in the non-influenced segment of the network, we have determined the number of additional links required to make them active immediately.

As per the working principle of the linear threshold propagation model, a node v can be activated if the given threshold is greater than or equal to $|\mathcal{N}_A(v)|/|\mathcal{N}(v)|$. Here $\mathcal{N}(v)$ and $\mathcal{N}_A(v)$ denote neighboring nodes of v and activated neighboring nodes of v , respectively. Let us make a non-influenced node v active, and $|\mathcal{N}_{A^+}(v)|$ additional activated nodes need to be connected with v . Here, $\mathcal{N}_{A^+}(v)$ denotes set of additional active nodes that are to be linked to make node v active. Thus, after adding $|\mathcal{N}_{A^+}(v)|$ additional links, new number of activated neighboring nodes of v will be $(|\mathcal{N}_A(v)| + |\mathcal{N}_{A^+}(v)|)$, and new number of neighboring node of v will be $(|\mathcal{N}(v)| + |\mathcal{N}_{A^+}(v)|)$. Therefore, Eq. 7.3 becomes valid after adding $|\mathcal{N}_{A^+}(v)|$ new links. Solving Eq. 7.3, we obtain Eq. 7.5 for computing the number of additional links required for every non-influenced node v to make them active.

$$\frac{|\mathcal{N}_A(v)| + |\mathcal{N}_{A^+}(v)|}{|\mathcal{N}(v)| + |\mathcal{N}_{A^+}(v)|} \geq \text{Threshold}, \quad (7.3)$$

$$|\mathcal{N}_A(v)| + |\mathcal{N}_{A^+}(v)| \geq \text{Threshold} \times (|\mathcal{N}(v)| + |\mathcal{N}_{A^+}(v)|),$$

$$|\mathcal{N}_A(v)| + |\mathcal{N}_{A^+}(v)| \geq (\text{Threshold} \times |\mathcal{N}(v)|) + (\text{Threshold} \times |\mathcal{N}_{A^+}(v)|),$$

$$|\mathcal{N}_{A^+}(v)| - \text{Threshold} \times |\mathcal{N}_{A^+}(v)| \geq (\text{Threshold} \times |\mathcal{N}(v)|) - |\mathcal{N}_A(v)|,$$

$$|\mathcal{N}_{A^+}(v)| \times (1 - \text{Threshold}) \geq (\text{Threshold} \times |\mathcal{N}(v)|) - |\mathcal{N}_A(v)|,$$

$$|\mathcal{N}_{A^+}(v)| \geq \frac{(\text{Threshold} \times |\mathcal{N}(v)|) - |\mathcal{N}_A(v)|}{1 - \text{Threshold}},$$

Table 7.1 Computation of number of additional links required for each non-influenced node presented in Fig. 7.5a

v	$\mathcal{N}(v)$	$ \mathcal{N}(v) $	$\mathcal{N}_A(v)$	$ \mathcal{N}_A(v) $	Threshold	$ \mathcal{N}_{A^+}(v) $
1	[2, 3]	2	[]	0	0.4	$\frac{0.4 \times 2 - 0}{1 - 0.4} = 1.33 = 2$
2	[1, 3, 6]	3	[6]	1	0.4	$\frac{0.4 \times 3 - 1}{1 - 0.4} = 0.33 = 1$
3	[1, 2]	2	[]	0	0.4	$\frac{0.4 \times 2 - 0}{1 - 0.4} = 1.33 = 2$

Table 7.2 Potential nodes from the non-influenced segment for the Dolphins social network presented in Fig. 7.15a

Nodes	Coverage	Timestep	Links required
(46, 52)	39	13	15 (46: 8, 52: 7)
(15, 38)	39	13	17 (15: 9, 38: 8)
(34, 38)	39	15	15 (34: 7, 38: 8)
		:	

$$|\mathcal{N}_{A^+}(v)| = \frac{(\text{Threshold} \times |\mathcal{N}(v)|) - |\mathcal{N}_A(v)|}{1 - \text{Threshold}} \quad (7.4)$$

For the partial influence propagation presented in Fig. 7.5a, Table 7.1 shows the computation of the number of additional links required to make every non-influenced node to active.

A very straightforward link recommendation strategy could be the non-influenced nodes to be linked one after another gradually according to their minimum number of additional link requirements. Here each non-influenced node v has to be linked with $|\mathcal{N}_{A^+}(v)|$ number of active nodes. After converting an individual node active, the LT propagation needs to execute and check for convergence. However, this approach might demand an ample amount of new links. In contrast to that, an evolutionary strategy works better for identifying few nodes from the non-influenced segment of the network. The identified set of nodes must have the ability to maximally influence the non-influence segment after influence propagation initiated from them. This set of nodes will be one end of the recommended links. Then link the identified nodes with the required number (i.e., the sum of missing links to activate these nodes) of the active node to make them active. As the ranking information is available, we have adopted the locally globally tuned biogeography-based optimization algorithm [7, 8] to recognize this set of nodes. Table 7.2 shows propagation statistics of few potential nodes against Fig. 7.15a. Here the best combination (shown in boldface in Table 7.2) is (46, 52) as they cover all remaining 39 nodes with the lowest (15 in this case) number of additional links requirement.

After obtaining the desired set of nodes from the non-influenced segment of the network, it is time to find another end of the links within the active nodes. Sometimes establishing a new link to an active node might not be possible. Linking an inactive

Table 7.3 Computation of threshold after adding a link to a non-influenced node

v	Timestep	$\mathcal{N}_A(v)$	$\mathcal{N}(v)$	$ \mathcal{N}_A(v) $	$ \mathcal{N}(v) $	Threshold before	Threshold after
4	5	[5, 7, 9]	[5, 6, 7, 9]	3	4	$3/4 = 0.75$	$3/(4+1) = \mathbf{0.60}$
5	4	[7, 8]	[4, 7, 8]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$
6	6	[4, 9]	[2, 4, 9]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$
7	2	[10, 9]	[4, 5, 8, 9, 10]	2	5	$2/5 = 0.40$	$2/(5+1) = 0.33$
8	3	[7, 10]	[5, 7, 10, 14]	2	4	$2/4 = 0.50$	$2/(4+1) = \mathbf{0.40}$
9	1	[10, 11]	[6, 7, 10, 11]	2	4	$2/4 = 0.50$	$2/(4+1) = \mathbf{0.40}$
10	0	[]	[7, 8, 9, 11, 12]	0	5	1.0	1.0
11	0	[]	[9, 10, 12, 15]	0	4	1.0	1.0
12	1	[10, 11]	[10, 11, 13, 15, 16]	2	5	$2/5 = 0.40$	$2/(5+1) = 0.33$
13	4	[12, 16]	[12, 14, 16, 17]	2	4	$2/3 = 0.67$	$3/(4+1) = \mathbf{0.60}$
14	5	[8, 13]	[8, 13, 17]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$
15	2	[11, 12]	[11, 12, 16]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$
16	3	[12, 15]	[12, 13, 15, 17]	2	4	$2/4 = 0.50$	$2/(4+1) = \mathbf{0.40}$
17	5	[13, 16]	[13, 14, 16]	2	3	$2/3 = 0.67$	$2/(3+1) = \mathbf{0.50}$

node to an activated node certainly lowers the threshold of the active node. In this situation, the propagation stops much before the expected propagation. Therefore for the activated node, it is essential to check whether a link can be added or not? Let the desired node be activated at timestep T_x . Therefore, the ratio of activated neighboring nodes and the number of neighboring nodes at $T_x - 1$ timestep must be greater than or equal to the supplied threshold value. Thus, the activation requirement can be checked using Eq. 7.5.

$$\frac{|\mathcal{N}_A^{(T_x-1)}(v)|}{|\mathcal{N}(v)| + 1} \geq \text{Threshold} \quad (7.5)$$

where $\mathcal{N}_A^{(T_x-1)}(v)$ is the activated neighboring nodes of v at timestep $T_x - 1$. The number of neighboring nodes after adding a missing node at time $T_x - 1$ is $|\mathcal{N}(v)| + 1$.

Table 7.3 shows the computation of the new threshold after establishing a link to a non-influenced node. The column ‘threshold after’ in this table presents the computed value. As the minimum threshold value required to activate a node is 0.40, nodes 7 and 12 do not allow establishing a link from them. Nodes other than 7 and 12 are suitable for consideration for establishing links (In Table 7.3 respective threshold after values are shown in boldface). The threshold value of nodes 10 and 11 will always be one as they are the seed nodes. In this particular experiment, the number of links required (1) is less than the number of seeds (2). Therefore without any beforehand computation, the missing link can be directly linked to either of the seeds (as the seed nodes having minimum timestep value). Thus, (2, 10) or (2, 11) are the recommended links in this case.

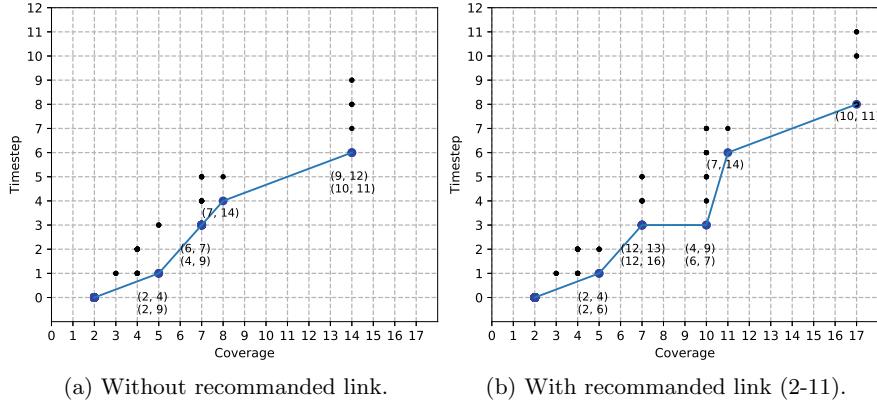


Fig. 7.12 Identification of initial non-dominated solutions for the network presented in Fig. 7.5

7.5 Experimental Setup and Result Analysis

The experimental study uses three different use cases that might occur during the identification of seeds and recommended links. This study uses three different social networks, namely (1) an experimental network used throughout this article, (2) the interaction network of Zachary's karate club, and (3) Dolphins social networks. In this study, the presentation of the network follows a few basic conventions. They are as follows: (1) Circle represents a node, (2) a line between two circles indicates a link, (3) white-colored nodes are non-influenced nodes, (4) influenced nodes are gray colored, (5) double-bordered nodes are the seeds of the propagation, (6) a blue node indicates the node becomes active in the current timestep, and (7) blue-dashed lines are recommended links. Throughout the study, we have used the linear threshold propagation model with a threshold value of 0.40.

Figure 7.12 presents the multi-objective seeds identification for the network presented in Fig. 7.5. X-axis of the plot shows the total influence coverage after the end of the propagation. Y-axis of the plot shows the required timestep to meet the maximum possible coverage. Blue dots are the non-dominated solutions, whereas black dots are the dominated solutions. The solid blue line is the Pareto front. However, as total influence coverage in terms of the number of nodes has a higher priority as compared to the timestep, solutions (9, 12) and (10, 11) are the best among all other Pareto optimal solutions. In this study, we have chosen seeds (10, 11). Figure 7.12a shows the maximum coverage of seeds without a recommended link. In this scenario without the recommended links, the best seeds are able to cover 14 nodes out of 17 nodes. The influence propagation took a maximum of six timesteps. Figure 7.12b shows the propagation after adding the recommended link (2, 11). After adding the recommended link (2, 11), the influence covers complete networks. The additional propagation completes in eight timesteps.

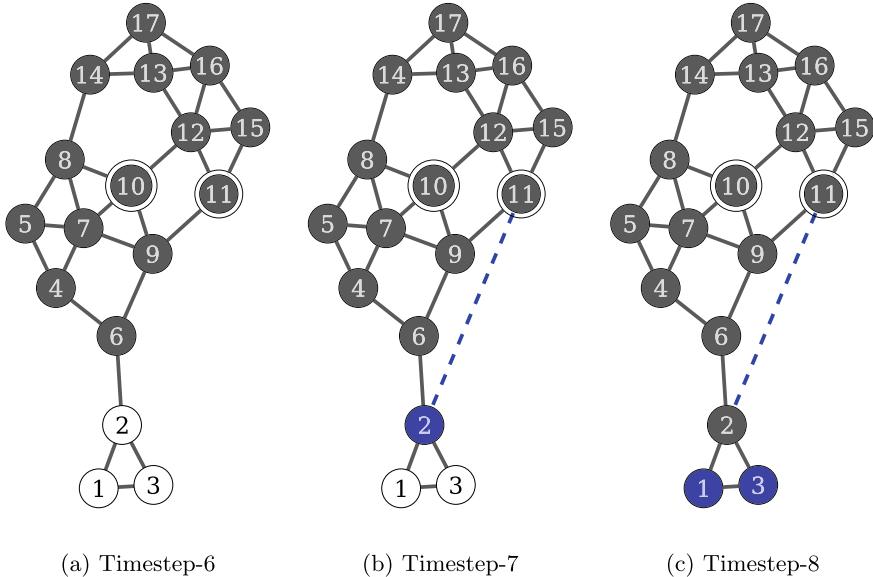


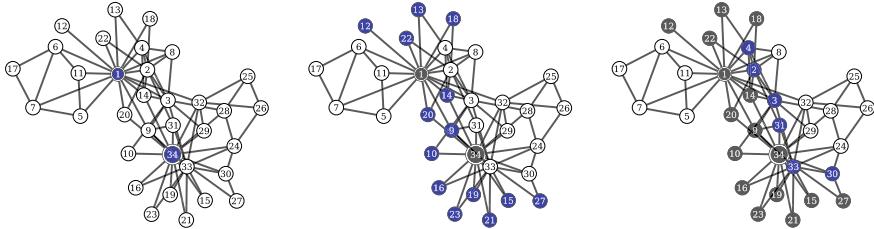
Fig. 7.13 Linear threshold propagation with recommended link on experimental network

Figure 7.13a shows the limited propagation with best seeds (10, 12) without the recommended link. Figure 7.13b, c shows the influence propagation with recommended link (2, 11). In Fig. 7.13b, node 2 becomes active at timestep 7 after adding link (2, 11), and at timestep 8 (Fig. 7.13c), nodes 1 and 3 become active. In this particular case, additional timesteps are required to maximize influence.

Figure 7.14 presents the influence propagation on the interaction network of Zachary's karate club. Table 7.4 shows the initial seeds and their respective influence coverage and timestep. Seeds (1, 34) are the only non-dominated solution present in this table (shown in boldface) as it covers maximum with the lowest timestep. Figure 7.14a, b presents the influence propagation without recommended link and with recommended link for the seeds (1, 34), respectively. In this particular experiment, only a single link (5, 34) is capable of propagating influence throughout the network. However, the added link does not demand any additional timesteps.

Table 7.4 Seeds with their coverage and timestep

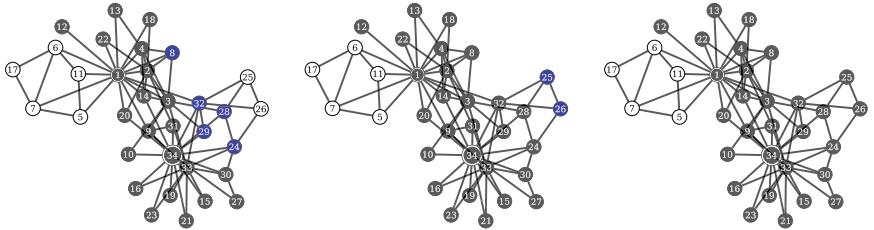
Seeds	Coverage	Timestep	Seeds	Coverage	Timestep	Seeds	Coverage	Timestep
(1, 34)	29	4	(4, 34)	29	6	(13, 34)	29	10
(1, 33)	29	5	(2, 33)	29	7	(18, 34)	29	10
(2, 34)	29	6	(8, 34)	29	8	(20, 33)	29	10



(a1) Timestep-0

(a2) Timestep-1

(a3) Timestep-2

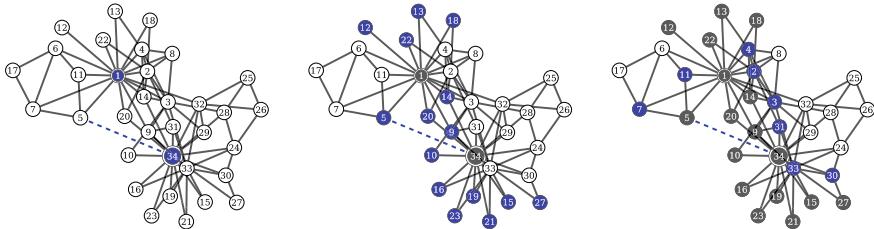


(a4) Timestep-3

(a5) Timestep-4

(a6) Final

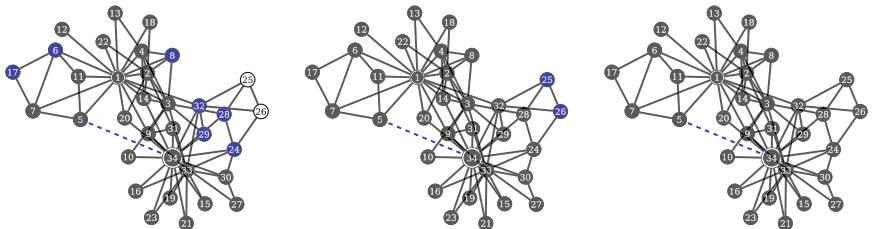
(a) Propagation without recommended links.



(b1) Timestep-0

(b2) Timestep-1

(b3) Timestep-2



(b4) Timestep-3

(b5) Timestep-4

(b6) Final

(b) Propagation with recommended links.

Fig. 7.14 Influence propagation on the network of Zachary's karate club under linear threshold model. A threshold value 0.4 is used

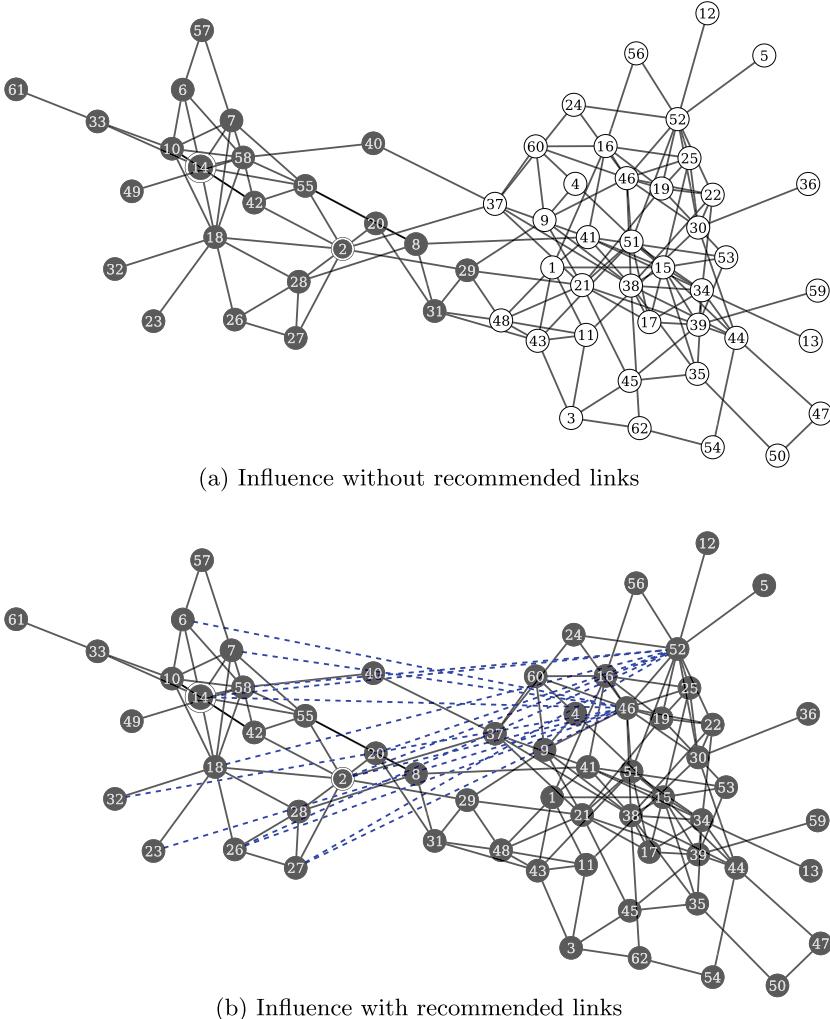


Fig. 7.15 Influence maximization on Dolphins social network under linear threshold model. A threshold value 0.4 is used as a parameter

For the Dolphins social network, the best seeds for LT propagation with a 0.4 threshold are (2, 14) and (2, 58). These seeds can influence 23 nodes out of 62 nodes in 11 timesteps. In this study, we have chosen (2, 14) as the candidate seeds. The propagation for seeds (2, 14) is presented in Fig. 7.15a. In Fig. 7.15a, the non-influenced section contains 39 nodes. For these 39 nodes, LGBBO has been executed and determined potential nodes set that maximally influence the remaining 39 nodes. Table 7.2 presents some solutions. We have chosen solution (46, 52) as they can influence the remaining 39 nodes with the lowest timestep. Most importantly, the solution requires

less number of missing links, i.e., 15 to make them active. Figure 7.15b shows the influence propagation after adding the recommended links.

7.6 Conclusion

In today's world, maximizing one's influence in social networks has huge ramifications. Identifying a set of seed nodes that are suitable for influence maximization, on the other hand, is extremely difficult. The identification process is considered NP-hard. Alongside that due to network structure and given propagation model, sometimes it is not possible to propagate influence that covers the entire network or at least a significant portion of the network. However, we observed that a little structural modification by adding a few missing links further allows information propagation in the non-influenced segment of the network. Determining the number of missing links required to establish smooth propagation is hard. Effective recommendation of such links is computationally NP-hard. In the view of maximizing propagation, our proposed two-stage approach identifies seeds and few missing links with a minimum timestep for propagation. In the first stage, we have invoked a multi-objective optimizer specially NSGA-II to find out the initial set of seeds along with coverage and timestep for the convergence. In the second stage, we have statistically calculated the number of missing links that are required to convert the status of the non-influence nodes. Using LGBBO, we have determined suitable nodes that maximally propagate influence in the non-influenced segment. Lastly, we have checked and established links to the active nodes. In the empirical analysis section, we have shown three use cases to validate the proposed strategy. The empirical analysis showed a good improvement in influence maximization after using the recommended seeds and recommended links.

References

1. Li, F., Du, T.C.: The effectiveness of word of mouth in offline and online social networks. *Expert Syst. Appl.* **88**, 338–351 (2017)
2. Hsiao, J.P.H., Jaw, C., Huan, T.C.: Information diffusion and new product consumption: a bass model application to tourism facility management. *J. Bus. Res.* **62**(7), 690–697 (2009)
3. Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. *ACM SIGMOD Rec.* **42**(2), 17–28 (2013)
4. Hao, L., Yang, L.Z., Gao, J.M.: The application of information diffusion technique in probabilistic analysis to grassland biological disasters risk. *Ecol. Model.* **272**, 264–270 (2014)
5. De, S.S., Dehuri, S.: Multi-objective biogeography-based optimization for influence maximization-cost minimization in social networks. In: International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making, pp. 11–34. Springer (2019)
6. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In: International Conference on Parallel Problem Solving from Nature, pp. 849–858. Springer (2000)

7. Giri, P.K., De, S.S., Dehuri, S.: A novel locally and globally tuned biogeography-based optimization algorithm. In: Soft Computing: Theories and Applications, pp. 635–646. Springer (2018)
8. Giri, P.K., De, S.S., Dehuri, S.: Adaptive neighbourhood for locally and globally tuned biogeography based optimization algorithm. *J. King Saud Univ.-Comput. Inf. Sci.* **33**(4), 453–467 (2021)
9. Ritzer, G., et al.: The Blackwell Encyclopedia of Sociology, vol. 1479. Blackwell Publishing, New York, NY (2007)
10. Ryan, B., Gross, N.: Acceptance and diffusion of hybrid corn seed in two Iowa communities. *Iowa Agric. Home Econ. Exp. Stn. Res. Bull.* **29**(372), 1 (1950)
11. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146 (2003)
12. Iacobucci, D.: Network models of the diffusion of innovations. *J. Mark.* **60**(3), 134 (1996)
13. Valente, T.W.: Network models and methods for studying the diffusion of innovations. In: Models and Methods in Social Network Analysis, vol. 28, pp. 98–116 (2005)
14. Kimura, M., Saito, K.: Tractable models for information diffusion in social networks. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 259–271. Springer (2006)
15. Kempe, D., Kleinberg, J., Tardos, É.: Influential nodes in a diffusion model for social networks. In: International Colloquium on Automata, Languages, and Programming, pp. 1127–1138. Springer (2005)
16. Li, Y., Fan, J., Wang, Y., Tan, K.L.: Influence maximization on social graphs: a survey. *IEEE Trans. Knowl. Data Eng.* **30**(10), 1852–1872 (2018)
17. Fonseca, C.M., Fleming, P.J., et al.: Genetic algorithms for multiobjective optimization: formulation discussion and generalization. In: ICGA, vol. 93, pp. 416–423. Citeseer (1993)
18. Zitzler, E., Laumanns, M., Thiele, L.: Spea2: improving the strength pareto evolutionary algorithm. TIK-Report 103 (2001)
19. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**(2), 182–197 (2002)
20. Zhang, Q., Li, H.: MOEA/D: a multiobjective evolutionary algorithm based on decomposition. *IEEE Trans. Evol. Comput.* **11**(6), 712–731 (2007)
21. Coello, C.A.C.C.: A short tutorial on evolutionary multiobjective optimization. In: International Conference on Evolutionary Multi-Criterion Optimization, pp. 21–40. Springer (2001)
22. Zitzler, E., Laumanns, M., Bleuler, S.: A tutorial on evolutionary multiobjective optimization. In: Metaheuristics for Multiobjective Optimisation, pp. 3–37 (2004)
23. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 57–66 (2001)
24. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 61–70 (2002)
25. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 420–429 (2007)
26. Gong, M., Yan, J., Shen, B., Ma, L., Cai, Q.: Influence maximization in social networks based on discrete particle swarm optimization. *Inf. Sci.* **367**, 600–614 (2016)
27. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 199–208 (2009)
28. Zhou, C., Zhang, P., Guo, J., Zhu, X., Guo, L.: UBLF: an upper bound based approach to discover influential nodes in social networks. In: 2013 IEEE 13th International Conference on Data Mining, pp. 907–916. IEEE (2013)

29. Zhou, C., Zhang, P., Guo, J., Guo, L.: An upper bound based greedy algorithm for mining top-k influential nodes in social networks. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 421–422 (2014)
30. Wang, Y., Cong, G., Song, G., Xie, K.: Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1039–1048 (2010)
31. Barbieri, N., Bonchi, F., Mancuso, G.: Topic-aware social influence propagation models. *Knowl. Inf. Syst.* **37**(3), 555–584 (2013)
32. Guo, J., Zhang, P., Zhou, C., Cao, Y., Guo, L.: Item-based top-k influential user discovery in social networks. In: 2013 IEEE 13th International Conference on Data Mining Workshops, pp. 780–787. IEEE (2013)
33. Goyal, A., Bonchi, F., Lakshmanan, L.V.: A data-based approach to social influence maximization. arXiv preprint [arXiv:1109.6886](https://arxiv.org/abs/1109.6886) (2011)
34. Zhou, C., Zhang, P., Zang, W., Guo, L.: Maximizing the cumulative influence through a social network when repeat activation exists. *Procedia Comput. Sci.* **29**, 422–431 (2014)
35. Goyal, A., Bonchi, F., Lakshmanan, L.V.: Learning influence probabilities in social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 241–250 (2010)
36. Saito, K., Nakano, R., Kimura, M.: Prediction of information diffusion probabilities for independent cascade model. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 67–75. Springer (2008)
37. Yang, J., Liu, J.: Influence maximization-cost minimization in social networks based on a multiobjective discrete particle swarm optimization algorithm. *IEEE Access* **6**, 2320–2329 (2017)
38. Lu, Z., Zhang, W., Wu, W., Kim, J., Fu, B.: The complexity of influence maximization problem in the deterministic linear threshold model. *J. Comb. Optim.* **24**(3), 374–378 (2012)

Chapter 8

Performance Analysis of State-of-the-Art Classifiers and Stack Ensemble Model for Liver Disease Diagnosis



Barnali Sahu, Supriya Agrawal, Hiranmay Dey, and Chandani Raj

Abstract Around the world, liver disease is one of the leading causes of death. The number of persons who suffer is steadily rising. It is vital to maintain a healthy liver to assist functions like digestion and detoxification. Fatty liver, cirrhosis, and hepatitis are some of the most prevalent liver problems that require medical attention. Because of the modest symptoms, it is also difficult to predict in the early stages. To address the issue, performance analysis of different heterogeneous machine learning algorithms are employed to discover the most suitable model for liver disease diagnosis. The proposed research is carried out in two phases, with data collected in Andhra Pradesh, India's North East. The classification algorithms like naïve Bayes, SVM, KNN, logistic regression, decision tree, and neural networks are implemented on the dataset in first phase. Different metrics are used to test the classifier's output. Additionally, in the second phase, the multiple classifiers are contrasted to an ensemble model in second phase, which combines the decisions of several models to assess if individual models or a combined model delivers superior accuracy and reliability.

8.1 Introduction

Health care is a very important aspect for every human being. Despite having some world's best clinicians, medical errors are happening all the time which would have prevented needless deaths. Patients who have chronic disease often find that these diseases are diagnosed late, and they are not managed well. So, there is a solid requirement to provide an effective and practical framework to foresee the result of such infection. Machine learning offers a principled approach to process the huge clinical data beyond the scope of human capability and converting it into diagnostic model which gives better insights to practitioners that are helpful in taking necessary steps and proper treatment rightfully.

B. Sahu (✉) · S. Agrawal · H. Dey · C. Raj

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

e-mail: barnalisahu@soa.ac.in

The liver is a key organ in the human body. It helps in digestion, synthesis of proteins, and in maintaining the metabolism of body. So, it is important to have a healthy liver, but nowadays, liver disease has become eminent and is the leading cause of death across the globe. It is a broad term that encompasses different types of damage to the liver. The widespread cause of this infection is due to unhealthy life style, smoking, and excessive alcohol consumption. The problems are not identified in the initial stages as the liver functions normally even if it is damaged partially. There are so many types of liver infections such as cirrhosis, hepatitis, and liver failure. Liver diseases can lead to a decline in liver function as well as permanent damage. Therefore, an early diagnosis is very crucial and will increase the survival rate of patients. Hence, an automatic classification tool is advantageous for diagnosing the infection and will be helpful in reducing the time delay caused due to unnecessary commuting between pathology lab and visiting doctors, and it would also be cost-effective. In the medical field, high accuracy is critical for disease identification, as it reduces the margin of diagnostic error. It can be solved by medical data mining, which can assist doctors in making judgments by analyzing massive amounts of data, detecting hidden structures in the data, and taking appropriate safeguards. This sparked our interest in conducting research for a medical dataset on liver disease diagnostics. There are several researchers who have already worked on the liver dataset such as Rajeswari et al. [1] employed NB, K Star and FT tree to classify liver disease and discovered that FT tree performed better. Ramana et al. [2] used various classification techniques like decision tree, SVM, naïve Bayes, KNN, and backpropagation and concluded that back propagation algorithm performs well.

To discriminate between a healthy liver and an infected liver, we used a number of heterogeneous classifiers in our research. According to the studies [1–9], feature selection based on correlation was not applied for the liver dataset, and ensemble approaches mostly consisted of bagging and boosting. Our approach, on the other hand, was used to develop a stacking ensemble model by performing feature selection using Pearson correlation and various anomalous classifier pairings. Our work is unique in this way. The study is divided into two sections. The first portion evaluates traditional classifiers, while the second evaluates an ensemble model's performance. The goal is to see which model produces the most accurate results for the liver dataset. The Pearson correlation coefficient method, which uses a filter-based approach to find the most relevant features, was chosen as the feature selection method. Finally, an ensemble model is built that combines the predictions generated by the various models. The ensemble model is a method of integrating weak learners into a powerful predictive model. Bagging, boosting, and stacking are some of the options for accomplishing this. We used stacking in this investigation because bagging and boosting combine homogeneous weak learners, but stacking combines heterogeneous weak learners. The following sections make up the paper: The literature review is shown in Sect. 8.2, the dataset description and methods for classification and feature selection are presented in Sect. 8.3, the experimental analysis and discussion is presented in Sect. 8.4, and finally Sect. 8.5 deals with the conclusion.

8.2 Literature Review

Rajeswari and Reena [1] had used the data mining algorithms of naïve Bayes, K star and FT tree to analyze the liver disease. Naïve Bayes had provided 96.52% correctness in 0 s. 97.10% accuracy was achieved by using FT tree in 0.2 s. K star algorithm had classified the instances about 83.47% accurately in 0 s. On the basis of outcomes, highest classification accuracy was offered by FT tree on liver disease dataset as compared to other data mining algorithms. Ramana et al. [2] had discussed various classification techniques for automatic medical diagnosis. The classification algorithms used were decision tree, SVM, naïve Bayes classification (NBC), KNN, and backpropagation. Observed parameters indicated that for accurate diagnosis, three common features of liver disease are very important. Karthik et al. [3] had applied intelligence techniques for diagnosis of liver disease. The classification algorithms were implemented in three stages. Artificial neural network was taken in first phase for classification of liver disease followed by second phase where rough set of rules were applied using algorithm learn by example for classification of liver disease which improved the accuracy. In third phase, the types of the liver disease were identified using fuzzy rules. Six rules were generated using LEM algorithm which showed an accuracy of 96.6%. Vohra et al. [4] had performed a technique to classify the liver patient. Classifiers used were MLP, random forest, SVM, and Bayesian network. First, all algorithms were applied on the original dataset to get the percentage of correctness. Second, feature selection method was applied to get the significant subset; then again, all algorithms were used to test the subset. Then, the outcome was compared for before and after feature selection. Following Feature Selection, the accuracy of J48, MLP, SVM, Random Forest, and Bayes Net is 70.669 %, 70.8405 %, 71.3551 %, 71.869 %, and 69.125 %, respectively. Montazeri Mitra [5] had proposed automatic classification tools as a diagnostic tool for liver disease. The classification algorithm was naïve Bayes, random forest, 1NN, AdaBoost, and SVM. Accuracy of the models were 55%, 72%, 64%, 70%, and 71% and area under ROC curve of 0.72, 0.72, 0.59, and 0.67 was 0.5. Random forest model was having highest level of accuracy. Random forest and naïve Bayes models had the largest area under ROC curve. Vijayarani and Dhayanand [6] had built a design to predict liver disease using classification algorithms. They had used naïve Bayes and support vector machine algorithms. They got the high classification accuracy values in case of SVM and low execution time in case of NB. Singh et al. [7] had performed detection of liver disease by applying QDA, LDA, SVM, and feedforward neural network (FFNN)-based approaches. They found out that on the accuracy rate basis, SVM outperformed other classification algorithms. Antony et al. [8] had implemented GA to decrease the dimensions of the original dataset which enhanced the correctness and performance of classifiers, namely J-48, naïve Bayes, and KNN. Orczyk et al. [9] had presented a relative learning of several feature selection techniques for propelled liver fibrosis determination. These algorithms were consolidated with chosen machine learning estimation which incorporates ensembles classifiers like J48 Pruned C4.5 decision tree, IbK KNN classifier, random forest one rule, and decision table classifier.

8.3 Materials and Methods

This section deals with the description of the dataset, the methods used for feature selection and classification for liver data.

8.3.1 Dataset Description

The Indian liver patient dataset is used in this study which comprise of records of 583 patients that has 416 liver patients and 167 non-liver patients. The dataset has been downloaded from UCI ML repository [10]. It has ten feature columns and one target column, with the liver patient being labeled 1 and the non-liver patient being labeled 2.

8.3.2 Feature Selection Using Pearson Correlation

Feature selection is used to minimize the number of input variables in order to improve a model's performance which also lowers down the computational cost of modeling. The technique used in this paper for feature selection is Pearson correlation analysis which helps to figure out how strong a link between two variable is. It is known as filter-based method as it filters out the irrelevant features. Equation 8.1 shows the formula for calculating the correlation coefficient.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (8.1)$$

where x_i is predictor variable values, y_i is response variable values, \bar{x} is mean of predictor variable values, and \bar{y} is the mean of response variable values. The value of r ranges from -1 to 1 , with 1 representing a positive correlation and -1 signifying a negative correlation while 0 indicates no association.

8.3.3 Classification Algorithms with the Defined Parameters for the Model

Classification is a supervised learning approach where a mapping function is estimated between input and output variables that is discrete. It categorizes unknown items into a discrete set of categories or classes. The concept is simple: to predict the unseen label/category for new data. The different algorithms used in this study

are naïve Bayes, support vector machine, K-nearest neighbors, logistic regression, decision tree, neural network, and stacking ensemble classifier.

Naïve Bayes: is a classification method based on a hypothesis that if a class label is given, then the attributes are independent, depending on some conditions. Bayes theorem allows to calculate posterior probability, i.e., it finds the probability of an occurring event where the probability of another event that has already occurred is given. The mathematical equation of Bayes' theorem is defined in Eq. (8.2).

$$P(C_i|X) = (P(X|C_i)* P(C_i))/P(X) \quad (8.2)$$

where $X = (X_1 \dots X_n)$ represents features and $i =$ possible classes.

Support Vector Machine: SVM is a classification technique that employs a supervised approach to determine the optimum hyperplane for classifying data. Each data point is plotted in n-dimensional space (where n denotes the number of features), with the value of each feature being the coordinate value. The SVM kernels are used to convert low-dimensional data to high-dimensional data because it is impossible to separate two classes via hyperplane for nonlinear data. In this study, the gamma value is set to 1, and else all parameters are default.

K-Nearest Neighbor: KNN is a method which classify instances based on their resemblance to others. It takes a cluster of labeled points and uses them to learn to label unknown test points. The entire dataset is used to train the model. The key deciding factor is the number of nearest neighbors, which is denoted by the letter K. For an even number of classes, the odd value of k is picked, and vice versa. K is set to the default value of 5 in this study.

Logistic Regression is used to predict binary outcomes such as true/false, yes/no, 1/0. It classifies records of a dataset based on the values of input fields. It is a mathematical model that models conditional probability using the logit function. It employs the sigmoid function, which transforms the independent variables into a probability expression that spans from 0 to 1 in relation to the dependent variable.

Decision Tree: DT is a supervised approach where recursive partitioning is used to categorize the data in decision trees. It maps out all the possible decision paths in the form of a tree. The attribute for the partitioning is chosen by checking various parameters like entropy and information gain.

MLP Neural Networks: MLP is a classical type of neural network which is comprised of one or more layers of neuron. It consists of one input layer, one output layer, and between these two contain one or more hidden layers. It is very flexible and suitable for predicting classification problems. All the computations are performed in the hidden layer only, and the default activation function used is rectified linear unit as ReLu which shows a linear nature for positive input and returns 0 for negative input.

Stacking Ensemble model: Stacking or stacked generalization is an ensemble model which uses a meta-classifier to merge the results of base level models (individual classifiers). The meta-learner takes the output of each individual predictors as

input and makes the final prediction. A separate model is trained to perform prediction. The architecture of stacking ensemble model consists of two or more base models and a metamodel. The entire original training set is fed to the base models, but in the case of metamodel, data which has not been applied to train the base models are fed to the meta-model. K-fold cross validation of the base models can be used to do this. Then, the metamodel is trained in isolation using out-of-sample data predictions from base models. Linear regression is used as a metamodel for regression problems, and logistic regression is used as a metamodel for classification problems in general.

8.3.4 Proposed Model

The model that has been proposed is based on supervised classification algorithm techniques where a fixed number of features are available. Using those features, the dataset has been analyzed and preprocessed. Primarily, the dataset is made ready by cleaning it which involves preprocessing like dealing with duplicate entries, handling null values, and converting categorical data into numerical data. Further, the preprocessed data is split into train and test sets. The feature selection has been performed to check which all features are important so that redundant attributes can be removed. All the individual classifier are applied with and without performing feature selection, and the accuracy are compared. The classification is then carried out in two phases. The state-of-the-art classifiers are trained in first phase, and the performance has been analyzed. In the second phase, the predictions from the base stacked models are given to the meta-classifier which then yields final predictions.

The flowcharts for the suggested technique are shown in Figs. 8.1 and 8.2. The entire procedure is divided into two parts, according to the flow chart. Stage 1 analyzed the performance of the individual classifiers, whereas stage 2 analyzed the performance of an ensemble model. The motivation behind incorporating this step is due to the advantages this model layout. There is a hypothesis that stacking performs better than any single model [11]. The meta-classifier used in this study is logistic regression, whereas the base models are naïve Bayes, SVM, KNN, logistic regression, decision tree, and neural networks.

8.4 Experimental Analysis and Discussion

The experiment is run on a PC with an i3/i5 processor, 8 GB RAM, and a 1 TB hard disk running Windows 10, using Python 3.8 as the programming language. The operations performed in the evaluation of model are shown in Figs. 8.1 and 8.2. The details are as follows:

The raw dataset is collected and made ready by cleaning it. In this study, 26 rows were found identical so those were managed, all the feature variables are integer

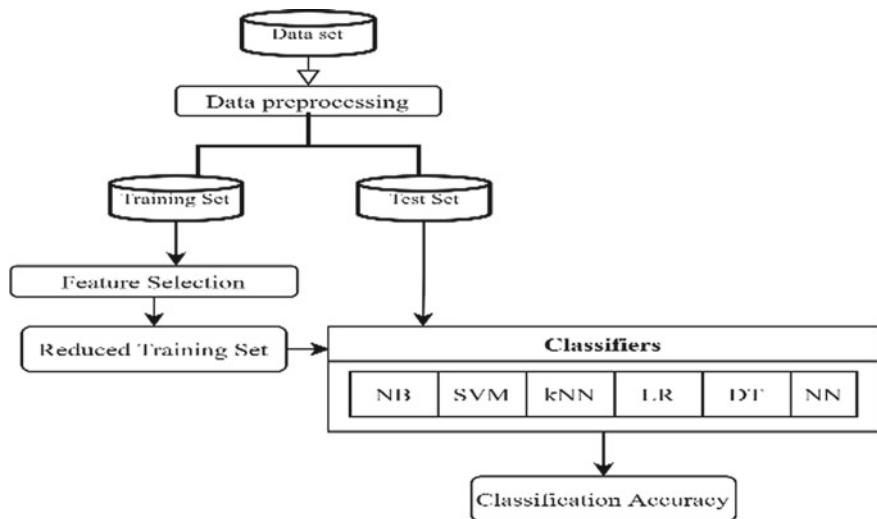


Fig. 8.1 Schematic representation of liver disease classification using state-of-the-art classifier



Fig. 8.2 Schematic representation of liver disease classification using stacked ensemble model

except an attribute gender which is a categorical variable as clearly visible in Table 8.1, so we have converted it into numerical attribute. Albumin and globin ratios have null values, so they are controlled by infusing them with their mean. The dataset has labels 1 and 2 for liver and non-liver patients, respectively, thus they are transformed to 1 and 0 for ease of use. The preprocessed data is then divided into two sets: training and testing, with 70% of the data going to training and 30% to testing. The correlation heat map is generated to visualize the relationship between every attribute as shown in Fig. 8.3. The heat map also shows the correlation coefficient values to know up to what degree the attributes are correlated. The value “1” indicates a strong correlation between two variables which is true as each variable always perfectly correlates with itself.

The models are trained with all of the attributes in the dataset at first, and the accuracy is measured. Then, we used the Pearson correlation approach to choose features. The four pairs of attributes are clearly visible in the heat map displayed in Fig. 8.3, i.e., (direct_bilirubin and total_bilirubin), (alamine_aminotransferase and aspartate_aminotransferase), (albumin and total_proteins), and (albumin and

Table 8.1 Comparing the accuracy of state-of-the-art classifiers before feature selection and after feature selection

Classifiers	Accuracy before feature selection (%)										Accuracy after feature selection (%)									
NB	59.87										61.13									
SVM	71.08										71.18									
KNN	70.40										70.40									
LR	72.67										73.17									
DT	67.13										69.40									
NN	70.39										72.42									

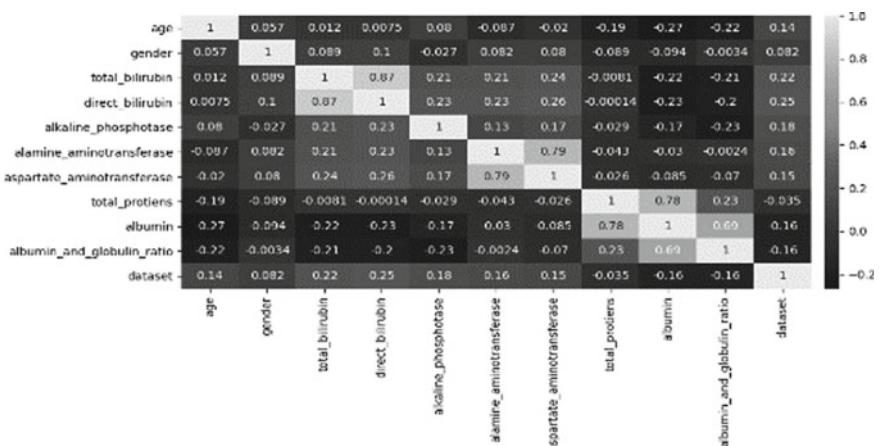


Fig. 8.3 Correlation heat map of all the attributes of liver dataset

albumin_and_globulin_ratio) possess redundancy. As a result, we eliminated any of them and trained the models with a smaller training set, then measured the accuracy. Table 8.1 compares the accuracy of the classifiers before and after feature selection.

After eliminating correlated attributes, relevant attributes were selected, but we got the same result as the classification accuracy is similar or increased by [1, 2]%, so we have proceeded by taking the reduced list of attributes. Though there is not much difference, but removing redundant attributes not only saves space but the execution time of complex machine learning algorithm also reduces. In order to improve accuracy, we performed model analysis of stacking ensemble model, and the values of precision, recall, f1-score, and accuracy are recorded. The accuracy of each model has been calculated by using a statistical method cross validation setting scoring value as accuracy and by taking cv value as 10. Table 8.2 shows the calculated value of these metrics for all the classifiers and stacked ensemble model. Every model's performance in proposed approach has been examined in detail by calculating the precision, recall, f1-score, and accuracy metrics value whose equations are defined in Eqs. (8.3, 8.4, 8.5, and 8.6), respectively.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (8.3)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (8.4)$$

$$\text{F1-score} = (2 \times \text{Precision} \times \text{recall})/(\text{precision} + \text{recall}) \quad (8.5)$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (8.6)$$

where TP = true positive, FP = false positive, TN = true negative and FN = false negative.

Logistic regression performed better out of all and achieved the highest accuracy of 73.17% as can be seen in Table 8.2 and Fig. 8.4. The second-best model which performed best is neural network model with an accuracy of 72.42%. The

Table 8.2 Classification results of all the classifiers and stacking model

Classifiers	Accuracy (%)	Precision	Recall	F1-score
NB	61.13	0.73	0.54	0.55
SVM	71.18	0.80	0.73	0.63
KNN	69.40	0.62	0.65	0.63
LR	73.17	0.62	0.68	0.63
DT	69.40	0.66	0.66	0.66
NN	72.42	0.58	0.70	0.60
STACKING	71.18	0.80	0.73	0.63

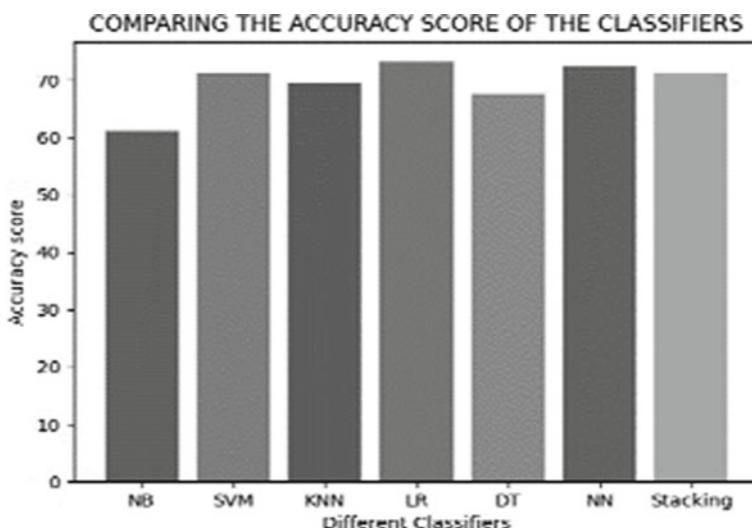


Fig. 8.4 Comparison of classification accuracies of state-of-the-art classifiers and stacking ensemble model

naïve Bayes, SVM, KNN, DT, and stacking gave the accuracy of 63.13%, 71.18%, 69.40%, 69.40%, and 71.18%, respectively. The stacking framework combines the output of single classifiers for producing high accuracy, and there are different ways in which aggregation can happen such as bagging, boosting, and stacking. Bagging helps to minimize variance in the data while boosting helps to minimize bias in data, so the purpose of choosing stacking was to manage both variance and bias which in turn would increase the predictive performance. We assumed that stacking ensemble model will uplift the accuracy score, but it fails to perform better than single model for the retrieved liver disease dataset. So it can be concluded that stacking ensemble model does not always guarantee to improve the model performance. Individual classifier in stacking framework has varied set of abilities as they use different learning algorithms and gives predictions independently, which helps in producing better results. However, the selection of classifiers for training the base models is also important. So, we can conclude though we have achieved the best accuracy in logistic regression model, but it is not always true. One classifier can perform well in one dataset, another classifier can outperform for different dataset.

8.5 Conclusion

The main focus of this research is to analyze the performance of different heterogeneous machine learning classification algorithm for liver disease diagnosis, pick the best suitable model which gives best accuracy out of all. It is found out that logistic

regression gave better results with an accuracy of 73.17% compared to all other classifiers. And naïve Bayes performed the worst with an accuracy of 61.13%. The time complexity is reduced with the feature selection method. Also, the classifiers were contrasted with an ensemble model which gave an accuracy of 71.18% so, it is concluded that the individual conventional classifier performed better for the retrieved dataset comparing to stacking classifier. A combination of more advanced algorithms can be used with the tuning of hyperparameters which can produce more customized results. Moreover, the studies can be extended for multi-class classification and predicting the stages of liver disease.

References

1. Rajeswari, P., Reena, G.: Analysis of liver disorder using data mining algorithm. *Glob. J. Comput. Sci. Technol.* **10**, 48–52 (2010)
2. Ramana, B., Surendra, P., et al.: A critical study of selected classification algorithms for liver disease diagnosis. *Int. J. Database Manage. Syst.* **3**(2), 101–114 (2011)
3. Karthik, S., Priyadarshini, A., et al.: Classification and rule extraction using rough set for diagnosis of liver disease and its types. *Adv. Appl. Sci. Res.* **2**(3), 334–345 (2011)
4. Vohra, R., Rani, P., et al.: Liver patient classification using intelligent techniques. *Int. J. Comput. Sci. Inf. Technol. (IJCSIT)* **5**(4), 5110–5115 (2014)
5. Mitra, M.: Machine learning models for predicting the diagnosis of liver disease. *Koomesh* **16**(1) (2014)
6. Vijayarani, S., Dhayanand, S.: Liver disease prediction using SVM and naïve bayes algorithms. *Int. J. Sci. Eng. Technol. Res. (IJSER)* **4**(4), 816–820 (2015)
7. Singh, A., Pandey, B.: Diagnosis of liver disease by using least squares support vector machine approach. *Int. J. Healthc. Inf. Syst. Inf. (IJHISI)* **11**(2), 62–75 (2016)
8. Antony, D., et al.: Dimensionality reduction using genetic algorithm for improving accuracy in medical diagnosis. *Int. J. Intell. Syst. Appl.* **8**(1), 67 (2016)
9. Orczyk, T., Porwik, P.: Liver fibrosis diagnosis support system using machine learning methods. In: Advanced Computing and Systems for Security, pp. 111–121. Springer, New Delhi (2016)
10. Dataset: <https://archive.ics.uci.edu/ml/machine-learning-databases/00225/>
11. Stacking: https://en.wikipedia.org/wiki/Ensemble_learning

Chapter 9

CryptedWe: An End-to-Encryption with Fake News Detection Messaging System



Anukampa Behera, Bibek K. Nayak, Saswat Subhadarshan, and Nilesh Nath

Abstract In the current pandemic situation while world has become overwhelmingly dependent on the online platform for communication, a big concern is raised over corroboration of privacy of data shared and receiving trusted information over popular messaging applications. Fake news circulation has become a prime concern whenever any emergency situation arises nationally as well as globally. In this work, a chatting engine model is proposed, which not only provides popular chatting software features with end-to-end encryption but also a monitored environment to control the spread of fake news. The application “CryptedWe” that is developed based on the proposed model reinforces safety over the privacy issues and makes the chat available only to the trusted peoples as well as restricts the forwarding of fake messages. This application is best for small to medium user groups where authenticity of the communication and privacy aspects are given the maximum priority.

9.1 Introduction

In the recent times the whole world passing through a very difficult situation on one hand fighting with deadly disease COVID-19 and its multiple variants, while on the other hand, many countries are in disarray, many are dealing with their unstable political and leadership scenario. In such circumstances it is very important that the people are well informed with the facts about the situation around them, various safety as well as precautionary measures needs to be taken, etc. But the rampant spread of fake messages through popular messaging applications and social media platforms has raised the concern of researcher’s throughput the world [1]. Many serious and fatal cases are reported due to the self-medication followed by people inspired by the fake treatments suggested; many riots and public agitation are also noticed, where people are incited with fake messages. Also serious threat to one’s privacy is reported by hacking of the private messages in messaging applications.

A. Behera (✉) · B. K. Nayak · S. Subhadarshan · N. Nath

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

e-mail: anukampabehlera@soa.ac.in

The remaining of the paper is organized as follows: in section the motivation for the proposed work along with the objectives are given. In Sect. 9.3, proposed chatting model is discussed followed by development of *CryptedWe* application and analysis in Sect. 9.4. In Sect. 9.5, the paper is concluded with the scope for future work.

9.2 Motivation and Objective

Recently when messaging applications are gaining popularity in a lightning speed, security of the data getting exchanged and the genuineness of the information propagated have become of utmost concern. The systematic study limitations in two very popular messaging applications namely, WhatsApp and Telegram has acted as the prime motivation behind the proposed work.

- Lack of fake news circulation control: WhatsApp does not have any control over spreading of misinformation or matters that might exploit human psychology. Attacks such as FakeApp can be used to spread fake statements in the name of legitimate users by misusing the “quote” feature available in WhatsApp groups.
- Media File Jacking attack: It exploits the mechanism of storage and retrieval of media files from a device’s external storage. In here malware is concealed inside any application seemingly harmless and once installed it keeps a watch on all the incoming files to replace the original attachments with fake ones [2].
- Lack of Anonymity: Telegram protocol uses conventional methods like centralized server-based user authentication which acts instrumental in removal of anonymity as the origin if the message can be traced and known to the server. But this becomes a serious bottleneck when privacy in a chat is claimed [3].

Keeping in consideration the security concerns the need of the hour is to develop a secured application with the following prime features.

- End-to-end encryption is implemented for security where none of the user communication is stored in the server that makes the communication hack-proof.
- Fake news detection engine that restricts circulation of such messages.

9.3 Proposed Messaging Application—*CryptedWe*

In tune with the objective of developing a full proof messaging application ensuring data security and restricting circulation of fake messages; in this paper, a messaging application that implements end-to-end security with added facility of fake message detection is proposed. Based on the model a chatting application “*CryptedWe*” is developed. This section is divided into two parts. In part—1, the encryption technique used in the proposed application is discussed and in part—2, the fake news detection method is discussed.

9.3.1 End-to-End Encryption

To ensure security if data in case of a highly confidential communication, end-to-end communication is a much sought after technique. Some of the widely accepted encryption techniques those widely used for messaging applications are: Triple Data Encryption Standard (DES) which is accepted worldwide as the recommended standard, but it is very time consuming [4]. Advanced Encryption Standard (AES) [5] is considered to be very effective to all most all cryptographic attacks and are available for different key and block length of 128, 192 or 256-bits. RSA security encryption technique [6] is considered as the most widely accepted asymmetric encryption algorithm based on prime factorization. Blowfish algorithm is another widely used symmetric encryption algorithm which is the most sought after encryption technique because of its speed and efficiency [7]. In Twofish encryption the keys can be scaled up to 256-bits having the requirement of only one key as it comes under symmetric encryption technique [8]. The principle of irreversibility is adopted by ECC Asymmetric Encryption Algorithm [9] which makes it too complex and difficult to gain back the original points once they are encrypted.

AES-256 encryption algorithm has been used for the purpose of end-to-end encryption in the proposed work which is described below.

AES-256 encryption technique:

In Advanced Encryption Standard (AES) algorithm all computations are performed on Bytes. It takes 128 bits of plain input text for processing as a 16 byte block which is further arranged in the form of a 4X4 matrix [10]. In a 256-bit keys encryption 14 rounds of operation is performed. The unique 128-bit round key that is obtained as the calculation performed on the original AES key is used in each of the round. These 14 rounds are completed under three phases of operation namely, the initial, main and final rounds where the same sub-operations are performed in various combinations. In AES 256, 13 rounds of operations are covered under main round and the last one is completed in the final round. The details of AES 256 workflow is depicted in Fig. 9.1.

The subsections performed in various rounds are described below:

1. **AddRoundKey:** This is the first and only operation performed in the initial round which never repeated in other rounds. In this round *RoundKey* is input which is Exclusive ORed with the input 16 byte block.
2. **SubBytes:** This is the Byte Substitution step where the input block is split into 16 individual bytes and is passed into a Substitution Box or S-Box lookup table.
3. **ShiftRows:** In this phase, shifting is applied to the 128-bit internal state of the cipher. Each row in the 4X4 matrix created for the internal state representation of AES is referred to as *row* in *ShiftRows* state. In this process (i) No shifting is applied to the first row, (ii) one position (one byte) is shifting toward left is applied on the second row, (iii) two positions shifting is applied on third row and (iv) Three positions shifting is applied to the fourth row. The result is a new 4X4 matrix but shifted with respect to each other.

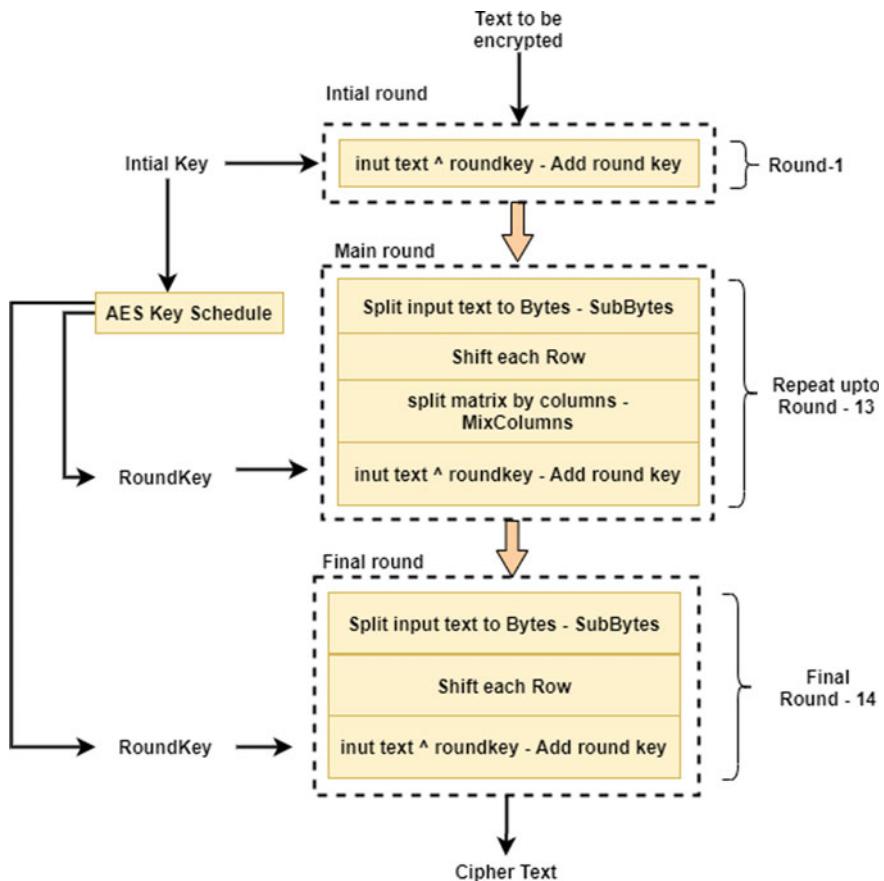


Fig. 9.1 AES-256 encryption work flow

4. **MixColumns:** In this phase, the columns of the 4X4 matrix are passed to a special mathematical function one by one. It produces four entirely new bytes for each column which replaces the original column values in the matrix. This process is repeated for 13 rounds only and skipped in the last round.
5. **AES Key Schedule:** It accepts the initial key as the input and produces desired number of round keys based on the type of AES encryption, i.e., in AES 256 it produces 14 round keys.
6. **Decryption:** The encryption process performed in reversed order in the decryption phase.

9.3.2 *Fake News Detection*

The main objective of fake news detection is to prohibit the circulation of malevolence messages like mob-lynching incidents to provoke unwarranted violence and agitation amongst public, especially in the current scenario circulation of false information regarding imposition and withdrawal of restrictions during lockdown, treatment measures for COVID-19, etc. are leading to hazardous situations. A detailed flow diagram adopted in the proposed work is shown in Fig. 9.2.

Step-1: Creation of fake news dataset: As there is no readily available dataset for fake news, the dataset need to be created first. In this concern, various fake news those are in circulation are collected with the help of web crawler from

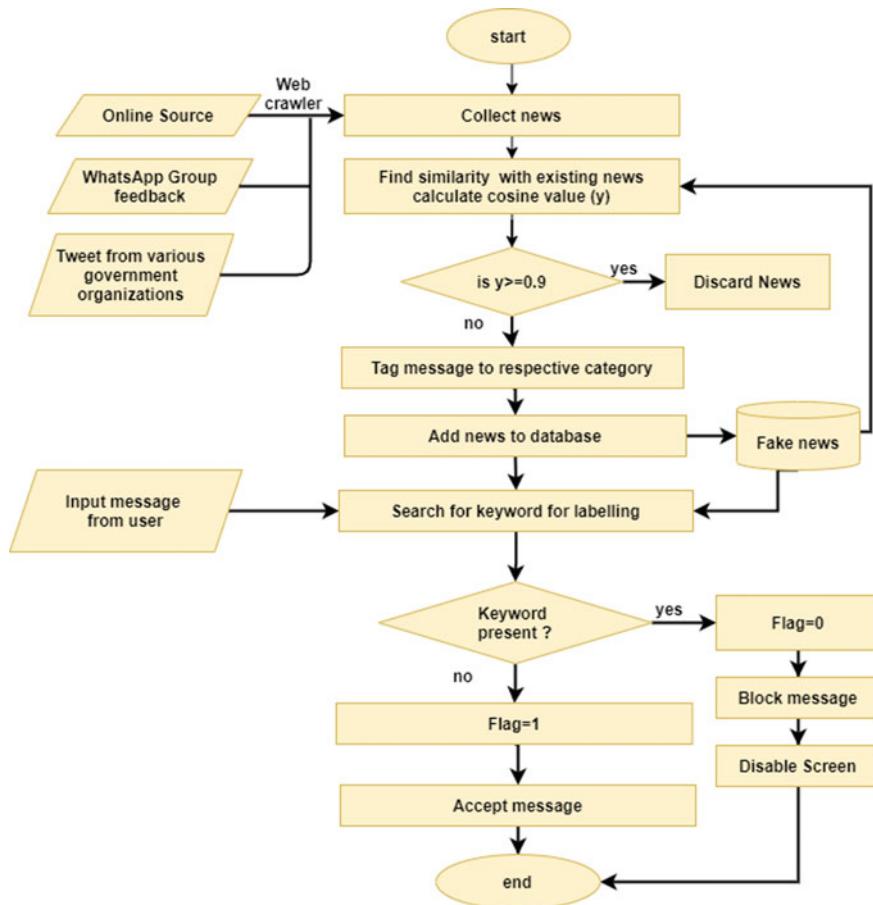


Fig. 9.2 Detail flow of fake new detection model

various online resource, tweets those declare a message as fraud by government and feedbacks are also collected from various WhatsApp groups.

Step-2: Discarding repeated or similar news: Calculation of the similarity between the newly found news and existing news with the help of cosine function is done. The cosine value is checked for a threshold value y . If y is more than 0.9 then the news is accepted otherwise it is rejected. A threshold value more than or equal to 0.9 represents the angle between two participating vectors to be at least 90 degrees which indicates they are dissimilar.

Step-3: Tagging of messages: Once a unique message is found it is tagged based on the category obtained from the source like political, social, health, entertainment, etc.

Step-4: Add news to the database: The tagged message is added to the existing database of fake messages.

Step-5: Label message: Once the user tries to send/forward a message using chatting application, the accepted message is checked with the presence of sequence of words matching with the existing database of fake news. If the match is found the label indicator is turned to be 0 which acts as a signal to block the message as well as disable the chatting screen of the user with “Fake News” alert. In the other hand if no match is found then the label indicator is turned to be 1 and then message is delivered to the corresponding receiver.

9.4 Application Development and Analysis

The proposed work is implemented using Java, where JDK 11.0.10 have been used for project layout and Java swing is used for UI building. The latest version Netbeans 12.3 which has updated features of background color adjustments, adding image in UI and, etc. is used as the IDE for experimentation purpose. Netbeans is preferred as it is relatively easy to build a UI using it and it also uses latest tools for UI formation as well as project deployment.

The implementation creates two databases; first one is used to stores the login/registration credentials only and second one is to store the fake news. None of the private messages of the user, his/her IP address are not stored in anywhere in “*CryptedWe*” application; neither they are sent to or shared with any other servers. Credentials like name, living area, phone number, email id and password which can be also used to retrieve account in case user forgets password. The application starts with a welcome page having two options, login or register. If a user have already signed up then he/she needs to login directly and can start chatting, but in the case of a new user he/she has to register first and fill up very common credentials. The second layer of security is provided by approving the registration through verified OTP sent to the given phone number.

After successfully logging in, user comes to chatting interface where he/she needs to connect to server. Once the connection is established, the user is ready to start chatting with other users. User has functionality like connect/disconnect from server

which will, respectively, show in server monitoring screen that how many online user are connected.

Once connected, message exchanged is encrypted using AES-256 encryption algorithm. First when user chats the plain text gets encrypted to cipher text in server frame and from server frame it gets converted to plain text featuring decryption technique. Being a symmetric algorithm; ciphers use the same key for decryption and encryption. The encryption and decryption process in server and respective clients are shown in Fig. 9.3.

Whenever a user wants to send any attachment or share any file a browse screen appears as shown in Fig. 9.4. Next with the fake news detection algorithm, if this is matched with a fake message then the entire screen is disabled with a “fake document” alert as shown in Fig. 9.5.

9.4.1 *Product Highlights*

The following are attributed as the major highlights of the product “*CryptedWe*”:

- The user interfaces has been kept simple that has made the application light weighted for lesser data consumption.
- Except user information no other data is stored in the database which makes the application work really faster and makes it secured.
- End-to-end encryption is implemented for data security.
- Fake news detection and alert is implemented prohibiting the user from forwarding any unsolicited message or wrong information.

9.5 Conclusion and Future Work

Amidst the recent pandemic situation a major paradigm shift have been noticed in the work culture to carry out majority of the job online. This makes most of the administrative and information technology based organizations to heavily depend on chatting software for their communication medium. In such a situation not only the security and privacy of the data communicated must be guaranteed but also the sanity of the information getting shared and forwarded must be preserved.

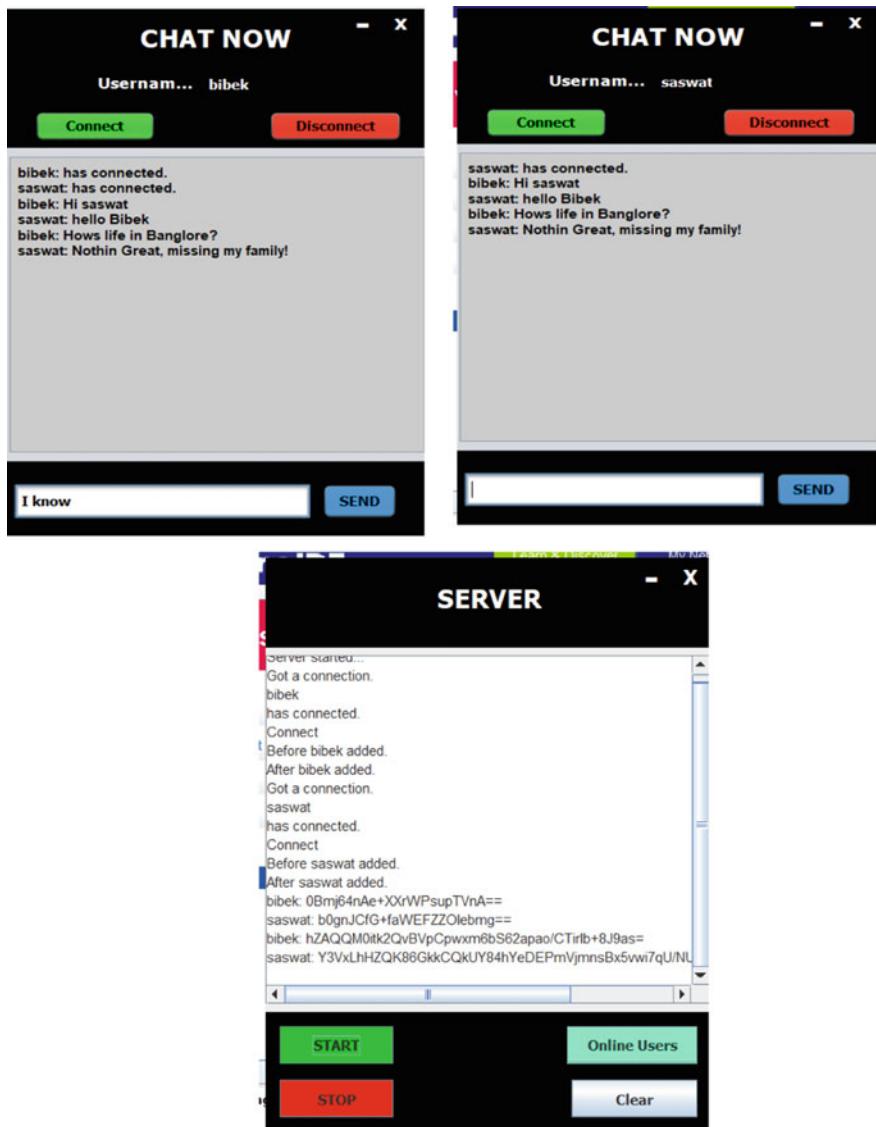


Fig. 9.3 Encryption in server frame and decryption in user frame

The proposed messaging model attains the objectives of creating a secured messaging application with scrutinized message forwarding. Security is ensured by applying end-to-end encryption to the messages sent and received catering to the need of keeping one's privacy intact while chatting. Based on the proposed model, in the chatting application “*CryptedWe*” also guarantees the privacy. Here, not only the messages sent/received are encrypted also none of the communication that is taking

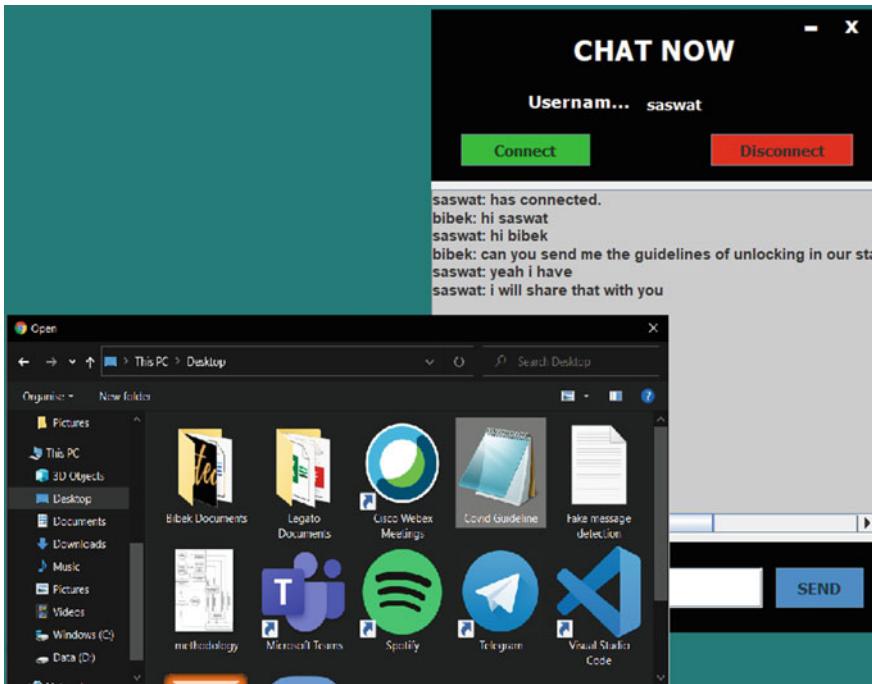


Fig. 9.4 Selecting files to be forwarded

place within users is stored in the server which prevents it from any data theft. At the same time the forwarding and attachment of fake messages are restrained using the proposed fake news detection algorithm. The entire chat session is disabled responding to the issue of forwarding fake messages.

Our future upgrade will include auto reporting of any anti-social activities planned over chat not by anyways tampering with privacy but with detection of series of words in sequence while typing the message.

Fig. 9.5 Chat screen getting closed with a fake alert



References

1. Nath, G., Adhi, G., Thandul, P.: An Attempt to Detect Fake Messages Circulated on WhatsApp (2020)
2. Xavier, K.: Cryptography Used in Whatsapp. ScienceOpen Posters (2020)
3. Anglano, C., Canonico, M., Guazzone, M.: Forensic analysis of Telegram Messenger on Android smartphones. Digital Invest. **23**, 31–49 (2017). ISSN 1742–2876. <https://doi.org/10.1016/j.diin.2017.09.002>
4. Agrawal, M., Mishra, P.: A comparative survey on symmetric key encryption techniques. Int. J. Comput. Sci. Eng. **4**(5), 877 (2012)
5. Ahmad, N., Wei, L., Jabbar, M.: Advanced encryption standard with galois counter mode using field programmable gate array. J. Phys. Conf. Ser **1019**, 012008 (2018). <https://doi.org/10.1088/1742-6596/1019/1/012008>
6. Mahajan, P., Sachdeva, A.: A study of encryption algorithms AES, DES and RSA for security. Glob. J. Comput. Sci. Technol. (2013)
7. Nie, T., Song, C., Zhi, X.: Performance evaluation of DES and blowfish algorithms. In: 2010 International Conference on Biomedical Engineering and Computer Science (2010). <https://doi.org/10.1109/icbeics.2010.5462398>
8. Rizvi, S.A.M., Hussain, S.Z., Wadhwa, N.: Performance analysis of AES and TwoFish encryption schemes. In: 2011 International Conference on Communication Systems and Network Technologies. (2011). <https://doi.org/10.1109/csnt.2011.160>

9. Bafandehkar, M., Yasin, S.M., Mahmod, R., Hanapi, Z.M.: Comparison of ECC and RSA algorithm in resource constrained devices. In: International conference on IT convergence and security (ICITCS), pp. 1–3 (2013). <https://doi.org/10.1109/ICITCS.2013.6717816>
10. Santoso, K., Muin, M., Mahmudi, M.: Implementation of AES cryptography and twofish hybrid algorithms for cloud. J. Phys. Conf. Ser. **1517**, 012099 (2020). <https://doi.org/10.1088/1742-6596/1517/1/012099>

Chapter 10

Enabling Data Security in Electronic Voting System Using Blockchain



M. Thangavel, Pratyush Kumar Sinha, Ayusman Mishra,
and Bhavesh Kumar Behera

Abstract Electronic voting is essential in the digital world for voters, who will make their choice of vote through smartphones or computers. Ballot composition, ballot casting, ballot recording, and tabulation are the four steps in the election process. Electronic voting technology aims to speed up ballot counting, minimize the expense of hiring employees to tally ballots manually and increase accessibility for disabled voters. Results can be reported and published within a short period. The major challenge in electronic voting is to maintain data confidentiality, verify data integrity, and ensure authenticity in the overall process. In this paper, blockchain-enabled electronic voting has been proposed, addressing security challenges such as confidentiality, integrity, and authenticity. The proposed work implementation results show the effectiveness of blockchain-based electronic voting scheme.

10.1 Introduction

Information system is used to enable control and decision-making in an organization by collecting, processing, storing, and distributing data from interconnected components. It is made up of hardware, software, and communication networks. For enabling voters to cast a secret ballot through Internet, electronic voting process is introduced through electronic media. The voting process in practice is paper-based voting, which is non-flexible, inconvenient, and time-consuming. It accelerates the entire voting process with less cost investment (workforce and printed materials required for conducting poll are reasonably reduced). This type of computer-mediated voting will be more flexible for the voters because it permits voters to poll from any part of the world through Internet.

M. Thangavel (✉)

School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh 466114, India

P. K. Sinha · A. Mishra · B. K. Behera

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

It also supports the authorities by enhancing the registration process, centralizing registration databases, and allowing voters to check their registration status before the vote. Because of electronic voting, there is a chance of reducing errors by neglecting vote tampering and reducing over-votes for decreasing the number of legitimate ballots not counted). Data security and handling data in a distributed environment for electronic voting are still challenging issues.

Blockchain is a decentralized, distributed ledger that tracks the origins of digital assets. It is a database that saves encrypted data blocks and then links them together to provide a single source of truth for the data. Because the asset is decentralized, the public has real-time access and transparency. By using a distributed consensus mechanism, algorithm based on encryption, point-to-point transmission, and smart contracts, blockchain may create trustworthy peer-to-peer communication for securing vital information. It also safeguards all data transmissions between devices.

To secure the electronic voting information system from any internal and external attacks, confidentiality, integrity, and authenticity need to be ensured. Blockchain has the potential to boost service delivery quality as it enhances data integrity and confidentiality. It makes the transactions secure and transparent. Decentralization and accompanied capabilities of the blockchain make it the ideal component to ensure security. The cryptographic algorithms used by this technology make data more private.

Our contribution is to develop an electronic voting information system using blockchain by ensuring data security by addressing confidentiality, integrity, and authenticity issues. The rest of the paper has been organized as follows: Sect. 10.2 discussed existing research works on the electronic voting information system. Section 10.3 explains the proposed framework of the electronic voting system using blockchain. Section 10.4 presents the experimental analysis of the proposed framework. Section 10.5 summarizes the work discussed and directs the further enhancements of the proposed framework.

10.2 Related Works

Researchers proposed various schemes for security enhancement in electronic voting systems by addressing confidentiality integrity and authentication issues using and without using Blockchain. Existing research works address the above-mentioned security issues separately but not in combined form. A detailed discussion of the existing system has been provided in this section.

Fusco et al. [1] proposed to enhance the total security of an electronic voting application by providing the solution with Shamir's secret sharing methodology using blockchain. Aspects like authenticity and non-repudiation are not given enough attention. Park et al. [2] address the difficulty of using blockchain-based crowdsourcing on a cloud computing platform. Information trustworthiness is raised through widespread replication of blockchains.

Daramola et al. [3] perform an in-depth investigation of blockchain-based electronic voting infrastructures. The proposed scheme is more resistant to DDOS attacks, which can be executed on a large scale during a national election and after a few other essential upgrades. As a result, it appears that blockchain-based electronic voting architecture is more immune to DDOS attacks than centralized systems. Yi [4] concerns various security elements of electronic voting such as confidentiality, integrity, authenticity, and non-repudiation. The synchronized model of voting records has been proposed based on distributed ledger technology (DLT) to avoid fraudulent votes.

Dhulavvagol et al. [5] made a comparison between Geth and Parity Ethereum. The proposed solution enables the construction of a secure blockchain network to establish peer-to-peer network security. Hjálmarsson et al. [6] address the limitations of the existing electronic voting systems. The proposed system showcases the advantages of distributed electronic voting systems that use “permissioned blockchain,” which assists in overcoming the limitations and hurdles of electronic voting systems and providing additional measures for increased throughput.

Mehboob et al. [7] address the difficulty in achieving widespread adoption of an e-voting system. The solution presented is a well-established real-world e-voting strategy based on the multi-chain platform. The development of an end-to-end verifiable e-voting scheme that increased blockchain technology’s resistance to the problem of “double voting.” Subhash et al. [8] concerned about the security vulnerabilities of the electronic voting system. The solution is proposed with a distributed ledger technology called blockchain. The benefits of the proposed scheme are more cost-effective and time-saving election process and better support for more complicated applications.

Yavuz et al. [9] focus on the lack of transparency, authentication, and provability in ordinary digital services and e-voting platforms. The benefits of introducing smart contracts to the Ethereum platform are numerous. Blockchain has been transformed into a bigger solution-base for many Internet-related concerns, and it has been improved. Chen et al. [10] discuss a blockchain-based data integrity verification mechanism for the Internet of things (IoT). With the usage of private blockchain, the system’s efficiency can be improved even more and time saved. Wang et al. [11] focus on employing blockchain technology for protecting health care records stored in the cloud. Blockchain provides a new way to protect the personal health records sharing system, and the benefits are that the user’s healthcare report is safe, and their privacy is increased.

Yatskiv et al. [12] address the issues of protecting the integrity of a video file. The benefit of using blockchain to protect video integrity includes increased privacy and video integrity and a reduction in rendering time through the usage of time-lapse video. Dong et al. [13] addressed the security vulnerabilities in the Internet of things (IoT), which comprises many small devices that share a large amount of data. In order to protect large amounts of valuable data stored in remote cloud services by IoT and to solve the problem of trusted third-party auditor (TPA) in auditing the integrity of data, a secure auditing scheme based on hyperledger is proposed.

Choi et al. [14] address nuclear power plants, which are vulnerable to cyber-attacks. The solution provided is to monitor the data integrity of programmable logic controllers (PLC) using blockchain, with the benefits being protecting nuclear power plants and their safety systems from cyber-attacks. Although blockchain can prevent data tampering, scalability problem occurs when storing supply chains directly on the blockchain. Blockchain Index Storage (BIS) is recommended by Kuo et al. [15]. BIS can reduce the cost of surveillance cameras and sensor nodes, adjusting buffer time to reduce costs. Sharma et al. [16] say that data is an invaluable asset, and an attack on data integrity can affect important organization decisions. The issue stands right in distributed environments as well. This study proposes a peer-to-peer network integrity verification architecture, which uses blockchain technology to improve the unstable traditional verification approach.

Rani et al. [17] address the problem of authentication and confidentiality of authenticated data through a financial wallet system using blockchain. Researchers propose by preventing the loophole of security breaches during the validation process or address lookup. The proposed work uses the AES algorithm for encryption and ZKP algorithm for confidentiality. Dang et al. [18] solve the issue regarding easy leakage of confidential data and attacks. Researchers proposed a multi-dimensional identity authentication system. In the proposed solution, users' identity features and cross-domain authentication are error-free and easier. Cui [19] address the security issue of the Internet of things (IoT). IoT multi-WSN authentication system based on blockchain has been proposed. High performance in computation, storage, and energy consumption is observed. Wu et al. [20] are concerned about the old authentication mechanism incompatible with the power terminal cross-domain authentication mechanism. Researches propose creating a master-slave blockchain-based cross-domain authentication architecture for power terminals.

Xiang et al. [21] are concerned about an effective authentication scheme in electronic health systems. Researchers propose blockchain-based management of identity and user authentication mechanisms with permissions. Wang et al. [22] believe complex processes and privacy leakage during cross-domain authentication of power terminals will become barriers for improving operational efficiency and user experience as the Internet of things (IoT) grows. Researchers propose a Blockchain-based identity identification system. Guo et al. [23] scheme can be applied to data-sharing platforms for verification on a pilot basis and can be optimized further. The study is about a blockchain-based solution that uses computationally hard jobs like the proof-of-work puzzle to verify networks' integrity and validity.

The limitations of the existing system are considered while designing the proposed system for securing data in electronic voting using blockchain.

10.3 Proposed System

The proposed system ensures data protection and reduces the amount of manual labor required with the following functionalities.

- (i) The e-voting system based on the blockchain is open, distributed, and decentralized. It can keep votes cast on a variety of mobile devices and PCs.
- (ii) The blockchain-based e-voting system allows voters to easily audit and validates their ballots.
- (iii) The voting database is self-contained and uses a distributed timestamp server over a peer-to-peer network.
- (iv) Voting on blockchain is a workflow in which voters' concerns about data security are minimal, removing e-property voting's limitless reproduction.

Figure 10.1 presents the proposed system architecture for electronic voting using blockchain. The overall process carried out by election commission and voter are listed as follows:

Election commission:

- The election commission will login after providing correct credentials.
- Perform candidate registration for respective positions.
- Start the election.
- Grant access to voters.
- End the election.
- After the election is over, the election commission can view and declare the result.

Voter:

(i) User Registration:

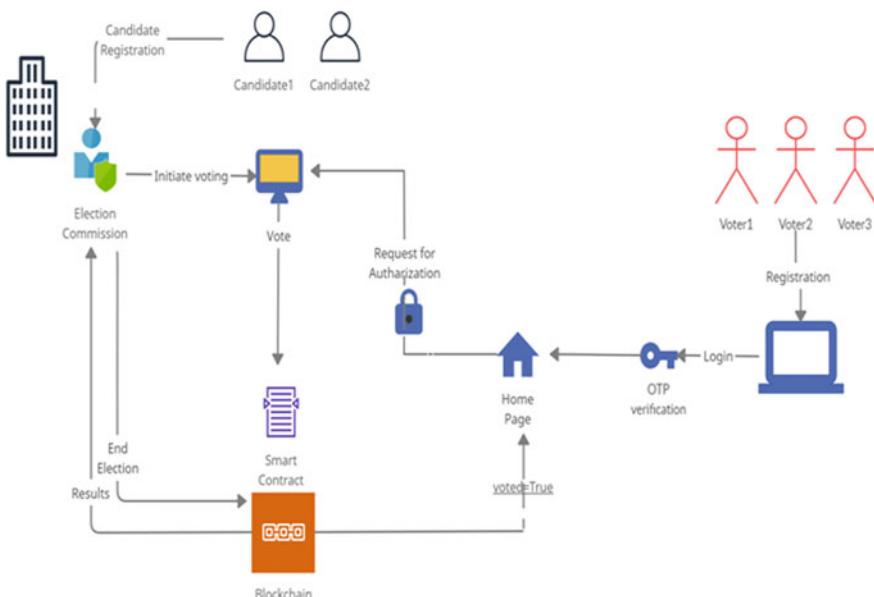


Figure 10.1 Proposed system architecture

This is the first process as a new user/voter. The form will ask users to enter various details like name, Aadhar number, DOB, etc.

- Ask the user for a valid username and password.
- Date of Birth checking is done, i.e., a user is only allowed to register if his/her age is above 18 years.
- The constituency of the user is taken from the Aadhar which the user enters.

(ii) **User Login:**

This attribute allows the user to login into the system by putting in the right credentials. After entering the right ID and password, the user is verified and logged into the system.

- The person registering should get a user id when they sign up
- The system should allow the login of only those users who have given valid credentials.
- The system performs authorization process, i.e., OTP is sent to the user's registered mail ID and
- After successful verification, the user gets access to the home page.
- The user must be able to log out.

(iii) **Vote Phase:**

Once the user login to the system, they will be able to vote after they get authorized by their respective constituency election commission.

- Check the status of the authorization.
- Request for authorization after checking the status.
- Cast vote after access granted by the election commission.

Figure 10.2 represents the control flow of the proposed system. The steps carried out in the electronic voting are as follows:

- Step1. The voter will register by giving details like name, email, phone number etc.
- Step2. The voter will have to login with an email, Aadhar number, and password.
- Step3. After login, OTP verification needs to be performed. An alphanumeric code will be sent to the registered email.
- Step4. After successful OTP verification, the voter will be redirected to the home page.
- Step5. Check for Authorization status.
- Step6. If authorization status is false, request the respective election commission for authorization.
- Step7. On successful authorization, the voter will be redirected to the voting page.
- Step8. The voter can cast their vote.
- Step9. After a successful vote cast, the voter will be redirected to the home page after 5 s.

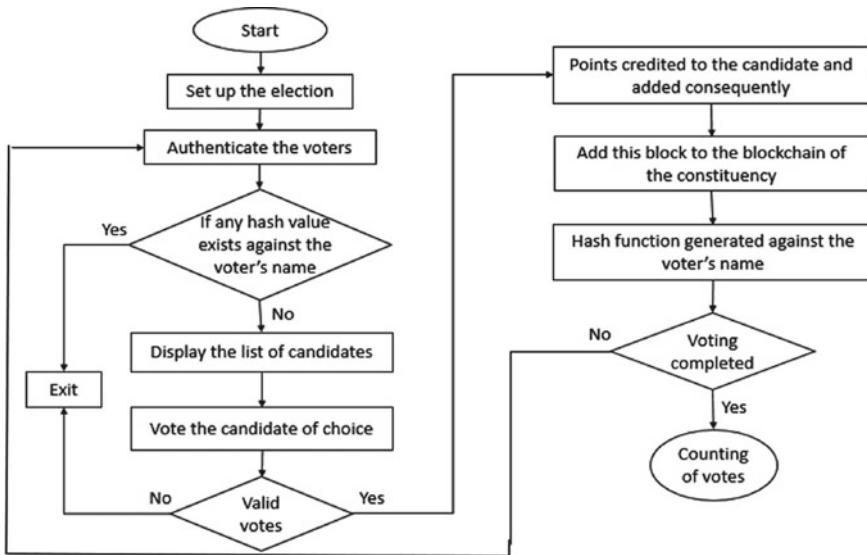


Figure 10.2 Proposed system—Control flow

Step10. The voting process is completed.

In the proposed electronic voting system, the security issues are handled as such in the voting process flow. After successful voting by the voter, the voting point relevant to the candidate will be increased and added as a block in the blockchain. After the voter does successful voting, the voting data is considered to be confidential. The voting data will be visible to the corresponding voter only for five seconds. Then, the voting count of the respective candidate will be increased and added as a block in the blockchain of the candidate. Since each vote casting will be performed in the same sequence, the attacker cannot view or modify the data stored in the block. Because of these action sequence, the data confidentiality and integrity has been ensured. Blockchain will be managed by the election commission only. After successful completion of the election, the election commission may verify the blocks of each candidate and declare the results of the election. The voter has a two-step verification process for ensuring the identity of the user. By this action sequence, authentication and data availability have been ensured.

10.4 Implementation

The proposed electronic voting system using blockchain has been implemented in Ganache—Ethereum blockchain environment. Figure 10.3 illustrates the design and development platform of the proposed system.

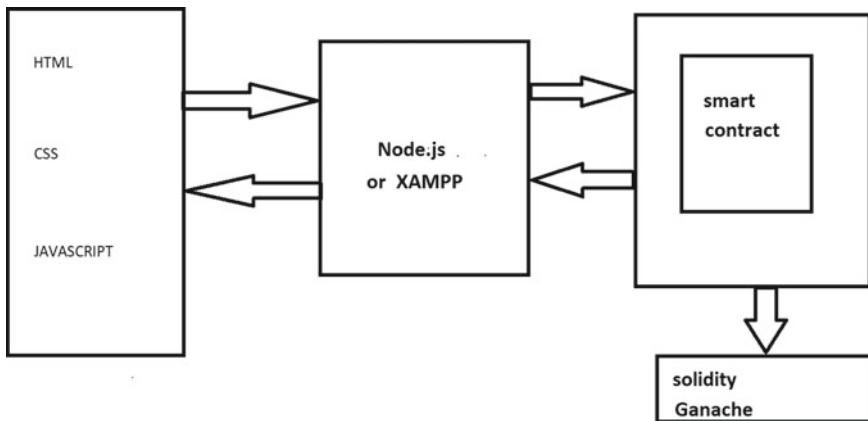


Figure 10.3 Proposed system—design and development platform

The Web-based electronic voting application has been designed with HTML, CSS, and JavaScript. Remix and Truffle IDE have been used to write smart contracts. Ganache has been used to process the blockchain concepts with the frontend Web design (Fig. 10.4). Metamask has been used as a wallet for performing the blockchain process (Fig. 10.5).

The smart contract of the functionalities of the proposed electronic voting system has been written in solidity language. The solidity code has been tested with the unit testing feature in remix IDE (Fig. 10.6).

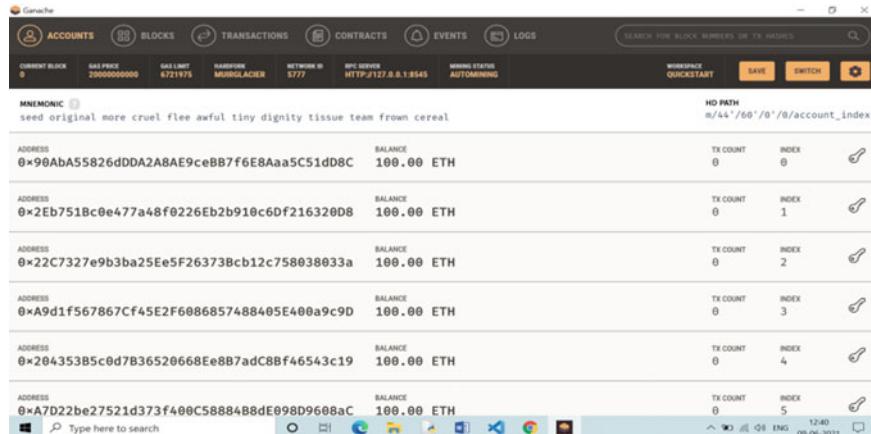


Figure 10.4 Ganache

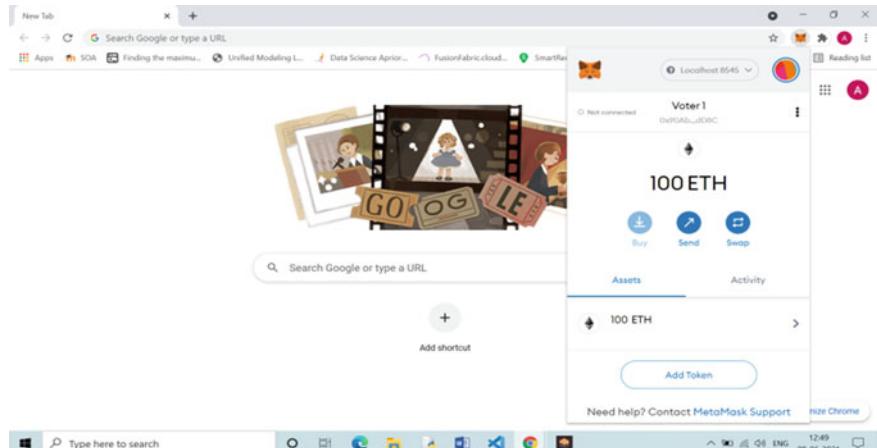


Figure 10.5 Metamask

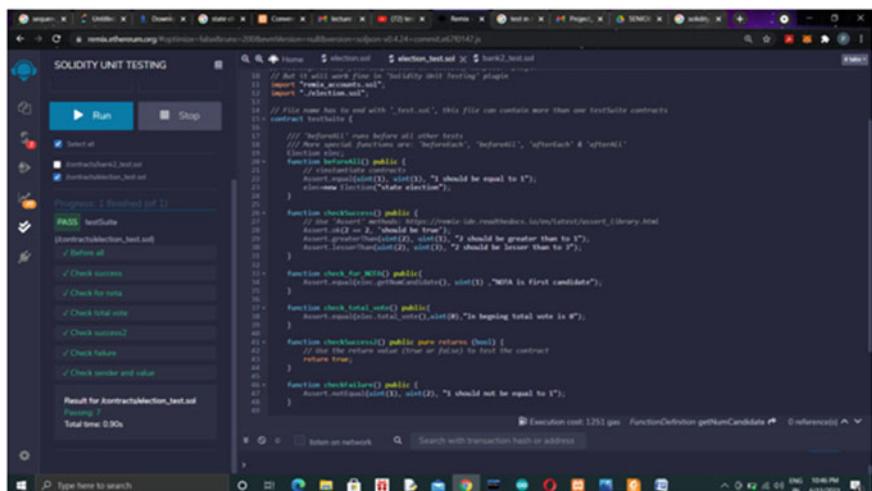


Figure 10.6 Solidity testing

10.5 Conclusion

Existing research focuses on only one feature of security out of the CIA (confidentiality, integrity, and authenticity) triad. As the proposed idea chosen is to work on an electronic voting application, the work attempts to enhance all the security features of information systems and distributed environments in-depth. The proposed system enhances all the security features, namely “confidentiality,” “integrity,” and “authenticity,” and makes the system scalable to a larger extent. The proposed model for

implementing the electronic voting application uses the blockchain technology to store the voting phase details, which will ensure confidentiality and cater to the need for “secret voting.” The processes carried out in a general election in India have adhered. All phases of the electoral processes are digitalized, except some which had to be done manually. The system is scalable enough, yet it can be extended to a mass level by utilizing better Ethereum networks, and the main blockchain instead of the private blockchain is used for development purposes. In the future, the application needs to be extended to a larger scale by using the “Ethereum Mainnet” network and “HyperLedger Fabric” for developing Blockchain and writing the smart contracts, respectively. Furthermore, the interactivity of the application needs to be improvised by using other advanced tools and technologies such as ReactJS and ExpressJS.

References

1. Fusco, F., Lunesu, M.I., Pani, F.E., Pinna, A.: Crypto-voting, a Blockchain based e-Voting System. In: 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (2018)
2. Park, J., Park, S., Kim, K., Lee, D.: CORUS: Blockchain-based trustworthy evaluation system for efficacy of healthcare remedies. In: IEEE International Conference on Cloud Computing Technology and Science (CloudCom), pp. 181–184 (2018)
3. Daramola, O., Thebus, D.: Architecture-centric evaluation of blockchain-based smart contract E-voting for national elections. *Informatics* **7**(2), 16 (2020)
4. Yi, H.: Securing e-voting based on blockchain in P2P network. *EURASIP J. Wirel. Commun. Network.* **137** (2019)
5. Dhulavvagol, P.M., Bhajantri, V.H., Totad, S.G.: Blockchain Ethereum clients performance analysis considering E-voting application. In: International Conference on Computational Intelligence and Data Science (2019)
6. Hjálmarsson, F.P., Hreiðarsson, G.K., Hamdaqa, M., Hjálmtýsson, G.: Blockchain-based E-voting system. In: IEEE 11th International Conference on Cloud Computing (CLOUD), pp. 983–986 (2018)
7. Mehboob, K., Arshad, J., Khan, M.: Secure digital voting system based on blockchain technology. *Int. J. Electron. Gov. Res.* **14**, 53–62 (2018)
8. Yadav, A.S., Urade, Y.V., Thombare, A.U., Patil, A.A.: E-voting using blockchain technology. *Int. J. Eng. Res. Technol. (IJERT)* **09**(07) (2020)
9. Yavuz, E., Koç, A.K., Çabuk, U.C., Dalkılıç, G.: Towards secure e-voting using Ethereum blockchain. In: 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–7 (2018)
10. Chen, Y., Wang, L., Wang, S.: Stochastic blockchain for IoT data integrity. *IEEE Trans. Net. Sci. Eng.* **7**(1), 373–384 (2020)
11. Wang, S., Zhang, D., Zhang, Y.: Blockchain-based personal health records sharing scheme with data integrity verifiable. *IEEE Access* **7**, 102887–102901 (2019)
12. Yatskiv, V., Yatskiv, N., Bandrivskyi, O.: Proof of video integrity based on Blockchain. In: 9th International Conference on Advanced Computer Information Technologies (ACIT), pp. 431–434 (2019)
13. Dong, G., Wang, X.: A secure IoT data integrity auditing scheme based on consortium Blockchain. In: 5th IEEE International Conference on Big Data Analytics (ICBDA), pp. 246–250 (2020)
14. Choi, M.K., Yeun, C.Y., Seong, P.H.: A novel monitoring system for the data integrity of reactor protection system using blockchain technology. *IEEE Access* **8**, 118732–118740 (2020)

15. Kuo, S.-S., Su, W.-T.: A blockchain-indexed storage supporting scalable data integrity in supply chain traceability. In: IEEE International Conference on Smart Internet of Things (SmartIoT), pp. 348–349 (2020)
16. Sharma, P., Jindal, R., Borah, M.D.: Blockchain-based integrity protection system for cloud storage. In: 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), pp. 1–5 (2019)
17. Rani, P.J., M. J.: Authentication of financial wallet system and data protection using blockChain. In: IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), pp. 1–5 (2020)
18. Dang, F., Gao, F., Liang, H., Sun, Y.: Multi-dimensional identity authentication mechanism for power maintenance personnel based on Blockchain. In: International Wireless Communications and Mobile Computing (IWCMC), pp. 215–219 (2020)
19. Cui, Z., et al.: A Hybrid BlockChain-based identity authentication scheme for multi-WSN. *IEEE Trans. Serv. Comput.* **13**(2), 241–251 (2020)
20. Wu, C., Lu, J., Li, W., Meng, H., Ren, Y.: Master-slave Blockchain based cross-domain trust access mechanism for UPIOT. In: 5th International Conference on Computer and Communication Systems (ICCCS), pp. 498–503 (2020)
21. Xiang, X., Wang, M., Fan, W.: A permissioned Blockchain-based identity management and user authentication scheme for E-health systems. *IEEE Access* **8**, 171771–171783 (2020)
22. Wang, X., Gao, F., Zhang, J., Feng, X., Hu, X.: Cross-domain authentication mechanism for power terminals based on Blockchain and credibility evaluation. In: 5th International Conference on Computer and Communication Systems (ICCCS), pp. 936–940 (2020)
23. Guo, S., Hu, X., Guo, S., Qiu, X., Qi, F.: Blockchain meets edge computing: A distributed and trusted authentication system. *IEEE Trans. Industr. Inf.* **16**(3), 1972–1983 (2020)

Chapter 11

Prediction of Used Car Prices Using Machine Learning



Dibya Ranjan Das Adhikary, Ronit Sahu, and Sthita Pragyna Panda

Abstract As the Indian auto-industry entered BS-VI era from April 2020, the value proposition of used cars grew stronger, as the new cars became expensive due to additional technology costs. Moreover, the unavailability of public transport and fear of infection force people toward self-mobility during the outbreak of Covid-19 pandemic. But, the surge in demand for used cars made some car sellers to take advantage from customers by listing higher prices than normal. In order to help consumers aware of market trends and prices for used cars, there comes the need to create a model that can predict the cost of used cars by taking into consideration about different features and prices of other cars present in the country. In this paper, we have used different machine learning algorithms such as k-nearest neighbor (KNN), random forest regression, decision tree, and light gradient boosting machine (LightGBM) which is able to predict the price of used cars based on different features specific to Indian buyers, and we have implemented the best model by comparing with other models to serve our cause.

11.1 Introduction

The used cars market is growing at a fast pace. With digitization and telecom revolution brought by Reliance JIO in 2016, it has broken the barrier of getting internet at high price. People today have enough Internet data to spend and which leads to the growth of online communities and YouTube channels in our country. Thanks to bloggers covering used cars market as a profession and car resellers getting contracts from all across the nation. Online sites and applications like cardekho.com, cars24.com, and many more have boosted the used cars market by giving customers as well as car sellers the information that they require and filled the gap. But, still as our country is large both in terms of population and area, there is still a long way to cover in this segment of market. During the period (2020–2025), the used automobile market in

D. R. Das Adhikary (✉) · R. Sahu · S. Pragyna Panda

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan, Deemed To Be University, Bhubaneswar, Odisha, India

e-mail: dibyadasadhikary@soa.ac.in

India is predicted to increase at a compound annual growth rate (CAGR) of 15.12% [1]. The reduction of the GST on used automobiles from 28% to (12–18%) is one of the market's major growth drivers. Even after the first lockdown of Covid-19 pandemic, the used cars industry bounced back 99% compared to 77% of the new cars market. With the rise of the organized and semi-organized sectors, the country's used car industry has evolved. In FY 2018, the pre-owned or used automobile market surpassed 4million units, indicating that the used car market is 1.3 times the new car market. With the implementation of the new BS-VI emission requirements by the Indian government, consumers will be unable to afford the technological costs of vehicles that satisfy the regulations. Due to their business approaches and the rising number of pre-owned car retail shops, the used automobile market in India is consolidated, with important players owning the majority share of the industry. In the local market, there are a variety of outlets and enterprises that are targeting the local people like people of their own town and nearby villages. The unorganized sector accounts for roughly 80% of the total used cars market.

The surge in demand for used cars made some car sellers to take advantage from customers by listing higher prices than normal. To help consumers aware of market trends and prices for used cars, there comes the need to create a model that can predict the cost of used cars. Our problem definition is all about creating a system that can help the common people aware of the market trends and the exact prices for the used cars that they are planning to buy. So, we have collected information about used cars from different sources in India, and we have created a model using machine learning algorithms to determine the approximate value of a used car. We have chosen machine learning as the iterative nature of machine learning aids it in making informed decisions. This will eventually save a lot of time of the people, and even, they will be aware about the market values.

11.2 Literature Survey

Though the use of machine learning in prediction is common, its application on the used car cost prediction is vary sparse. Listiani [2] in his master thesis first used machine learning to predict the cost of used car in a car leasing application. Gongqi et al. [3] proposed a new model using back propagation neural network and nonlinear curve fit to predict the cost of the used car. The result drawn from the experiment proved that the proposed model is able to predict the cost quite accurately. Pudaruth [4] used various supervised machine learning technique to prediction the cost of the used car. However, the dataset used in this research is very limited. Noor et al. [5] proposed a vehicle price prediction system using machine learning techniques. Another similar approach was proposed by Monburinon et al. [6]. In this approach, they used several regression models like multiple linear regression, random forest regression, and gradient boosted regression trees to predict the cost. Nabarun et al. [7] have proposed a supervised machine learning-based prediction technique to evaluate the cost of used car. The results drawn from the experiment show a training accuracy

of 95.82% and a testing accuracy of 83.63%. Similarly, Gagic et al. [8] have proposed a machine learning-based ensemble approach for used car price prediction. A data-driven unsupervised machine learning-based methodology had been proposed by Cerquitelliet. al. [9]. However, this proposed approach is used to estimate the cost of used heavy trucks, and instated of a traditional dataset, it uses the real-time telematics data. A supervised machine learning model using linear regression technique to predict the cost of used car had been proposed by Kiran [10]. The final result of the proposed approach shows an accuracy of 90% and a root mean square error of about 10%.

11.3 Proposed Model

Some great initiatives and works have been done on the prediction of used car prices using machine learning algorithms. The needs of these models are widely accepted nowadays. Researchers have used models like KNN, random forest, support vector machine (SVM), multiple regression, artificial neural networks to train their models. Studies have focused on performing work by collecting data on the used cars from their media houses, articles and basically firms which has data about their country. Addressing this problem actually develops some insight about the country's stance on used cars market from which dataset has been made. It also directs some possible investment scenarios that can be done for future. Therefore, we have collected data of our country's used cars and used the most significant traits that an Indian buyer will look upon.

We thought of implementing different machine learning models to do a comparison among the models and will implement the one with best results while training. After a thorough study, we choose KNN, decision tree, random forest, and LightGBM [11] for this work. LightGBM is recent technique which has been applied to various fields of computation. A lot of work has already been done on this problem using the common algorithms, so LightGBM has been chosen to implement on our test dataset because of its great efficiency, and it takes least amount of time to train the model checking a variety of parameters and discarding unnecessary data. A flowchart showing the necessary steps to be carried out while addressing our prediction of used cars price problem is shown in Fig. 11.1.

The core idea of this manuscript is to use different supervised machine learning models using the dataset, train them first and then to depict the best model to predict the prices of unknown test dataset. The features of the dataset [12] are as follows: *Name, Location, Year, Kilometers Driven, Fuel Type, Transmission, Owner Type, Mileage, Engine, Power, Seats, and Price.*

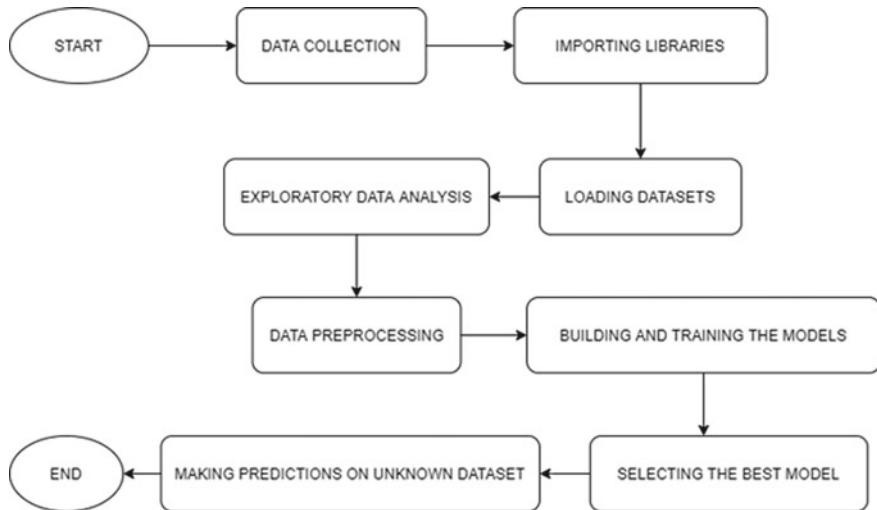


Fig. 11.1 Workflow showing the necessary steps of the proposed model

11.3.1 Exploratory Data Analysis and PreProcessing

Machine learning models rely mostly on datasets. So, selection of good dataset needs to be done before training and testing a model. Datasets give us approximately exact scenario of real world as it contains lots of features in it which actually happens to be best parameters while addressing a problem. We used '*Used Cars Price Prediction*' dataset that we have taken from Kaggle [12]. It contains 6019 data for training the models and 1234 data for testing and predicting the unknown price for the cars which is quite a moderate value for analyzing the result through training and testing of our models.

The dataset is made focusing the Indian market. Data have been collected by various sources from all across the country. It helped in giving an estimated idea about the current market price for used cars. Training dataset contains the prices of used cars which help us to train our model and then implement on our test dataset where price of used cars are to be calculated. A sample dataset with four features is displayed in Table 11.1.

Table 11.1 Sample dataset with only few features

Name	Location	Year	Kilometers driven
Honda Jazz V	Chennai	2011	46,000
Nissan Micra Diesel XV	Jaipur	2013	86,999
Maruti Ciaz Zeta	Kochi	2018	25,692

Exploratory data analysis is done to get an insight about the dataset. The graphical representations used in this process help to identify the outliers in the dataset. These outliers need to be either removed or replaced with suitable values such that they will not deviate the results while training a machine learning model. We have used basic steps to do the exploratory data analysis part. We started by describing the content of the dataset like total number of columns and rows present, mean, standard deviation, etc. We checked whether there are any missing rows of data or whether mistakenly same data have been inserted to rows. We used functions to get the unique number of data present in some of the specific columns, and those are *Name*, *Location*, *Fuel Type*, *Transmission*, *Owner Type*, *Seats*, and *Year*. We used a number of graphs to plot the data. Keeping price as our basic constraint, we have plotted graphs with respect to it. The price is chosen as basic constraint because our problem statement is to predict the prices of used cars once our model will be trained.

The rows that have null values in them are found and dropped from the dataset to avoid ambiguity while training the model. The cars which are extremely high priced have been removed as they will deviate our model if retained. This is done because we focus on developing this model for a moderate search engine where majority of the audience will be people searching economic or just basic premium cars. We aren't focusing for highly luxurious cars as the potential buyers aren't our target audience. Normally, the cars that are being resold are usually driven for around 999 km to 70,000 km. So, we drop all rows that contain cars being driven less than 999 km and greater than 70,000 km.

The dataset was formed taking values of cars which are being listed till 2019. And the electric vehicle segment hasn't made a great debut in India. We decided to drop cars which are powered by electricity. There are also very less to no charging stations present in India by which it won't attract audience as there will be issues that people will face and might not get instant solutions.

We created new variable named 'Car Age' which basically stores the difference between 2020 and the year of car. It gives us an idea of the car age.

Name present in dataset contains very long name as it contains detailed brand name along with model name. We used a variable to store only the first two words of the brand name. By this, we narrowed down the brand name, and it can be used extensively to depict relations and process data. We did some further preprocessing of dataset to remove most outliers as possible and dropped the unimportant columns.

11.4 Model Descriptions

The problem statement that we are addressing here is a regression problem. We know that the output variable in a regression problem is a real or continuous value. This real value or continuous value can be a real or floating point value, in our case prediction of price is continuous in nature, and it is our objective to get best prices for our test data of used cars. The skill of a regression predictive model should be expressed as an error in the predictions because it predicts a quantity. The root mean squared

Fig. 11.2 KNN (actual values of prices vs. predicted values of prices)



error, abbreviated as RMSE, is a popular approach to measure the performance of a regression predictive model.

The k-nearest neighbor algorithm is a simple, easy to implement and can be used to solve both regression problem and classification problem. But here, we are dealing with a regression problem and trying to predict the value of a feature which is continuous in nature. KNN regression is a non-parametric method that intuitively approximates the relationship between independent variables and continuous outcomes by averaging data in the same neighborhood [13]. Because KNN is based on feature similarity, the most significant component is determining the value of k. This is a process known as parameter tuning, and it is critical for improved accuracy.

Keeping all important points for better results of our problem and when used in our problem, we got a RMSE of 6.72. We can visualize from Fig. 11.2 that there is quite a deviation of actual and predicted prices for used cars. Hence, we didn't use this model to our final test dataset to predict prices of used cars.

Decision tree is a statistical, non-parametric supervised learning which classifies data into classes and to represent the results in a tree-like structure which determine each course of action. Each branch of the tree indicates a probable occurrence or consequence to a decision. This model classifies data in a dataset by running it through a query structure from root to leaf, which represents one class. The attribute that plays a major role in categorization is represented by the root, and the class is represented by the leaf [14]. They can fit complex datasets and show the user how a choice is made. The more sophisticated the decision criteria are the more accurate the model becomes. Decision trees learn from data to approximate a sine curve with a series of if–then–else decision rules. The RMSE value obtained for decision tree is 3.45 which is better than KNN's RMSE value obtained earlier, but still, deviation is being observed in Fig. 11.3.

Random forest is a method that works by generating numerous decision trees during the training phase, with random forest selecting the majority of the trees' decisions as the final decision. Random forest is a machine learning technique that can be used to perform a number of tasks, such as regression and classification. A

Fig. 11.3 Decision tree graph (actual values of prices vs. predicted values of prices)



random forest model consists of numerous small decision trees known as estimators; each of which generates its own predictions. The random forest model combines the estimators' predictions to provide a more precise result [15].

While implementing this model, we have taken the number of estimators to be 1000 and the random state value to be 42. This model took really much more time compared to other models used, and the RMSE value obtained is 2.69 which is quite good; the graph plot of the result is display in Fig. 11.4.

Ensemble techniques are a machine learning methodology that combines numerous base models to create a single best-predictive model. Boosting is a method of combining numerous simple models into a single composite model. The total model becomes a stronger predictor when additional simple models are included.

We used LightGBM to predict our used car prices after training as the final model. Few years back Microsoft announced its gradient boosting framework called LightGBM, and today, it is $6 \times$ times faster than XGBoost. LightGBM stands for lightweight gradient boosting machines. To determine a split value, it employs a novel methodology known as gradient-based one-side sampling (GOSS). Because

Fig. 11.4 Random forest graph (actual values of prices vs. predicted values of prices)



of its great speed, LightGBM is prefixed as ‘Light’. LightGBM is widely used because it can manage massive amounts of data while using little memory [11].

We have split our training and testing data into 75:25 ratio. This was chosen after it gave better RMSE values for our model against 80:20 ratio which is basically a good step to start training a model. The core parameters used in LightGBM model consist of objective type which is regression here. We used traditional gradient boosting decision tree (gbdt) as the algorithm that we want to run in our model. Number of leaves is kept near the default value that is 30 in this case. The bagging fraction and feature fraction of the model are used to avoid over-fitting of data and speed up the process by using appropriate values in it. Then, the early stopping round is set at 50 such that training until the validation scores don’t improve for 50 rounds. We got root mean squared log error (RMSLE) to be 2.35 which is quite good and the least among the RMSEs concluded from other models like random forest regression whose RMSE value is 2.69; decision tree model’s RMSE value is 3.45, and KNN’s RMSE value is 6.72.

As we can see from Fig. 11.5, our actual value of price and predicted price value are quite closer to the line of best fit. And indicate that neither our model is over-fitted nor under-fitted. It is really fast in nature. Also, it can be depicted from the graph that the cars having price range under 40 lakhs are having their actual and predicted price to be really close and doesn’t deviate much. And as discussed earlier, our target audience is normal people whose price ranges to purchase cars are mostly less than 20 lakhs, so our model is good to solve our cause.

After the implementation of model in test dataset, our model calculated the price for different cars depending on model, car age, etc., and price has been allocated to the test dataset against their respective car models. A sample set of predicted price is displayed in Table 11.2.

Figure 11.6 gives us an insight about the RMSE values obtained from using different machine learning models for our problem statement.

Fig. 11.5 LightGBM graph (actual values of prices vs. predicted values of prices)



Table 11.2 Sample set of predicted price

Name	Price
Honda Jazz V	450,000
Nissan Micra Diesel XV	350,000
Maruti Ciaz Zeta	995,000

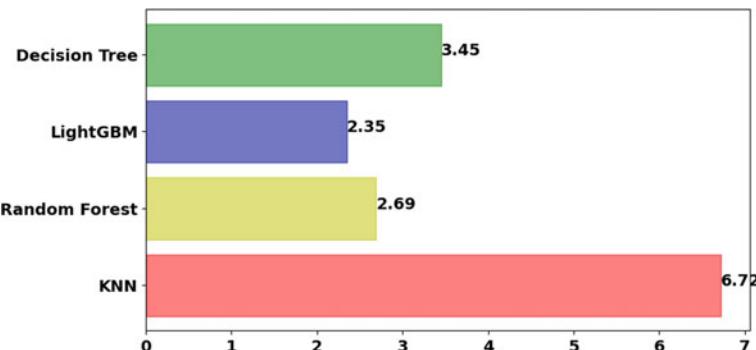


Fig. 11.6 RMSE values versus models used

11.5 Conclusions and Future Work

In this paper, we used different machine learning models, namely random forest regression, decision tree regression, and LightGBM regression, but ultimately, we got the least RMSE value in LightGBM method along with least time taken to train and test the model and hence implemented it to our problem. We stored the new price for the test dataset and compared it with our given training model, and those appeared to be well within the scope without any such huge deviation. In future, we will add more real time up to date data to provide a better accuracy, and this process will be repeated in a timely manner. Further, we will be using advanced algorithms such as fuzzy logic, genetic algorithms to handle the ambiguity in the dataset, and we also like to link it to the search engines to reach a wider section of the audience. Our work is in its first stage where we addressed the problem from our basic knowledge. With time, we will update it.

References

1. Sheth, A., Krishnan, S., Samyukktha, T.: India Venture Capital Report 2020 (2020)
2. Listiani, M.: Support vector regression analysis for price prediction in a car leasing application. Unpublished. <https://www.ifis.uni-luebeck.de/~moeller/publist-sts-pw-andm/source/papers/2009/list09.pdf> (2009)
3. Gongqi, S., Yansong, W., Qiang, Z.: New model for residual value prediction of the used car based on BP neural network and nonlinear curve fit. In: 2011 Third International Conference

- on Measuring Technology and Mechatronics Automation, vol. 2, pp. 682–685. IEEE (2011)
4. Pudaruth, S.: Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol.* **4**(7), 753–764 (2014)
 5. Noor, K., Jan, S.: Vehicle price prediction system using machine learning techniques. *Int. J. Comput. Appl.* **167**(9), 27–31 (2017)
 6. Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., Boonpou, P.: Prediction of prices for used car by using regression models. In: 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 115–119, IEEE (2018)
 7. Pal, N., Arora, P., Kohli, P., Sundararaman, D., Palakurthy, S.S.: How much is my car worth? A methodology for predicting used cars' prices using random forest. In: Future of Information and Communication Conference, pp. 413–422. Springer, Cham (2018)
 8. Gegic, E., Isakovic, B., Keco, D., Masetic, Z., Kevric, J.: Car price prediction using machine learning techniques. *TEM J.* **8**(1), 113 (2019)
 9. Cerquitelli, T., Regalia, A., Manfredi, E., Conicella, F., Bethaz, P., Liore, E.: Data-driven estimation of heavy-truck residual value at the buy-back. *IEEE Access* **8**, 102409–102418 (2020)
 10. Kiran, S.: Prediction of resale value of the car using linear regression algorithm. *Int. J. Innov. Sci. Res. Technol.* **6**(7), 382–386 (2020)
 11. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T.Y.: Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* **30**, 3146–3154 (2017)
 12. <https://www.kaggle.com/avikasliwal/used-cars-price-prediction>. Last accessed 11 June 2021
 13. Subramanian, D.: A simple introduction to K-Nearest Neighbors Algorithm. *Towards Data Science* (2019)
 14. Dakou, E., D'heygere, T., Dedecker, A.P., Goethals, P.L., Lazaridou-Dimitriadou, M., De Pauw, N.: Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquat. Ecol.* **41**(3), 399–411 (2007)
 15. Ao, Y., Li, H., Zhu, L., Ali, S., Yang, Z.: The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *J. Petrol. Sci. Eng.* **174**, 776–789 (2019)

Chapter 12

Complexity Classification of Object-Oriented Projects Based on Class Model Information Using Quasi-Opposition Rao Algorithm-Based Neural Networks



Pulak Sahoo, Ch. Sanjeev Kumar Dash, Satchidananda Dehuri,
and J. R. Mohanty

Abstract On the basis of the surveys conducted over a period of time, it is realized that the top four reasons for failure of software projects are functionality issues, failure to meet deadline, quality issues, and budget overruns. Customer goodwill, which is extremely important for continued business, is greatly influenced by timely delivery of quality software products. A rational project schedule with allocation of appropriate resources is essential for achieving these goals. This requires a complexity assessment of the projects by the organization so that a project team of suitable size and skill set can be decided at a higher level. This work proposed a classification approach of object-oriented projects using the information present in unified modeling language (UML) class models. The classification approach combined the best attributes of quasi-opposition-based Rao algorithms (QORA) and artificial neural networks (ANNs) is named as QORA-ANN. The classification of a project into three classes, namely large scale, medium scale, and small scale, has been done using the training data consisting of a number of contemporary object-oriented projects executed in various IT firms. We conduct an extensive evaluation of QORA-ANN, radial basis function network (RBFN), and logistic regression on the project dataset and make a comparative study. It is found that the result obtained by QORA-ANN approach is promising than RBFN and logistic regression.

P. Sahoo · Ch. Sanjeev Kumar Dash (✉)

Department of CSE, Silicon Institute of Technology, Bhubaneswar, India

P. Sahoo

e-mail: pulak.sahoo@silicon.ac.in

S. Dehuri

Department of Information and Communication Technology, Fakir Mohan University, Vyasa Vihar, Balasore, Odisha 756019, India

J. R. Mohanty

School of Computer Engineering, KIIT Deemed To Be University, Bhubaneswar, India

12.1 Introduction

Due to increased competition in recent times, IT firms are giving utmost importance to customer satisfaction and retention. The timely delivery of software projects along with its functionality and quality is considered vital for this. One way to achieve this goal is through careful project planning, activity scheduling, and suitable resource allocation. In order to perform these activities properly, there is a requirement to conduct a high level complexity assessment of the project. This enables the firm to determine a project team of suitable size and skill set to execute the given project.

One of the most popular languages to represent an object-oriented project is the unified modeling language (UML). The behavioral and architectural facets of a project can be represented effectively through various UML models. Past studies have already established that the UML models produced with appropriate features can be applied for evaluating the complexity of an object-oriented project [1–8]. The class UML model contains the class components of the project and the inter-relationships between them. The functionality of the project is divided between the classes which make them appropriate inputs for determining the project complexity. This work proposed an automated classification approach of object-oriented projects using the facts present in UML class models. The classification of a project into three categories, namely large scale, medium scale, and small scale, was done by applying appropriate soft computing techniques. The techniques employed were optimal ANN with quasi-opposition-based Rao (QORA-ANN) algorithm. The model was trained based on real-project data consisting of 31 contemporary object-oriented projects executed in various IT firms. The results achieved by applying the suggested method confirmed the validity and accuracy of this approach. Remainder parts of this work are documented below. Section 12.2 recounts some highly relevant work done in proposed area. Section 12.3 explains the suggested project complexity determination method. Section 12.4 holds the details of the proposed complexity determination method applied on recently undertaken case study projects and represents the obtained outcomes. Section 12.5 provides the final remarks along with the scope for extending this work in future.

12.2 Relevant Work

This section provides explanation of a number of recognized project classification and estimation processes relevant to this work. The main purpose of these approaches is usage of details available in UML models for classification and estimation purpose. Papers [1–7] had focused their attention on studying UML models including the class models for classification and estimation of software projects. Costagliola et al. [1] offered an approach called the class point (CP) method. The CP method estimated the development effort of an OOS by using the class UML models to categorize the complexity of the system. Firstly, the CP1 measurement was applied to obtain

the preliminary product size using the methods present and the services requested in the classes. This was revised later by applying the CP2 measurement that used the class attributes in addition to other two attributes. The UML points approach was proposed by Kim et al. [2] to evaluate the development effort of a project and classify them by integrating use case points (UCPs) and class points (CPs). The UCPs are obtained from the system level requirements captured in the use case models. Sahoo et al. [3–6] proposed a number of improvements to class point and use case point approaches among other enhancements to estimate and classify software projects accurately. These approaches were applied to the testing stage of the product development yielding better accuracies. A number of refinements to the CP approach were suggested by Satapathy et al. [7]. He incorporated application of soft computing techniques (using MLP and RBFN) for optimization of the effort parameters. The results obtained demonstrated that the MLP approach gave better performance than the RBFN approach.

Rao et al. [9, 10] had designed a new set of algorithms with the name “Quasi-Oppositional-based Rao algorithms.” When compared their results with basic Rao algorithms, differential evolution (DE), particle swarm optimization (PSO), genetic algorithm (GA), teaching–learning-based algorithm (TLBO), multi-objective genetic algorithm (MOGA), Jaya algorithm, real-coded GA (RCGA), and self-adaptive multi-population (SAMP) Rao algorithms, their result was superior.

12.3 Proposed Method

This work presents an innovative complexity classification approach of object-oriented projects using the facts present in its class models. The classification of the projects was done into three groups, namely: large scale, medium scale, and small scale. The artificial neural network with QORA-ANN structure was used for the classification. This structure examined the search space for some probable ANNs reached at a best possible network during the process of evolution. The model was trained based on real data collected for 31 contemporary object-oriented projects (shown in Table 12.1) executed in different IT firms. Given below is the process flow for this approach. The 31 project dataset used in this work contained below attributes collected from class UML models of the projects. These attributes were utilized by class point method proposed by Costagliola et al. [1] to calculate the size of a project in class points metric. They were as follows: (1) number of attributes, (2) number of external methods, and (3) number of requested services. The fourth attribute (output) was the class labels for each project which was one of the below values. The values were as follows: (1) 1—small scale (completed within 35 man-days), (2) 2—medium scale (completed between 36 and 70 man-days) (3) 3—large scale (took more than 70 man-days). The abovementioned project dataset was divided into training and test instances. The training instances were utilized to train the QORA-ANN-based model. Then, the test instances of projects were used to verify the accuracy of classification. The predicted complexity category of the projects in the test set was compared with

Table 12.1 Case study projects with class model attributes and complexity categories

Project id	No. of attributes (NOA)	No. of external methods (NEM)	No. of services requested (NSM)	Complexity of the project (3-large scale, 2-medium scale, 1-small scale)
1	37	19	30	2
2	58	29	24	3
3	32	19	8	1
4	51	21	15	2
5	14	22	15	1
6	46	39	15	2
7	76	54	29	3
8	55	44	11	2
9	26	13	16	1
10	11	7	8	1
11	41	20	13	2
12	18	19	12	1
13	50	20	14	2
14	44	15	8	2
15	93	76	48	3
16	34	55	23	2
17	30	12	8	1
18	40	48	31	2
19	23	20	11	1
20	61	48	31	3
21	49	44	26	2
22	30	32	17	2
23	81	46	32	3
24	26	13	16	1
25	24	31	19	2
26	94	52	28	3
27	34	23	17	2
28	11	14	8	1
29	53	37	19	2
30	22	19	8	1
31	110	67	36	3

actual complexity category obtained based on the man-days required to complete the projects. The prediction accuracy was computed accordingly.

12.3.1 Artificial Neural Network

The ANN with a single middle layer is displayed in Fig. 12.1. The initial layer of the ANN receives the inputs x_1, x_2, \dots, x_n sending them to the middle layer. The j th hidden layer neuron is computed by the product of the input and the associated weight added with the bias. Then, activation function is applied to calculate the output as follows:

$$a_i = \sigma \left(\sum_{i=1}^n (w_{j,i} * x_i) + b_j \right) \quad (12.1)$$

The output layer is calculated by multiplying weights and activations from the middle layer with addition of the bias (Eq. 12.2). The ANN structure produces an output (\hat{y}) in the final layer. The gap between target output and expected r value is computed next. The error function is then propagated back to regulate the model arguments using applying the rule of gradient descent.

$$\hat{y} = \sigma \left(\sum_{i=1}^m (a_i * w_i) + b \right) \quad (12.2)$$

$$\text{error} = \text{abs}(\text{actual} - \hat{y}) \quad (12.3)$$

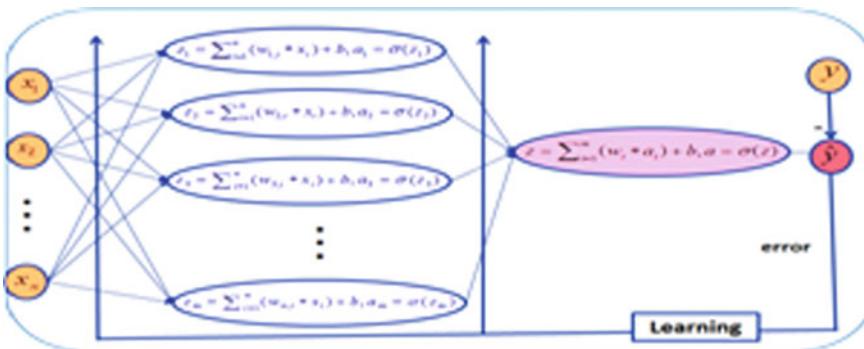


Fig. 12.1 Architecture of ANN

12.3.2 Quasi-Oppositional-Based Rao Algorithm

QORA [9, 10] is innovative optimization algorithms based on population that use the concept of quasi-opposition-based learning in Rao algorithms. The idea is to increase the search space and improve the speed of convergence [9]. These simple techniques can solve complex problems without needing human intervention. Figure 12.2 illustrates the high-level algorithm.

Assuming that $f(w)$ is the objective function requiring optimization. Let the population contains a set of candidate solutions with m propose variables each. Each candidate solution is associated with a fitness/error value. Better solution has lower error signal. Let $f(w)_{\text{best}}$ be the best and $f(w)_{\text{worst}}$ be the worst solution in the

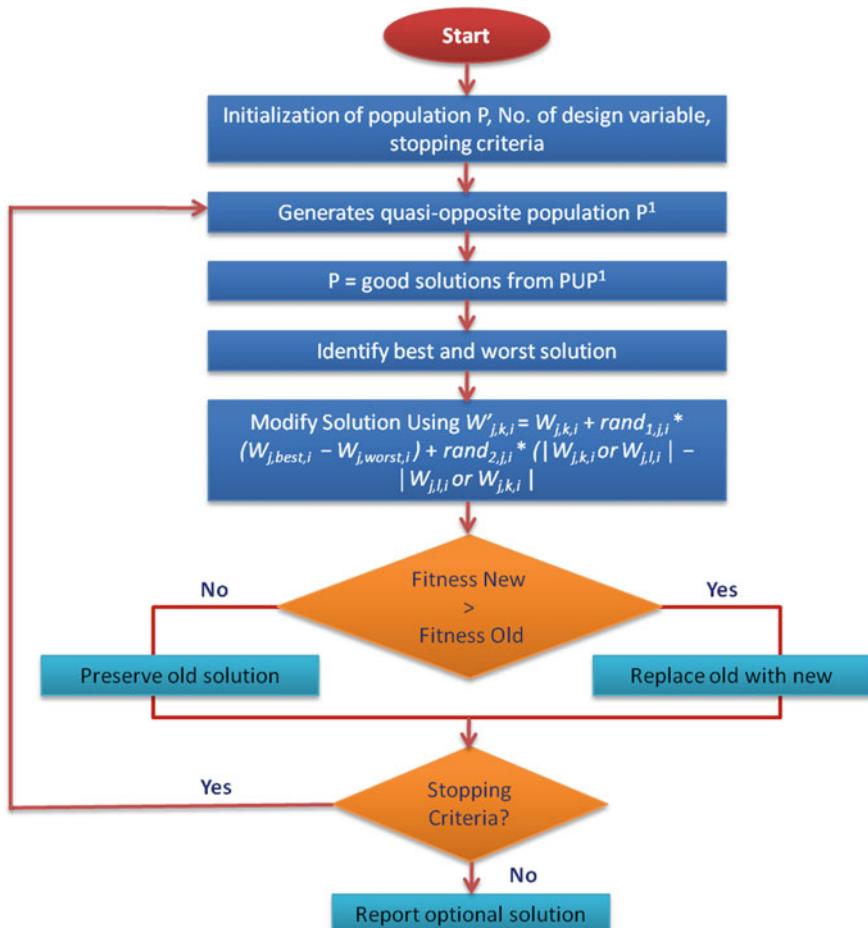


Fig. 12.2 Flow of QORA

population for a given iteration. Equations 12.4, 12.5, and 12.6 show the present value of a candidate solution at the i th iteration.

$$W'_{j,k,i} = W_{j,k,i} + \text{rand}_{1,j,i} * (W_{j,\text{best},i} - W_{j,\text{worst},i}) \quad (12.4)$$

$$\begin{aligned} W'_{j,k,i} &= W_{j,k,i} + \text{rand}_{1,j,i} * (W_{j,\text{best},i} - W_{j,\text{worst},i}) \\ &\quad + \text{rand}_{2,j,i} * (\lvert W_{j,k,i} \text{ or } W_{j,l,i} \rvert - \lvert W_{j,l,i} \text{ or } W_{j,k,i} \rvert) \end{aligned} \quad (12.5)$$

$$\begin{aligned} W'_{j,k,i} &= W_{j,k,i} + \text{rand}_{1,j,i} * (W_{j,\text{best},i} - \lvert W_{j,\text{worst},i} \rvert) \\ &\quad + \text{rand}_{2,j,i} * (\lvert W_{j,k,i} \text{ or } W_{j,l,i} \rvert - W_{j,l,i} \text{ or } W_{j,k,i}) \end{aligned} \quad (12.6)$$

where

$W_{j,k,i}$ is the value of variable j in k th solution ($k = 1, 2, 3, \dots, n$) at iteration i .

$W'_{j,k,i}$ is the updated value of variable j in k th solution at iteration i .

$W_{j,\text{best},i}$ and $W_{j,\text{worst},i}$ are the j th variable values of the best and worst solutions at iteration i .

$\text{rand}_{1,j,i}$ and $\text{rand}_{2,j,i}$ are the random values in $[0, 1]$.

The expression “ $W_{j,k,i}$ or $W_{j,l,i}$ ” (in Eqs. 12.5 and 12.6) illustrates the comparison of fitness between the k th solution and the arbitrarily drawn l th solution. Based on their fitness values, a swap of the information takes place ensuring communication between the candidate solutions. Then, the opposition-based learning feature is included in Rao algorithms [9] in order to get better convergence rate. Then, a population opposite to the present one is generated in each of iteration. Based on (i) the center of the search space (a), (ii) the opposite value of $W_{j,k,i}$ (b) and (iii) upper and lower limits of variable j (W_j^U) and (W_j^L) a quasi-opposite solution $W_{j,k,i}^q$ is generated as given below:

$$W_{j,k,i}^q = \text{rand}(a, b) \quad (12.7)$$

$$a = \frac{w_j^L + W_j^U}{2} \quad (12.8)$$

$$b = W_j^L + W_j^U - W_{j,k,i} \quad (12.9)$$

12.3.3 ANN + QORA-Based Classification

The suggested hybrid ANN + QORA [10] approach is described here (Fig. 12.3). The base model for this approach is the $(n \times m \times 1)$ ANN (Fig. 12.1). Sliding window approach is used to establish the final set of input for the model structure. The window size is the number of input neurons (n). The size of output layer is fixed as one. The middle layer size is decided experimentally. The error signal is computed at the output layer from the associated weight ($w_{11}, w_{12}, \dots, w_{1n}, \dots, w_{mn}$) and bias vector for a given input vector (x_1, x_2, \dots, x_n) and target (y). The purpose is to determine the best weight and bias vectors with minimal error using QORA for the given ANN. The initial population (P_1) contains some probable solutions. Then, using Eq. 12.7, the quasi-opposite population P_2 is determined. It is recommended to use a fit solution pool as initial population to improve the convergence rate. The QORA takes into account both P_1 and P_2 to obtain a near optimal population (P). To fine tune the process of searching, search operators (in Eqs. 12.4 and 12.5 or 12.6) are used. The global minima are reached after successive iterations.

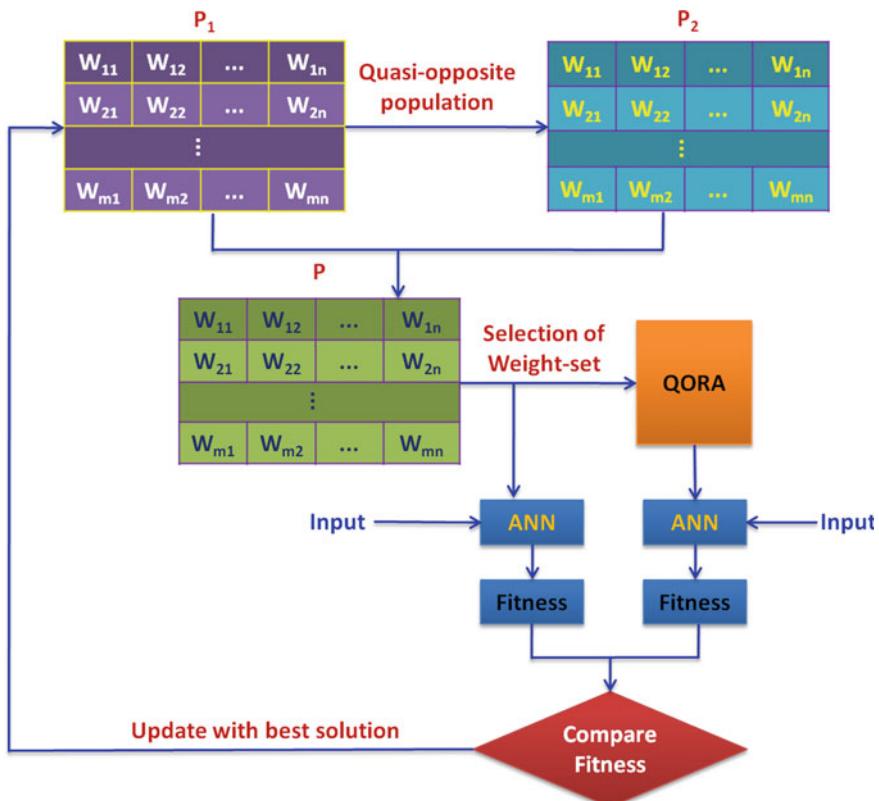


Fig. 12.3 QORA-ANN training

Table 12.2 Project complexity classification results obtained by QORA-ANN, RBFN, and logistic regression approaches

Model	Accuracy	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
QORA-ANN	93.75	0.906	0.0746	0.1877	16.7588	39.7584
RBFN	81.25	0.7197	0.125	0.3536	28.0971	74.8868
Logistic regression	90.62	0.8588	0.0945	0.2148	21.247	45.4933

12.4 Implementation and Experimental Study

As mentioned above, a recent dataset of 31 projects was used in this work to train the QORA-ANN model. Each instance of the dataset contained three attributes taken from class UML models of the project as suggested by Costagliola in his class point-based estimation approach. The fourth attribute is being the class labels containing one of the three complexity categories, namely large scale, medium scale, and small scale. In this work, 80% of projects was used for the purpose of training, and remaining 20% projects was used as test data to check the classification accuracy.

12.4.1 Case Study Project Dataset

12.4.2 Results Analysis

The soft computing techniques applied on the case study projects dataset produced below results. This work compared the QORA-ANN, RBFN, and logistic regression approaches. The accuracies of project complexity categorization obtained by the above approaches are specified in “Accuracy” column of Table 12.2.

The results obtained by applying the proposed soft computing approaches confirmed that all three of them produced reasonable project complexity classification accuracies. However, the QORA-ANN approach produced the best classification accuracy of 93.75% (Fig. 12.4).

12.5 Concluding Remarks

A rational project schedule with allocation of appropriate resources is vital for a timely delivery of software projects. In order to achieve this, a sensible and accurate project complexity assessment is required so that a project team of suitable size and

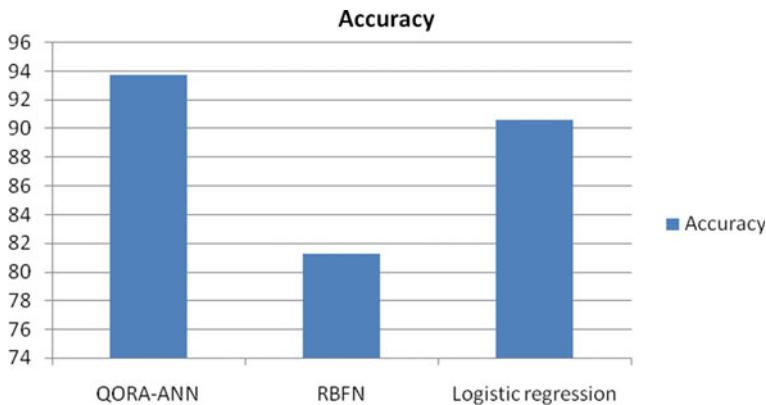


Fig. 12.4 Accuracy of different model

skill set can be allocated to the project. This work proposed an innovative classification approach based on the information present in UML class models. The classification approaches consisting of QORA-ANN, RBFN, and logistic regression techniques were applied was on a data consisting of 31 contemporary projects executed in various IT companies. The results obtained validated that the QORA-ANN approach is the most appropriate one producing an accuracy of 93.75%.

References

- Costagliola, G., Ferrucci, F., Tortora, G., Vitiello, G.: Class point: an approach for the size estimation of object-oriented systems. *IEEE Trans. Software Eng.* **31**(1), 52–74 (2005)
- Sang Eun, K., Lively, W., William, M., Simmons, D.B.: An effort estimation by UML points in early stage of software development. *Softw. Eng. Res. Pract.* 415–421 (2006)
- Sahoo, P., Mohanty, J.R.: Early test effort prediction using UML diagrams. *Indonesian J. Electr. Eng. Comput. Sci.* **5**, 220–228 (2017)
- Sahoo, P., Mohanty, J.R.: Early System Test Effort Estimation Automation for Object-Oriented Systems, pp. 325–333. *Information and Decision Sciences*, Springer (2018)
- Sahoo, P., Mohanty, J.R.: Test effort estimation in early stages using use case and class models for web applications. *Int. J. Knowl.-Based Intel. Eng. Syst.* **22**(1), 215–229 (2018)
- Sahoo, P., Mohanty, J.R.: System Test Effort Estimation Using Class Model: A Case Study, pp. 239–250. *Smart Intelligent Computing and Applications*, Springer (2020)
- Satapathy, S.M., Kumar, M., Rath, S.K.: Class point approach for software effort estimation using soft computing techniques. In: International Conference on Advances in Computing, Communications and Informatics, IEEE, pp. 178–183 (2013)
- Haykin, S.S.: *Neural Networks and Learning Machines*. Prentice Hall, New York (2009)
- Rao, R.V., Pawar, R.B.: Quasi-oppositional-based Rao algorithms for multi-objective design optimization of selected heat sinks. *J. Comput. Design Eng.* **7**(6), 830–863 (2020)
- Dash, C.S.K., Behera, A.K., Nayak, S.C., Dehuri, S.: QORA-ANN: quasi opposition based Rao algorithm and artificial neural network for cryptocurrency prediction. In: 2021 6th International Conference for Convergence in Technology (I2CT), pp. 1–5. IEEE (2021)

Chapter 13

Mood-Based Movie Recommendation System



Soumya S. Acharya, Nandita Nupur, Priyabrat Sahoo, and Paresh Baidya

Abstract Mood-based movie recommendation system provides a platform which makes a simple task of choosing a movie, more efficient. Here, the person has to choose the mood according to which the system can recommend a movie based on their choices. As we all know, searching and deciding which movie to watch take much more time than actually watching the movie. So, through this application, they are able to make a better movie choice in much less time. This movie recommendation system is the one that considers the mood of the user for extracting out few specific genres for further selection. The system then asks for user's some favorite movies in the chosen genre so that it can recommend a movie based on their choices. Here, we have used Python and flask Web framework (and different libraries) to build the Web application, and the dataset used is scrapped from IMDb's Web site and combined from different dataset available on Kaggle.

13.1 Introduction

In this anxious world, people are enough stressed regarding their job, relationship where watching movie is one way to release it, but the selection process is itself an additional stress. In this fast era, where even the pizza delivery can't take more than 30 min. Then, why waste so much time looking for the perfect movie. Other than that people don't know about all the movies present on the Internet. There is also a deficit of a proper platform where we can actually find movies to watch. At last, we fall exhausted without any output. Recommendation systems are described as the software-based techniques and tools that offer recommendations for items [1]. Nowadays, many companies are actually dependent on their recommendation engines' effectiveness, in order to attract new users and even manipulating them to spend more time with their products. Indeed, there are properties might influence the user event, like scalability and robustness for a recommendation system [2]. The

S. S. Acharya · N. Nupur · P. Sahoo · P. Baidya (✉)

Computer Science and Engineering, Siksha 'O' Anusandhan Deemed To Be University,
Bhubaneswar, Odisha, India

e-mail: pareshbaidya@soa.ac.in

destiny is a feasible possibility to initiate summarizes and customized RS covering the emotional aspect accompanied by the capacity to enforce multi-agent system (MAS), among different domains. Such retrieval machine is essential that together with the project of statistics collection can also contain in selective filtering as in step with the hobbies and feelings precipitated through the statistics within side the challenge [3, 4]. It allows users' objectives, social network attributes, and networking behavior to be embedded into the RS programmer. For conducting the users' objective, MAS must be dependent on selective statistics [5]. In this implementation, we have discussed about a machine learning model that is developed considering the user's choices and curates a list of movies based on genre and plot of the movie. Each movie will be given a score based on the parameters and will be ranked accordingly. The system will generate top 20 recommendations along with movie title, description, IMDb reading, and year of release for user's discretion. Yet people do not have efficient and sophisticated application so that they can easily find their requirements. Through this application, they can achieve an efficient output in optimal time.

13.2 Literature Survey

There have been many recommendation engines proposed in the recent years. Many of these propositions have been published, but very few have been implemented as a stand-alone platform, providing specific recommendations.

13.2.1 Existing Systems

In 2011, Quijano-Sanchez et al. provided a movie recommender system which targeting the group of people integrating with the social network Facebook. Its name was Happy Movie. In [6], the method uses three properties user's personality, social trust, and past experience. Menezes and Tagmouti proposed an emotion-based movie recommender system. The system's primary goal was to be adaptive and provide the suggestions to users. It employed a mix of collaborative filtering and content-based approaches [7]. Hong et al. proposed a singular approach to degree users' choice on film genres. Matrix factorization framework was used for regularization [8]. In 2014, Rajenderan proposed a movie recommendation system that takes users emotions for generating recommendations [9]. There are many such proposals revolving around the application of the hybrid recommenders for getting better movie recommendations. But, the problem arises while predicting or recognizing the correct emotion of the user at given time, as the proposed systems actually focus on collecting the data through reading the user emotions. Here, the main concern is of data security. Most of the proposed systems are completely focused on extracting information through acute observation of the users, by following each and every minute fickle in his emotional state while going through his/her daily chore or while going through a

particular content. This can actually create a breach in users' privacy, and even if we fortify the system to make it completely secure, at the end, most of our work would actually be concentrated in making the system secure rather than focusing on making it accurate.

13.2.2 *Proposed Systems*

Our proposed system actually solves the problem of data security and gives freedom to the user to select a preference rather than suggesting based on his/her previous activity. Since there was a clear struggle between data privacy and recommendation accuracy in the proposed systems, we decided for following a different approach. The proposed system actually lets the user decide his/her mood or current emotional status and displays a set of curated genre list for each selection. Using the conventional handwritten rules for features in machine learning, we collected data from various surveys, conducted by us, and publicly available studies on human emotions and their preferences in particular emotional states. So, each emotion has actually a unique set of genres associated with it and is provided to the user to choose from. Then, the system asks for user's favorite movie in that particular genre, similar to which he/she would prefer to watch. The data related to the provided input then undergoes feature extraction using text processing and assigning score to each extracted feature based on the importance. Then, these scores are compared with each of the available data in the database to find similar elements, using cosine similarity. Finally, the system displays a set of 20 (twenty) movies based on the similarity, as recommendations. We will further see the detailed implementation of the system.

13.3 Recommendation Systems and Approaches

There are two main techniques followed for building recommendation systems, content-based filtering and collaborative filtering. The approach which has been developed in the recent years is the hybrid approach. Hybrid approach has been proven to be quite successful in making accurate predictions at times, but it increases the cost of the systems. This prediction is computed employing a range of prognostic models that have a typical characteristic, generally using user-provided ratings/evaluations [10]. The most important and actually critical aspect for the process of building the recommendation engine is the collection and integration of data. In this paper, our aim is making limited recommendations although the best quality is reflected with assurance that the results are more suitable.

Content-based filtering, This approach is based on user interactions and preferences on each platform, previously visited podcasts, and information about different types of interactions (for example, based on users choosing genres, and podcasts they already like. Collaborative filtering, it is another widely used technique. This

technique actually uses the collaborative data of all the users, collecting information from the metadata of many of the users who have similar tastes and choices in items, and deriving recommendations for a single user [11]. The similarity is not limited to the user's taste, and the similarity between various elements can also be considered. When we have a lot of information about users and elements, the system will give more effective suggestions. It can provide high-precision recommendations to a large number of users in time [12]. The user similarity information includes scores assigned by two users. Store the data that determines the similarity of users in the database so that similar users can be intercepted when making recommendations.

13.4 Implemented Methodology

In this application, people can not only find their best choice but also in minimal time. They will be getting recommendations based on their mood which is the most important thing. They will be able to get specific genres to choose that will be suitable for their current mood, and according to that the system will generate recommendations. It will also ask for the favorite movie on the chosen genre similar to which it will recommend. So, people will get the best choices possible. We have incorporated content-based approach for building up the movie recommendation engine. There are many underlying processes, starting from the data collection to giving recommendations, but we can broadly categorize the processes in three categories. As discussed earlier, in order to maintain the data privacy and uphold the security factor of the users, we are letting the users decide their mood/current emotional state. These are a set of handwritten rules to decide the genres that should be under particular moods. These rules are made through references from various surveys and large-scale studies on the topic of genre selection by users in different emotional states. We also conducted some of our own surveys consisting of basic questions regarding a person's choices when it comes to movies. Figure 13.1 shows the result of a survey that we conducted

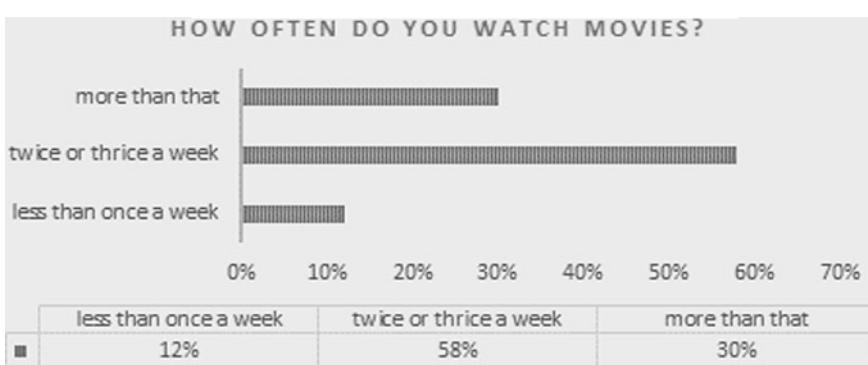


Fig. 13.1 Survey—how often watch movies

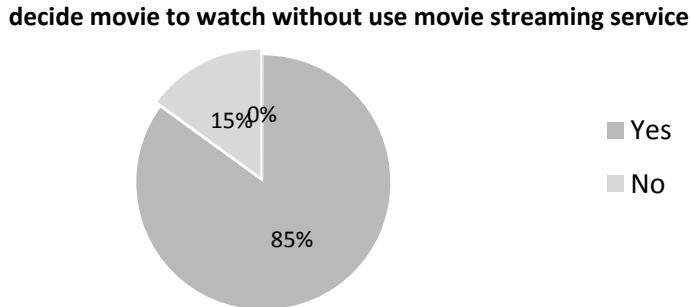


Fig. 13.2 Survey—decided movie to watch without use movie-streaming service

about the frequency of watching movies by an individual. It was found out that 58% of the total participants preferred to watch movies twice or thrice a week. Figure 13.2 shows that 85% of the people decide on a movie before watching, rather than moving ahead with the recommendation by the movie-streaming platform. This indicated the need of a stand-alone system to provide recommendations to a user, in turn saving their precious time.

Document Vectorization: Vectorized documents allow us to perform various tasks, such as finding related documents, categorizing them, and grouping them. Now, there are various approaches to vectorize a document. The most common being the Bag of Words. But, we refrained from using this approach as it does not account for noise in the dataset. Like, the most commonly used word in English language is ‘the’ which makes up 7% of all written or spoken words. Nothing could be derived from a text just because it contains the word ‘the’. Therefore, we chose Tf-Idf, a technique that prioritizes the relative importance of the words. This proves to be quite useful in this scenario.

Term Frequency (TF), the term frequency is the probability of a word occurring in the document.

$$\text{tf}_{w,d} = \frac{n_{w,d}}{\sum_k n_{w,d}}. \quad (13.1)$$

Here,

w word

d document

$\text{tf}(w, d)$ term frequency of word ‘ w ’ in document ‘ d ’

$n(w, d)$ total count of word ‘ w ’ in document ‘ c ’.

Document Frequency (DF), it is the importance measurement of the document in the entire corpus. The difference of DF from TF is that term frequency is the frequency counter of term ‘ w ’ in document ‘ d ’, where document frequency is the frequency

of document ‘ d ’ in which the term ‘ w ’ occurs. We consider the case where a term contains a document at least once.

Inverse Document Frequency (IDF), IDF of a word is the logarithmic ratio of the occurring documents of the word and total number of documents.

$$\text{idf}(w) = \log\left(\frac{N}{df_t}\right) \quad (13.2)$$

For a rare word, IDF (without logarithmic terminology) is 1000 and for a common word $\text{IDF} = 1$. Here, we see a much larger difference, thus dominating the Tf-Idf. But, considering the logarithmic value, for a common word $\text{IDF} \sim 0$ and for the rare word $\text{IDF} = \log(1000) \sim 6.9$.

Cosine Similarity, cosine of the angle between the two vectors X and Y is computed for imputing similarity. If the vectors are close to each other, then smaller will be the angle and larger will be the cosine.

$$\text{Similarity}(x, y) = \frac{xy}{|x| * |y|} \quad (13.3)$$

Functional Model: The functional model is tried to be made as user-friendly as possible. Then, secondary objective was to make it faster without the use of APIs. The current prototype system uses a database of about 75,000 movies. Further, the database is segregated into genre-based chunks. In the preprocessing, the genre-wise datasets have also been processed by applying Tf-Idf, on each, and the resultant sparse matrices are stored in the disk. The primary dataset, the genre-wise datasets, and the genre-wise sparse matrices are accessible by the system for use. Next, the implemented system takes the users’ mood as input, and then displays a set of genres exclusive to a particular mood. After selecting the genre, the user is then prompted to enter a movie similar to which he/she would like to watch. Then, the system checks for availability of data related to the movie; if absent, it prompts the user to select another sample movie if he/she want. If present, the system uses text processing on the sample movie’s description and extracts features accordingly. The features are assigned scores based on the importance (as stated in Tf-Idf). Simultaneously, the system loads the sparse matrix of the chosen genre and performs cosine similarity on it to find the similarity index of each movie in comparison to every other movie. Then, the system searches for the sample movie provided by the user in the obtained similarity matrix. It extracts the required row and then sorts it in descending order in terms of the similarity with the sample movie. The system then picks the top 20 items, excluding the sample movie itself, and displays it to the user.

An important change has been made in the cosine similarity function. The original pre-defined function of the Scikit-learn library preprocesses each class in the background. This consumption is much valuable CPU time [13]. Therefore, we have modified the cosine similarity function to provide the similarity matrix without

preprocessing each class. The modified function normalizes the sparse matrix and performs matrix multiplication of the transpose of the provided sparse matrix with itself. This method is proven to take much less time compared to the conventional pre-defined function [14]. The result is the similarity matrix having the similarity scores of each item with respect to every other item.

13.5 Conclusion and Future Scope

In this paper, we designed and implemented a mood-based movie recommender system. Since data security is a major concern for every user, the proposed model collects minimum data from the user. We applied Tf-Idf for text processing and a modified approach to calculate the cosine similarity, increasing its low latency applications. We encourage the researchers to further implement and improve the idea through: (1) implementing an intuitive system for predicting the mood. (2) using hybrid recommender for better results. (3) implementing spell checker and auto-recommender while searching.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **17**(6), 734–749 (2005)
2. Ghuli, P., Ghosh, A., Shettar, R.: A collaborative filtering recommendation engine in a distributed environment. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I). IEEE (2014)
3. Bhatt, B.: A review paper on machine learning based recommendation system. *Int. J. Eng. Dev. Res.* (2014)
4. Debnath, S., Ganguly, N., Mitra, P.: Feature weighting in content-based recommendation system using social network analysis. In: Proceedings of the 17th International Conference on World Wide Web. ACM (2008)
5. Wakil, K., et al.: Improving web movie recommender system based on emotions. (*IJACSA*) *Int. J. Adv. Comput. Sci. Appl.* **6**(2) (2015)
6. Quijano-Sanchez, L., Recio-Garcia, J.A., Diaz-Agudo, B.: Happymovie: A facebook application for recommending movies to groups. In: 2011 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp. 239–244 (2011)
7. Ho, A.T., Menezes, I.L., Tagmouti, Y.: E-MRS: Emotion- based movie recommender system. In: Proceedings of IADIS e-Commerce Conference. University of Washington Both-ell, USA, pp. 1–8 (2006)
8. Nie, D., Hong, L., Zhu, T.: Movie recommendation using unrated data. In: 2013 12th International Conference on Machine Learning and Applications (ICMLA), pp. 344–347 (2013)
9. Rajenderan, A.: An Affective Movie Recommendation System (2014)
10. Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: Recommender systems handbook. Springer, pp. 257–297 (2011)
11. Melville, P., Mooney, R.J., Nagarajan, R.: Content-boosted collaborative filtering for improved recommendations. In: AAAI/IAAI, pp. 187–192 (2002)

12. Costa, H., Macedo, L.: Emotion-based recommender system for overcoming the problem of information overload. In: Highlights on Practical Applications of Agents and Multi-agent Systems, Springer, pp. 178–189 (2013)
13. <https://medium.com/data-science-at-microsoft/how-we-reduced-our-text-similarity-runtime-by-99-96-e8e4b4426b35>
14. <http://na-o-ys.github.io/others/2015-11-07-sparse-vector-similarities.html>

Chapter 14

Covid-19 and Awareness of the Society: A Collection of the Important Facts and Figures Related to the Global Pandemic



Prabhat Kumar Sahu, Parag Bhattacharjee, and Nikunj Agarwal

Abstract Our article COVID-19 AWARENESS is created to provide latest and correct information regarding the current pandemic situation. It provides the public with statistics of active, recovered, and death cases country-wise, all around the world. It also provides them with latest news regarding the pandemic every 24 h. It also provides helpline numbers with search functions for the people to call for help. Our site also provides guidelines on preventive measures along with the steps that a person should follow if they are infected with the coronavirus. The other guidelines are displayed on the Web site in an attractive and responsive way. We also provide a search function for helpline numbers which searches on the basis on the name of a states or a part of it. Our article provides all the necessary information to the public that are required to be healthy and safe in these difficult times for the humanity because of the pandemic.

14.1 Introduction

COVID-19 is a global pandemic caused by the novel coronavirus. Basically, they are a group of RNA viruses, as shown in Fig. 14.1 that attack the protein cells of the body and causes diseases in mammals and birds. The human heart is most vulnerable to this virus as it attaches itself to the receptors of the protein cells. It can cause mild illnesses in humans like cough, cold, headache, fatigue, but can also cause lethal diseases like SARS, MERS, and COVID-19. In animals such as cows, bulls, and pigs, they can cause diseases like diarrhea [1].

This article gives detail information about COVID-19 which is caused by the spread of coronavirus starting from the year 2019. This virus first started spreading from the city of Wuhan, in China. The World Health Organization (WHO) declared an international public health emergency regarding the novel coronavirus on January 30, 2020, and then later on declared a global pandemic on March 11, 2020. As of

P. K. Sahu (✉) · P. Bhattacharjee · N. Agarwal

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan Deemed To Be University, Bhubaneswar, Odisha, India

e-mail: prabhatsahu@soa.ac.in

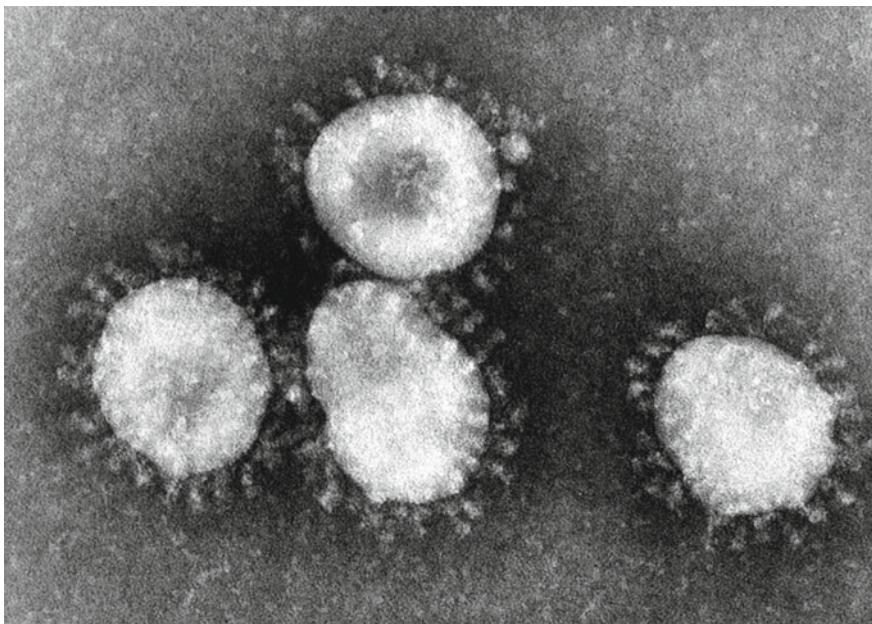


Fig. 14.1 Electron micrograph of avian coronavirus

June 6, 2021, confirmed covid cases have crossed 172 million, with confirmed deaths crossing 3.7 million, making this pandemic, one of the deadliest (if not the deadliest) pandemics in history.

This article is a Web application made by us to help provide latest information of current pandemic to the public [2]. Firstly, the information about the number of cases active, recovered, and deaths is collected from the Internet and sorted so that they can be accessed easily later on. Then, the data are stored in a file and used in the 3D world globe to show the statistics of different countries in this pandemic. Other data are also collected such as news from respected news sites regarding the current situation. We have solved various myths regarding coronavirus that have been confusing a lot of people and represented in an attractive representation so that people would be interested to read it. The Web application also provides the different helpline numbers of different states of India with an easy search facility so that users can easily find their respective state's number for call for help. The application also provides information regarding symptoms of the virus, preventive measures to avoid getting infected, and also the steps to take if one gets infected with the virus. All this is represented in attractive way to make the users more interested in it. In future, we will also add information regarding the vaccines and other important info we left to add.

14.2 Problem Definition and Objectives

The problem definition is creating awareness among the public through the Web application. The aim is to provide every kind of information regarding COVID-19 and removing the myths regarding coronavirus that has been misleading a lot of people in the society. The aim is also to provide helpline information of different states to the public with user-friendly search for easy retrieving of information [3, 4].

The objectives for the Web application are as follows:

- To provide latest news regarding the COVID-19 pandemic from authentic resources.
- To provide latest statistics in a visually attractive map about the number of active cases, recovered, and deaths due to the COVID-19 in each country.
- To provide helpline numbers with easy search and call facilities.
- To provide information about the symptoms, the precautions to be taken, the myths surrounding this situation, and other useful information.

The motivation behind selection of this article is as follows:

- The situation of humanity in this global pandemic named COVID-19 which has caused widespread panic and various others myths which is scaring the public and forcing them to take wrong steps.
- To spread awareness and removing all the myths surrounding this situation and providing people correct steps to prevent being infected and providing them with helpline numbers of different states with easy search function.

14.3 Literature Survey

14.3.1 Existing System

Examples of existing system for COVID-19 awareness provided to the public is done by WHO, Indian Government, etc.

- WHO SITE (<https://covid19.who.int/>)

It is the site created by the World Health Organization to provide all the people in the world information about various diseases that exists in the world currently. Above site URL directs us specifically to the Web site created by WHO to provide various details regarding COVID in the world scale as shown in Fig. 14.2.

- MYGOV SITE (<https://www.mygov.in/covid-19>)

It is the site created by the Indian Government to provide statistics and other details regarding COVID in India specifically. This site was specifically created to provide information regarding COID situation in every state of India. The above URL directly



Fig. 14.2 Screenshot of WHO COVID-19 site

takes to the COVID-19 dashboard of the government site MyGov. This site is the product of National Informatics Center (NIC) India as shown in Fig. 14.3.

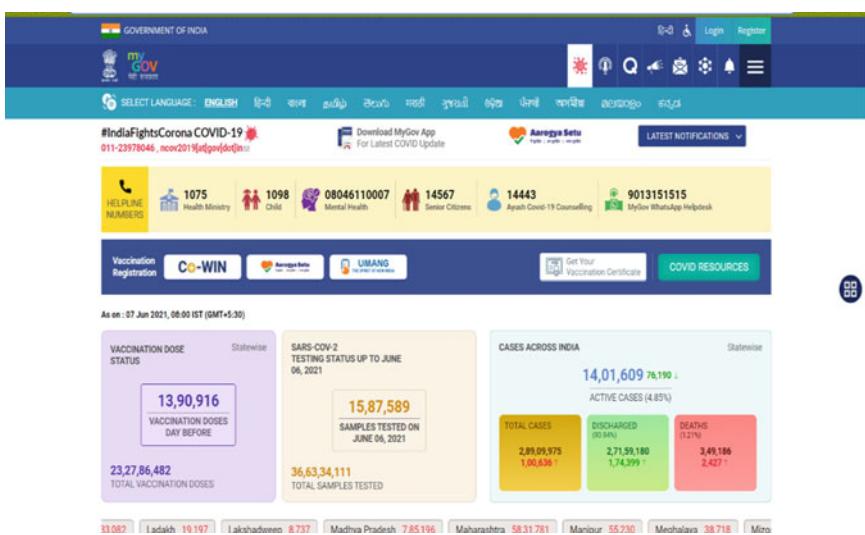


Fig. 14.3 Screenshot of MyGov site

14.3.2 Proposed System

- All the statistical data shown on the map is scrapped from Wikipedia.
- All the latest news is scrapped from google search.
- This collection of data is done using bs4 and happens every 24 h and is stored in a file in json format using an API created with flask.
- Taking the file as input and using Python analytical tools, namely numpy, pandas, and plotly, we create an interactive choropleth world map visualization with latest statistics and save it to an html file.
- An interactive search function is provided in the list of helpline numbers which is created through JavaScript.
- Rest all static data are displayed in a responsive and interactive manner.

Advantages of Proposed System

- User can get a wide variety of data related to covid at one place (basically would not require to go through 3–4 different Web sites which are dedicated toward a particular section.)
- User-friendly ui/ux.
- Visually attractive Web site.
- Gives updates every 24 h.
- Not much manual work is required, saves time and effort.

Limitations of Proposed System

- State-wise covid statistics are not present for our country.
- International helpline numbers are not available.
- Since the application is hosted free of cost, advertisements are popping-up.

14.3.3 Feasibility Study

It plays an important role in analyzing the system [5]. It helps people/groups to decide whether they should move forward with the idea to design the article and till what extend can they modify/improvise upon it. Study of article feasibility can be done on different grounds such as study of technological feasibility, feasibility in organization, and economy.

Technical Feasibility: It decides whether the technology which is proposed to be used, while implementation of the article idea is available or not and if this technology is going to have long-term support or be deprecated in near future. It also determines if the technology can be integrated with the system and the degree of flexibility of the system (if it can work with alternating components or not).

Organizational and Culture Feasibility: It decides whether the proposed idea, if implemented will suite the current working environment, be it an organization or an institution, and whether it adheres by the rules and regulations of the system

or not. Another very important aspect of this feasibility is to decide whether the idea or the solution will meet the expectations of the people on the receiving end (whether the users find the implemented idea/solution easier to work with or not). The technological tools used in our article are very much used in the software industry, which are very flexible and hence can be easily integrated in the system.

Economic Feasibility: This category of study of feasibility majorly deals with the investment going in and the output of the proposed work. Investment can be not only in monetary terms but also other important aspects like providing resources (including manpower), time consumption in design, and implementation. Then, it decides whether the output can recover the cost of all inputs or not. Hence, effectiveness of the idea is decided. In this article, there is no monetary investment, just our efforts act as the input, and compared to the amount of practical knowledge and experience gained, we can say that the system is effective [6, 7].

14.4 Design Steps and Algorithm

14.4.1 *Design Steps*

First of all, we need to store all the static data in the front end. We will use JavaScript to create search facility for state-wise covid helpline numbers. We will also scrap and store all statistics and news in files and create an API for this purpose. Next, we will create map and plot the data for effective visualization.

14.4.2 *Algorithm*

- STEP 1. All the statistical data shown on the map is scrapped from Wikipedia.
- STEP 2. All the latest news is scrapped from google search.
- STEP 3. This collection of data is done using bs4 and happens every 24 h and is stored in a file in json format using an API created with flask.
- STEP 4. Taking the file as input and using Python analytical tools, i.e., NumPy, pandas, plotly, we create an interactive choropleth world map visualization with latest statistics and save it to an html file.
- STEP 5. An interactive search function is provided in the list of helpline numbers which is created through JavaScript.

Pseudocode for scraping and saving statistics and news

Get the data from Wikipedia [1], parse the data, append data to a dictionary, and dump it to a json file.

Pseudocode for creating world map from saved statistics

Read the country codes, merge them with the statistics, create a layout for the map, plot the map using the merged statistics, and save it to an HTML file.

14.5 Results and Discussions

This is one of the most important chapters of an article report. It actually describes or proves whether the design requirements and the expectations from the initial idea have been achieved or not. This section describes several aspects such as evaluation of design, degree of success (in terms of achievement or output), and feasibility conditions evaluation. The user interface has been described in Table 14.1.

14.5.1 User Interface

14.5.2 Outputs/Screenshots

See Figs. 14.4, 14.5, 14.6, 14.7, 14.8, 14.9, and 14.10.

Table 14.1 Description of user interface

Screen name	Description
Statistics	Shows an interactive 3D world map visualization with latest covid statistics (country-wise) on hover
Newsfeed	Shows a list of latest news related to covid with date as well as link to corresponding articles
Helpline	Shows state-wise helpline numbers in a tabular format with quick search and calling facilities
Myth-busters	Answers frequently asked questions and myths in an accordion format
If I'm sick	Shows points to prevent the spread of covid-19 if someone is sick
Symptoms	Shows the common symptoms and spreading ways of COVID-19 virus
Prevention	General guidelines for protection from COVID-19 pandemic

Symptoms of COVID-19



Ways through which COVID-19 spreads



Fig. 14.4 Screenshot of symptoms tab of our Web site

14.6 Socioeconomic Issues

14.6.1 Practical Relevance

- Our article COVID AWARENESS has a lot of practical relevance; it helps the public know a lot about the current pandemic and helps them know about necessary information so that they can take the right steps.

Steps to prevent the spread of COVID-19 if you are sick



Fig. 14.5 Screenshot of sickness tab of our Web site

Guidelines for prevention from COVID-19



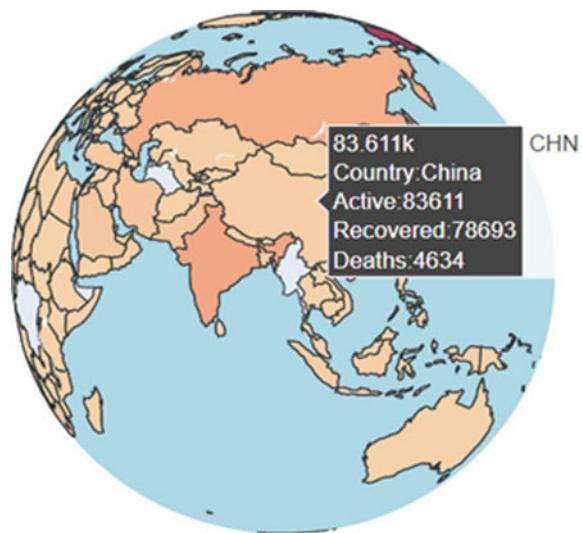
Fig. 14.6 Screenshot of prevention tab of our Web site

Frequently asked question about COVID-19

How deadly is Coronavirus ?	Can covid19 be passed through food ?
Coronavirus has Infectivity but Mortality.	Can eating chicken cause covid19 ?
Will I die if I catch the disease ?	Which mask protects against covid19?
Who is getting sick ?	Can pets/ animals carry covid19?
Can eating garlic kill this virus ?	Do hand-dryers kill covid-19?
Will warm weather kill covid-19?	Can an infected person recover ?
How long can the virus survive ?	Can UV light kill the new coronavirus?

Fig. 14.7 Screenshot of myth-busters tab of our Web site

Fig. 14.8 Screenshot of interactive world map with latest corona statistics



Latest NEWS of COVID-19



Fig. 14.9 Screenshot of latest news related to coronavirus

- Our article is created to provide all the latest statistics related to the pandemic and display them in an attractive way so that the users get interested in viewing them.
- This article would also help remove the misconceptions created by rumors regarding the virus that have been causing a lot of bad decisions in the public. Due

Central Helpline Number for COVID-19: +91-11-23978046, Toll Free no: 1075	
Helpline Numbers across India	
States	Helpline Number
Andhra Pradesh	0866-2410978 ☎
Arunachal Pradesh	9436055743 ☎
Assam	6913347770 ☎
Bihar	104 ☎
Chhattisgarh	104 ☎
Goa	104 ☎
Gujarat	104 ☎
Haryana	8558893911 ☎

Fig. 14.10 Screenshot of helpline tab with search and call facilities

to our article, the public will be able to remove their doubts and listen to doctors more keenly.

14.6.2 *Global Impact*

- Our article spreads awareness across the world, specially in our country India because it provides awareness to the public about the pandemic and helps in removing the rumours regarding it.
- It also helps in providing public latest details on the active, recovered, and death cases and along with it latest news regarding the pandemic that is happening all around the world.
- This also helps in providing the public latest information easily.
- The helpline search function of our article helps the citizen of India to get their state helpline numbers that makes it easy for people to get help.
- Our article also provides necessary guidelines to people so that they can be safe from the virus, and if infected, they can follow the steps to become cured.

14.7 Conclusion and Future Scope

The main objective of this Web application is to present a variety of data relative to the global pandemic situation in a user-friendly, visually appealing, and organized way with regular updates. The statistics are presented in a 3D world map plot so that people can easily search and compare stats of different countries. The helpline numbers can be easily searched for using our search function and can be directly called from the Web site. All other static data is also presented in a responsive and interactive way. The main aim is to provide a reliable source which people can refer to when required so that the people stay aware regarding this pandemic situation.

The system can be horizontally scaled, i.e., it can be made to expand/contract according to user requests rate. Details regarding vaccination and slot booking can be included. Access speed can be optimized. Design can be modified. Mailing system for concerned authorities can be introduced.

References

1. Richardson, S., Hirsch, J.S., Narasimhan, M., Crawford, J.M., McGinn, T., Davidson, K.W., Northwell COVID-19 Research Consortium: Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. *Jama* **323**(20), 2052–2059 (2020)
2. Grant, M.C., Geoghegan, L., Arbyn, M., Mohammed, Z., McGuinness, L., Clarke, E.L., Wade, R.G.: The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): a systematic review and meta-analysis of 148 studies from 9 countries. *PloS One* **15**(6), e0234765 (2020)
3. Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., Liu, L., Shan, H., Lei, C.L., Hui, D.S., Du, B.: Clinical characteristics of corona virus disease 2019 in China. *N. Engl. J. Med.* **382**(18), 1708–1720 (2020)
4. Castro, M.C., Kim, S., Barberia, L., Ribeiro, A.F., Gurzenda, S., Ribeiro, K.B., Abbott, E., Blossom, J., Rache, B., Singer, B.H.: Spatiotemporal pattern of COVID-19 spread in Brazil. *Science* **372**(6544), 821–826 (2021)
5. Perondi, B., Miethke-Morais, A., Montal, A.C., Harima, L., Segurado, A.C.: Setting up hospital care provision to patients with COVID-19: lessons learnt at a 2400-bed academic tertiary center in São Paulo, Brazil. *Braz. J. Infect. Dis.* **24**, 570–574 (2021)
6. Carmona, M.J., Quintão, V.C., Melo, B.F., André, R.G., Kayano, R.P., Perondi, B., Miethke-Morais, A., Rocha, M.C., Malbouisson, L.M., Auler-Júnior, J.O.: Transforming operating rooms into intensive care units and the versatility of the physician anesthesiologist during the COVID-19 crisis. *Clinics* **12**, 75 (2020)
7. World Health Organization: Public Health Surveillance for COVID-19: Interim Guidance, 7 August 2020. No. WHO/2019-nCoV/Surveillance Guidance/2020. World Health Organization (2020)

Chapter 15

Implementation of Blockchain-Based Cryptocurrency Prototype Using a PoW Consensus Mechanism



Danish Raza, Pallavi Nanda, and Sudip Mondal

Abstract The term “blockchain” has arisen as a dynamic technological force in a variety of government and private-sector processes. The technology creates a data structure with built-in security features. It is based on encryption, decentralization, and consensus principles, which maintain transaction trustworthiness. The fundamental target of this blockchain-based work is to give data with respect to digital currency advancement utilizing PoW, a productive agreement component. The information and ideas presented are based on the blockchain’s features, which provide changeable and unchangeable advanced records created without the use of a central repository and, in most cases, without the use of a central authority. They enable a local area of clients to see and write exchanges in a common record in that local area, with the purpose that no exchange may be changed after distribution under normal circumstances of the blockchain network. This article provides a study of high-level specialist overview of blockchain technology.

15.1 Introduction

Blockchains [1] are distributed, tamper-proof digital ledgers that do not have a single repository and are tamper-proof and resistant to falsification. Blockchains allow a number of participants to keep transactions in a shared ledger within a community as long as the blockchain network is active, with the effect that no transaction can be altered once it has been recorded [2]. In modern cryptocurrencies, which are electronic cash secured by cryptographic processes rather than by a centralized repository or authority, the blockchain notion was integrated with a range of other technologies and computer concepts in 2008 [3–5]. First ever cryptocurrency based on a blockchain was Bitcoin. The Bitcoin blockchain links information that represents electronic cash to a digital address [6]. Digitally signing and transferring ownership of content to another user is possible with Bitcoin, and the Bitcoin blockchain records

D. Raza · P. Nanda · S. Mondal (✉)

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

e-mail: sudipmondal@soa.ac.in

all transactions for all authorized users to independently verify their legitimacy [7]. The Bitcoin blockchain is cooperatively maintained and managed by a distributed group of persons. The blockchain is impenetrable against further initiatives to manipulate the ledger, like as altering blocks or fabricating transactions, thanks to this and cryptographic methods [8]. A distributed group of people maintains and manages the Bitcoin blockchain collectively. The blockchain is impenetrable against further initiatives to manipulate the ledger, like as altering blocks or fabricating transactions, thanks to this and cryptographic methods [9]. The use of blockchain technology has created a lot of excitement, but the technology is not well-understood by the general public. It is not a miracle, and it is not going to solve all of your problems. There is a natural urge to apply modern technology to every industry and in every way conceivable, as it does with every technological advances. This article contains the knowledge needed to gain a rising awareness of the technology and ensure correct application. Blockchain technology, which is named after the extensive usage of cryptographic functions, underpins modern cryptocurrencies [10]. Public and private keys are used by users to digitally sign and transact securely within the system. Using cryptographic hash functions, users can overcome such hurdles in digital currency blockchain networks that include mining, in the hopes of being rewarded with a specific amount of money [11, 12]. On the other hand, blockchain technology may be more generally helpful than cryptocurrencies [13]. In this article, we emphasize on the bitcoin usage case because that is the most popular technique of the current technology; nevertheless, awareness in other domains is developing. We chose to move on with a proposal that firmly establishes cryptocurrencies supported by blockchain as the currency of the future, taking into account the immense potential of the cryptocurrency industry and its clear benefits over present financial systems (banks and fiat currency).

15.2 Overview of the Transaction Execution

- **Transaction:** A transaction that modifies the state of a blockchain record. The transaction can be a monetary value exchange or the execution of an intelligent contract agreement, depending on the application.
- **Block:** It consists of a block header and a block information block. Block metadata data such as the Merkle tree root hash, previous block hash, date, and block form can be found in a block header; however the data is composed of a number of genuine trades that have occurred.
- **Merkle tree (root hash):** A hashing calculation is used to hash all of the transactions in the block. The hash esteems are then linked pairwise and hashed again until there is only one hash esteem left. This value is known as the Merkle tree root hash esteem.
- **Block hash:** We get it by continuously hashing the block header to get the block's unique identity.

- Previous block hash: It is the hash of the block before this one in the chain. The current block's parent is the block before it, and vice versa. Using the previous block hash value in the block header ensures that the blockchain record is permanent.
- Timestamp: It shows when the block was made.
- Block variant: Here, we can see what kind of blockchain conventions are being used.
- Mining: An exchange block is a grouping of valid exchanges that is broadcasted to the rest of the company.
- Genesis block: This is the most important section of the document. All of the blocks in the chain are connected to the block that started the chain and so on. Among the things that are contained in the beginning block are the organization attributes, the contract convention, the entry control permissions, the hash function, the block age stretch, and block size.

15.2.1 The Execution Stream Comprises the Accompanying Advances

15.2.1.1 Transaction Proposition

The user hashes the share information using a hash function to ensure that the data is trustworthy later. The hashed data is then encoded with the client's private key to provide client verification, and the scrambled yield is referred to as the advanced mark of that transaction. The organization is informed of the exchange of data and the mark.

15.2.1.2 Transaction and Block Approval

Each full hub in the organization approves the exchange by performing two undertakings: (a) client confirmation by unscrambling the advanced mark utilizing the public key of the proposing client, and (b) information uprightness by hashing the exchange information and contrasting it and the decoded signature.

The organization's block generators or miners receive the permitted transaction. In accordance with an agreement, a designated miner verifies the authenticity of the transactions and then combines them into a defined block size. The miner keeps a record of the Merkle root hash value. The Merkle root computes a productive cycle to examine an exchange in a block based on the sum of the hashes of a large number of exchanges. A hub only needs the hash advantages of the Merkle way linking the interchange to the Merkle root to validate if an agreement is preserved for a block. As a result, a hub that does not hold a complete copy of the record might confirm an exchange by requesting the path without having to obtain the entire block, reducing correspondence costs. When a hub uses Merkle root to confirm an exchange in a

block of n exchanges, it only needs $\log_2 n$ hash values as opposed to n hash values if Merkle root is not used. The block hash is generated after the Merkle root hash value is computed. The excavator notifies the organization of the stumbling stone. The approving hubs validate the block's legitimacy by checking the following: (1) the block hash, (2) the block timestamp is bigger than the previous block's timestamp, (3) the block height and size esteems, (4) the previous block hash worth, and (5) the authenticity of the block's numerous exchanges. The major chunk of the data is appended to each approved hub's own clone.

The reproduction of the record in a blockchain removes the challenges of quality requirements and data ownership by a gathered together specialized co-op, in addition to the concerns of a weak link and high organization idleness. The record's duplication should ideally be predictable and easily available between hubs. Data transfers, on the other hand, may be delayed or lost in a dispersed framework where an organization section may emerge. As a result, maintaining high consistency while yet being accessible is a difficult undertaking. As a result, a solution should be found. The blockchain network uses the monotonic prefix consistency [14] replication technique (MPC). Every hub in the organization is connected to n other hubs, which are connected to n more, producing a hub chain of increasing importance.

15.2.2 Single-Ledger-Based Architecture

The Ethereum [15] platform presented this design in 2013, and the members of the organization are addressed by companions (or hubs). A hub could be basic, comprehensive, or mining. A customer connects to the blockchain using the RPC, and an integration administrator connects to an external system using the RPC. An exterior framework is utilized when an exchange's permission is based on external facts such as the current weather, the cost of an offer market, or the cash conversion scale. If the external framework appears to be shady, the transaction's legitimacy is called into question. The following is the transaction execution stream in this engineering:

1. The exchange payload is created and hashed by the customer.
2. The hashed payload's computerized mark is created.
3. The organization is informed of the exchange payload and advanced mark.
4. The validators approve the trade, which is then communicated to the miners.
5. A chosen miner creates a block of significant exchanges.
6. The organization is informed about the block.
7. The block chains confirm the block, and the record is renewed.

For example, this concept can be applied to transportation, production network and project management, computerized content ownership, accounting, and energy trading. We cannot utilize it for private applications such as medical services or education because anyone may join, the flow of information is public, and there is no admittance control.

15.3 Workflow and Their Implementation

See Fig. 15.1.

15.3.1 Ensuring Immutability Using Hashing (SHA-256)

is_chain_valid()—This api checks if the new block mined is valid or not by checking the proof of work and verifying the previous hash key of the block. A hash is also a part of a block. It is similar to a fingerprint that is unique to each block. It uniquely identifies a block and all of its contents, similar to a fingerprint. As a result, once a block is generated, every modification within the block will substantially alter the hash. As a result, the hash is quite useful for detecting changes in intersections.

For example: Loop over ever block in the chain

```
{
```

Check if the previous_hash field of the block = hash of the previous block.

Check if $(\text{proof of current block}^2 - \text{proof of previous block}^2)$ has 4 leading 0.

```
}
```

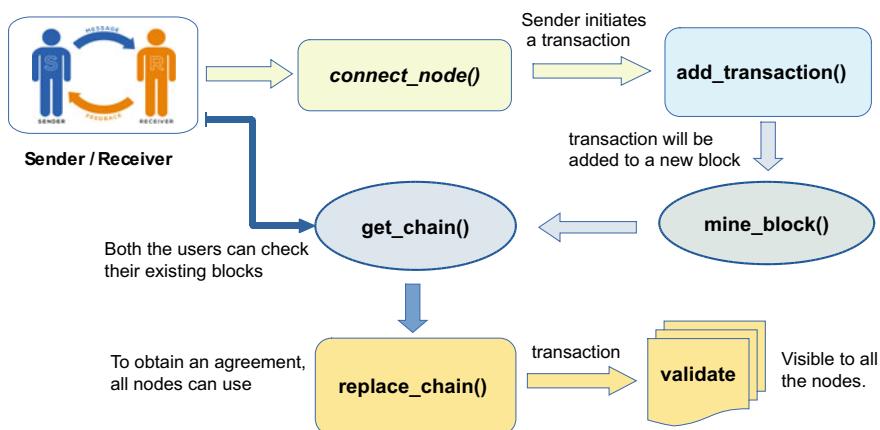


Fig. 15.1 Systematic diagram of the transaction workflow

15.3.2 Mining I.e mine_block()

Users can mine or create a new block in our DaP coin blockchain using this api. This function uses proof of work algorithm.

To protect vulnerable usage of computer resources as well as other attacks on the system, including such denial-of-service assaults as well as other service abuses including spamming, a system must conduct a considerable number of calculations by convincing the service requester to complete some costly work. In a blockchain network, the PoW [16, 17] consensus mechanism requires network mining nodes to prove that the work they have done and submitted qualifies them to add new transactions and blocks to the blockchain. To add the next block to the blockchain, proof of work nodes are picked in accordance with their computational or hashing capacity. Using their computational power, the nodes compete with one another. It operates by calculating hash values and validating transactions until another hash value contains the desired number of leading zeros. A nonce is the number that generates the hash with the specified number of trailing zeros. The miners have to spend a lot of money on hardware and electricity to execute these calculations, whereas the verification of this mathematical challenge is simple and quick.

For example: Check if the SHA-256 encoded form of (new_block proof² – previous_block proof²) starts with four zeroes:

if the above is true, then add this block to the chain in the current node and give the miner a reward.

15.3.3 Decentralized P2P I.e connect_node()

This function helps a particular user (or node) to connect with other users (or nodes), which makes our DaP blockchain network decentralized and P2P. They will need to make a POST request using the nodes JSON file that we have made available along with necessary changes.

Nodes.json: It is required while connecting a particular node to other nodes in the network. It looks like

```
{
  "nodes": [http://127.0.0.1:5001,
             http://127.0.0.1:5002]
}
```

Peer to peer (P2P) is a decentralized communications technology that consists of a set of computers or devices (nodes) that also can collaboratively store and distribute data. Each node in this system functions as a separate peer. Because the communications are done without the need of a central server, the middleman is eliminated, and

all nodes have equal power to do the same activity. P2P technology can be classified as structured, unstructured, or hybrid peer-to-peer networks and is suited for a variety of use cases. P2P architecture is the foundation of most cryptocurrencies because it is decentralized. It allows anyone to transact anywhere around the world without the need for intermediaries or central servers, etc.

15.3.4 Transparency I.e get_chain()

This function will allow users to view the existing chain in a particular node of our DaP blockchain. Our blockchain is fundamentally made of a list of dictionaries. So, this function will just display the entire list of dictionaries (blocks) to the user.

15.3.5 Transactions I.e add_transaction()

This function helps users to transact DaP coins between them. They will need to make a post request using the transactions JSON file we have made available, along with necessary changes.

Transaction.json: It is required to transact DaP coins in our blockchain. It looks like

```
{  
    "sender": " ",  
    "receiver": " ",  
    "amount": " "  
}
```

15.3.6 Consensus Mechanism I.e replace_chain()

This function replaces the chain with the longest chain among all of the nodes, i.e., does consensus.

If: ((length of current chain < greatest length of chain among all the nodes)) AND (is_chain_valid is True).

then: replace the current chain with the longest chain in the network.

else: display the current chain as the rightful chain.

Consensus protocol forms the backbone of a blockchain network. It helps all the nodes in the distributed network to verify transactions and come to a common agreement about the present state of the distributed ledger. In this way, a decentralized network achieves reliability and establishes trust between unknown parties (peers) without a central governing agency.

15.4 Relevance in the Real World

Onboarding, record keeping, customer screening, data management, security, privacy, and transaction and trade processing are some of the practical applications of blockchain in the financial services business. In the insurance industry, there are also costly errors, scams, and false claims. According to the FBI, fraud in non-health insurance industries has resulted in the theft of more than \$40 billion. By taking on the time-consuming task of validation, blockchain can speed up and safeguard the claims process. Customizable, self-executing smart contracts based on blockchain technology can assist parties in maintaining ownership rights and complying with privacy regulations. Smart contracts can be used for the benefit of entertainment industries. The entertainment industry will benefit from blockchain's traceable real-time reporting network.

Cryptocurrencies have already spawned a thriving business. They are owned by more transparent institutions that are responsible for overseeing all digital coin exchanges around the world. Cryptocurrencies are decentralized, allowing users to trade currency across national borders. Technology will enable a financial revolution, allowing everyone to be more financially connected, empowered, and enabled. Because cryptocurrencies and blockchain do not require physical infrastructure or operations, the expenses of transacting with them are modest. All blockchain and cryptocurrency transactions are tracked on a distributed ledger because they are automated and digital. The finest element is that it cannot be influenced by individuals or businesses, dramatically reducing the risk of fraud and corruption.

It will only be a matter of time before these cryptocurrencies make their way into our lives, transforming them for the better, with an emphasis on economic process and inclusivity. Thanks to the wonderful possibilities that cryptocurrencies bring to the table, millions of individuals will now be able to invest, send money across borders, save money, and establish a business.

15.5 Conclusion

The cryptocurrency model described in this paper demonstrates how a PoW-based blockchain network works in a straightforward and practical manner. It has all of the essential characteristics of a blockchain-based transaction system, such as decentralization, consensus, and the ability to see and record transactions. Through connecting

to the local server, several users can transact with one another. Although practically all blockchains have scalability concerns, they can be effectively employed in a private zone for ultra-fast micro-transactions of coins or tokens.

We intend to use our blockchain not only for crypto transactions, but also for non-fungible tokens (NFT) in the future. It is a type of digital asset or token that lets us represent real-world assets such as art, music, in-game stuff, and movies. So, in a nutshell, they allow us to tokenize real-world items by storing the data in a blockchain, which certifies a digital asset's uniqueness and proves its ownership history. They are often encoded with the same underlying logic as cryptocurrency or tokens and are bought and traded online, often with the help of cryptocurrency. NFTs are different from cryptocurrency or fiat money because they cannot be traded or exchanged for one another. Consider that \$1 is always the same as \$1, or that 1 bitcoin is always the same as 1 bitcoin, but NFTs are impossible to exchange for or equal.

References

1. Maesa, D.D.F., Mori, P.: Blockchain 3.0 applications survey. *J. Parallel Distrib. Comput.* **138**, 99–114 (2020)
2. Yaga, D., Mell, P., Roby, N., Scarfone, K.: Blockchain Technology Overview (2019). arXiv preprint [arXiv:1906.11078](https://arxiv.org/abs/1906.11078)
3. Swan, M.: *Blockchain: Blueprint for a New Economy*. O'Reilly Media, Inc. (2015)
4. Bashir, I.: *Mastering Blockchain: Distributed Ledger Technology, Decentralization, and Smart Contracts Explained*. Packt Publishing Ltd. (2018)
5. Wang, Y., Han, J.H., Beynon-Davies, P.: Understanding blockchain technology for future supply chains: a systematic literature review and research agenda. *Supply Chain Manage. Int. J.* (2019)
6. Romano, D., Schmid, G.: Beyond bitcoin: a critical look at blockchain-based systems. *Cryptography* **1**(2), 15 (2017)
7. Niranjanamurthy, M., Nithya, B.N., Jagannatha, S.: Analysis of Blockchain technology: pros, cons and SWOT. *Clust. Comput.* **22**(6), 14743–14757 (2019)
8. Miller, R.: Rules for Radicals-Settling the Cyber Frontier. *IntellectualCapital.com*, vol. 1 (1999)
9. Pilkington, M.: Blockchain technology: principles and applications. In: *Research Handbook on Digital Transformations*. Edward Elgar Publishing (2016)
10. Kimani, D., Adams, K., Attah-Boakye, R., Ullah, S., Frecknall-Hughes, J., Kim, J.: Blockchain, business and the fourth industrial revolution: whence, whither, wherefore and how?. *Technol. Forecast. Soc. Change* **161**, 120254 (2020)
11. Narayanan, A., Bonneau, J., Felten, E., Miller, A., Goldfeder, S.: *Bitcoin and cryptocurrency technologies: a comprehensive introduction*. Princeton University Press (2016)
12. Catalini, C., Gans, J.S.: Some Simple Economics of the Blockchain (No. w22952). National Bureau of Economic Research (2016)
13. Jaag, C., Bach, C.: Blockchain technology and cryptocurrencies: opportunities for postal financial services. In: *The Changing Postal and Delivery Sector*, pp. 205–221. Springer, Cham (2017)
14. Girault, A., Gössler, G., Guerraoui, R., Hamza, J., Seredinschi, D.A.: Monotonic prefix consistency in distributed systems. In: *International Conference on Formal Techniques for Distributed Objects, Components, and Systems*, pp. 41–57. Springer, Cham (2018, June)

15. Bogner, A., Chanson, M., Meeuw, A.: A decentralised sharing app running a smart contract on the ethereum blockchain. In: Proceedings of the 6th International Conference on the Internet of Things, pp. 177–178 (2016)
16. Liu, D., Camp, L.J.: Proof of work can work. In: WEIS (2006)
17. Gervais, A., Karame, G.O., Wüst, K., Glykantzis, V., Ritzdorf, H., Capkun, S.: On the security and performance of proof of work blockchains. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 3–16 (2016)

Chapter 16

Employing Deep Learning for Early Prediction of Heart Disease



Abdul Aleem, Ayush Raj, Rahul Raj Sahoo, and Amulya Raj

Abstract Numerous people get affected by heart diseases due to their daily inappropriate living habits. Prediction of heart disease at an earlier stage becomes crucial to prevent the disease or in the treatment of the disease. However, predicting the heart condition accurately as per symptoms is challenging, even for experienced doctors. The most demanding job is to anticipate the illness accurately. The medical or health sector is generating a large amount of data about heart disease every year. The easy availability of data in medical and healthcare fields and the accurate analyzing techniques for the medical data of earlier patients make it possible to predict various diseases. Machine learning algorithms may be used with medical information to forecast cardiac illness successfully. Developing the predictor of heart disease using machine learning algorithms is more beneficial than conventional approaches for accurate predictions of disease. This article proposes a deep learning model better than the other existing machine learning techniques. The dataset utilized for prediction is the heart dataset available at <https://www.kaggle.com/>. The proposed model has been compared with k-nearest neighbors (kNN), logistic regression (LR), support vector machine (SVM), Naïve Bayes (NB), decision tree (DT), random forest (RF), and artificial neural network (ANN). The comparative analysis of various machine learning techniques establishes the proposed model as the most precise heart disease predictor.

16.1 Introduction

People aim for a luxurious and comfortable life. To fulfill their wish for comfort, they work hard like robots and earn enough money to live the desired life. Nevertheless, they forget to take care of their health. People's busy schedules and inappropriate lifestyles cause various changes in their daily habits, which are detrimental to their health, causing them to suffer from diseases at a very early age. People do not pay

A. Aleem (✉) · A. Raj · R. R. Sahoo · A. Raj

Department of Computer Science and Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

e-mail: abdulaleem@soa.ac.in

much attention to the kind of food they eat. Nonetheless, after all this negligence, when people fall ill, they take medication without consulting a doctor, which makes people more prone to diseases.

World Health Organization (WHO) states that a large population worldwide dies due to heart disease every year (17.9 million as per <https://www.who.int>). Among all fatal illnesses, it is the most prevalent cause of death. The heart is a significant body part that helps in the flow of blood throughout the body. Any dysfunction in the heart could fail other body parts as well. Medical practitioners undertake several surveys on heart diseases to collect information on heart patients, their symptoms, and the development of the condition. Patients with prevalent illnesses that exhibit typical symptoms are increasingly being reported.

The emergence of machine learning enables computers to think more intelligently than humans. Machine learning uses various techniques to analyze past data and come up with better decision making. These techniques could help medical professionals predict heart disease, which is a major concern for human beings around the world. The digital data related to health is enormous, providing difficulty for the medical experts in effectively identifying ailments early on. Hence, modern computational approaches such as machine learning algorithms have been established to identify significant patterns of data and hidden information that can be utilized to decide critically. The easy availability of medical data and the application of various machine learning techniques opens the door for disease prediction. However, the major challenge to date is the accuracy of such prediction, as a wrong prediction may cause even the loss of life.

This article aims to study the usage of numerous machine learning algorithms for predicting heart disease at an early stage and then proposes a deep learning model for the most accurate prediction of heart disease. Numerous machine learning techniques, like, KNN, NB, DT, SVM, RF, ANN, and LR, will be assessed for disease identification. Early diagnosis will help in the speedy recovery and control the situation. The overarching aim of this study is to correctly predict the existence of heart disease with a few tests and attributes. The comparative study applies various machine learning algorithms on the heart dataset available at Kaggle and analyzes which algorithm performs better and gives the most accurate result compared to others to predict heart disease at a very early stage. The attributes of the dataset are analyzed so that results provide more or less reliable findings. The research aims to help health professionals with quick and accurate detection of cardiac disease. The proposed deep learning model is the top-performing machine learning model for predicting heart disease.

16.2 Related Work

Many research works have already been carried out related to heart disease prediction. Different works achieved different accuracies. Shah et al. [1] proposed several machine learning algorithms like NB, kNN, DT, and RF. For this system, the highest

accuracy was achieved in the case of kNN for $K = 7$. Rajdhan et al. [2], in their study, proposed a model implementing machine learning algorithms like DT, LR, RF, and NB. The highest accuracy was achieved in the case of the RF algorithm. Lutimath et al. [3] proposed a model for predicting heart disease using algorithms like NB and SVM, among which SVM shows better accuracy in the prediction of heart disease. Performance of both the algorithm was evaluated using root mean squared error (RMSE), sum of squared error (SSE), and mean absolute error (MAE).

Nikhar and Karandikar [4], in their work, proposed a model to predict cardiac disease using algorithms like NB and DT algorithms, out of which DT algorithm performed better than the NB algorithm. Parthiban and Srivatsa [5] proposed a model of whether a diabetic patient has the chance of getting heart disease or not using a diabetic diagnosis. Two algorithms used for the purpose were NB and SVM. The inference drawn from these two algorithms was that SVM showed higher accuracy in predicting the chances of getting heart disease for patients diagnosed with diabetes.

Apart from these, various traditional machine learning algorithms, viz. NB, DT, RF, SVM, kNN, etc., have been used in numerous research works [6–10]. Aljanabi et al. [11] have reviewed all such research works that employ machine learning techniques to predict heart disease. The main motivation of all the research done in these existing articles is to use several machine learning algorithms to develop a heart disease prediction system using the datasets available and analyze which algorithms have better accuracy in predicting heart disease. The research in this article moves a step further and applies neural networks to predict heart disease and measures its effect in terms of accuracy.

16.3 Proposed Work

This article proposes a prediction model based on the deep neural network technique which predicts heart disease from the dataset available. This article also presents the comparative result of the proposed DNN model with several machine learning algorithms used to predict heart-related disease, viz. kNN, SVM, NB, LR, DT, RF, and ANN. The proposed work is effective in predicting whether one person had a heart-related disease or not. Various machine learning algorithms were used so that analysis of all the used algorithms can be done using which it could be determined which algorithm was more accurate in predicting the result. The metrics used for comparison are precision, recall, f1-score, and accuracy percentage. These are well-known standard metrics used for the comparison of machine learning algorithms.

16.3.1 *Dataset Exploration and Analysis*

The dataset used for developing the heart disease predictor is collected from Kaggle. The dataset consists of medical data of the different patients of different age groups.

Three hundred three rows are representing the medical data of 303 different patients. Each row has 14 columns, among which the first 13 columns represent the patient's medical information and the last column consists of information on whether the particular patient has heart disease or not. The medical characteristics of the patients considered are the patient's age, sex, maximum heart rate, fasting blood sugar level, resting blood pressure, etc. This dataset is bifurcated into the training portion and testing portion. Table 16.1 shows information about dataset features.

The attributes of the dataset have been analytically visualized for their significance. Figure 16.1 shows a plot for the frequency of heart disease versus the ages of the patients. Dataset consisted of information about patients of ages lying between 29 and 77 years. Furthermore, this plot shows how many patients of a particular age have heart disease and how many do not have heart disease.

Table 16.1 Dataset features description

S. No.	Attribute	Description
1	Trestbps	Blood pressure while resting
2	Restecg	Electrocardiographic measures while resting
3	Thal	Type of defect (normal/fixed/reversible)
4	Age	Patient's age (years)
5	Cp	Type of chest pain
6	Sex	Male/female
7	Oldpeak	Exercise-related ST depression (yes/no)
8	Exang	Angina due to exercise (yes/no)
9	Fbs	Fasting sugar in blood (high/low)
10	Chol	Cholesterol content in serum (mg/dl)
11	Slope	ST segment slope with respect to peak exercise
12	Ca	Fluoroscopy-colored major vessels count (0–3)
13	Thalach	Maximum count of heart beats
14	Target-output	No disease or heart disease (0/1)

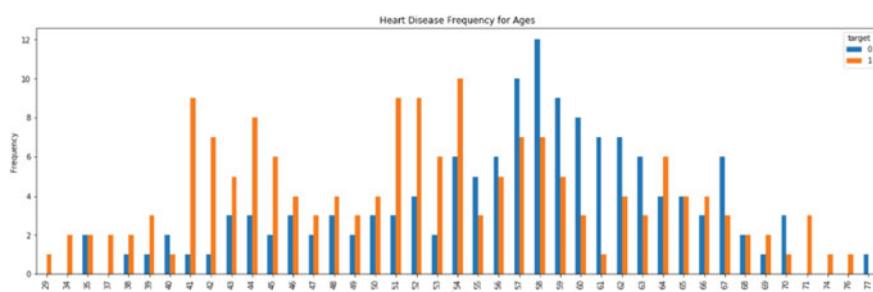


Fig. 16.1 Heart disease frequency versus ages

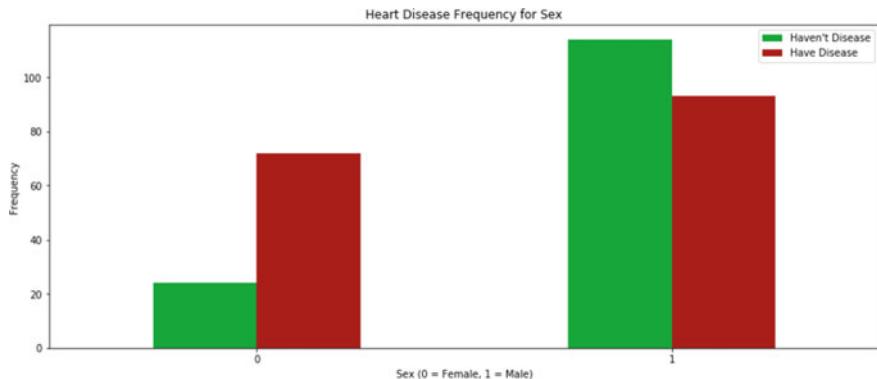


Fig. 16.2 Heart disease frequency versus sex

Similarly, Fig. 16.2 shows the plot for heart disease frequency versus gender of the patients highlighting the number of male and female patients who have heart disease and who do not have heart disease. Figure 16.3 shows the frequency of heart disease concerning the ST segment slope for peak exercise, having values as 0 for up sloping, 1 for flat, and 2 for downsloping. This shows the number of patients having or not having heart disease when slope parameter values are up sloping, flat, or downsloping.

The other attributes assessed in the dataset are fasting blood sugar level and the type of chest pain. Figure 16.4 shows the plot for heart disease frequency versus fasting blood sugar level with the values 1 for $fbs > 120$ mg/dl and 0 for $fbs < 120$ mg/dl. The plot shows the number of patients having or not having heart disease when their fasting blood sugar is greater or less than 120 mg/dl. Figure 16.5 shows the plot for heart disease frequency versus chest pain type with the values 0 for asymptomatic, 1 for atypical angina, 2 for non-anginal pain, and 3 for typical angina.

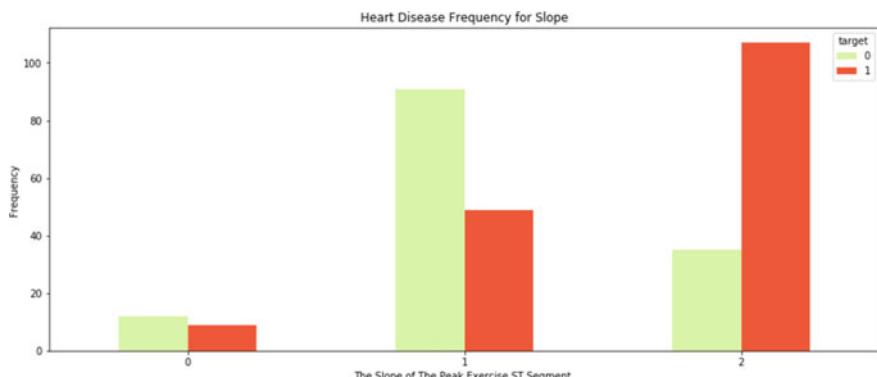


Fig. 16.3 Heart disease frequency versus slope

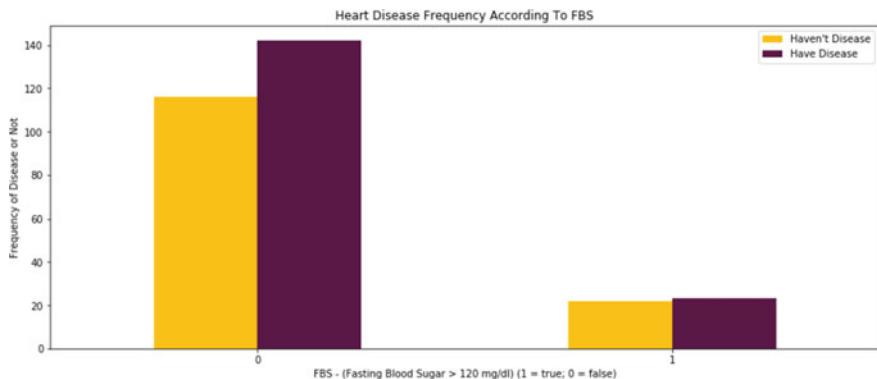


Fig. 16.4 Heart disease frequency versus FBS

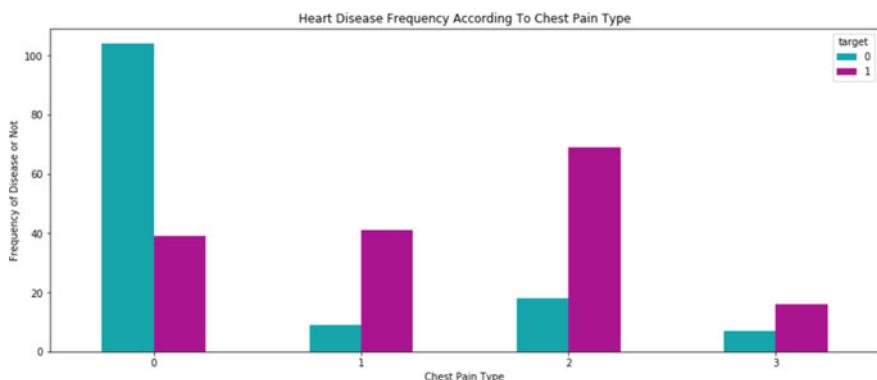


Fig. 16.5 Heart disease frequency versus chest pain type

This plot shows the number of patients having or not having heart disease regarding the stated values for the chest pain type.

16.3.2 Proposed Deep Neural Network Model

This article proposes a prediction model with better accuracy and a deep learning model with a configuration optimized for better results. The model based on deep learning predicts the occurrence of heart disease for patients based on the inputs. The workflow of the proposed model is depicted in Fig. 16.6. The workflow has six steps, namely dataset preparation, preprocessing data, DNN model configuration, train the model, test it, and evaluate results.

The preparation of the dataset required just loading it from the .csv file. The preprocessing of data required removing rows with missing or incomplete values.

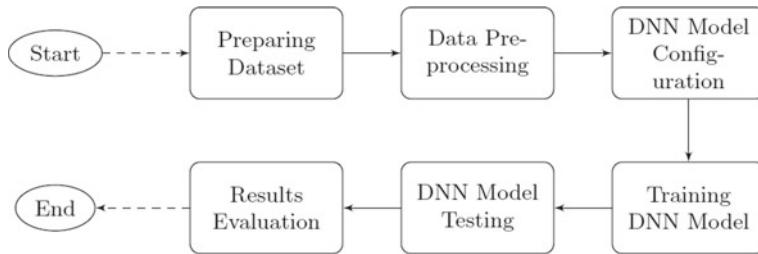


Fig. 16.6 Workflow of the DNN model (proposed)

The main task has been the configuration of the proposed DNN model. Three hidden layers have been designed along with the input and output layers for the proposed model. The complete connection format has been configured for connecting nodes of every layer. Figure 16.7 shows the configuration structure of the proposed model.

The input layer has been configured with thirteen neurons, according to the thirteen inputs. The initial two hidden layers have 270 neurons each, whereas the last hidden layer has 128 neurons. The **tanh** activation function has been used by all the hidden layers for activating neurons. A single neuron is configured in the output layer as per the two target outputs of yes and no. The sigmoid activation function is configured for the output layer. The whole DNN model is compiled using Adam optimizer to minimize the error.

The proposed DNN model has been trained for 100 epochs. The whole dataset has been divided into two parts with 80–20% division, 80% of the dataset used for training the models, and later the rest 20% used to test the model's accuracy. Various models like NB, kNN, DT, RF, LR, SVM, and ANN have been created and trained with the same training dataset and tested against the testing dataset. The ANN model utilized has been configured with a single hidden layer having 300 neurons.

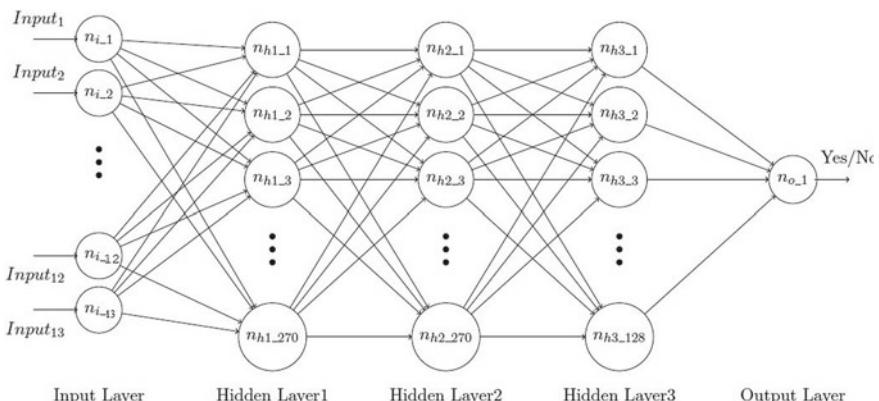


Fig. 16.7 Proposed DNN model configuration structure

The execution of traditional machine learning algorithms has been done at least 100 times, corresponding to 100 epochs of the proposed model. The accuracy of the individual models has been assessed after training and testing.

16.3.3 Implementation Details

All the models have been implemented on a laptop PC installed with Windows 10 64-bit OS, having an Intel i5 processor, 8 GB RAM, and 500 GB memory. Spyder IDE with Anaconda configuration has been used for implementing all the models using Python programming language. Keras library of Python is used for neural network implementation of ANN and proposed DNN model. RF, SVM, DT, kNN, NB, LR, ANN, and DNN (proposed) are the eight models that have been implemented. The Matplotlib functionality visualized the various models' accuracy. The confusion matrix has been computed using Scikit-learn (sklearn) library.

16.4 Results and Analysis

Results are the very crucial part of the analysis phase. It helps in determining whether the performed task can be declared to be successfully completed or not. The comparison has been made based on the values of well-known metrics for machine learning: precision, recall, f1-score, and accuracy percentage. Table 16.2 has recorded the value of these metrics for the various models that have been implemented in our research of heart disease prediction.

Among the traditional approaches, RF and kNN obtain the best accuracy of 88.5%. Other traditional algorithms like NB, LR, and SVM show little less accuracy than previously stated algorithms. DT shows the lowest accuracy among all the algorithms used for predicting heart disease. Surprisingly the ANN model gives an accuracy of

Table 16.2 Accuracy of algorithms

Algorithms	Precision	Recall	F1-score	Accuracy (in %)
Logistic regression	0.88	0.88	0.88	86.89
Random forest	0.89	0.91	0.89	88.52
K-nearest neighbors	0.89	0.91	0.89	88.52
Naive Bayes	0.88	0.88	0.88	86.89
Support vector machine	0.88	0.88	0.88	86.89
Decision tree	0.82	0.79	0.81	78.69
Artificial neural network	0.85	0.85	0.85	85
Proposed deep neural network	0.92	0.92	0.92	92.31

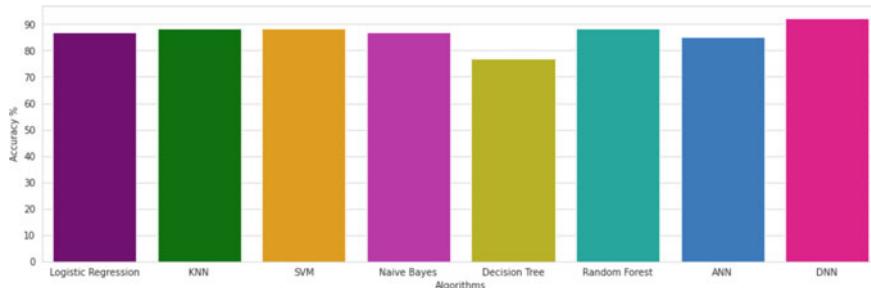


Fig. 16.8 Accuracy comparison of algorithms

85%, only stating the importance of optimal configuration of a neural network model to be a deep learning model. Finally, the proposed model outperforms all the other algorithms in terms of these metrics, which have been used for comparative analysis. Figure 16.8 shows the graphical representation of the training phase accuracy obtained by various algorithms to predict heart disease. The graphical representation quickly visualizes the comparison of accuracies, which is helpful for researchers in the area of machine learning.

16.5 Conclusion and Future Scope

Using machine learning algorithms is a vast and innovative step in predicting heart diseases and some other diseases. Predicting heart disease utilizing numerous machine learning algorithms like SVM, LR, kNN, NB, DT, RF, ANN, and DNN has been done. These algorithms have been applied to the heart dataset available at www.kaggle.com. The outcome could successfully predict whether a patient has heart disease according to details like age, gender, resting blood pressure (RPB), cholesterol, chest pain type, number of major vessels, etc., thus saving doctors precious time. Based on the prediction, patients can get the help they need to recover with proper diagnoses. Early detection of these diseases prevents the worst case from rising later, and the treatment cost can be reduced. The proposed DNN model yields the best accuracy rate among all the algorithms, which is slightly greater than 92%. This establishes that the deep learning models are way ahead of other models for predicting heart disease. The future scope of the proposed work is endless. More accurate and improvised data about patients can be obtained from the hospital, further increasing accuracy. Data mining techniques can also be used in combination to extract more accurate and hidden data utilizing artificial intelligence, thus making results more accurate and reliable for the human being, saving time and cost.

References

1. Shah, D., Patel, S., Bharti, S.K.: Heart disease prediction using machine learning techniques. *SN Comput. Sci.* **1**, 345 (2020). <https://link.springer.com/article; https://doi.org/10.1007/s42979-020-00365-y>
2. Rajdhan, A., Sai, S., Agarwal, A., Ravi, D.: Heart disease prediction using machine learning. *Int. J. Eng. Res. Technol.* **9** (2020)
3. Lutimath, M.N., Chethan, C., Pol, B.S.: Prediction of heart disease using machine learning. *Int. J. Eng. Res. Technol.* **8** (2019)
4. Nikhar, S., Karandikar, A.M.: Prediction of heart disease using machine learning algorithms. *Int. J. Adv. Eng. Manage. Sci. (IJAEMS)* **2** (2016)
5. Parthiban, G., Srivatsa, S.K.: Applying machine learning methods in diagnosing heart disease for diabetic patients. *Int. J. Appl. Inform. Syst. (IIAIS)* **3** (2012)
6. Ramalingam, V.V., Dandapat, A., Raja, M.K.: Heart disease prediction using machine learning techniques. *Int. J. Eng. Technol.* **7(2.8)**, 684–687 (2018)
7. Bindhika, G.S.S., Meghana, M., Reddy, M.S., Rajalakshmi: Heart Disease Prediction Using Machine Learning Techniques (2020). <https://www.irjet.net/archives/V7/i4/IRJET-V7I4993.pdf>
8. Jindal, H., Agrawal, S., Khera, R., Jain, R., Nagrath, P.: Heart disease prediction using machine learning algorithms. In: IOP Conference Series: Materials Science and Engineering, vol. 1022, 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) held at Rajpura, India on 24th October (2020)
9. Kogilavan, S.V., Harsitha, K., Jayapratha, P., Mirthula, S.G.: Heart disease prediction using machine learning techniques. *Int. J. Adv. Sci. Technol.* **29**(3s), 78–87 (2020)
10. Mohapatra, S.K., Behera, S., Mohanty, M.N.: A comparative analysis of cardiac data classification using support vector machine with various kernels. In: 2020 International Conference on Communication and Signal Processing (ICCSP), IEEE, pp. 515–519 (2020)
11. Aljanabi, M., Qutqut, M., Hijjawi, M.: Machine learning classification techniques for heart disease prediction: a review. *Int. J. Eng. Technol.* **7**(4), 5373–5379 (2018). <https://doi.org/10.14419/ijet.v7i4.28646>

Chapter 17

Detection of COVID-19 Cases from Chest Radiography Images



Aniket Kumar, Nishant Niraj, Venkat Narsimam Tenneti,
Brijendra Pratap Singh, and Debahuti Mishra

Abstract The COVID-19 epidemic continues to have a devastating influence on the global population's well-being and economy. One of the most important advances in the fight against COVID-19 is thorough screening of infected individuals, with radiological imaging using chest radiography being one of the most important screening methods. Early studies revealed that patients with abnormalities in chest radiography images were infected with COVID-19. Persuaded by this, a variety of computerized reasoning and simulated intelligence frameworks based on profound learning have been suggested, with promising results in terms of precision in differentiating COVID-infected individuals. COVID-Net, a neural system configuration custom-fit for the recognition of COVID-19 instances from chest radiography photographs that is open source and accessible to the general public, is presented in this study. Many techniques have been used for the detection of COVID-19, but here we are going to focus on the chest radiography technique with the application of machine learning and image processing concepts.

17.1 Introduction

The outbreak of COVID-19 pandemic occurred in December 2019 at a place named ‘Wuhan’ in China. Till now more than 5 crore people have been infected by the virus including more than 10 lakh deaths worldwide. Detection of COVID-19 infection is the most challenging part because this virus shows symptoms almost 14 days after infecting a person. There are multiple methods for the detection of COVID-19 out of which one is using CT-scan or chest X-ray report [1–5]. Here, we are going to use this technique along with some machine learning algorithms to differentiate between normal, pneumonic and COVID-19 patients. The primary motivation behind

A. Kumar · N. Niraj · V. N. Tenneti · B. P. Singh (✉) · D. Mishra

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India
e-mail: brijendrasingh@soa.ac.in

D. Mishra

e-mail: debahutimishra@soa.ac.in

this work is looking at the current situation of the whole world we realized that we need to do something to at least facilitate faster detection of COVID-19 so that the diagnosis can be done properly, thereby helping people cure faster. Lockdowns and curfews will be of no use if the infected patients are not cured or if the presence of virus is not detected inside a community or population. We introduce COVID-Net in this work, which we use to detect and analyze COVID-19 instances using chest radiography images. We know that for any machine learning algorithm to work properly we need to train the system first with a huge number of training data and then test it for some new data. In this work, we use a split value of 0.1, which splits the training and testing data in 9:1 ratio (i.e., out of ‘16,756’ images each time 90% images will be used for training purpose, and the rest 10% will be used for testing purpose). Post training and testing we do the analysis on how COVID-Net applies the algorithm to make predictions, based on which doctors and medical experts can do further studies and perform PCR testing on those patients whose report will show COVID-19 positive as per the prediction of our system.

17.2 Background

This section describes in brief the existing methodology used worldwide for COVID-19 detection, its pros and cons, and it also describes the methodology proposed by us for the detection of COVID-19 and its advantages and feasibility over the existing system. COVID-19 is pneumonia of unknown ideology, which first occurred in December 2019 at a place called Wuhan in China. ‘International Committee on Taxonomy of Viruses’ named the virus as ‘severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)’. The ‘World Health Organization’ analyzed the disease and named it COVID-19 or novel coronavirus. Currently, most of the healthcare agencies are using ‘Polymerase Chain Reaction (PCR)’ and antibody testing for detection of COVID-19. PCR test is done by directly collecting nasopharyngeal swab samples of the infected or suspected patients within a population, wherein the antigen present in the patient’s blood is subjected to testing [6–9]. Problem with polymerase chain reaction is discussed here. PCR tests are extremely labor intensive and time taking. Collecting swab samples from every person and sending it to the testing laboratory takes more than 72 h for the final reports to come. And during those 72 h whichever person is coming in contact with that suspected person has a probability of catching the infection because viral infections spread from person to person very easily. This COVID-19 virus is so strong that it can spread from one infected person to up to four normal people at a time. Also in countries like India, there are very few laboratories which are performing PCR tests and thus looking at the present scenario we surely need a secondary solution to this.

17.2.1 Problems with Serology Testing

Serology testing, which can also be called antibody testing, is also being used to detect the occurrence of COVID-19 among a population. This form of testing is applicable only when a person is already attacked by coronavirus, and the immune system in the person's body develops antibodies to fight against the disease causing antigens. This test can confirm that how many people in a particular community actually had this disease without showing any symptoms and it will also confirm that now there is antibody to fight against COVID-19 in that person's body which can also be transferred in the form of a vaccine to a healthy person's body so that he/she never catches the infection in near future. Serology is only effective after the patient has recovered from COVID-19, when antibody is already present in his blood. But the main problem here is that the scientists or medical experts still does not know that whether these antibodies are actually protecting the person from future infection or not and if it protects also, then for how long, because this virus has different forms and the antibody of one form may not fight against the antigen of another form.

17.3 Proposed System

The limited availability of viral testing equipment, as well as the time-consuming nature of these tests, has pushed radiology to the forefront of detection. Their report is increasingly becoming a critical factor in treatment selection. The speed with which COVID-19 spreads is dependent on our ability to reliably identify contaminated patients with a low rate of false negatives. Early discovery of the sickness allows for the use of all ongoing care necessary by the affected people, such as seclusion to prevent the disease from spreading. According to a study conducted at Wuhan's Department of Radiology, 'deep learning algorithms' can be utilized to distinguish COVID-19 from community-acquired pneumonia. CT scans may be performed at any local pathology lab or hospital, making this procedure far faster and more viable than the PCR method. After receiving the plates, a rudimentary classification can be done by looking at patterns in the CT image. 'ground glass opacities,' 'consolidations' and 'crazy paving pattern' are specific features observed in a 'chest radiography' image or 'CT scan' that distinguish COVID-19 from other kinds of pneumonia as in Fig. 17.1. Ground glass opacity, a hazy opaque area seen between the bronchial vessels and vascular structures, is seen in the majority of COVID-19 patients on chest radiography images. The tissues and alveolar spaces in the lungs are normally filled with air, and when the presence of water and other fluids are found instead of air in chest X-ray plates, that condition is known as lung consolidation. When a liner pattern is found imposed on the hazy opaque area found between the bronchial vessels vascular structures, it is termed as crazy paving pattern. The detection process can be made even faster by using machine learning to train the computer system with a huge amount of exiting data ('X-ray' plates or 'CT scan' images of normal, pneumonic and

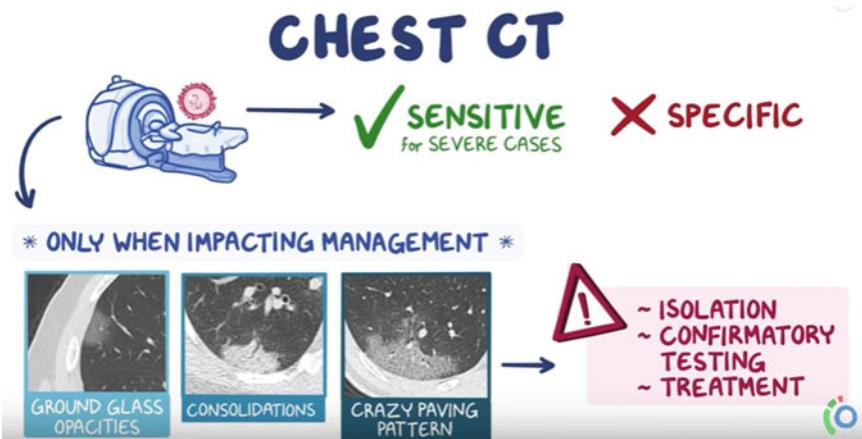


Fig. 17.1 Symptoms for COVID-19 that are found in chest CT plates: presence of ground glass opacities, consolidations and crazy paving pattern

COVID-19 patients), where the system will find out the pattern in the images based on the training algorithm and use the same to test some new data and classify it as either normal or pneumonia or COVID-19. With the growing interest in screening a large number of novel coronavirus or COVID-19 cases and the rise in false negatives in PCR testing, the need for a simple COVID-19 screening tool based on radiological images (such as chest X-rays) is becoming more important. In this case, artificial intelligence (AI) and deep learning (DL) provide rapid, robotized and persuasive ways for identifying anomalies and focusing critical highlights of the changed lung parenchyma, which may be identified with explicit COVID-19 infection signs. In this work, we used a synergic structure technique in which a human-driven system is combined with a machine-driven design to detect COVID-19 faster utilizing chest radiography.

17.3.1 System Analysis and Design

This describes as in Fig. 17.2 the design and functionality of the proposed system and also how the datasets have been generated and implemented in a convoluted neural network model. The following applications and libraries are required for the development and working of the work: Python 3.9, Tensorflow, OpenCV 4.5.2, Matplotlib, Numpy, Scikit-Learn, Pandas, Jupyter and PyDicom. Our system has been developed to classify the patient into following three types: (i) Normal (no disease), (ii) Not COVID-19 (may be any other form of pneumonia) and (iii) COVID infection.

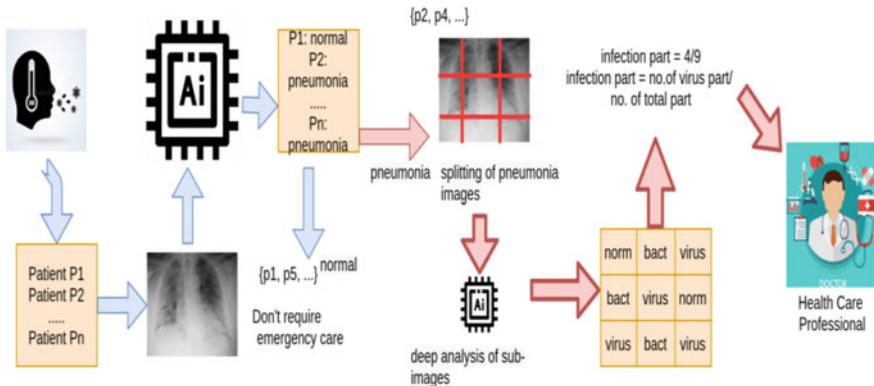


Fig. 17.2 Workflow diagram

17.4 Implementation Detail

The initial training of COVID-Net was done on ‘ImageNet’ and post that the Adam streamlining agent was used to train on COVIDx dataset in such a way that system has the capability to minimize the training (learning) rate when the training deteriorates over a certain period of time. Many accompanying hyperparameters have been used for the preparation where ‘learning rate = $2e-5$, no. of epochs = 10, cluster size = 8, patience = 5 and factor = 0.7’. A group of re-adjusting methodology is also been presented to advance better dissemination of every type of contamination. We have also used Tensorflow and Keras learning library for the construction of COVID-Net.

17.4.1 Generation of Dataset

The COVIDx dataset which been primarily used to assess COVID-Net comprises a sum of ‘16,756 chest radiography’ pictures across ‘13,645 patient cases’. For generating the COVIDx dataset, we have consolidated two openly accessible datasets (i) covid-chestxray-dataset, (ii) ‘rsna-pneumonia-detection-challenge’. Both these datasets are available in the Internet for research and exploration purpose. The currently available datasets have been very useful for research purpose, but we need more data to train the system so that it functions more accurately. The COVIDx dataset consists of only ‘76’ radiography pictures collected from ‘53’ COVID-19 patients. Also, there are around ‘8066’ patients who do not have pneumonia (normal) and ‘5526’ patients who have some other form of pneumonia [10–13]. COVID-19 infections study has been also carried out in [14–16] using deep learning, support vector machine and convolutional neural network.

17.4.2 Algorithm and Pseudo Code

Machine learning and artificial intelligence have experienced enormous growth in the last few years, and it is still growing. Now, computers are gaining human intelligence through extensive training and testing, and they are now capable to perform all sort of human activities in much less time compared to humans. A computer is also more accurate than a human just because it is a machine, and it only works based on the input data and not based on emotions. Among all the major fields, here we are going to discuss about image processing which is technically known as computer vision. The world of computer vision is actually quite vast. It includes image and video processing using OpenCV and Tensorflow, image analysis, creation of games, media, animations, etc. Deep learning approaches using computer vision over a specific algorithm known as convolutional neural network can be used to produce the product.

Mutation operation: Here, a ‘donor vector (V_a)’ is created for each ‘target vector (X_a)’ in the population as: [10]

$$V_{a,g} = x_{r1,g} + F(x_{r2,g} - x_{r3,g})$$

In this function, generation is represented by g , scaling factor is represented by F . F amplifies the difference vector and lies within $[0, 1]$. $r1, r2$ and $r3$ are random numbers from $[1, NP]$ such that $r1 \neq r2 \neq r3 \neq a$.

The best vector of population can also be used to produce ‘mutant vector’ as: [10]

$$V_{a,g} = x_{\text{best},g} + F(x_{r1,g} - x_{r2,g})$$

Crossover operation: Here, the crossover can be exponential or binomial. In both, the ‘trial vector (U)’ is created combining the mutant vectors and target vectors according to predefined conditions. Binomial crossover is performed as: [10]

$$U_{b,a,g} = \begin{cases} V_{b,a,g} & \text{if } \text{rand}_{a,b}[0, 1] \leq C_R \text{ or } b = b_{\text{rand}} \\ X_{b,a,g} & \text{otherwise} \end{cases}$$

where C_R is crossover rate in the range $[0,1]$. $a = 1, 2, \dots, NP$ and $b = 1, 2, \dots, D$. b_{rand} are a randomly selected variable of the mutant vector which ensures that the trial vector is not simply a replica of target vector [10].

In exponential crossover given in algorithm 1, we choose a random variable initially and then choose e consecutive components circularly from mutant/donorvector. The probability with which i th element is replaced in $\{1, 2, \dots, e\}$ decreases exponentially as value of i increases [10].

Algorithm 1: Exponential crossover pseudocode taken from [10]

- 1: $a \leftarrow 1$
- 2: for a start from 1 to NP

```

3: do
4:    $b \leftarrow$  select randomly in  $[1, D]$ 
5:    $U_{b,a,g} \leftarrow X_{b,a,g}$ 
6:    $e \leftarrow 1$ 
7:   while rand  $[0, 1] \leq C_R$  and  $e \leq D$ 
8:     do
9:        $U_{b,a,g} \leftarrow V_{b,a,g}$ 
10:       $b \leftarrow (b + 1) \bmod D$ 
11:       $e \leftarrow e + 1$ 
12:    end while
13:  end for

```

The unit testing was done on each dataset containing chest X-ray images by changing the value of Train: Test split variable and comparing the obtained output with the actual data present in the csv file. It is done to ensure that each image is correctly classified as either normal or pneumonia or COVID-19 based on the structural findings on the chest X-ray plates. After the successful completion of unit testing, we perform integration testing combining all the three datasets and comparing the obtained result with the actual data present in the csv file. The final result shows the total number of patients who are normal and who are suffering from pneumonia and COVID-19.

17.5 Result and Discussion

For assessing the adequacy of COVID-Net, we perform in detail investigation in order to show signs of improvement comprehension of its location execution and dynamic conduct. The section illustrates the quantitative and qualitative analysis for generating the accurate output. For examining the trained COVID-Net, we registered the testing precision, just as affectability and positive predictive value (PPV) for every type of contamination. The correctness intricacy (as far as number of parameters) and computational unpredictability (as far as number of duplicate gathering (MAC) activities) are displayed in Table 17.1. The point to be noted is that COVID-Net successfully maintains a decent balance among exactness and computational details by accomplishing ‘92.4%’ precision on testing with the requirement of only ‘2.26 billion’ MAC activities for performing expectation cases.

Table 17.1 Performance of COVID-Net on COVISx test dataset

Params (M)	MACs (G)	Accuracy (%)
111.6	2.26	92.4

Table 17.2 Sensitivity for each infection type

Sensitivity (%)		
Normal	Non-COVID-19	COVID-19
95	91	80

Table 17.3 Positive predictive value (PPV) for each infection type

Positive predictive value (%)		
Normal	Non-COVID-19	COVID-19
91.3	93.8	88.9

17.5.1 Qualitative Analysis

Our further research and investigation on the way COVID-Net handles expectations utilizing GS Inquire, which is a logic strategy that has been appeared to give great bits of knowledge into how profound neural systems go to their choices. The basic variables recognized in some model CT pictures of COVID-19 positive cases are present. Tables 17.1, 17.2 and 17.3 show the performance, sensitivity and positive predictive value.

17.5.2 Result

The output obtained from training and testing the existing datasets using split value as 0.1 is shown here. Out of the total data, 13,653 numbers of data are used for training, and 1510 numbers of data are used for testing to give the final classification of normal/pneumonia/COVID-19 in the test_count and train_count array. Figure 17.3 shows the snapshot of the result after the program execution. Print statements in source code to display the final statistics:

```
E:\SDP\COVID-Net>python create_COVIDX_v3.py
Data distribution from covid-chestxray-dataset:
{'normal': 0, 'pneumonia': 33, 'COVID-19': 267}
Key: pneumonia
Test patients: ['8', '31']
Key: COVID-19
Test patients: ['19', '20', '36', '42', '86', '94', '97', '117', '132', '138', '144', '150', '163', '169']
test count: {'normal': 0, 'pneumonia': 5, 'COVID-19': 31}
train count: {'normal': 0, 'pneumonia': 28, 'COVID-19': 236}
test count: {'normal': 885, 'pneumonia': 594, 'COVID-19': 31}
train count: {'normal': 7966, 'pneumonia': 5451, 'COVID-19': 236}
Final stats
train count: {'normal': 7966, 'pneumonia': 5451, 'COVID-19': 236}
Test count: {'normal': 885, 'pneumonia': 594, 'COVID-19': 31}
Total length of train: 13653
Total length of test: 1510
```

Fig. 17.3 Dataset training and testing results which shows the number of individuals who are normal or suffering from pneumonia or COVID-19

- (i) print('Final stats')
- (ii) print('Train count:', train_count)
- (iii) print('Test count:', test_count)
- (iv) print('Total length of train:', len(train))
- (v) print('Total length of test:', len(test)).

17.6 Socioeconomic Issues

17.6.1 Practical Relevance

Starting from December 2019, till now more than 10 M people have been affected with COVID-19 globally. Since most people experience respiratory issues, the detection process of the infected ones takes time which causes severe impact on the population and controlling the spread becomes difficult. The economic impact has been huge, causing more and more unemployment. In such a situation of devastation, early detection of the disease becomes highly essential. The existing process of PCR testing makes a delay in detection because of multiple samples being tested simultaneously and nature of the infection being studied. We do not have enough medical facilities and equipment for the prognosis to go smoothly.

Also being a viral infection, chances of spreading increase with time. It is near to impossible to accommodate and provide medication to a large number of people. Through Chest Radiography images that are captured has shown a typical character for the COVID-19 cases. The aim of this procedure is early detection of COVID-19 and to distinct only those cases with high chances of being positive. Chest radiography is affordable and feasible. The machinery needed to carry out this process can be easily set up at any pathology lab. Only certain precautions of sanitization need to be taken since this is contagious, and one patient can contaminate another. Once the result from similarity in symptoms and imaging is generated and the patient is found positive, his/her report is further forwarded for specific infection test.

17.6.2 Global Impact

Since the pandemic is affected globally, the use of chest radiography technique can also be implemented worldwide. With increase in the size of dataset with more number of training data, detection will become accurate. Certain general characteristics like ground glass opacities, consolidations and crazy paving pattern can be analyzed and made the base of diagnosis until viruses go evolution. COVID-19 test kits with advancement in technology will shortly be a competition in the market. But with such usage of proposed system and machinery in time of need will reduce cost of diagnosis along with enhancement of the existing technology.

17.7 Conclusion and Future Scope

Our main focus is to fasten the procedure for detection of the disease so that early diagnosis is possible resulting in decrease of the average mortality rate in the world. In countries like India, we know that the population is too high and we do not have feasibility to accommodate every individual in a hospital in case the person is detected positive for COVID-19. Also, we do not have a proper testing kit to conduct vigorous testing within a population. PCR test reports are taking more than 3 days to come, and by that time, an infected person is infecting thousands of others.

Our work is also subject to more improvements in near future. We can use better image processing technologies to give better results. Also, as our training dataset will become larger with more data in future, the accuracy of the detection algorithm will increase. Recently, we came to know that there are different forms of COVID-19 infection and each country is having a different form of viral DNA which is infecting people, so if this prediction methodology is implemented globally then our dataset will contain all forms of COVID-19-based data which would really help in detecting coronavirus of a different form in a particular population.

References

1. Radiology Assistant Chest CT: <https://radiologyassistant.nl/chest/lung-hrct-basic-interpretation>
2. Butt, C., Gill, J., Chun, D., Babu, B.A.: Deep learning system to screen coronavirus disease 2019 pneumonia. *Appl. Intel.* (2020)
3. Wikipedia: en.wikipedia/wiki/COVID-19_testing
4. Serology: <https://www.cdc.gov/coronavirus/2019-ncov/lab/serology-testing.html>
5. Stephanie Stephanie, M.D. et al.: Determinants of Chest Radiography Sensitivity for COVID-19: A Multi-Institutional Study in the United States (2021)
6. Radiological Society of North America: RSNA Pneumonia Detection Challenge. <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data>
7. Radiology Assistant Lung Disease: <https://radiologyassistant.nl/chest/chest-x-ray/lung-disease>
8. Ng, M.-Y., Lee, E.Y., Yang, J. et al.: Imaging profile of the COVID-19 infection. *Radiol. Find. Literat. Rev.* (2020)
9. Su, J., Vargas, D.V., Sakurai, K.: Attacking convolutional neural network using differential evolution. *IPPSJ T Comput. Vis. Appl.* **11**, 1 (2019). <https://doi.org/10.1186/s41074-019-0053-3>
10. Singh, D., Kumar, V., Vaishali, et al.: Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks. *Eur. J. Clin. Microbiol. Infect. Dis.* **39**, 1379–1389 (2020). <https://doi.org/10.1007/s10096-020-03901-z>
11. Martínez Chamorro, E., Díez Tascón, A., Ibáñez Sanz, L., Ossaba Vélez, S., Borruel Nacenta, S.: Radiologic diagnosis of patients with COVID-19. *Radiología (English Edition)* **63**(1), 56–73 (2021), ISSN 2173-5107. <https://doi.org/10.1016/j.rxeng.2020.11.001>
12. Yasin, R., Gouda, W.: Chest X-ray findings monitoring COVID-19 disease course and severity. *Egypt J. Radiol. Nucl. Med.* **51**, 193 (2020). <https://doi.org/10.1186/s43055-020-00296-x>
13. Rousan, L.A., Elobeid, E., Karrar, M., et al.: Chest x-ray findings and temporal lung changes in patients with COVID-19 pneumonia. *BMC Pulm. Med.* **20**, 245 (2020). <https://doi.org/10.1186/s12890-020-01286-5>

14. Das, A., Mohapatra, S.K., Subudhi, A., Mohanty, M.N.: Classification of COVID-19 infected X-ray image using deep learning techniques. *Ann. Romanian Soc. Cell Biol.* **2736** (2021)
15. Mohapatra, S.K., DebarchanMohanty, M., Mohanty, M.N.: Corona virus infection probability classification using support vector machine. *Int. J. Adv. Sci. Technol.* **29**(8s), 3093–3098 (2020)
16. Das, A., Mohanty, M.N.: Covid-19 detection from X-ray images using convolutional neural network. *Int. J. Adv. Sci. Technol.* **29**(8s), 3099–3105 (2020)

Chapter 18

Monitoring the Heart Rate—An Image Processing Approach



Samuka Mohanty, Sumit Pal, Shubhrajit Parida, and Manosmita Swain

Abstract Nowadays, in this age of innovation, human life is getting easier as new gadgets and devices are coming out with the capability complementing their working lives. There are gadgets that can monitor their health parameters constantly irrespective of time or location. The majority of deaths in the world happen due to cardiovascular diseases (CVD). Due to the increase in population, people ignore their health most of the time because of workload. For example, atrial fibrillation is one of the most common diseases and can be deadliest at times. In atrial fibrillation, the tempo or the time between each heartbeat will vary. This system is capable of generating medical reports without any intervention of doctors, and also, patients get an added advantage of having the option to have personal medical feedback from an experienced doctor. In this age, the cost of basic health monitoring devices has decreased the communication gap between the doctors and patients, so our project intends to reduce the cost of monitoring patients regardless of location. So, in this work, web page designing techniques and Python programming are implemented together to make things platform-independent to monitor the patient.

18.1 Introduction

From smallest beings like ants to the largest beings like whales, heart is the most important organ; it has the responsibility for pumping blood through the blood vessels of the circulatory system. At rest, the heartbeat of a normal healthy person is close to 72 beats per minute. The beats per minute can go down to 30–40 beats per minute for athletes. So, while exercising, the heart rate rises, and if a person maintains this level of activity, their heart rate falls during rest, which is a really good indicator.

In today's fast-paced society, the human body is becoming more susceptible to heart-related disorders, with an increase in the number of deaths as a result. Coronary artery disease, congenital heart defects (disorders acquired before birth), arrhythmias (heart rhythm issues), and other cardiac diseases are among the most well-known.

S. Mohanty (✉) · S. Pal · S. Parida · M. Swain

Department of Computer Science and Engineering, Siksha O Anusandhan (Deemed To Be University), Bhubaneswar, Odisha, India

These are only a few instances, but heart diseases encompass a wide spectrum of disorders that affect a person's heart.

However, in the current situation, cardiovascular diseases represent the greatest concern; accounting for 31% of all global deaths, with 17.9 million fatalities owing to CVDs. Good healthcare necessitates enough financial resources, which are a major issue and a major setback in India when it comes to preventing deaths.

As we progress through life, newer technologies are introduced and become a part of our daily lives. Every part of the human body is monitored by machines. However, in rural locations, these gadgets are difficult to come by and may be prohibitively expensive for most individuals. People who are financially secure can monitor their health on a regular basis, but others who are not so financially secure may be unable to attend hospitals or monitor their health due to their hectic schedules. The cost of basic health monitoring devices has decreased, and the communication gap between the doctors or medical personnel and patients is also decreasing. So, in this report, the capability of fast and easy access through web page and Python is implemented together as optimization techniques to reduce error and make things platform-independent to monitor the patient in real time.

The heart rate monitoring system aims for a digital setup for those people who are unable to pay for the equipment. Faster, cheaper and devices with better accuracy are being produced. There are machines to monitor every aspect of a human body. These are very easy to get our hands on as they could be found at most of the pharmaceutical stores.

18.2 Literature Survey

Some of recent papers follow very similar concepts like both models are made with ArduinoUno, and they evaluate the heartbeat rate. So, in these papers, it is modeled in such a way that, when the heart rate does not fall within the set safety range, an alert signal is generated using Arduino [1, 2]. The technique of using a Raspberry Pi to wirelessly communicate data related to heart parameters through a phone has been investigated [3].

In this paper, a prototype was developed to measure the people or patient's body temperature, blood pressure (BP), pulse rate, electrocardiogram (ECG), and glucose level using multiple sensors. The prototype can also track the patient's location, and all of the data is stored and processed in the microprocessor [4, 5]. For PPG-based heartbeat monitoring, the concept applied is IoT, i.e., Internet of Things. A wireless heartbeat monitoring system is created and integrated with GSM technology in this article. This has the potential to be an important component of a personal healthcare system as described in [6]. Parameters such as body temperature, sweat rate, and heart rate are detected and sent to a computer so that the person's health state can be monitored remotely and data transmitted to a doctor via GSM technology in [7]. This paper describes a system that uses numerous sensors to monitor the patient's heartbeat and body temperature, as well as the ability to track the patient's location using

GSM technology. In this scenario, numerous sensors were employed to obtain the patients' real-time heart conditions, as well as IoT and Raspberry Pi kit technologies [8]. This study describes a robotic mechanism that uses an integrated three-axis truss manipulator structure to recognize distinct items based on color characteristics and conduct various operations on them, such as object sorting and item grabbing. They carried out this intelligent recognition and intelligent handling-based research using ‘machine vision’ and OpenCV [9, 10]. Recently, a lot of research directions are available for accuracy analysis of ECG using preprocessing techniques, feature extraction and classification [11].

18.3 Proposed System

This system is built using HTML and CSS for the user interface, Python OpenCV for image processing and Python Flask to connect the algorithm to the main Python script.

18.3.1 *Flask*

Python-based Flask is a micro-Web framework. This is a template engine that assists us in compiling modules, packages, and libraries to aid in the development of Web applications. The first step is to obtain the Flask code along with a specifier, which is needed to inform the interpreter of where the CSS and other static folders are located. Following that, a route to the homepage and output page is created. If the server is locally hosted, the index page will be the first page to load when the IP is entered. The homepage features a form that collects user information as well as the video path and duration to be processed. The main Python is called from the output page. The Python script outputs the heartbeats every minute and saves the graph locally. In the output page, the heart rate per minute is presented along with all of the user data.

18.3.2 *OpenCV*

OpenCV is a package that has collection of programming functions that are primarily intended for real-time computer vision. Intel was the first to develop it. In order to determine the heartbeat rate, OpenCV plays a significant role in this project. For real-time operation, OpenCV includes GPU acceleration. For image processing and filtering, OpenCV is employed in this model.

18.3.3 Procedure

A single frame image from the input is taken by placing the fingertip on top of a small light source; when blood goes through our veins, the light flickers/dims. As a result, blood represents a heartbeat as it flows through the body. In this case, the heartbeat rate is calculated by counting how frequently the light intensity dimmed over the time period supplied by the user. In this phase, the number of frames has been calculated with respect to the video file input.

With all the above necessary data and variables initialized, the video will be edited through Python to fit the user's time constraints. This is done to reduce processing time, as processing the entire video rather than just the time range will take longer and slow down the entire system.

The video must now be transformed to gray-scale, as the data it contains in its colorized version is useless. In Python, a video is saved as a matrix, with a matrix for each frame. As a result, the matrix's dimension is the same as the video file's dimension.

For accurate findings, the graph must be smoothed out in order to fix it. If the data is transmitted to be processed without being corrected, the heart rate will be incorrect. This is because the HRM will be calculated using a threshold in the future phases, and if the readings are below the threshold, it will be counted as a beat. However, if the graph is skewed, certain values may not even appear below the threshold, thus missing the beats, while others may be tallied as a single beat.

Batches are used to correct the data. The first batch will be used as a reference, and the remaining batches will be corrected in accordance with the initial reference batch. As a result, a median is computed in the first batch, which is done by taking the peak and trough of the batch graph.

$$\text{median}(m_{\text{ref}}) = \frac{\text{peak} + \text{trough}}{2} \quad (18.1)$$

This median is then used to calculate the deviation over the remaining batches. Now that we know the median, the following step is to determine the median and deviation of the following batch's median in comparison with the reference batch's median. The next batch's peak and trough are determined once more, as well as the median. The deviation is calculated by subtracting the reference batch's median from the current batch's median.

$$\text{deviation}(\delta) = m_{\text{ref}} - m_{\text{cur}} \quad (18.2)$$

A multiplying factor is calculated using the following formula.

$$m_{\text{factor}} = \frac{60}{\text{video length}} \quad (18.3)$$

Using this multiplying factor, the heart rate per minute is calculated by using the following formula.

$$\text{HRM} = \text{ceiling}\left(\frac{c}{2}\right) \times m_{\text{factor}} \quad (18.4)$$

18.4 Design Steps

The goal of this model is to minimize the use of hardware, reduce the cost, and most importantly omit the communication gap between the patient and the medical personnel. This model's frontend has two web pages, one is for data collection, and other is for presenting the data along with heartbeat rate report.

- Step-1: First thing the user will interact with is the data registration page which is made using HTML and CSS. In this web page, the user will enter the required information namely—first and last name, email, phone number, gender, address, illness (optional), additional details (optional) and most importantly the video path and time of the video that is to be processed, these details are sent to the Python script, and this is done with the help of flask which is used to connect/route HTML pages along with Python scripts.
- Step-2: After getting the video address and time, these parameters are sent to the Python script which will calculate and return the heartbeat rate and the plot.
- Step-3: With the video path, the video is accessed and stored in a variable using Open Source Computer Vision Library (OpenCV).
- Step-4: Then, the video is checked if it is same or greater than the time provided by the user for processing. If true, then the video is trimmed accordingly, and the trimmed video is used further for processing. Else if it turns out to be false, the script ends after throwing error message to the user.
- Step-5: Now that the foundation is made, the parameters of the trimmed video are captured such as the number of frames, frames per second and total length.
- Step-6: In this step, the video is converted from RBG colorized to gray-scaled. This is done because the data required is the intensity of light in each frame, and the change in intensity depicts the heartbeat rate. In Python, a gray-scaled frame is stored in form of a 2-D matrix of dimensions same as resolution of the video. In this matrix, each cell value is the white value of that pixel ranging from 0 to 255.
- Step-7: Here, the gray-scaled data is iterated, and sum of all the cells of one frame are stored in an array which will be of size same as the number of frames the trimmed videos have.

- Step-8: With the data we had right now a graph is subplotted, in this plot, the dipping values represent the heartbeats. As shown in Step-6, whenever the white light values dip, at that moment the heart pumped blood. Whenever blood passes through the arteries, the light passing through the finger will flicker/dim down.
- Step-9: Due to environmental noise or camera adjusting light, the graph could appear to be tilting along the y-axis. In order to stop this, graph is straightened out with respect to the y-axis.
- Step-10: Another subplot is plotted to represent the outcome of Step-9's algorithm.
- Step-11: Now to remove further noise from the graph, it is filtered using median filter algorithm with a factor of 9.
- Step-12: Another subplot is plotted to represent the outcome of step-11.
- Step-13: Now that we have the necessary data to calculate the heartbeat rate, in order to calculate the plot is converted to a square wave.
- Step-14: Another subplot is plotted to represent the outcome of step-13.
- Step-15: From this square wave, the number of changes is calculated, i.e., whenever value changes from 0 to 1 or 1 to 0, counter is incremented.
- Step-16: The counter in step-15 is now halved, and ceiling value is considered. With the video time, a multiplying factor is calculated.
- Step-17: The counter is multiplied with the factor which gives the heartbeat per minute.
- Step-18: These values are returned along with the complete graph.
- Step-19: Finally, the user detail HTML page is rendered through Flask where all the details with the plot and heartbeat rate per minute are displayed.

18.4.1 Result Analysis

After giving the desired input, the system correctly calculated the heartbeat rate and displayed graphs along with the time (in terms of BPM). The program then swiftly processes the video input and analyzes heartbeat for faster processing. It then processes the saturation changes of light and plots a graph. To decrease noise from the preceding generated signal, it filters (filtering parts of the signal that are exceeding threshold) and rectifies the graph to generate an impulse signal. The impulse signal thereby generated helps the system to measure heart rate, and finally, after processing the impulse graph, the system generates a report with patient details, date and time of report generation. For further details, a report is tabulated based on observations made for five consecutive days taken in morning as well as evening hours to find out the variations. This tabulation indicates the patient's heart rate after any vigorous/normal activity done by the patient and resting heart rate. From the above tabulation, we get a perspective of the varying heart rate for both morning and evening hours with the state of the patient during the measurement of heart rate.

This data collected in Tables 18.1 and 18.2 is of a person in his early 20s. Different conditions state different details of the person's state when he took the fingertip video.

Table 18.1 Observations for concurrent days 20 s in morning

Days	Morning			
	Heartbeat rate	Machine HRM	Condition	Error
Day1	93	93	Heavy activity	0
Day2	96	96	Heavy activity	0
Day3	102	101	Heavy activity	-1
Day4	93	92	Heavy activity	-1
Day5	87	87	Moderate activity	0

Table 18.2 Observations for concurrent days 20 s in evening

Days	Morning			
	Heartbeat rate	Machine HRM	Condition	Error
Day1	84	85	Moderate activity	+1
Day2	78	78	Resting	0
Day3	78	78	Resting	0
Day4	81	81	Moderate activity	0
Day5	84	84	Moderate activity	0

Resting conditions states that the person was on resting condition and has not moved from his place for at least past 10 min. Moderate activity states that the person has been doing some mild activity which includes walking or standing. Strenuous activity states that the person was doing some rigorous work which may include jogging, running, etc. The average maximum heart rate of a person in his early 20s is 200, and resting heart rate ranges from 60 to 100 bpm for a normal healthy person. So, the above data shows that the person is normal and will lead a healthy life.

18.5 Conclusion

Nowadays, taking care of one's health is critical, especially because we tend to neglect living a healthy lifestyle as a result of the hectic lives we all lead in order to make ends meet. However, maintaining this simple yet vital aspect is a difficult task. Having to visit the doctor every now and then may be inconvenient, but it may also reduce the risk of critical cases. Medical facilities may also be out of reach for persons living in remote places. Essentially everything can now be achieved from the comfort of one's own home, thanks to technological advancements. As a result, this system is capable of generating medical reports without the intervention of doctors, and patients benefit from the option of receiving personal medical feedback from an experienced doctor. If any abnormalities are discovered ahead of time, critical situations can be avoided. The model's goal is to reduce the cost of patient monitoring regardless of location.

In future, this study can be enhanced by taking more parameters of a patient's body into consideration to further improve the result of this model.

References

1. Mallick, B., Patro, A.K.: Heart rate monitoring system using fingertip through Arduino and processing software. *Int. J. Sci. Eng. Technol. (IJSETR)* **5**, 234–238 (2016)
2. Lakshmi, J.M.B., Hariharan, R., Devi, N., Sowmiya, N.: Heart beat detector using infrared pulse sensor. *Int. J. Sci. Res. Dev. (IJSRD)*, **3**(9), 214–220 (2015)
3. Das, S.: The development of a microcontroller based low cost heart rate counter for health care systems. *Int. J. Eng. Trends Technol.* **4**(2), 134–139 (2015)
4. Ufoaroh, S.U., Oranugo, C.O., Uchechukwu, M.U.: Heartbeat monitoring and alert system using GSM technology. *Int. J. Eng. Res. Gen. Sci.* **3**(4), 245–250 (2015)
5. Rao, P.V., Akhila, V., Vasavi, Y., Nissie K.: An IoT based patient health monitoring system using Arduino Uno. *Int. J. Res. Inf. Technol. (IJRIT)*, **1**(1), 314–319 (2017)
6. Chandra, R., Rathee, A., Verma, P., Chougule, A.: GSM based health monitoring system. In: Proceedings of IRF International Conference, Pune (2018). ISBN 978-93-84209-04-9
7. Prasad, M.B.: GSM based health care monitoring system. *Int. J. Innov. Technol. Exploring Eng. (IJITEE)*, **8**(2), 231–238 (2018). ISSN 2278-3075
8. Yeole, M., Daryapurkar, R.: Design and implementation of wireless heartbeat measuring device for remote health monitoring. *Vishwakarma J. Eng. Res. (VJER)*, **1**(2), 891–899 (2018). ISSN 2456-8465
9. Atiya, S.U., Madhuri, K.S.: GSM based patient monitoring system using biomedical sensors. *Int. J. Comput. Eng. Res. Trends. ISSN (O) 2349-7084*, **3**(9), 620–624 (2016)
10. Dubey, R.K., Mishra, S., Agarwal, S., Sharma, R., Pradhan, N., Saran, V.: Patient's health monitoring system using Internet of Things (IoT). *Int. J. Eng. Trends Technol. (IJETT)* **59**(3), 155–158 (2018). ISSN 2231-5381
11. Mohapatra, S.K., Mohanty, M.N.: ECG analysis: a brief review. *Recent Adv. Comput. Sci. Commun. (Formerly: Recent Patents Comput. Sci.)*, **14**(2), 344–59 (2021)

Chapter 19

Evaluation of Optimal Feature Transformation Using Particle Swarm Optimization



Dibyasundar Das, Suryakant Prusty, Biswajit Swain, and Tushar Sharma

Abstract Feature reduction is one of the essential steps for machine learning applications. It reduces redundancy in the feature set, which reduces the computational cost in the learning phase. The success of the reduction stage depends on the size of the feature selected and the separability of the transformed matrix. In most of the work, the feature transformation matrix is determined mathematically and most of the time depends on eigenvectors and eigenvalues. If the feature space is high, it is difficult to calculate the eigen matrix in smaller devices. Hence, this work proposes a method to generate an optimum transformation matrix using heuristic optimization approach that leads to better classification accuracy with less feature size. This study uses Bhattacharyya distance as an objective function to evaluate the separability of the transformed feature matrix. Moreover, to select a proper subset of features, a penalty parameter is added with respect to number of features in transformation matrix. This work proposes a modified version of particle swarm optimization that can satisfy the objective. The study shows the ability to learn a transformation matrix with competitive results in classification task. The proposed method is evaluated on various freely available public datasets, namely Fisher's IRIS, Wine, Wisconsin Breast Cancer, Ionosphere, Sonar, MNIST, NIT-R Bangla Numeral, and ISI-K Bangla Numeral datasets.

19.1 Introduction

Machine learning algorithms have proven to be reliable tools for solving many critical problems in this modern era [1]. It has already been applied to a wide range of computer vision applications like pathological brain detection [2], CAD system for blood smear analysis [3], and heart disease prediction [4], to mention a few. The learning capability of such algorithms comes from the ability to extract distinctive traits from past data. The feature is the characteristic of an instance that helps to

D. Das (✉) · S. Prusty · B. Swain · T. Sharma

Department of Computer Science and Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

e-mail: dibyasundardas@soa.ac.in

distinguish the class. However, some of the features are irrelevant to the current goal and contains redundant information [5]. This redundant information adds to the computational cost of training and prediction in machine learning algorithms. Hence, a feature reduction is necessary step for a preprocessing stage to any machine learning algorithm.

The advantages of feature reduction are handling class imbalance problems, noisy data, data shift problems (covariance shift), and reducing the operational cost of feature handling and processing. The feature reduction techniques can be of two types such as feature selection [6] and feature transformation [7]. In any case, the choice of the feature count is one of the hyperparameter settings to achieve better generalization in machine learning applications, which can only be possible by experiment. Hence, this study proposes a feature transformation scheme based on PSO and Bhattacharyya distance, which requires no prior knowledge of the number of features and gives an optimal feature transformation matrix.

The study uses various freely available UCI datasets like Iris, Wine, and character datasets like MNIST, NITRKL Bangla Numeral to evaluate the performance of the proposed scheme. The contribution of the paper can be summarized as follows:

1. The first contribution involves a new feature transformation scheme based on feature transformation and Bhattacharyya distance for feature reduction.
2. A modified PSO-based optimization strategy that learns the transformation matrix and automatically predicts the optimal size of features.

The rest of the paper is organized as follows. Section 19.2 briefly reviews preliminaries to the proposed model and presents the proposed model and explains its workflow. Section 19.3 gives the Experimental details with obtained results. Finally, Sect. 19.4 provides the concluding remarks of our work.

19.2 Proposed Model

See Fig. 19.1.

19.2.1 *Formulation of the Optimization Problem*

Let's assume a dataset X of size $m \times n$. The aim is to transform X into a dimension d where $d \ll n$. Hence, we can formulate a transform matrix (T) of size $n \times d$ such that $X' = X*T$ where X' represents the transformed feature matrix. However, T must preserve the feature property and increase classification accuracy. Hence, the transformed feature (X') must be well separable to classify the samples effectively. T can be chosen any matrix of size $n \times d$ which chooses T as an NP-hard problem. Here, we aim to find an optimum, transform matrix that reduces the number of features and improves classification accuracy. In such a case, the size of T in most works is

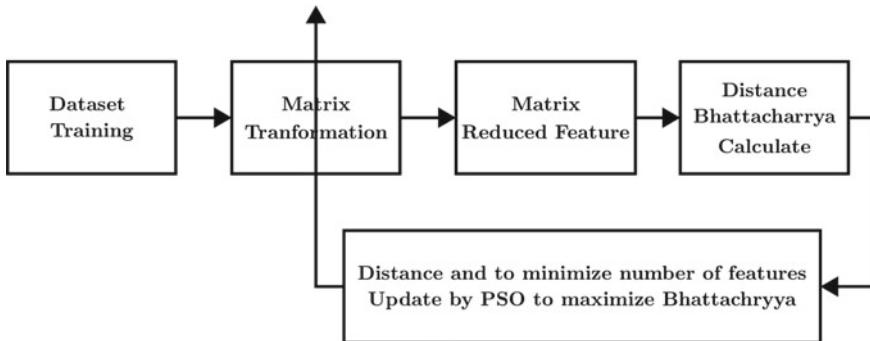


Fig. 19.1 Overall block diagram of the proposed model

set as a hyperparameter. But, the choice of size of T is an optimization problem in itself. Hence, here, the study addresses the optimal choice of size of T that can be obtained by an optimization algorithm. Therefore, our optimization problem can be re-stated, find the optimum transform matrix, and optimum size of transform matrix to increase the classification accuracy. Here, this work uses a modified particle swarm optimization (PSO) algorithm to achieve the objective and to evaluate the separability of the transformed features using Bhattacharyya distance.

19.2.2 Bhattacharyya Distance

Bhattacharyya distance [8, 9] is a class separability measure, developed by Anil Kumar Bhattacharyya, which is more reliable than Mahalanobis distance [10]. This is mainly because it grows to depend upon the difference between standard deviation which is a more generalization form of Mahalanobis distance measure. The Bhattacharyya distance between two classes can be defined as

$$b_{i,j} = \frac{1}{8}(\mu_i - \mu_j)^T \left[\frac{\Sigma_i + \Sigma_j}{2} \right] (\mu_i - \mu_j) + \frac{1}{2} \ln \frac{|(\Sigma_i + \Sigma_j)/2|}{|\Sigma_i|^{1/2} |\Sigma_j|^{1/2}}, \quad (19.1)$$

where μ represents class mean and Σ represents class covariance matrix. As Bhattacharyya distance is defined between two classes, the extension of it to a multi-class problem with class size L can be defined as

$$B = \sum_{i=1}^{L-1} \sum_{j>i}^L b_{i,j} \quad (19.2)$$

This considers that the total contribution for a given set of features is the sum of pairwise Bhattacharyya distance of classes for a given classification problem.

Here, this study uses this property of Bhattacharyya distance to evaluate the feature transformation matrix that is optimally determined by the PSO algorithm.

19.2.3 Modified Particle Swarm Optimization

PSO [11] is one of the most efficient heuristic optimization methods introduced in 1995 by James Kennedy and Russell Eberhart with inspiration from the study of biologist Frank Heppner on behavioral analysis of bird flocks. It follows the behaviors of a flock of birds searching for food or shelter. None of the birds know the optimum location to find food or shelter. Hence, they try to follow the bird which is nearest to the objective. An optimization problem is more or less similar to this concept. Hence, PSO initializes a group of random particles (solutions) and then searches for an optimum solution by moving in the direction of the best particle. The direction in which a particle move is decided by looking into global optima and previously obtained local optima by particle. A brief flow chart of PSO is given in Fig. 19.2. The PSO is designed to work on fixed-size vectors; however, our optimization problem needs to work on the variable size of matrices; hence, PSO operation is modified and is given as an algorithm in Fig. 19.3.

From algorithm it can be seen that it maps the size of the new population as a new transformation to the global best and local best size. This helps to randomly generate a new transformation matrix and test for the various dimension of features. Here, the study uses the Bhattacharyya distance to evaluate the separability of newly generated features and rate the population based on distance measures and several features. Hence, the objective function is defined as

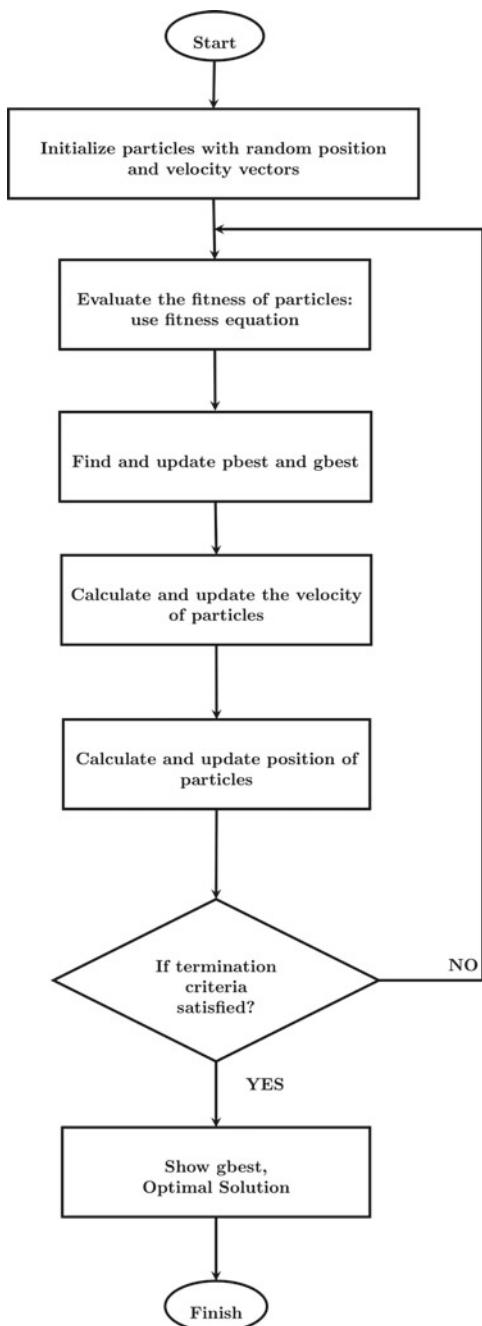
$$\text{obj}_i = \frac{d_i}{\sum_{i=1}^{\text{pop_size}} d_i} - \frac{s_{z_i}}{\sum_{i=1}^{\text{pop_size}} s_{z_i}} \quad (19.3)$$

where obj_i is the objective of i th particle, d_i is Bhattacharyya distance for the i th particle, s_{z_i} is the reduced dimension size for the i th particle, and pop_size represents the size of the population chosen for the PSO algorithm. This objective is defined to maximize the Bhattacharyya distance and minimize the number of features. The experiments are carried out on various freely available datasets and the following section gives the details of the results.

19.3 Results and Discussion

The proposed model is simulated in MATLAB 2018b in a Linux machine with a Core i7 processor and 16 GB RAM. The study uses various freely available datasets (UCI and character datasets) to evaluate our model and a description of the datasets used is given in Table 19.1. First, the dataset is divided into two parts training and testing

Fig. 19.2 Overall block diagram of the proposed modified-PSO model



Algorithm 1: Feature reduction by PSO & Bhattacharyya Distance

Input : Input training data X of size $m \times n$, M_{max} =Maximum limit to feature size
output : C : Coefficient matrix of size $m \times d$ where $d << n$ and $d \leq M_{max}$

Initialization: pop_{size} : Population size,
 max_{iter} : Maximum iteration,
 c_1 : Local velocity,
 c_2 : Global velocity

```

1 for  $i \leftarrow 1$  to  $pop_{size}$  do
2   |  $t \leftarrow$  random value in between 1 to  $M_{max}$ ;
3   |  $pop_i \leftarrow$  random matrix of size  $n \times t$ ;
4   |  $vel_i \leftarrow$  random matrix of size  $n \times t$ ;
5   |  $sz_i \leftarrow t$ ;
6 end
7 for  $i \leftarrow 1$  to  $pop_{size}$  do
8   |  $d_i \leftarrow Bhattacharyya_{dist}(X \times pop_i, L)$ ;
9 end
10 for  $i \leftarrow 1$  to  $pop_{size}$  do
11   |  $obj_i \leftarrow \frac{d_i}{\sum_{i=1}^{pop_{size}} d_i} - \frac{sz_i}{\sum_{i=1}^{pop_{size}} sz_i}$ ;
12 end
13  $global_{max} \leftarrow pop_j$  where  $j$  is the location of maximum  $obj$  value. ;
14  $global_{max\ size} \leftarrow sz_j$  where  $j$  is the location of maximum  $obj$  value. ;
15 for  $i \leftarrow 1$  to  $pop_{size}$  do
16   |  $local_{max}^i \leftarrow pop_i$ ;
17   |  $local_{max\ size}^i \leftarrow sz_i$ ;
18 end
19  $iter \leftarrow 1$ ;
20 while  $iter < max_{iter}$  do
21   for  $i \leftarrow 1$  to  $pop_{size}$  do
22     |  $t \leftarrow \frac{local_{max\ size}^i + global_{max\ size}}{2}$ ;
23     |  $r_{mat} \leftarrow$  random matrix of size  $n \times t$ ;
24     |  $t_p \leftarrow pop_i \times r_{mat}$ ;
25     |  $r_{mat} \leftarrow$  random matrix of size  $n \times t$ ;
26     |  $t_v \leftarrow vel_i \times r_{mat}$ ;
27     |  $r_{mat} \leftarrow$  random matrix of size  $n \times t$ ;
28     |  $t_{Gp} \leftarrow global_{max} \times r_{mat}$ ;
29     |  $r_{mat} \leftarrow$  random matrix of size  $n \times t$ ;
30     |  $t_{Lp} \leftarrow local_{max}^i \times r_{mat}$ ;
31     |  $t_v \leftarrow t_v + c_1.rand().(t_{Lp} - t_p) + c_2.rand().(t_{Gp} - t_p)$ ;
32     |  $t_p \leftarrow t_p + t_v$ ;
33     |  $o \leftarrow Bhattacharyya_{dist}(X \times pop_i, L)$ ;
34     if  $t_p$  is better solution by  $o$  then
35       | update respective variables;
36     end
37   end
38   |  $iter \leftarrow max_{iter}$ ;
39 end
40  $C \leftarrow global_{max}$ ;
41 return  $C$ 

```

Fig. 19.3 Algorithm for modified PSO

Table 19.1 Details of the dataset used to evaluate the proposed model

Name of the dataset	Samples	Classes	Training size	Testing size
Fisher's IRIS (D1)	150	3	120	30
Wine (D2)	178	3	142	36
Wisconsin Breast Cancer(D3)	569	2	456	113
Ionosphere (D4)	351	2	281	70
Sonar (D5)	208	2	167	41
MNIST (D6)	70,000	10	60,000	10,000
NIT-R Bangla Numeral (D7)	4345	10	3476	869
ISI-K Bangla Numeral (D8)	23,392	10	19,392	4000

set in a ratio of 80:20, respectively. Then the training set is used to find optimum feature transformation matrix and training a single hidden layer feedforward neural network with extreme learning machine algorithm with (number of feature + number of classes)/2 number of hidden nodes. The learned transformation matrix and trained model are used to find the accuracy of the testing set. The simulation is repeated ten times and the average results are reported in Table 19.2. For our experiment, we have set the global velocity constant (C_1) and local velocity constant (C_2) to 2.

Figure 19.4 shows the convergence graph of the maximization objective overall sample size. Though the Bhattacharyya distance is gradually increasing with an increase in iteration, however, after a certain point the increase in distance does not contribute to the increased inaccuracy. The obtained average accuracy over 20 runs is summarized in Table 19.2.

Table 19.2 shows that the proposed model learns a feature transformation matrix that can reduce the number of features significantly for neural network classification. Moreover, the newly generated matrix can achieve competitive accuracy. By using the proposed model, the learning process can achieve a balance of feature size and accuracy of classification using a neural network.

Table 19.2 Detailed of the results obtained

Dataset	Testing Acc. (%) (no reduction)	Testing Acc. (%) (with reduction)	Selected attributes	Total attributes
D1	96.67	96.67	1–2	4
D2	100.00	94.44	6–7	14
D3	99.12	92.04	6–10	30
D4	92.86	87.14	7–10	34
D5	80.49	79.27	1–3	60
D6	97.33	95.44	60–82	784
D7	90.45	92.05	90–100	784
D8	93.00	92.23	150–165	784

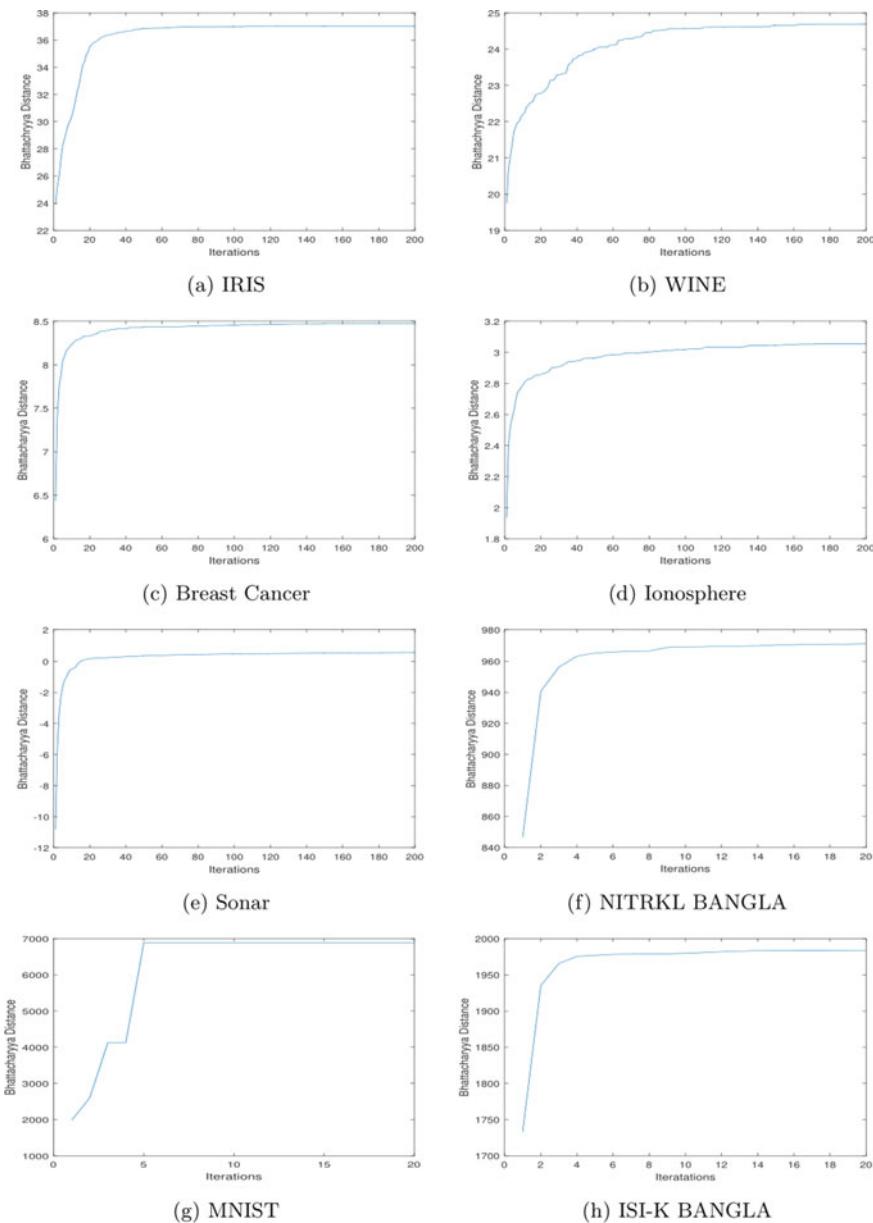


Fig. 19.4 Convergence graph of the various datasets

19.4 Conclusion

The feature selection algorithm is one of the important steps in machine learning approach to reduce unnecessary computational cost. The feature transformation matrix is useful to reduce the size of feature. However, the choice of number of features is generally set to random value and optimized empirically. This article presents a feature transformation matrix learning methodology using a novel modified PSO. The proposed optimization strategy finds optimal transformation matrix and finds the minimal feature size suitable for classification. The model is tested on various publicly available datasets and shows competitive results with proposition of optimal feature size for classification. In future, the model that can provide an explanation of the choice of transformation matrix will be studied.

References

1. Nayak, D.R., Dash, R., Majhi, B.: Automated diagnosis of multi-class brain abnormalities using MRI images: a deep convolutional neural network based method. *Pattern Recogn. Lett.* **138**, 385–391 (2020)
2. Kumar, R.L., Kakarla, J., Isunuri, B.V., Singh, M.: Multi-class brain tumor classification using residual network and global average pooling. *Multimedia Tools Appl.* **80**(9), 13429–13438 (2021)
3. Mishra, S., Mishra, S.K., Majhi, B., Sa, P.K.: 2d-dwt and Bhattacharyya distance based classification scheme for the detection of acute lymphoblastic leukemia. In: 2018 International Conference on Information Technology (ICIT), 2018 International Conference on Information Technology (ICIT), pp. 61–67 (2018)
4. Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M., Kwak, K.S.: A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inform. Fusion*. **63**, 208–222 (2020)
5. Kianat, J., Khan, M.A., Sharif, M., Akram, T., Rehman, A., Saba, T.: A joint framework of feature reduction and robust feature selection for cucumber leaf diseases recognition. *Optik* **240**, 166566 (2021)
6. Aghdam, M.H., Ghasem-Aghaee, N., Basiri, M.E.: Text feature selection using ant colony optimization. *Expert Syst. Appl.* **36**(3, Part 2), 6843–6853 (2009)
7. Maćkiewicz, A., Ratajczak, W.: Principal components analysis (PCA). *Comput. Geosci.* **19**(3), 303–342 (1993)
8. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **35**, 99–109 (1943)
9. Kailath, T.: The divergence and bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **15**(1), 52–60 (1967)
10. De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L.: The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* **50**(1), 1–18 (2000)
11. Abualigah, L.M., Khader, A.T., Hanandeh, E.S.: A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *J. Comput. Sci.* **25**, 456–466 (2018)

Chapter 20

Brain Image Classification Using Optimized Extreme Gradient Boosting Ensemble Classifier



Abhishek Das, Saumendra Kumar Mohapatra, and Mihir Narayan Mohanty

Abstract The abnormal development of brain cells is termed as brain tumor that can cause various neurological disorders. The in time detection of tumors can save the life of a person suffering from this dangerous disease. Various imaging techniques are being used to visualize the present condition of the brain so that the treatment will be followed accordingly. Magnetic resonance imaging (MRI) is considered one of the most utilized biomedical imaging techniques. After getting such images of the brain, the next task is the detection of the tumor. The automation in this problem field using machine learning algorithms leads to faster detection in comparison to manual observation. For this purpose, we have used the extreme gradient boosting-based ensemble classifier for brain MRI image classification. The classifier is well optimized by varying the inherent parameters of the classifier and the best score is observed with 96.1% classification accuracy. The training and validation losses are also decreased and recorded as 0.0069 and 0.1214 with proper parameter tuning.

20.1 Introduction

The most dangerous disease can be handled properly with prior knowledge of symptoms and proper diagnosis. Brain tumors also have some external as well as internal symptoms. Primary symptoms in brain tumors should not be ignored to avoid serious damage to health. The symptoms of brain tumors include, but are not limited to, continuous headache, memory loss, cognitive changes, motor deficit, seizures, personality change, visual problems, changes in consciousness, nausea or vomiting, sensory deficit, papilledema, depression, and anxiety[1–3]. After facing many of these symptoms, the next task is the diagnosis of the disease. This step includes brain scanning in different forms as these symptoms come under neurological problems.

A. Das · M. N. Mohanty (✉)

ITER, Siksha ‘O’ Anusandhan (Deemed to be University), Bhubaneswar, Odisha 751030, India

S. K. Mohapatra

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India
e-mail: saumendrakumam.sse@saveetha.com

Various techniques are adopted for the observation of the internal organs. Biomedical imaging is one of them. Ultrasound (US), X-ray, computed tomography (CT) scan, positron emission tomography (PET), and magnetic resonance imaging (MRI) are coming under biomedical imaging. MRI is mostly the preferred technique due to no use of ionizing radiation [4]. It helps in observing the current condition of brain tumors, but with the application of artificial intelligence, the diagnosis process can be further improved. Application of randomly generated graph-based convolutional neural network (CNN) has been done for brain MRI classification [5]. GeLU and ReLU activation were used in their work. The overall accuracy was 95.17%. A three-dimensional CNN model has been designed to classify brain image datasets developed with 12,000 CT images [6]. A thresholding step has been adopted before the CNN model. Transfer learning-based techniques are gaining the attention of researchers in recent years in the field of brain image classification [7]. Authors have used a pre-trained Google Inception V3 model in that work and obtained an F1-score of 0.913. Ensemble of two CNNs has been utilized in [8] by Ertosun and Rubin. The use of two CNNs may confuse the final decision if they predict opposite to each other. Therefore, using more than two numbers of base models is advisable. Authors have obtained 96% accuracy for glioblastoma multiforme (GBM) grade IV and lower grade glioma (LGG) classification, whereas 71% accuracy was for grade II and grade III types of glioma classification. Dataset enlargement by generative adversarial network (GAN) and classification by CNN have been applied in [9] that provided 88.82% accuracy in brain MRI classification. A combination of decision tree and bagging has also been used in the field of brain histology slides classification [10]. The features were extracted using an artificial neural network (ANN) model before classification and provided the F1-score of 0.648. Classical machine learning techniques can also perform better in comparison to deep learning-based models with proper modification [11]. Gauss derivation theorem has emerged as a new direction in the field of brain MRI classification [12]. Morphological segmentation [13] and edge detection [14, 15] have been proposed that can be used as preprocessing steps.

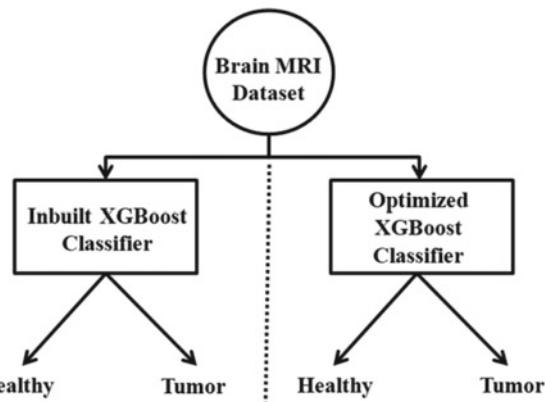
In this work, we have used extreme gradient boosting (XGBoost) model developed by Chen et al. [16] and proposed an optimized algorithm for brain MRI classification to detect the presence of a tumor. Optimization in the model performance is done by fitting the model with suitable parameters to compete with state-of-the-art methods.

The rest of the paper is organized as follows. Section 20.2 describes the proposed method in detail. Section 20.3 provides a detailed analysis of results obtained in the proposed method. Section 20.4 concludes the work.

20.2 Proposed Method

In this work, we have designed an XGBoost ensemble model optimized by varying the inherent parameters for brain MR image classification to detect whether the MR image contains a tumor or it is healthy. The block diagram of the proposed model is shown in Fig. 20.1.

Fig. 20.1 Block diagram of the proposed method



20.2.1 XGBoost Classifier

XGBoost is considered as an effective ensemble learning-based classifier due to its fast and accurate performance. The ensemble structure is formed by using several decision tree models. Trees are added to the parent model depending upon the training to fit the correct predictions and to eliminate the errors resulting from the previous trees. Gradient descent optimization is the method used to overcome the loss function. The loss gradient minimizes as the model is properly trained, hence named gradient boosting.

The XGBoost classifier from the Python repository is used in its initial condition to check the performance. The performance of such an inbuilt model is evaluated in terms of cross-entropy loss, also known as log loss as mentioned in Eq. (20.1). Less the value of log loss, more accurate is the classification.

$$\text{log loss} = y \ln(p) + (1 - y) \ln(1 - p) \quad (20.1)$$

where $p = \frac{1}{1+e^{-x}}$, y is the actual label in $\{0,1\}$, and p is the probability score evaluated by the model.

The system performance of the inbuilt XGBoost classifier is studied and variation in inherent parameters is done to decrease the log loss as described in the next subsection.

20.2.2 Optimized XGBoost Classifier

XGBoost classifier is tunable in terms of the number of trees and learning rate. The subsample ratio of the training instances and feature size in terms of cosample_bytree also have effects in optimizing the model. Cosample_bytree represents the subsample

ratio of values in columns when each tree is constructed. The XGBoost classifier is optimized using the Algorithm 1.

Algorithm 1. Optimization of XGBoost

Input: D=Dataset

Initialization

1. Load (XGB=XGBoost Classifier)
2. N=tree size, eta= learning rate

Training

3. **For** N=100 to 2000
 - a. **If** N=100
eta=0.3
Train XGB with D
 - b. **elseIf** N=500
eta=0.05
Train XGB with D
 - c. **elseIF** N=2000
eta=0.05
subsample=0.5
cosample_bytree=0.5
Train XGB with D
 - d. **end of If**
 - e. **Output:** Learning Curves
 4. **End of For**
 5. **Output:** Classification results
-

The objective function of XGBoost as mentioned in Eq. (20.2)

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

where $\Omega(f) = \gamma T + \lambda \frac{1}{2} \|\omega\|^2$ (20.2)

where l is the loss function, t represents the number of trees, Ω represents the complexity of the model that depends on sample rate and the number of features.

Variation in the previously mentioned hyperparameters is done to obtain a remarkable performance in comparison to the inbuilt XGBoost classifier, as well as to compete with the state-of-the-art methods.

20.3 Results and Discussion

The proposed model is verified using the brain image dataset available publicly at the online platform Kaggle [17]. The dataset has been prepared with a total of 253 MRI images of the brain including 155 numbers of tumor-contained brain images and 98 healthy brain images. A sample of the dataset is provided in Fig. 20.2.

20.3.1 XGBoost Classifier

The XGBoost classifier is having 100 numbers of trees. The learning rate is fixed at 0.3. Subsample value and cosample_bytree are 1 by default. The log loss obtained for both training and validation in such conditions is shown in Fig. 20.3.

Training loss of 0.0619 and validation loss of 0.1936 were obtained in the initial condition that can be observed from Fig. 20.3.

20.3.2 Optimized XGBoost Classifier

The XGBoost classifier is then optimized by varying the number of trees from 100 to 500, and the variation is observed in terms of learning curves as shown in Fig. 20.4.

A decrease in log loss values is observed by increasing the number of trees from 100 to 500. Training loss of 0.0503 and validation loss of 0.1612 were obtained in that condition and can be observed from Fig. 20.4.

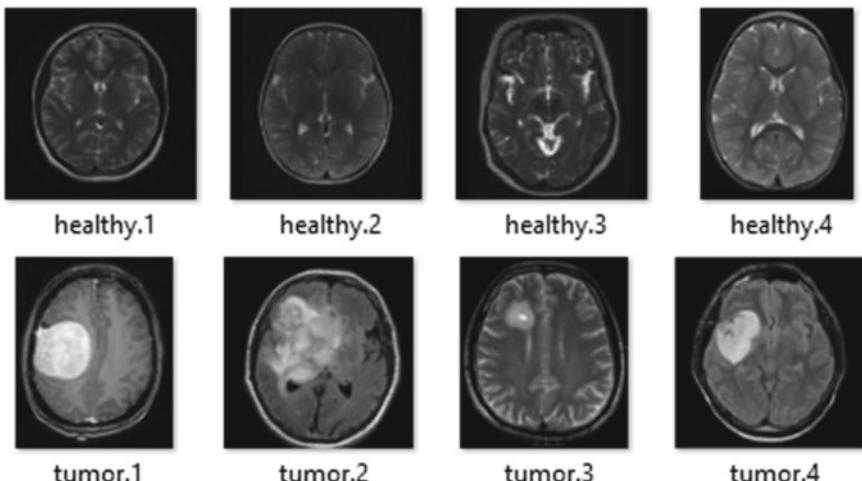


Fig. 20.2 Sample of the dataset

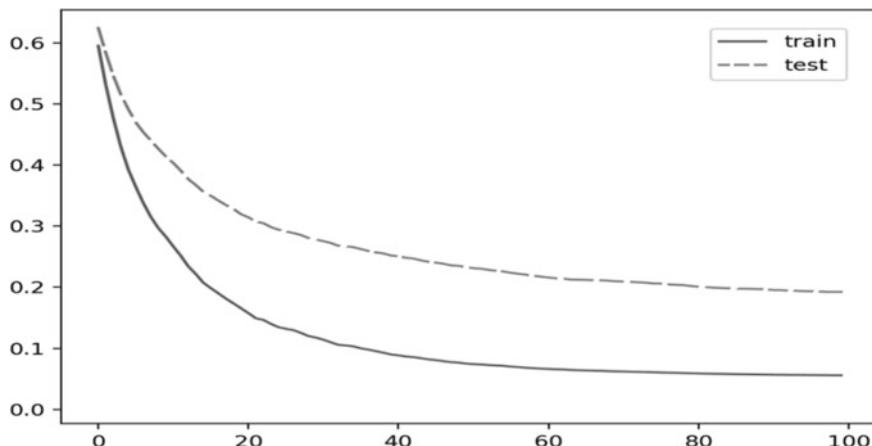


Fig. 20.3 Learning curves obtained from inbuilt XGBoost classifier

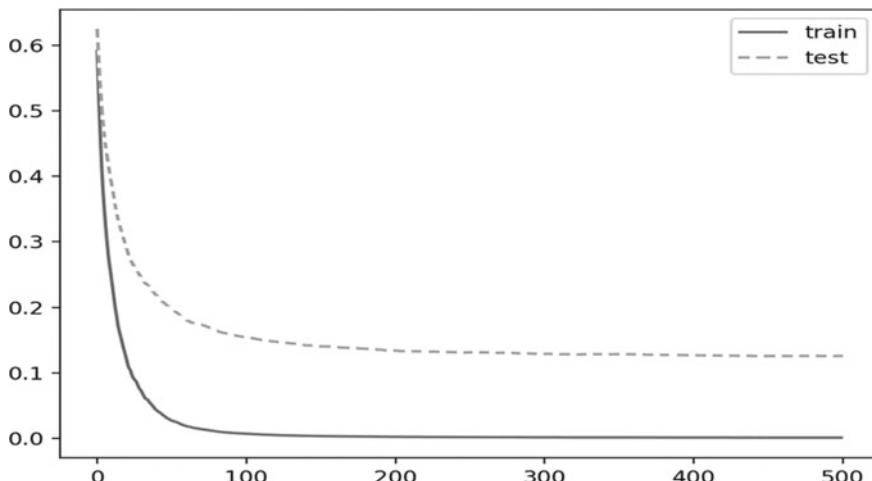


Fig. 20.4 Learning curves with tree size = 500

The learning rate is then changed to 0.05 to slow down the learning, and it is observed that the learning curves are now much better than the previous condition as mentioned in Fig. 20.5.

Both the subsample and the cosample_bytree parameters are varied to 0.5, and the number of tree sizes is also increased to 2000; the log loss values for training and validation obtained in this condition are shown in Fig. 20.6.

From Fig. 20.6, it is observed that a competitive performance is obtained in this condition. The log loss values for training and validation are very less, i.e., 0.0069

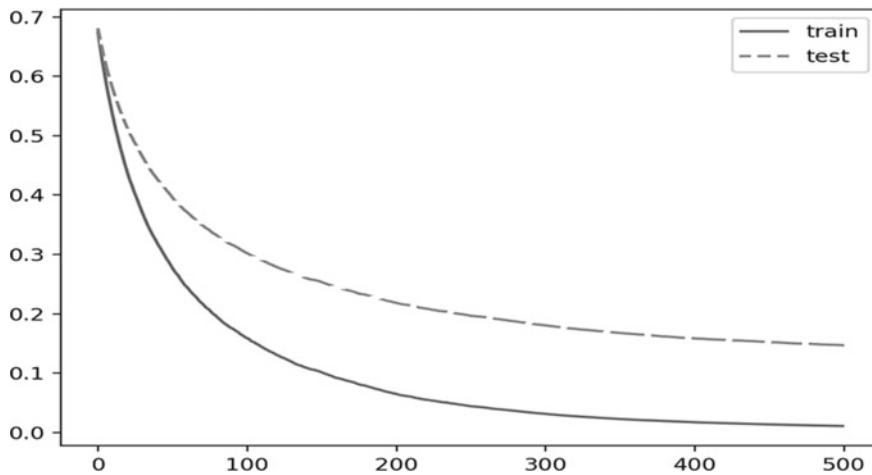


Fig. 20.5 Learning curves with tree = 500 and learning rate = 0.05

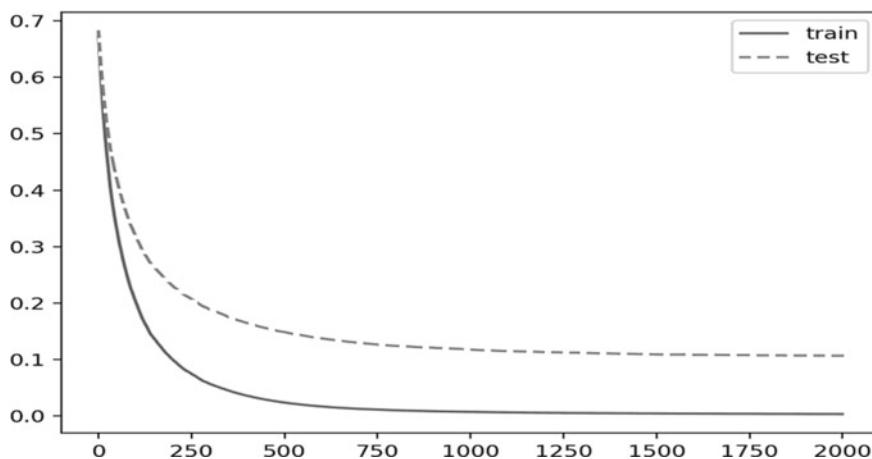


Fig. 20.6 Learning curves obtained by the proposed model

and 0.1214, respectively. The accuracy obtained with a variation in the number of trees is shown in Fig. 20.7.

An increase in accuracy is observed when there is an increase in the number of decision trees in the ensemble model. The accuracy values are 94.5, 95.3, and 96.1% for 100, 500, and 2000 numbers of tree size. The final improved result is obtained due to the optimization in the XGBoost classifier that satisfies the objective of this work.

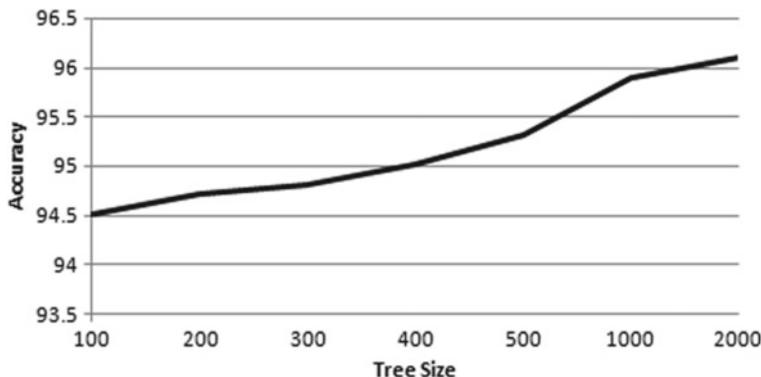


Fig. 20.7 Accuracy plots with varying tree size

20.4 Conclusions

Various machine learning techniques are being developed in the field of biomedical image analysis for faster and accurate classification and diagnosis. In this work, we have proposed the optimization in the XGBoost classifier concerning certain parameters for improved performance in the field of brain MRI classification for the detection of the tumor. The proposed algorithm is proved to be a better classification model with very less log loss values of training and validation and a higher value of accuracy. The work is to be further analyzed with deep learning-based ensemble models in the future for different biomedical image classifications.

References

- Chandana, S.R., Movva, S., Arora, M., Singh, T.: Primary brain tumors in adults. *Am. Fam. Physician* **77**(10), 1423 (2008)
- Marku, M., Rasmussen, B.K., Dalton, S.O., Johansen, C., Hamerlik, P., Andersen, K.K., Meier, S.M., Bidstrup, P.E.: Early indicators of primary brain tumours: a population-based study with 10 years' follow-up. *Eur. J. Neurol.* **28**(1), 278–285 (2021)
- Tibbs, M.D., Huynh-Le, M.P., Reyes, A., Macari, A.C., Karunamuni, R., Tringale, K., Burkeen, J., Marshall, D., Xu, R., McDonald, C.R., Hattangadi-Gluth, J.A.: Longitudinal analysis of depression and anxiety symptoms as independent predictors of neurocognitive function in primary brain tumor patients. *Int. J. Radiat. Oncol. Biol. Phys.* **108**(5), 1229–1239 (2020)
- Guerquin-Kern, M., Lejeune, L., Pruessmann, K.P., Unser, M.: Realistic analytical phantoms for parallel magnetic resonance imaging. *IEEE Trans. Med. Imaging* **31**(3), 626–636 (2011)
- Huang, Z., Du, X., Chen, L., Li, Y., Liu, M., Chou, Y., Jin, L.: Convolutional neural network based on complex networks for brain tumor image classification with a modified activation function. *IEEE Access* **8**, 89281–89290 (2020)
- Ker, J., Singh, S.P., Bai, Y., Rao, J., Lim, T., Wang, L.: Image thresholding improves 3-dimensional convolutional neural network diagnosis of different acute brain hemorrhages on computed tomography scans. *Sensors* **19**(9), 2167 (2019)

7. Ker, J., Bai, Y., Lee, H.Y., Rao, J., Wang, L.: Automated brain histology classification using machine learning. *J. Clin. Neurosci.* **66**, 239–245 (2019)
8. Ertosun, M.G., Rubin, D.L.: Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. In: AMIA Annual Symposium Proceedings (Vol. 2015, p. 1899). American Medical Informatics Association (2015)
9. Ge, C., Gu, I.Y.H., Jakola, A.S., Yang, J.: Enlarged training dataset by pairwise GANs for molecular-based brain tumor classification. *IEEE Access* **8**, 22560–22570 (2020)
10. Huda, S., Yearwood, J., Jelinek, H.F., Hassan, M.M., Fortino, G., Buckland, M.: A hybrid feature selection with ensemble classification for imbalanced healthcare data: a case study for brain tumor diagnosis. *IEEE access* **4**, 9145–9154 (2016)
11. Naga Srinivasu, P., Srinivasa Rao, T., Dicu, A.M., Mnere, C.A., Olariu, I.: A comparative review of optimisation techniques in segmentation of brain MR images. *J. Intell. Fuzzy Syst.* (Preprint) 1–12 (2020)
12. Srinivasu, P.N., Rao, T.S., Balas, V.E.: Volumetric estimation of the damaged area in the human brain from 2D MR image. *Int. J. Inf. Syst. Model. Des. (IJISMD)* **11**(1), 74–92 (2020)
13. Jyoti, A., Mohanty, M.N., Kumar, M.P.: Morphological based segmentation of brain image for tumor detection. In: 2014 International Conference on Electronics and Communication Systems (ICECS) (pp. 1–5). IEEE (2014)
14. Behera, S., Mohanty, M.N., Patnaik, S.: A comparative analysis on edge detection of colloid cyst: a medical imaging approach. In: Soft Computing Techniques in Vision Science (pp. 63–85). Springer, Berlin, Heidelberg (2012)
15. Mallick, P.K., Satapathy, B.S., Mohanty, M.N., Kumar, S.S.: Intelligent technique for CT brain image segmentation. In 2015 2nd International Conference on Electronics and Communication Systems (ICECS) (pp. 1269–1277). IEEE (2015)
16. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining (pp. 785–794) (2016)
17. Chakrabarty, N.: Brain MRI Images for Brain Tumor Detection (2019). Available: <https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection>

Chapter 21

Concolic-Based Software Vulnerability Prediction Using Ensemble Learning



**Swadhin Kumar Barisal, Pushkar Kishore, Gayatri Nayak,
Ridhy Pratim Hira, Rohit Kumar, and Ritesh Kumar**

Abstract Detecting software vulnerabilities are critical for limiting the damage caused by hostile exploits and program failures. This frequently necessitates the accurate identification of susceptible execution routes. However, as software complexity has increased, it has become notoriously difficult to find such susceptible paths by exploring the entire program execution space. Therefore, concolic testing is used in this paper as one of the ways to deal with this problem. Here, the observations and discoveries are collected from experimenting and implementing concolic testing. First, several trained classifier models like random forest, support vector machine, stochastic gradient descent, and AdaBoost are tested against a test dataset created by randomly selecting 30% of the data from the original dataset. Then, multiple classifier models help predict whether a program is faulty or benign. After testing out several classifier models, an ensemble is done on the top 3 highest accuracy classifiers. Overall, 87% accuracy is achieved with an F1-score of 85.1%. This result indicates that 87% of the program's labels are accurately detected by our proposed model while higher F1-score represents the proposed model's balanced detection.

21.1 Introduction

Vulnerabilities in software provide a persistent threat to software industries. Severe security flaws and their widespread impact on many users have pushed software security into the public and media spotlight in recent years. Securing software systems have become exceedingly difficult due to the rising complexity of software.

S. K. Barisal (✉) · G. Nayak · R. P. Hira · R. Kumar · R. Kumar
Department of Computer Science and Engineering, Siksha 'O' Anusandhan Deemed To Be
University, Bhubaneswar, Odisha, India
e-mail: swadhinbarisal@soa.ac.in

G. Nayak
e-mail: gayatrinayak@soa.ac.in

S. K. Barisal · P. Kishore
Department of Computer Science and Engineering, National Institute of Technology, Rourkela,
Odisha, India

According to the previous research, there are between 5 and 20 flaws per 1,000 lines of software code [1]. As a result, it is essential to be aware of program pathways that contain vulnerabilities. Code-similarity-based detection can detect multiple clone types and achieve higher detection accuracy. However, they are inefficient due to the higher false-negative rate and analytical complexity. On the other hand, code-pattern detection can achieve higher code coverage and work in a faster way [2, 3]. However, the lack of run-time information and low code coverage deems it ineffective. In the case of within-project prediction, the prediction of defective program modules is accurate, but expansion is poor. In cross-project defect prediction, dataset resources are effectively integrated for promoting new project development practices. However, the disadvantage is excessive code feature extraction granularity. The major drawback is the lack of extensive training data required for training the model.

Software vulnerabilities are flaws or errors in a program's code that make the entire code base vulnerable to malicious attacks [4–8]. There are multiple approaches to dealing with this issue, and different approaches work in different situations. Concolic testing [9, 10] is one of the ways to deal with this problem. Here, the whole technique is discussed in this article, as well as the observations and discoveries collected from experimenting with and implementing concolic testing, and then combining the results with multiple classifier models to evaluate if a program is faulty or benign.

21.2 Motivation

To deal with this issue, a concolic testing approach is paired with machine learning to make vulnerability prediction much faster and automated, hence overcoming the disadvantages of the previous systems and making this system more reliable. Since most works use manually crafted test cases and current testing methods are typically ineffective and sluggish to uncover flaws. Concolic testing solves this issue by automatically producing test cases with a high level of coverage. As most of the hardware necessary for this work is affordable and needs a one-time expenditure, it has been determined that this work is financially feasible. Apart from that, all of the software tools utilized in work are free and open-source; therefore, there is no additional cost. Furthermore, Krishna et al. [11] stated that the issues present in the code would lead to bugs, and no enhancement will happen for open-source programs. Thus, we can collect the behavior that defines the issues in the program and accordingly convert them into training features for finding vulnerable programs. The impetus for this work stems mainly from the need to improve the speed, accuracy, and reliability on vulnerability prediction.

21.3 Contributions

This work discusses using the concolic testing and ensemble technique to classify software modules as faulty or benign. This paper has four significant contributions like (i) to create the dataset by analyzing programs using a concolic tester tool named KLEE, (ii) to apply the random forest (RF), support vector machine (SVM), stochastic gradient descent (SGD), and AdaBoost techniques for designing a vulnerability classifier, (iii) more specifically, we design an ensemble model utilizing the features extracted from concolic testing, and (iv) to compare the accuracy achieved using code-based models with the state-of-the-art work.

21.4 Basic Concepts

This section discusses the basic concepts required for understanding the work done in this article.

21.4.1 *Concolic Testing*

Concolic execution examines a program's code in order to produce test data automatically that can execute all possible paths. However, there had been attempts to address this issue utilizing path constraints [9] or augmenting test paths with concrete values in concolic execution [12]. Concolic testing is one such approach that traverses all of the routes in a program decision tree using concolic and physical execution. A concolic testing strategy is a function that determines whether to use random testing or concolic execution and if concolic execution is used, then which program route to use. After all of the pathways have been located through concolic testing, the susceptible pathways are identified utilizing preferred machine learning approaches. Once these susceptible routes have been identified, the code is updated to make it bug-free, ensuring that any changes are thoroughly tested.

21.4.2 *Vulnerability*

One of the main reasons for security issues is software vulnerabilities. Hackers with advanced skills can utilize software flaws to carry out various malicious actions, including stealing users' personal data and shutting down critical equipment [13].

21.4.3 Support Vector Machine

In SVM, the training data is represented as points in problem space divided into two classes. It performs well in high-dimensional environments and only uses a few training points in the decision function, making it memory-friendly. A subset of the data points determines the maximum-margin separator. Support vectors are the data points in this subset. Since we use the support vectors to determine which side of the separator a test case is on, it will be computationally beneficial if just a tiny fraction of the data points are support vectors. We must solve the following optimizations to determine the largest margin separator. The optimization relations are shown in Eq. (21.1).

$$\begin{aligned} w \cdot x_c + b &> +1 \text{--- positive cases} \\ w \cdot x_c + b &< -1 \text{--- negative cases} \\ \text{And } \|w\|_2 &\text{ is as small as possible} \end{aligned} \quad (21.1)$$

21.4.4 Random Forest

Random forest is an ensemble learning approach. It corrects the tendency of decision trees to overfit during their training. It follows three basic steps. First, step is random selection of samples, second step is to construct a decision tree, and finally, it does voting for every predicted result.

21.4.5 AdaBoost

Boosting is a type of ensemble machine learning technique that combines the predictions of many weak learners. A weak learner is a model that is relatively basic but has some ability on the dataset. For example, long before a practical algorithm could be devised, boosting was a theoretical concept, and the AdaBoost (adaptive boosting) method was the first effective implementation of the concept. The AdaBoost algorithm employs very short (one-level) decision trees as weak learners introduced to the ensemble sequentially. Each model in the series seeks to correct the predictions made by the model preceding it. This is accomplished by balancing the training dataset to focus more on training cases where the previous models failed to predict correctly.

21.4.6 Stochastic Gradient Descent (SGD)

SGD algorithm is a simple optimization approach. The parameter values $\{wt, t = 1, \dots\}$ are selected randomly at each iteration. This algorithm does not remember all data of the previous iterations and process data on the fly. It processes only one sample at a time which makes it faster. It fits easily to system as a single training sample is executed every time.

21.5 Proposed Model's Methodology

This work uses two particular techniques to achieve the ideation proposed. One of them is concolic testing that combines concrete execution and symbolic execution, then the results are paired to create a dataset. The other is ensemble model where this dataset is provided to the machine learning techniques for training classifiers, and prediction is made on test data to validate the model's effectiveness. Figure 21.1 represents the architecture of our proposed model.

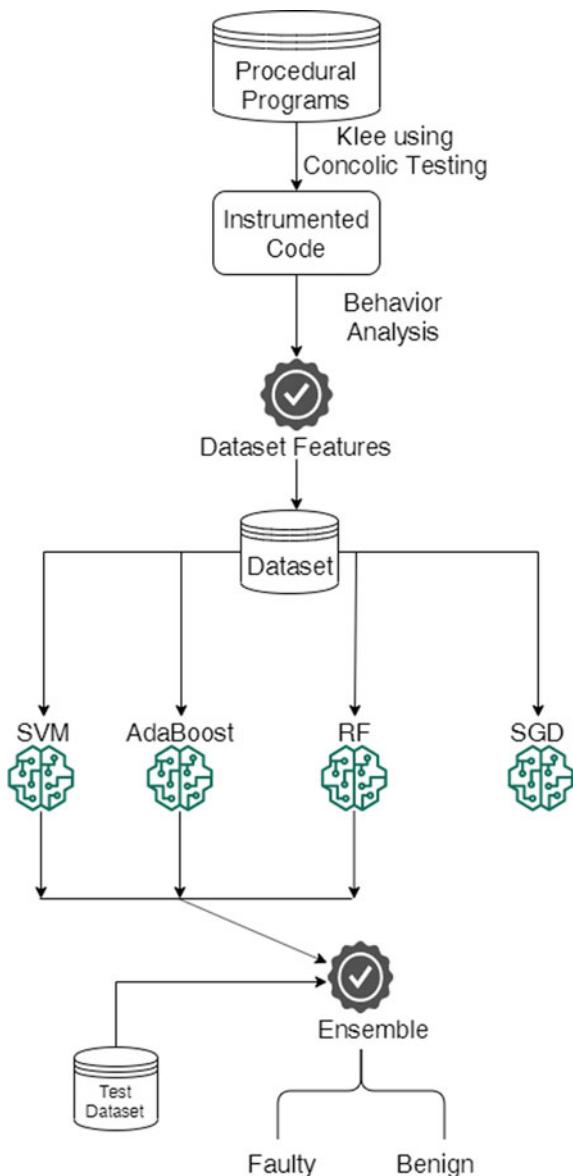
21.5.1 Detailed Description

This section outlines the whole work's process, from conception to execution. With the combination of concolic testing and machine learning, this research presents a new way to detect software vulnerability. In first phase, a curated set of c programs with moderate conditions is selected, and concolic testing is done. For concolic execution, a concolic execution tool known as KLEE [9] is used inside a Linux container, which is again created using Docker. The set of c programs is then given as input to KLEE, which provides us the coverage and execution details of the programs. Then, concrete execution is done over the same set of programs using curated test cases having random test values.

In the second phase, the output from concolic testing is studied. A dataset with appropriate parameters is created to train the classifier models, which will eventually be used to predict vulnerable code by labeling them either faulty or benign.

In the third phase, several trained classifier models are tested against a test dataset created by randomly selecting 30% of the data from the original dataset. Original dataset contains 200 programs written in C language. After testing several classifier models like RF, SGD, SVM, and AdaBoost, an ensemble is done on the top three highest accuracy classifiers. The maximum voting technique decides the final decision or label of the program. Overall, 87.6% accuracy is achieved, with a precision of 80%, recall of 90.9%, and F1-score of 85.1%.

Fig. 21.1 Proposed architecture of our approach



The entire experiment and the research that goes along with it prove to be highly productive and exciting, opening up a plethora of new avenues for creative thinking in the realm of software security and testing.

21.6 Experimental Results

We present our model in detail in this section and emphasize our proposed approach's experimental results. The ensemble technique is used for detecting whether the program is vulnerable or not.

21.6.1 Experimental Setup

Datasets have been separated into training and testing data in the ratio of 70:30, respectively, even though many studies have employed a random sample size to minimize bias and variance.

21.6.2 Performance Measure Parameters

The entire experimentation process may be broken down into three parts. The first is the testing step, which entailed creating a collection of C programs with a modest number of conditions and executing them using concolic and concrete execution. The second stage is the data organizing stage, which entails analyzing the output data from testing and creating a dataset with the needed parameters for training a classifier. The final stage is a machine learning based, in which the dataset created in the previous phase is used to train many classifier models, which are then tested against test data to see if the program is defective or not. Once our model is trained, we can provide it with any test data, and it will help us predict vulnerable code. This will reduce the amount of human intervention required to complete the task.

21.6.3 Performance Measure Evaluation

Table 21.1 provides the performance measure of the proposed ensemble model. Table 21.1 represents the performance measure of our ensemble model and derived from the combination of RF, SVM, and AdaBoost.

We apply SGD too for vulnerability prediction but unconsidered for ensemble due to its poor accuracy. Out of the three techniques considered for ensemble, AdaBoost

Table 21.1 Performance evaluation of our proposed model

Performance parameters	Accuracy (%)	Precision (%)	Recall (%)	TNR (%)	FPR (%)	FNR (%)	F1-score (%)
Values	87.6	80	90.9	84.4	15.6	9.1	85.1

has the highest accuracy and SVM has the lowest. In case of SVM, standard scaling is done to reorganize the features in a definite range and poly degree SVM is applied. Above combination of the models is effective since ensemble works effectively for the combination of strong and weak learners. The FPR value is higher and implies that some benign programs are termed faulty. But, the lower FNR ensures that least amount of faulty programs will be executed on the system. Since, the overall objective of our work is to secure the system from vulnerable programs, thus lower FNR is much objective fulfilling than comparatively a bit higher FPR. The accuracy of the detector is 87.6% and F1-score is 85.1%. The higher F1-score ensures a better harmonic mean between precision and recall.

Table 21.2 describes the result of programs, which are tested by our proposed model. There are seven columns representing the feature of the dataset like (i) INS—Instructions represent the number of executed instructions, (ii) EXT—Execution time represents the total time consumed by KLEE for analyzing the program, (iii) ICOV—Instruction coverage represents the instruction coverage in the LLVM bitcode, (iv) BCOV—Branch coverage represents the branch coverage in the LLVM bitcode, (v) IC—Instruction count represents the total static instructions in the LLVM bitcode, (vi) TSolver—Solvetime represents the time spent in the constraint solver, (vii) LOC—Lines of code represents the amount of code present.

Apart from that, AL represents actual label, PL represents predicted label, F represents faulty, and B represents benign dataset. Out of 15 programs, the highlighted ones are misclassified and increase count of FP and FN. However, overall the accuracy is higher in case of vulnerability prediction.

Table 21.2 Result of ensemble for tested input programs

Program name	Ins	Ext (s)	Icov (%)	Bcov (%)	IC	TSolver	LOC	AL	PL
Regexp_iterative	327,240	1.78	100	94.12	167	25.2	43	F	F
Masked_error1	74	0.01	96.3	83.33	54	65.7	22	F	F
Pointer_struct6	422,066	311.37	83.67	83.33	147	95.66	45	F	F
Pointer_struct5-11	18	0.01	18	8.33	100	0	36	B	B
mcdc	51,806,526	1276.79	78.95	50	38	38.16	15	F	F
bubble4	20	0.04	22.99	12.5	87	0	22	F	F
Addition	66	0.01	100	100	47	64.67	18	B	B
Pointer_symbolic	159	0.02	100	100	63	75	24	F	B
Regexp74	15	0.01	7.04	3.12	213	0	63	F	F
Branch	26	0.01	100	100	23	73.98	15	B	B
Arraysimple5	1938	1.83	100	100	115	39.87	31	F	B
Pointer_struct5-9	18	0	18	8.33	100	0	36	B	B
Regexp_small	4864	0.07	97	90	135	71.66	34	F	B
Pointer_wp1	132	0.01	100	100	60	73.51	26	B	F
Masked_error1	74	0.01	96.3	83.33	54	65.75	22	F	F

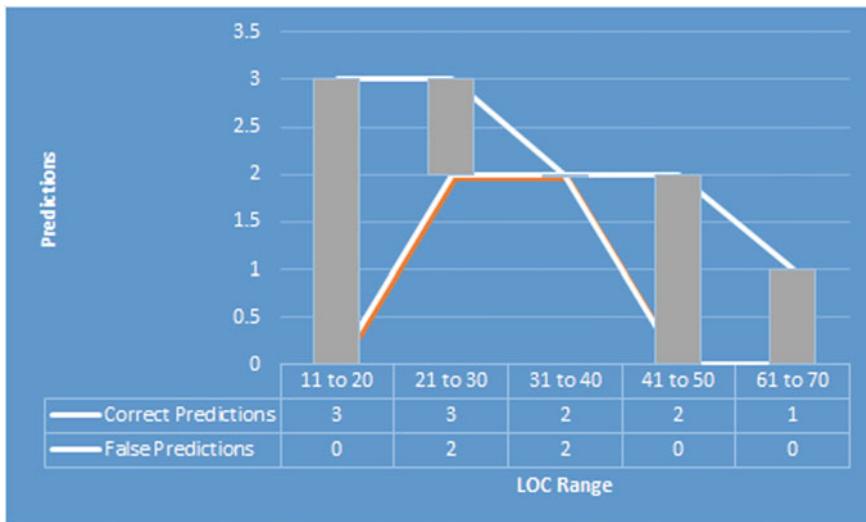


Fig. 21.2 Error in predictions for different LOC range

Figure 21.2 depicts that we are able to achieve better fault detection for the programs where LOC is higher or lower. Accuracy dips for the medium LOC value and indicates that we need to instrument the code properly or provide more samples for improving it.

21.7 Comparison with State-of-the-Art Works

Rizwan et al. [14] studied 14 non-graphic classifier performance indicators that are often employed in SFP investigations. These findings motivate researchers to focus in determining the best performance measure for evaluating models. They conclude that the F-measure and the G-mean¹ are both good candidates for evaluating SFP models with careful examination, as there is a danger of erroneous data in certain cases.

Kaur et al. [15] wanted to discover the best category, and thus the best classifier, for fault classification. This study applies twenty-one classifiers from five open-source applications. MATLAB's classification was used to test multiple classification models.

Jiang [16] studied many techniques and worked on data from a public repository, revealed the merits and disadvantages of performance evaluation methodologies and concludes that selecting the “best” model requires consideration of project cost factors, which are unique to each development environment.

21.8 Conclusion and Future Work

It is apparent from the data and results of the studies that the model described in this article is feasible and valid. With an accuracy of 87 percent, the model is a unique combination of concolic testing and machine learning, making it a fast and reliable type of software testing. Apart from that, concolic execution decreases the number of inputs needed to detect bugs and defects in the code.

In future, using an AI-assisted extension that automates the whole process during development, testing time may be slashed in half. It might even be extended to locate and correct the precise code block where a fault exists. Another option is to look at how to enhance the vulnerability prediction's localization and interpretation.

References

1. Libicki, M.C., Ablon, L., Webb, T.: The defender's dilemma: charting a course toward cybersecurity. Rand Corporation (2015)
2. Li, J., He, P., Zhu, J., et al.: Software defect prediction via convolutional neural network. In: Proceedings of the 2017 IEEE International Conference on Software Quality, Reliability and Security (QRS), IEEE, pp. 318–328, Prague, Czech Republic, July 2017
3. Zhao, L., Shang, Z., Zhao, L., et al.: Siamese dense neural network for software defect prediction with small data. *IEEE Access* **7**, 7663–7677 (2018)
4. Li, Z., Zou, D., Xu, S. et al.: VulDeePecker: a deep learning based system for vulnerability detection (2018). <https://arxiv.org/abs/1801.01681>
5. Li, Z., Zou, D., Xu, S. et al.: SySeVR: a framework for using deep learning to detect software vulnerabilities (2018). <https://arxiv.org/abs/1807.06756>
6. Xiaomeng, W., Tao Z., Runpu, W., Wei, X., Changyu, H.: CPGVA: code property graph based vulnerability analysis by deep learning. In: Proceedings of the 2018 10th International 14 Security and Communication Networks Conference on Advanced Infocomm Technology (ICAIT), IEEE, pp. 184–188, Stockholm, Sweden, August 2018
7. Dam, H.K., Pham, T., Ng, S.W., et al.: A deep tree-based model for software defect prediction (2018). <https://arxiv.org/abs/1802.00921>
8. Kishore, P., Barisal, S.K., Vaish, S.: Nitrcst: a software security tool for collection and analysis of kernel calls. In: TENCON 2019–2019 IEEE Region 10 Conference (TENCON), pp. 510–515 (2019)
9. Boonstoppel, P., Cadar, C., Engler, D.: RWset: attacking path explosion in constraint-based test generation. In: International Conference on Tools and Algorithms for the Construction and Analysis of Systems. Springer, Berlin, Heidelberg. pp. 351–366 (2008)
10. Barisal, S.K., Behera, S.S., Godbole, S., Mohapatra, D.P.: Validating object-oriented software at design phase by achieving MC/DC. *Int. J. Syst. Assur. Eng. Manage.* **10**(4), 811–823 (2019)
11. Krishna, R., Agrawal, A., Rahman, A., Sobran, A., Menzies, T.: What is the Connection Between Issues, Bugs, and Enhancements? In: 2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE-SEIP), pp. 306–315 (2018)
12. Su, T., Pu, G., Fang, B., He, J., Yan, J., Jiang, S., Zhao, J.: Automated coverage-driven test data generation using dynamic Concolic execution. In: 2014 Eighth International Conference on Software Security and Reliability, IEEE. (SERE), pp. 98–107 (2014)
13. Barisal, S., K., Dutta, A., Godbole, S., Sahoo, B. and Mohapatra, D. P.: BOOMPizer: Minimization and prioritization of CONCOLIC based boosted MC/DC test cases. *Journal of King Saud University-Computer and Information Sciences.* **34**(1), 1–20 (2022)

14. Rizwan, M., Nadeem, A., Sindhu, M.A.: Analyses of classifier's performance measures used in software fault prediction studies. *IEEE Access* **7**, 82764–82775 (2019)
15. Kaur, I., Kaur, A.: Comparative analysis of software fault prediction using various categories of classifiers. *Int. J. Syst. Assur. Eng. Manage.* **12**(3), 520–535 (2021)
16. Jiang, Y., Cukic, B., Ma, Y.: Techniques for evaluating fault prediction models. *Empirical Softw. Eng.* **13**(5), 561–595 (2008)

Chapter 22

Very Short-Term Photovoltaic Power Forecast by Multi-input Long Short-Term Memory Network



Sasmita Behera, Debasmita Mohapatra, Aman Kaushal, and Shrabani Sahu

Abstract Electricity generation from solar power is indispensable in the scenario of environmental and energy crises. Progressively increased share of installing solar photovoltaic (PV) technology widely has occurred for gradual reduction of cost and short gestation period. But besides this, sunlight has diurnal unavailability and irregular tendency due to its weather dependence. To utilize it in a large scale, accurate forecasting of solar power is necessary for assuring the economic and stable operation of the smart grid and participation in bidding. In this paper, long short-term memory (LSTM) network, a type of recurrent neural network, has been used to forecast the PV output power using multiple weather inputs and power. Rigorous testing of multiple-input LSTM networks has been done, and a selection of dominant weather inputs is carried out. The LSTM has been optimized based on the number of neurons, number of layers and also optimized concerning the loss function used. The LSTM network obtained at the end is fully optimized for the solar power data for the prediction of 10 min ahead power forecast.

22.1 Introduction

Photovoltaic (PV) power generation from the sun has got recognition to adapt in the past several years due to environmental concerns, simple and easy to install. However, the PV power depends on the weather. The power trade, management of operational cost and power market participation see reflections of fluctuations of this when it is a grid level large dispatch [1]. For the various aforementioned reasons stated to take maximum benefit from PV power, precise and fast forecast of this uncontrolled source in short and very short-term ahead is very important. Some of the researches for the forecast of PV power involve historical data from power only for time-series prediction, and some involve weather variables. For forecasting PV

S. Behera (✉) · D. Mohapatra · A. Kaushal

Department of EEE, Veer Surendra Sai University of Technology, Burla 768018, India

S. Sahu

Department of EE, Veer Surendra Sai University of Technology, Burla 768018, India

power for different seasons in a year, the physical model equations and feed-forward neural network (NN) have been used in comparison for a quarter-hour sampling [2]. The NN has proven better with five number time-series inputs including weather and power, but the inference about the selection of neurons and epochs is not exactly presented. Many researchers have contributed towards the forecast of solar irradiation and then calculate power [3]. In such an application, random forest regression has shown a better forecast of irradiation for the multi-time step ahead based on hourly data. However, as such the computational effort involved raises the costs. For a better forecast of power, the consecutive hourly power inputs and weather forecast for the next day are utilized in a NN. But, in this approach, the error in the weather forecast may introduce an error in the power forecast [4]. A similar two-step approach by autoregression models is presented for the long-term prediction of power [5]. But, in a shorter prediction time and sampling, the accuracy of the technique may suffer. In [6], NN is used to predict monthly solar radiation by weather variables for longer historical data. The training and the network structure for the improvement of performance are discussed in detail. Long short-term memory (LSTM) [7], a type of recurrent neural network (RNN), can learn the sequential dependence as used for irradiation prediction. Only single-layered LSTM networks have been compared with networks like least squares regression (LR) and back propagation neural networks (BPNN) [7]. A direct power prediction method by regression analysis determining a correlation of power on historical data of weather has been used in [8]. An RNN has been tested for power prediction for a day ahead based on only weather data [9]. For all time frames of forecast NN with autoregression has been successfully modified for power prediction by temperature, irradiation and some weather-derived input [10]. Therein the PV model has been predicted by the ANN. A convolutional neural network (CNN) with LSTM has been used with and without using weather data for an hour ahead and long-term forecasts of PV power in the latest research [11]. Furthermore, extreme learning machine for short-term PV power forecast has been proposed recently [12].

To overcome the problems left open as discussed, the application of NNs is an emerging technique for forecasting using historical data. For a very short-term prediction for 10 min ahead or down below, the accuracy and time of run for training and prediction are to be balanced. Involving historical data of both weather parameters along with power information gives a better result. However, the involvement of a greater number of parameters is computationally expensive and sacrifice time. This paper suggests 10 min ahead power generation forecast of a PV system by a NN model. In this paper, a proper study of the application of LSTMs to time-series forecasting of power and how to balance the accuracy and computation involved has been done. The present study adds to [7] by providing methods to decrease the error from LSTM networks. The focus has been on two parameters of LSTM: neurons per layer and the number of layers while training the network.

The paper has been divided into two parts, Sect. 22.2 and Sect. 22.3. Section 22.2 is focused on the methodology and architectures of LSTM. Section 22.3 shows the results obtained, and it gives the correlation of various inputs, the effect of removing the non-dominant inputs and shows how the LSTM network is optimized to extract

maximum efficiency with minimum error and comparative study of results with increasing epochs. Finally, the conclusions are drawn.

22.2 Methodology

The methodology is depicted in Fig. 22.1. To implement the prediction, first of all, the 10 min sampled data is collected for a year to train for a whole seasonal variation for different weather parameters and power from available resources for a particular location. Then, the data is checked for missing data, such as unrecorded power, and such entries are excluded. The data is then scaled down as the parameters are measured in a different range to appropriately give importance to changes in data in samples during training of the network. The Pearson correlation coefficient is determined to check the dominant parameters that affect power for that location. This reduces the computational effort further. Then, dividing the data into training and testing the LSTM network is trained and then tested for different network hyperparameters. The network structure and functioning are further detailed in this section. The ADAM optimizer is used for the training of the LSTM. The loss function, scaling method and Pearson correlation for the multi-input LSTM network are further discussed.

22.2.1 LSTM

LSTM is a particular type of RNN, which has the capability to track long-term dependence [7]. It represents a structure of a chain of blocks called a unit. An LSTM unit normally consists of a cell and gates, i.e. input, output and forget gates as illustrated in Fig. 22.2. X_T is the input vector. C and H stand for memory and output, respectively, and the subscripted notation corresponding to current cell T and previous cell $T - 1$.

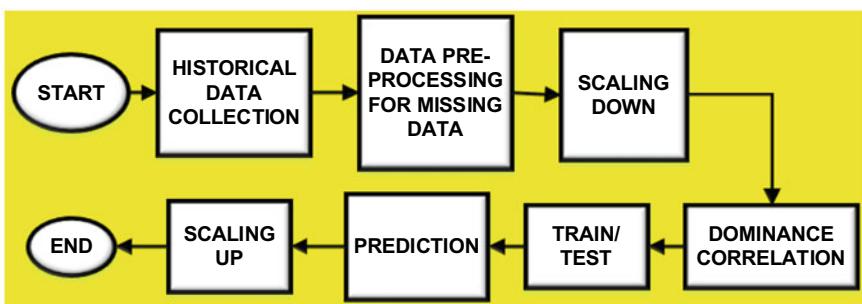


Fig. 22.1 Schematic of the forecast

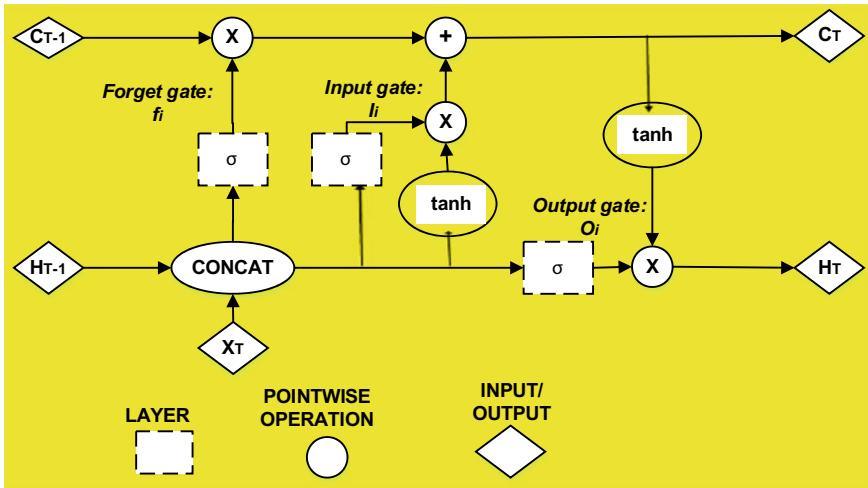


Fig. 22.2 Building block of an LSTM network

The cell keeps track of the dependencies of the input sequence. The input and output gates control the flow of the new input sequence and the level to which the memory in the cell is utilized to determine the activation of the LSTM units, respectively. The forget gate controls what part of the input data sequence remains in the cell. Using various optimization algorithms like gradient descent or ADAM, etc., the LSTM units can be trained to calculate the change in the weights of the network in proportion with the derivative of errors with respect to the existing weights. It is trained in a supervised fashion. In this study, the ADAM optimizer has been used because it is easy to implement, it is quite computationally efficient and works well with large datasets and large parameters.

Various hyperparameters of the network like training epochs, batch size, lookback and which optimizer to choose to need proper tuning and modification for the LSTM to process the data efficiently and without too much overfitting.

22.2.2 Multiple-Input LSTM

LSTMs combine simple deep neural network architectures with clever mechanisms to learn what parts of history to ‘remember’ and what to ‘forget’ over long periods. Further, for multivariate inputs for time-series forecasting, it is very easy to give the inputs. The LSTM learns patterns of data over long sequences and makes good time-series forecasting. Here, multiple inputs have been taken, i.e. irradiation, rainfall, relative humidity, temperature, etc., to predict the solar power.

22.2.3 ADAM Optimizer

Adaptive moment estimation (ADAM) is a replacement of the classical stochastic gradient descent method for iterative weight update as the LSTM is trained with a dataset. The adaptation improves training [7].

22.2.4 Error Functions

In most of the NNs, error functions/loss functions are defined based on error that is the subtraction of the predicted result and the actual result. The error transmits back to front from the output through inner layers, and the weights and biases are adjusted so that the error is decreased to a minimum. The gradient of the loss function is calculated by optimization functions with respect to weights, and the weights are modified until the desired accuracy is achieved iteratively. Adaptive moment estimation (ADAM) is an optimization algorithm useful for noisy data that updates network weights iteratively to optimal with adjustment of learning rate. It is used here for training the LSTM. The mean squared error (MSE) loss function is used in this paper to progress the training. The rest two discussed here are used as performance indicators.

The MSE averages the squares of the errors of n samples in training.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - Y'_i)^2 \quad (22.1)$$

where Y_i = True output.

And Y'_i = Predicted value.

The MSE is indicative of the quality of the training, the nearer to zero the better. It is always non-negative.

The root-mean-square error (RMSE) aggregates the magnitudes of the errors in predictions into a single measure. RMSE is sensitive to outliers or missing data as the square root of such error included in averaging dominates.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - Y'_i)^2} \quad (22.2)$$

RMSE is always positive, and an accurate fitting is indicated by 0. For the index to be comparable, the scale of the data must be in the similar range.

The mean absolute error (MAE) is a positive value that computes the average of magnitude of difference in the forecast for samples. It is a measure of the average accumulation of forecast error. The lower the value the better.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_i - Y'_i| \quad (22.3)$$

22.2.5 Scaling

Deep learning NN models map the relation of input variables to an output variable. Input variables have different scales according to magnitude. These differences in scales in the input increase the complexity of the problem, reduce sensitivity to low scale input values and conclude to the large error function. So, the data is normalized, and all the values are brought down to a range of 0 and 1.

$$X_{\text{Scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (22.4)$$

where x = input to be scaled, x_{\min} = lowest value of x and x_{\max} = highest value of x .

22.2.6 Pearson Correlation

It computes the linear correlation r between two variables x and Y the PV power expressed in $[-1, +1]$. Its magnitude indicates the dominance, and the sign indicates a direct or inverse relation. It is a bivariate correlation.

$$r = \frac{N \sum xY - \sum x \sum Y}{\sqrt{\left[N \sum x^2 - (\sum x)^2 \right] \left[N \sum Y^2 - (\sum Y)^2 \right]}} \quad (22.5)$$

Here, N counts all the samples (Table 22.1).

Table 22.1 Raw inputs acquired from MERRA 2 and NREL

S. No	UT Time (min)	Temperature (F)	Wind Direction (degree)	Rainfall (cm)	Shortwave irradiation (W/m ²)	Pressure (bar)	Wind speed (km/s)	Power (MW)
0	10	299.72	247.74	9.8033	1.8492	974.60	1.83	0.31
1	20	299.77	249.00	10.1930	3.0531	974.70	1.88	0.54
2	30	299.83	250.20	10.5827	4.2679	974.80	1.93	0.72
3	40	299.90	251.93	10.9342	5.6703	974.89	2.00	0.79
4	50	300.00	254.09	11.2475	7.3762	974.98	2.09	0.90

22.3 Results

The data has been taken from MERRA 2 (Modern-Era Retrospective analysis for Research and Applications, Version 2). The data is acquired by NASA and has a resolution of 10 min. These are available since January 1980 and are regularly updated with approximately one month of delay. The PV generation data samples with the 10-min intervals for 1 year (2006) have been collected, with other variables temperature, relative humidity, pressure, wind direction, rainfall and shortwave irradiation of a 25 MW PV power plant, New York (41.65 N latitude and -73.35 W longitudes). It has a total of 52,540 values. The data was divided into training (95%, 49,913 values) and testing (5%, 2627 values). Then, scaling was done on the raw data. However, the correlation with temporal information such as date, month and time has not been considered. To forecast the PV generation, the proposed multi-input LSTM network is used. The training was first done with seven inputs. Readings with 20 and 50 epochs, respectively, have been shown in Table 22.2.

To further simplify the network, decrease prediction time and training time, the dominant input among the four inputs was found. This was done by finding the correlation of all inputs with each other. To find the correlation, Eq. (22.5) is used. The power has the highest correlation with power, shortwave irradiation, relative humidity and temperature. Later on, referring to Fig. 22.3 (correlation of power with different parameters as the last set of bars), the weakest correlation such as wind direction, pressure and rainfall was dropped, the training was then done with

Table 22.2 Errors of one layer with seven inputs LSTM network

Epoch	Layer	Neuron/layer	Loss function	MSE	RMSE	MAE
20	1	50	MSE	3.886×10^{-5}	0.00623	0.004423
50	1	50	MSE	2.0304×10^{-5}	0.00450	0.003464

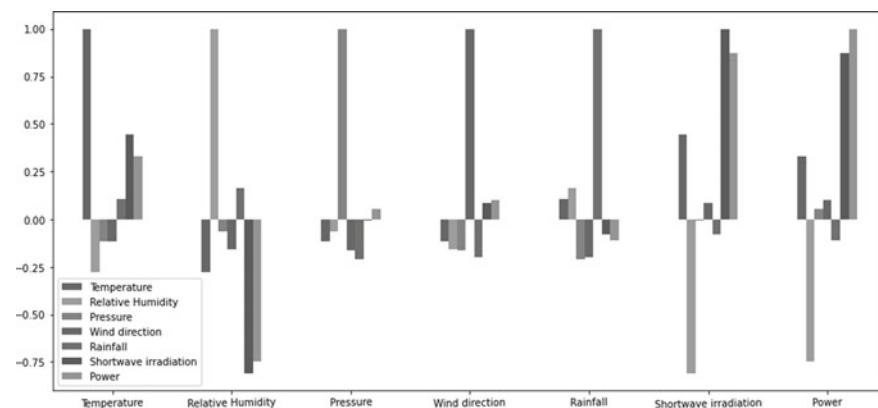


Fig. 22.3 Correlation of power with other parameters

four inputs as shown in Table 22.3, and better results were found, i.e. loss function (MSE) was minimized. Here, four inputs, namely temperature, relative humidity, shortwave irradiation and power, are taken. For the single layer, the increase of epoch shows a reduction of the loss function. 12% and 24% decrease in MSE for the increase in epochs from 20–50 to 50–60, respectively. Increasing epochs have shown improvement but with an increase in run time (290 μ s for 20 epoch to 300 μ s/sample for 60 epoch). As for the short-term forecast, time is important, and the selection of 50 epochs would be a good choice [7]. This will help the network to be re-trained and checked in the future after gathering historical measurements in future. Surprisingly, with an increase in layers, the LSTM performance degraded by 10 times. Tables 22.2 and 22.3 show a comparative view of various error indexes of the LSTM used along with the magnitude and loss function of the error. Dropping the non-dominant inputs reduced the time as well as errors for the same network.

It is well known that if the number of epochs is increased then it may lead to overfitting of the data. Overfitting of data results in high accuracy in training data and very low accuracy in testing data. But overfitting was not observed in our LSTM model. The convergence of loss function for training is shown in Fig. 22.4. The MSE has less decrement after 20 epochs.

From the above results, the best model was applied to forecast power, and the results were promising. The 10 min ahead predicted power graph is shown in Fig. 22.5 for January after the training for 10 min sampled data on a yearly data with the best

Table 22.3 Errors of four inputs LSTM network

Epoch	Layer	Neuron/layer	MSE	RMSE	MAE
20	1	50	3.912×10^{-6}	0.00185	0.00110
50	1	50	3.450×10^{-6}	0.00185	0.00101
60	1	50	2.619×10^{-6}	0.00161	0.00090
20	2	50	8.155×10^{-6}	0.00285	0.00225
50	2	50	1.657×10^{-5}	0.00407	0.00232

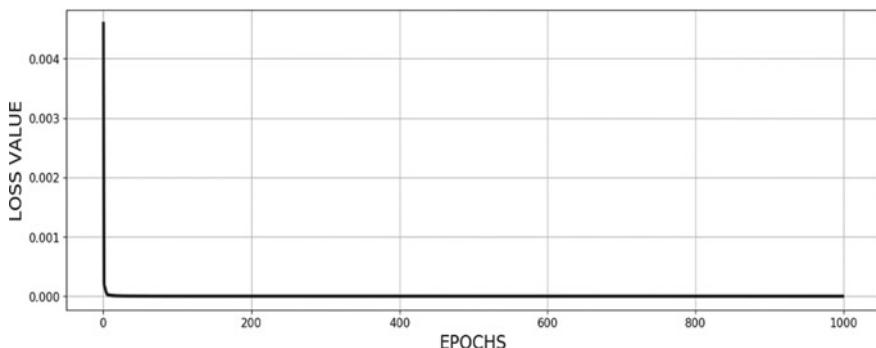


Fig. 22.4 Change of MSE loss function values with epochs

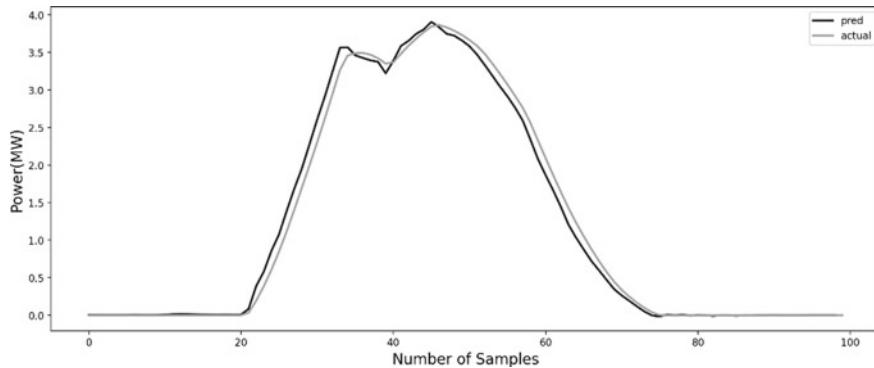


Fig. 22.5 Power prediction graph for a day in January 2006

network derived. The accuracy can be observed. Similarly, for eight consecutive days with different weather conditions, the power prediction is shown in Fig. 22.6, where the consistency of performance in testing is observed. The RMSE for the testing phase for a day in the forecast is found to be 0.0016, and it is better than a recent work for a similar short time frame where the value was 0.0383 [12].

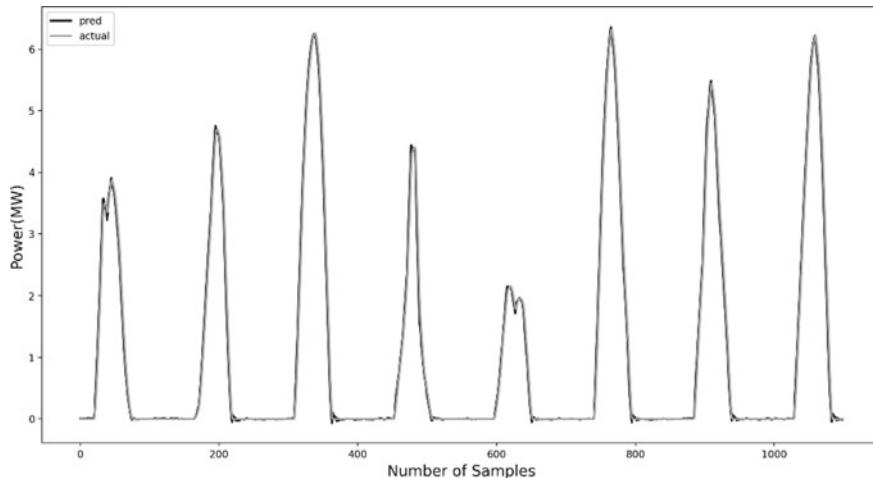


Fig. 22.6 Power prediction graph for eight consecutive days

22.4 Conclusion

In this work, multiple-input LSTM has been used for PV power prediction based on satellite data. After rigorous training and testing, it is found that the multiple-input LSTM network with a single layer, 50 neurons yields the best result. It is also observed that if pressure, wind direction and rainfall are removed, which are among the non-dominant inputs, then better accuracy in the reduced time is achieved. Also, it is found that a higher number of epochs beyond 50 increase the training time though there is an effect on the error, so 50 epochs is a good option. At last, the trained network is used to predict the power generated, and the results are plotted. It is found that for the best results only four inputs, i.e. temperature, shortwave radiation, relative humidity and power, were needed.

References

1. Wan C., Zhao J., Song Y., Xu Z., Lin J., Hu Z.: Photovoltaic and solar power forecasting for smart grid energy management. *CSEE Jou. Pow. Energy Syst.* **1**(4), 38–46, (2015)
2. Huang, Y., Lu, J., Liu, C., Xu, X., Wang, W., Zhou X.: Comparative study of power forecasting methods for PV stations. In: *IEEE International Conference on Power System Technology*, pp. 1–6 (2010)
3. Benali, L., Notton, G., Fouilloy, A., Voyant, C., Dizene, R.: Solar radiation forecasting using artificial neural network and random forest methods: application to normal beam, horizontal diffuse and global components. *Renew. Energy* **132**, 871–884 (2019)
4. Tao, C., Shanxu, D., Changsong, C.: Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement. In the 2nd IEEE Int. Symposium on Power Electronics for Distributed Generation Systems, pp. 773–777 (2010)
5. Bacher, P., Madsen, H., Nielsen, H.A.: Online short-term solar power forecasting. *Sol. Energy* **83**(10), 1772–1783 (2009)
6. Azadeh, A., Maghsoudi, A., Sohrabkhani, S.: An integrated artificial neural networks approach for predicting global radiation. *Energy Convers. Manage.* **50**(6), 1497–1505 (2009)
7. Qing, X., Niu, Y.: Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy* **148**, 461–468 (2018)
8. Kudo, M., Takeuchi, A., Nozaki, Y., Endo, H., Sumita, J.: Forecasting electric power generation in a photovoltaic power system for an energy network. *Elect. Engg. Jpn* **167**(4), 16–23 (2009)
9. Yona, A., Senju, T., Saber, A.Y., Funabashi, T., Sekine, H., Kim, C.H.: Application of neural network to 24-hour-ahead generating power forecasting for PV system. In: *2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century*, pp. 1–6 (2008)
10. Sansa, I., Missaoui, S., Boussada, Z., Bellaaj, N.M., Ahmed, E.M., Orabi, M.: PV power forecasting using different artificial neural networks strategies. In *IEEE 1st International Conference on Green Energy ICGE 2014*, pp. 54–59 (2014)
11. Lee, W., Kim, K., Park, J., Kim, J., Kim, Y.: Forecasting solar power using long-short term memory and convolutional neural networks. *IEEE Access* **6**, 73068–73080 (2018)
12. Behera, M.K., Nayak, N.: A comparative study on short-term PV power forecasting using decomposition based optimized extreme learning machine algorithm. *Eng. Sci. Technol. Int. J.* **23**(1), 156–167 (2020)

Chapter 23

I Hardly Lie: A Multistage Fake News Detection System



Suchintan Mishra, Harshit Raj Sinha, Tushar Mitra, and Manadeepa Sahoo

Abstract Over the recent years, fake news detection has been a topic of constant question and research among the society of cognitive engineers and many cynical enthusiasts in the field. With social media at easy disposal, it becomes but imperative to determine the factual truthfulness of the information. Rampant fake news sellers are a menace in society, threatening the very base of democracy and the freedom of speech. In this paper, we try to provide a brief understanding of the current methodologies in fake news detection, along with proposing a multistage fake news detection system. All the phases that have been designed are ordered in increasing order of complexity and efficiency. If fake news detection does not provide results with an acceptable level of accuracy or satisfactory results at any phase, we have designed for a subsequent phase up in the hierarchy which is assured to provide better accuracy.

23.1 Introduction

In the present day, information has become the most powerful tool to deal with any problem around us. Be it any political issue or any matter of public interest. The correct decision can only be taken only by the aid of correct information. However, with technological advancement in every field of our daily lives, misinformation has become almost an irreplaceable aspect. Misinformation is so scrupulously curated and disseminated amidst common people along with any other true information that it has become almost impossible to differentiate between the two. There is a serious need to strategically combat the menace caused due to fake news, since it can erode law and order in society. There have been many novel attempts and significant advancements in this field that aim to detect fake news based on its sources of

S. Mishra · H. R. Sinha (✉) · T. Mitra · M. Sahoo

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan Deemed To Be University, Bhubaneswar, Odisha, India

S. Mishra

e-mail: suchintanmishra@soa.ac.in

generation and means of propagation amidst the common public or even at the content level of the news. This paper presents a multistage fake news detection approach.

23.2 Problem Definition and Objectives

The easy access to social media has opened wider doors for widespread intentionally fabricated news. The very recent trends also suggest news channels fueling up these unchecked facts for easy points and climbing high in the ladder of competition [1]. Thus, it is extremely essential for us as a consumer to evaluate the news that is being circulated widely and then believe on that. Hence, that is why we have tried to develop a multistage fake news detection system that tries to add up to the momentum that this field of development is constantly gaining. The multistage fake news detection system works in solving the issue of detecting fake news in phases. All the phases are sequentially arranged in increasing order of complexity and improved efficiency. If the result at any stage is not satisfactory, there is always a provision to test it up higher in the next level [2].

23.3 Existing Systems

Current automatic fake news detection systems can be broadly classified into two broad categories: content-based methods and propagation-based methods. A majority of the existing work in the field of content-based fake news detection is majorly based upon extracting features from multiple datasets. They primarily focus upon understanding the nature of information on a much granular level:

- (a) Understanding the nature of the news, that is, application of natural language processing techniques.
- (b) Reliability and trustworthiness of the news source.
- (c) Information extracted from trusted social networks [2].

Existing approaches have traditionally been simulated over many state-of-the-art classification algorithms such as k-nearest neighbors (k-NN), random forests (RF), support vector machine with RBF kernel (SVM), and XGBoost (XGB). Current day automatic fake news detection provides a considerable amount of accuracy with only 40% of false positives [3]. The output could be considered as a numeric score of truthfulness. Any news could be just classified as true or fake; however, upon perceiving intently, many news pieces are a mix of true or false claims. The presence of such news in datasets, clearly mean higher chances of misclassification, hence lower accuracy [2, 4]. Models relying on social context for fake news detection find it often very challenging to mine fake news before it has reached its third stage. By this time, we lose much time and usually, it becomes quite late in fake news detection in such cases. In such cases, the sooner the better; because the more time lost, more

people are at a chance of getting exposed to fake news. Also, once any information gets appealed to the human cognition, it becomes next to impossible to discard it as false information, a phenomenon quoted as *Semmelweis reflex*.

23.4 Proposed System

The multistage fake news detection system works in three phases. As we walk you through all three stages simultaneously, we also are constantly trying to work upon the entire system of fake news detection to enhance the accuracy.

Stage 1: Check Your News

This stage is as easy and simplified as it is named. We check the input from the user from the data available online. This progresses in a Greedy Fashion, searching and finding matches for available news, online over a list of verified sources of news. We owe credits to news firms for using their website for web scraping and using the news available on their website. We sequentially extract data from each of these websites and then use Natural Language Processing techniques such as tokenization, stemming and lemmatization followed by vectorization to pre-process the text data that we have newly extracted. This text data is matched with the news input by the user. This is done using the text matching technique and a record of precision is kept in a note. That is total true matching found. At the end of it, a percentage of precision is reported. Firstly, a first-hand method to check the correctness of news with the ease of noncomplex algorithm assuring high accuracy since any recent news which we are checking would be easily caught, because it must be over the internet. Secondly, this stage saves us the task of performing a news verification job. It is so because we have used very reliable sources which use extensive research and verification methods for the data collection on their part.

The first step is the tokenization of the collected data. Let us try to understand the term tokenization very naively. It could be either sentence-level tokenization or word-level tokenization [5]. We have used word tokenization which is basically a type of white space tokenization. The second step is removal of stop words. Stop words are common terms or results of the process of tokenization that usually contain no special information. In language, they are basically terms that do not add any extra information about the sentence from the NLP point of view. These terms must be ignored or else they considerably add up to the size of our database. We have used the standard list of stop words provided by the *nltk* library in Python for the English language. After removal of stop words, we now have a filtered list of only necessary tokens to be used for further usage [6, 7]. The third step is the vectorization of each of the news. Text vectorization refers to conversion of text data into numerical data. We have used the *Normalized Tf-IDF vectorization* technique to convert the text into vectors. *Tf*-refers to the term frequency. It is the number of times a particular term has occurred in the news. *IDF* refers to the inverse document frequency. Document frequency (DF) is the total number of news containing a particular term t. In order to

calculate inverse document frequency, we divide the total news with the document frequency. Thus, the document frequency for a particular term can be shown as in Eq. (23.1).

$$\text{IDF} = \log(N/\text{df}(t)) \quad (23.1)$$

Now, the Tf-IDF score for a news is: $\text{TF} * \text{IDF}$. The normalized TF-IDF score is the unit vector of Tf-IDF vector. After vectorization, we go for similarity checking. The similarity measure used here is the *cosine similarity*. For two normalized vectors, the cosine similarity is defined the dot product of the two vectors [8]. We only consider the news for which the cosine similarity is greater than equal to 15%. This is an experimental value and is found that the set of news filtered after this is substantially good in number and is highly related to our test data. A list of all such news is created and stored. Following the step of finding the cosine similarity, we proceed further toward the concluding steps of this stage. For each of these news that we have made a list, we convert each of these news into its root form. Technically this process is termed as lemmatization. In linguistics, *Lemmatization* refers to converting words into their root form [6]. Many studies involving text retrieval suggest lemmatization to yield more relevant documents than stemming [7].

The final step involves parts of speech tagging. In linguistics, *Part of Speech Tagging* refers to marking tokens according to what part of speech does it belong vis-à-vis noun, pronoun, verb, adverb etc. With this, we have created a weightage array which assigns numerical values to different parts of speech depending upon their relevance in a sentence in offering meaning. For instance, nouns have larger weightage and tell more about a context of a sentence than preposition or pronoun. Thus, they must be accompanied by a weightage in order to represent the relative importance of each matching [8]. After this, we go for negative statement checking. That is, now that we know that we have a list of news that are relevant for our test news. We must now check if they convey a similar meaning or something totally different. That is whether both the news, one is our test news and the other one from the list that we created from cosine similarity, convey a similar meaning or both are absolutely contradictory in nature. We use an XOR operation for the purpose and negate the results. That is, if both the statements are affirmative and negative, then the result is true. In other case, the result is false, and that we can say there is absolutely no match. After calculations from all these stages, we find a numerical value which is a comparison percentage. The comparison percentage is our final result of stage 1. It is a precision measure, which is a ratio of total matches found between the test news and the list of news created by the total relevant news compared. The threshold value for this stage has been set to 55%.

Any news generating a value beyond 55% can be marked as a probably true news and below as may or may not be false; the verification of which could be done in stage 2. Along with that, we also display the top news that have the highest similarity

with the test news. This not only gives the user a proof of accuracy but also suggests the points of dissimilarity with the test news and the actual published news.

Stage 2: Walk Over What's Available

As stated earlier we go on to the next stage only when we don't receive any satisfactory results in the previous stages. So we shall first discuss what could be the possible reason to move ahead in stage 2. The one basic reason is if the precision in stage 1 didn't pass the threshold value. At that point, we are not at a stage to dismiss the news as fake. It is because of another possibility that the news we are testing is not present in our search space. In essence, the test news is considerably old. Also, some news is based upon some historical facts, or news published considerably back in time. Since older news is not checked in stage 1, it is a possibility that the lower accuracy is due to a lack of relevant news to be matched. This requires the news to be matched with a wider collection of news that spans over a substantial amount of time. Much like a database of news, that contains a large but quantifiable collection of news that uses a finite amount of time to go through. Thus Stage 2 uses a database of news using MySQL. We insert, update and delete data into the table. The target news is checked with news in the table in this stage. Again the similar data pre-processing techniques as in Stage 1 are applied and matching is found. If considerable accuracy is found, we can say that the news is true, else this means that the news is very old and needs to be checked in Stage 3. The threshold value for stage 2 is set at 40% of accuracy. Any news producing a value of accuracy of 40% or more can be dismissed as probably true. In any other case, it can be said that the news was fake.

Stage 3: Train, Test and Evaluate

As the name suggests, Stage 3 uses Model training to classify whether the news is true or fake. The type of models used is supervised in nature because of a fully labelled dataset. We have used binary classification algorithms here to do that. Let us understand the dataset used next. The dataset that we have used is based on the US Presidential election news dataset. The entire dataset is divided into two sub-datasets labelled as True.csv and Fake.csv containing true and fake news respectively. Each of these datasets contained columns labelled as Title (the title for the news), Text (the body of the news) Subject that is News, and the Date of Publishing. In both of the news datasets, a new column was introduced that each bearing binary-valued 0 for fake news and 1 for true news. The dataset was uploaded in CSV format. Since it is a fully labelled dataset, the input and output variables were separated, after which the train test split was performed. It is noteworthy here that 80% of the dataset was reserved for the training of the model and the rest 20% for testing purposes. The best results were found for the 80–20% split after trying with various combinations. Now we shall go through the models used in this particular stage in details. A *decision tree* algorithm involves segmenting the predictor space into several simpler regions. Decision trees can be applied to both regression and classification problems and are sometimes called the *CART Algorithm (Classification and Regression Tree)*. In a classification type problem, each of these segments is assigned different class labels. For a classification tree, we predict that each observation belongs to the most

commonly occurring class of training observations in the region to which it belongs. A decision tree grows by recursive binary splitting. However unlike the Regression tree which uses Residual Sum of Squares or RSS as a criterion for binary splitting, here it is not helpful since we have class labels as output variables. A natural alternative to RSS is the classification error rate which has been mathematically represented in Eq. (23.2). This is simply the fraction of the training observations in that region that do not belong to the most common class

$$E = 1 - \max(p_{mk}) \quad (23.2)$$

Here, p_{mk} represents the proportion of training observations in the m th region that are from the k th class. However, classification error is not sufficiently sensitive for tree-growing, and in practice, two other measures are preferable. The Gini index is defined by a measure of total variance across the K classes, which has been mathematically shown in Eq. (23.3).

$$G = \sum p_{mk}(1 - p_{mk}) \quad (23.3)$$

The Gini index takes on a small value if all of the p_{mk} 's are close to zero or one. For this reason, the Gini index is referred to as a measure of node purity—a small value indicates that a node contains predominantly observations from a single class. An alternative to Gini Index is cross entropy, and both are almost same. Before proceeding further, let us understand what ensemble learning is. It is a technique in which multiple weak learners are trained simultaneously in order to produce a single strong learner so as to enhance the accuracy of prediction. They are primarily of three types: bagging or bootstrap aggregating, boosting, and stacking. The second supervised classification algorithm that we use is *Random Forest Classifier* which is a specific type of ensemble learning. Unlike decision trees, since random forests do not sample over the same features, rather they split on a small subset of features; the final outcomes have very little correlation with them. In standard bagging type situations, the data split is such that tow-third of the data is kept for bootstrapping and the rest one-third is used for testing. However, for the one particular dataset, we used similar 80–20% ratio of training and testing data gave the best results [7]. The next supervised model used is *XGBoost*. It is a method that goes through cycles iteratively to add models into an ensemble. It begins by initializing the ensemble by a weak learner or a base model whose predictions are very naïve. With subsequent iterations of the algorithm, the errors are addressed. The gradient in XGBoost stands for gradient descent which is used in the loss function to determine the parameters. XGBoost has a number of parameters that can substantially alter the accuracy of prediction such as *n_estimators*, *early_stopping_rounds*, and *learning_rate*. This ensures the XGBoost model predicts with higher accuracy. The value we used was equal to 0.05. The final supervised model we use is *Logistic Regression*. It is much similar to linear regression; however, the cost function used here is much more complex.

The accuracy of first two stages are totally dependent upon the threshold values. These threshold values were fixed at 55% for stage 1 and 40% for stage 2 depending upon the accuracy of the stages. The threshold values are experimental and have been derived after testing upon a number of already tagged true and fake news. For the first stage, the average accuracy was around 83% for threshold value 55%. The accuracy fell for any other threshold values substantially less or greater than 55%. Similarly, for stage 2, the average accuracy was 79% for the threshold value fixed at 40%. In Stage 3, we have used several accuracy tests to determine the best model of the four supervised learning algorithms used. First, we plotted a confusion matrix for each of the models. A confusion matrix is a measurement for accuracy for classification type algorithms. As the name suggests, it is a matrix containing values. For a typical binary classification type problem, a confusion matrix contains four values that are true positive (TP), false positive (FP), true negative (TN), and false negative (FN) [9]. We have tried to graphically show the outcomes of each confusion matrix. The graphical confusion matrix for decision tree classifier has been shown in Figs. 23.1, 23.2, 23.3, and 23.4 for P -values (0%, 33%, 66%, 100%), respectively.

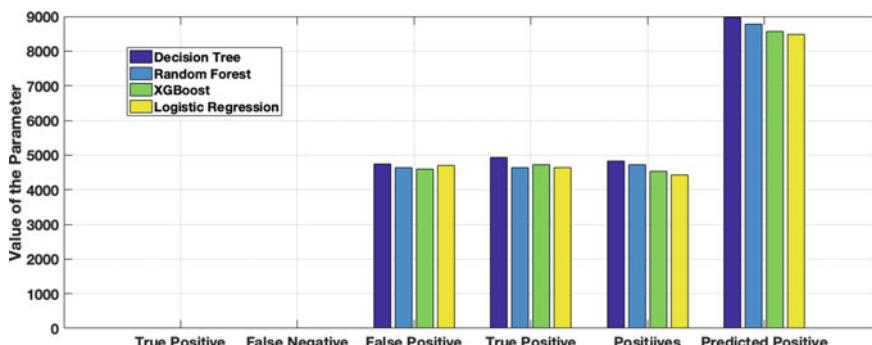


Fig. 23.1 Graphical confusion matrix for decision tree at P -value 0%

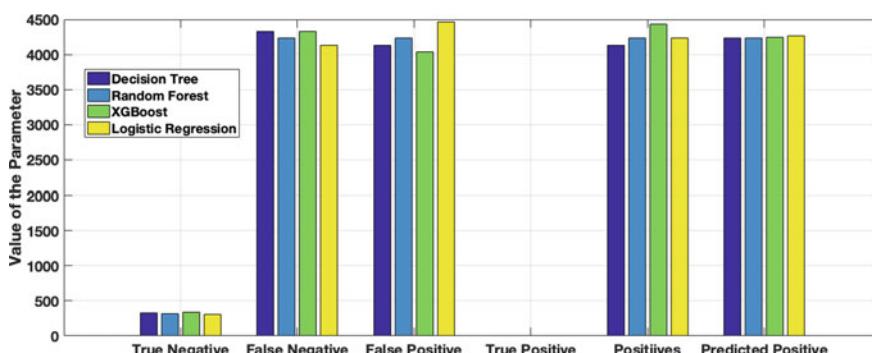


Fig. 23.2 Graphical confusion matrix for decision tree at P -value 33%

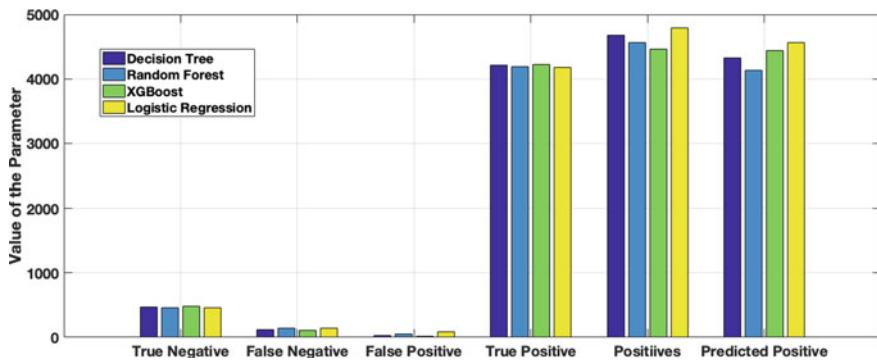


Fig. 23.3 Graphical confusion matrix for decision tree at P -value 66%

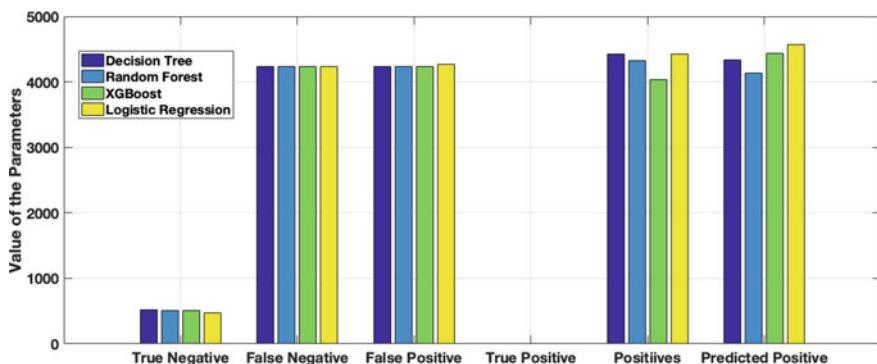


Fig. 23.4 Graphical confusion matrix for decision tree at P -value 100%

23.5 Conclusion and Future Scope

The division of the entire system into multiple stages provides fast way to identify fake news helps with making the whole process fast. Many new and recent news might not need going to stages 2 and 3 of the system because such news can be easily detected in stage 1. Stage 1 of the multistage system provides a real-time mechanism to detect fake news. It checks current news online to tell whether the target news is fake or real. The simultaneous display of top similar news on the Internet also adds to the credibility of Stage 1. The multistage fake news detection system considers the entire news as one and searches for similar news online from verified sources. These are the news sources which have been pre-tagged as authentic and are reputed as well as authorized national level media houses. Stage 2 provides an extensive methodology to verify considerably older news. The third stage uses machine learning techniques to detect fake news. The highest accuracy received was 99.7% for XGBoost algorithm.

References

1. Zhou, X., Jain, A., Phoha, V.V., Zafarani, R.: Fake news early detection: a theory-driven model. *Dig. Threats Res. Pract.* **1**(2), 25. Article 12 (2020). <https://doi.org/10.1145/3377478>
2. Reis, J.C.S., Correia, A., Murai, F., Veloso, A., Benevenuto, F.: Supervised learning for fake news detection. *IEEE Intell. Syst.* **34**(2), 76–81 (2019). <https://doi.org/10.1109/MIS.2019.2899143>
3. Oshikawa, R., Qian, J.: A survey on natural language processing for fake news detection. ArXiv preprint arXiv **1811**, 00770 (2018)
4. Wang, W.Y.: Liar, Liar pants on fire: a new benchmark dataset for fake news detection (2017). arXiv preprint [arXiv:1705.00648](https://arxiv.org/abs/1705.00648)
5. Traylor, T., Straub, J., Snell, N.: Classifying fake news articles using natural language processing to identify in-article attribution as a supervised learning estimator. In: 2019 IEEE 13th International Conference on Semantic Computing (ICSC) (pp. 445–449). IEEE (2019)
6. Saquete, E., Tomás, D., Moreda, P., Martínez-Barco, P., Palomar, M.: Fighting post-truth using natural language processing: a review and open challenges. *Exp. Syst. Appl.* **141**, 112943 (2020)
7. Balakrishnan, V., Yemoh, L., Ethel: Stemming and lemmatization: a comparison of retrieval performances. In: Proceedings of SCEI Seoul Conferences, 10–11 Apr 2014, Seoul, Korea (2014)
8. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
9. Chen, Z., Freire, J.: Proactive discovery of fake news domains from real-time social media feeds. In: Companion Proceedings of the Web Conference 2020, pp. 584–592 (2020)

Chapter 24

Software Design for Mobile Phone Auto-Silencer



Lambodar Jena and Dipti Pratima Minz

Abstract It is probably human instincts to overlook certain minor necessities, claiming to the fact of “busy life”. Mobile phones have become a pivotal point of our lives; however, it can be an inconvenience at times such as during a class, a meeting, a funeral or some other similar situations. The application, called the Phone Auto-Silencer, is an important element in the integration of knowledge and a small contribution for a busy schedule. The Phone Auto-Silencer can assist in setting a specific time interval whether it is tomorrow, any day of a week or a month. The app will automatically put the phones in silence mode for you according to your scheduled time as well as switch it back to its original mode once the particular time interval is over. The application is created using Android Studio, with Java as a frontend and XML as the backend.

24.1 Introduction

Smartphones have become an important asset in our lifestyles due to its multifunctional uses and sophisticated features [1]. Calendars, maps, timetables, cameras, and phone calls are all now computerised in a single, small package that used to be done by various software tools. While your Android phones can alert you of incoming calls, notifications, text messages or important reminders, there are certain circumstances when you need to use audible alerts, and there are certain situations when you need to place your phones in vibrate mode or silent mode [2, 3]. However, as the saying goes, “easier said than done”. Most often we are being told to keep the phones in silent mode or to switch off the phone to avoid disturbances and distractions during some ongoing important assemblage. In such a dilemma, the Phone Auto-Silencer app for Android can be of help. This app will switch your sound profile automatically

L. Jena (✉)

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (K L University), Vaddeswaram, Vijayawada, Andhra Pradesh, India

D. P. Minz

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

based on where you are or what time of the day it is according to the specific time interval scheduled by the user [4]. So, the purpose of this project is the implementation of mobile auto-silent services through Google Web services and Android API [5]. In this, the user simply needs to enter the details of the duration of the required time schedule of when they want their Android phones to be in silent mode.

24.2 Problem Statement

The entire application is designed to provide user flexibility in usability.

- To create a simple minimalistic UI.
- To enable users to get their phones in silent mode to avoid unnecessary interruptions.
- Also enable users to attend their phones when switched back to audible mode automatically.
- To emphasize on attractive icons and fonts to make the simple application elegant.

24.3 Literature Survey

In the existing system of our phones, we have the options to set the phones in silent, vibration or audible mode manually. When it is time for classes or offices, the user needs to check the mode of the phone and switch it to the required mode. Then later, the current mode needs to be switched back to the original.

Disadvantages of existing system:

1. Users have to switch the modes of sound profiles.
2. It requires human interface frequently.
3. It does not switch back to its original mode on its own.
4. It does not support dynamically.

24.4 Proposed System

In our proposed system, user can schedule any number of events for any day. They can also keep the event enabled or disabled for the next same day and time because the event does not get deleted on its own after the scheduled time is over. The purpose of this project is to provide a dynamic system to the user. This app will put the phones in their required mode when chiming or ringing of phones can be a distraction, and then later, it will automatically change back the sound profile to audible mode when the scheduled time is over.

Advantages of proposed system:

- i. It can repeat for series of days and time, so it requires less human interface than the existing system.
- ii. It can automatically switch to the required modes according to the scheduled time.
- iii. It does support dynamically.

To provide convenience over getting the phones in silent mode or audible mode as and when required will help people avoid unnecessary distractions due to chiming or ringing of phones. It will also help the users to get their required notifications and calls when the app will automatically switch back.

24.5 Pros and Cons of the System

Phone Auto-Silencer can help the user schedule any number of events for any day. They can also keep the event enabled or disabled for the next same day and time because the event does not get deleted on its own after the scheduled time is over. The purpose of this project is to provide a dynamic system to the user. This app will put the phones in their required mode when chiming or ringing of phones can be a distraction, then later it will automatically change back the sound profile to audible mode when the scheduled time is over.

Advantages of this application:

- i. It can repeat for series of days and time, so it requires less human interface than the existing system.
- ii. It can automatically switch to the required modes according to the scheduled time.
- iii. It does support dynamically.

Yet, it also has certain limitations such as

- i. The app will not set the specific time interval by itself.
- ii. The user has to take out some time before the scheduled time to set the specific time.

24.6 Requirement Specifications of the System

The requirement of the application is a description of what the system needs to be capable of doing in order to fulfill the project demands. The requirement when combined with the use case diagram is enough to start building a system to represent those needs. In this chapter, we discussed different functional and non-functional requirements along with various UML diagrams.

Functional requirements

1. SCHEDULED TIME: Sets the phone either in silent mode or audible mode according to the scheduled time. Also, the event remains until the user deletes it.
2. ADD EVENT: The user can create any number of events for any number of days and for any day.

Non-functional requirements

The application does not demand from or harm any safety of the user. The application is harmless and will not require any personal information of the user. The application will only set the user's phone in silent/vibration mode as per the settings of the user. It will also get the phone back in audible mode when the specific time interval of the scheduled time is over.

24.7 UML Diagrams

The proposed system is developed by following object-oriented programming (OOP) design approach [6, 7]. The UML diagrams depicted in Figs. 24.1, 24.2 and 24.3 represent the design aspect of the system.

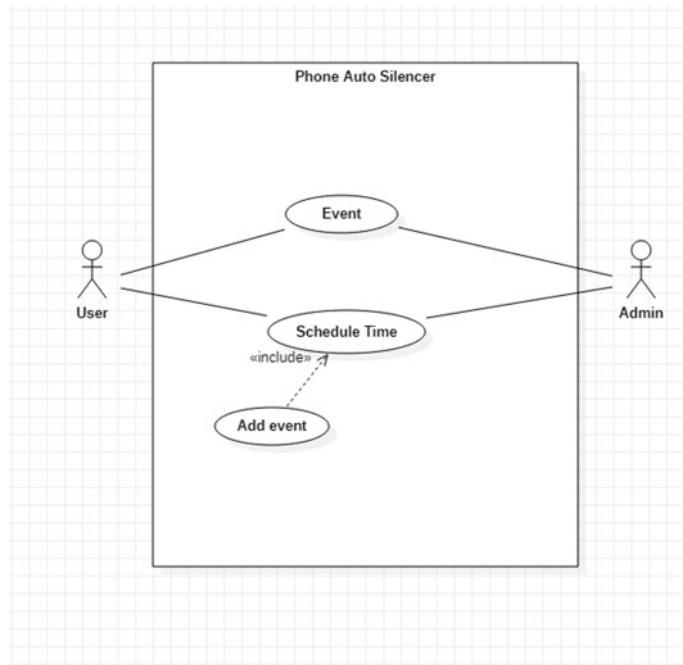


Fig. 24.1 Use case diagram

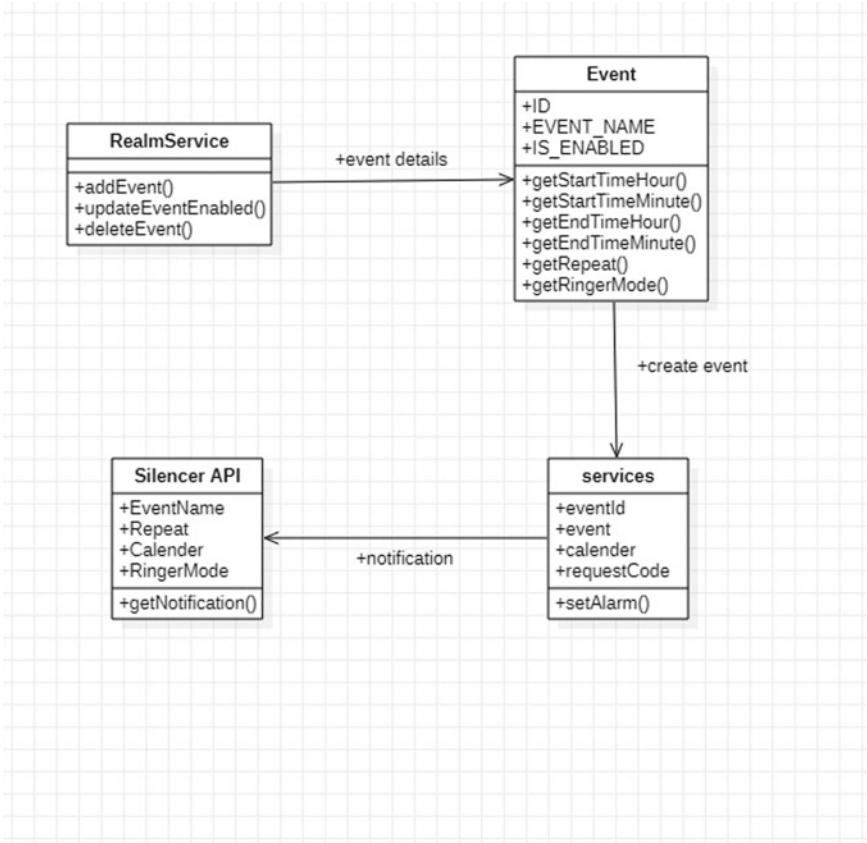


Fig. 24.2 Class diagram

24.8 Results and Discussions

The software has undergone the best possible fine-tuning efficiency, performance and accuracy level. The first step in user interface (UI) designing is to determine how the output is to be produced and in what form. Samples of the input and output are also presented in Figs. 24.4, 24.5 and 24.6. Figure 24.4 represents the user interface, Fig. 24.5 represents the time scheduling for auto-silencer of phone, and Fig. 24.6 represents the event setting for which auto-silencer is required.

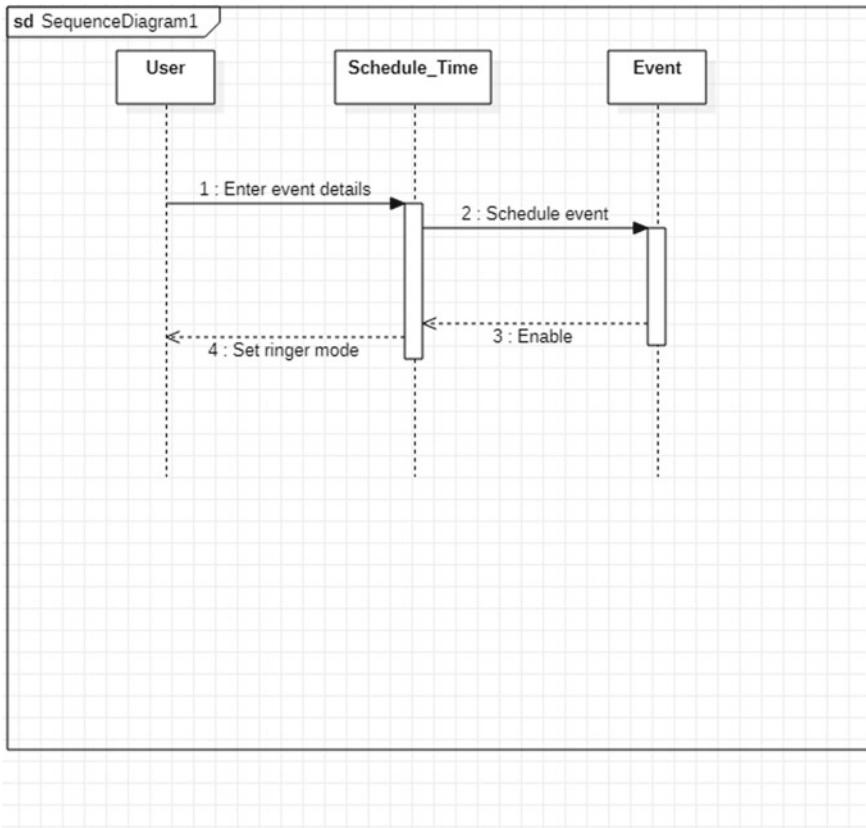


Fig. 24.3 Sequence diagram

24.9 Conclusions and Future Work

This work has been successfully completed, and it is working fine fulfilling the desired functions. The application can be modified further to be more efficient and more human independent, yet it provides the basic functionality of the objective undertaken. The system did perform well among the test audience and fulfilled the requirements of the users. It, according to the analysis, did satisfy the constraints that were on the path of its development. Thus, the phone auto-silencer system does work completely well and functions as per required.

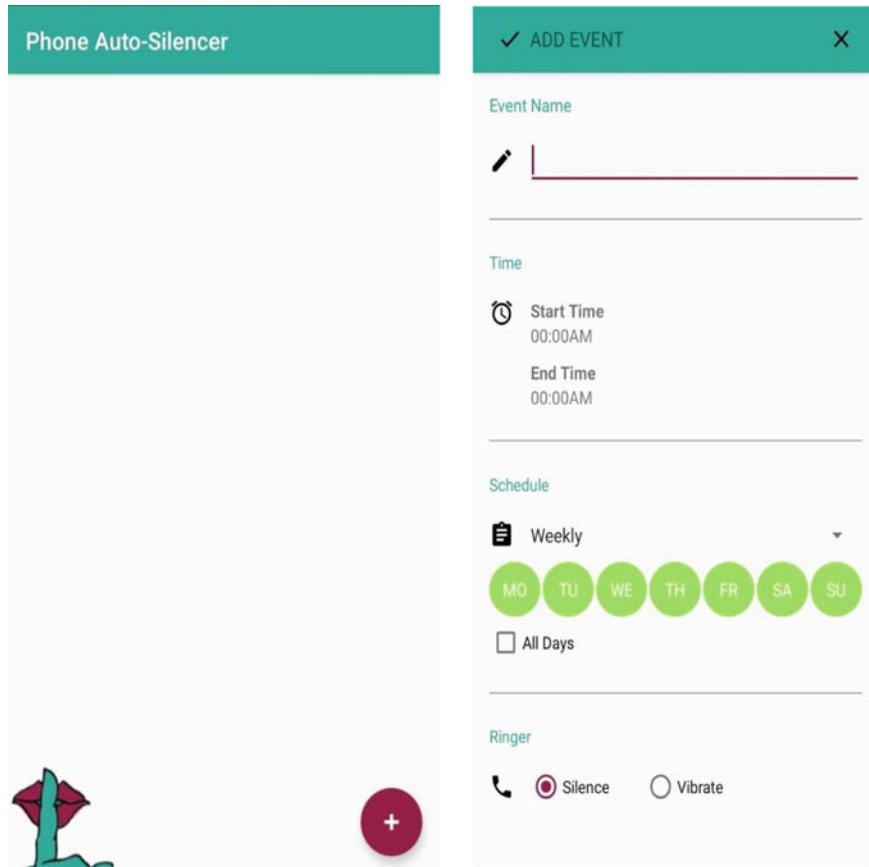


Fig. 24.4 User interface

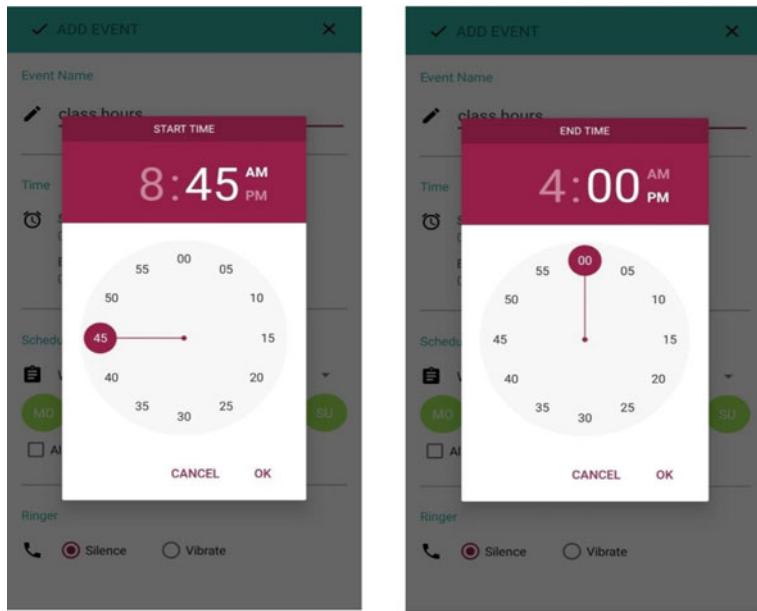


Fig. 24.5 Time schedule

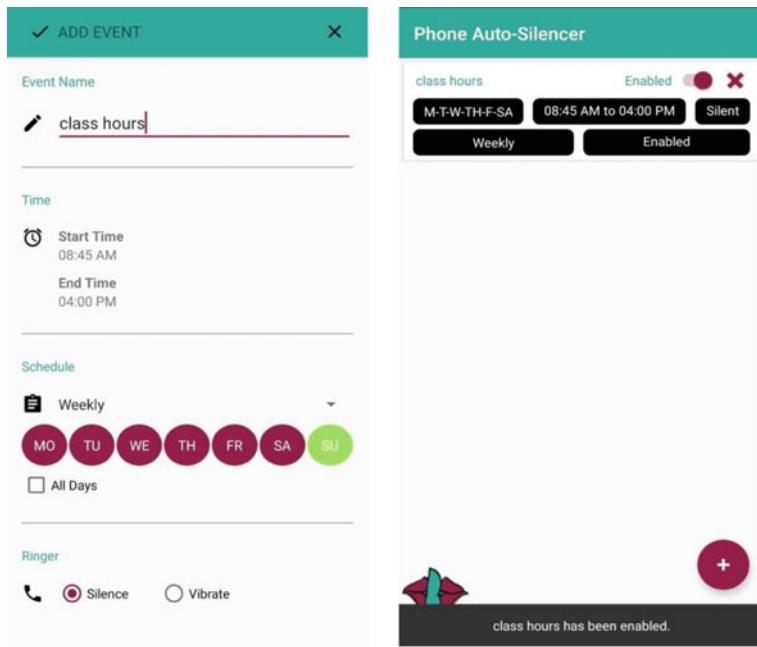


Fig. 24.6 Event for auto-silencer

References

1. Lee, R., Kathryn, Z.: Americans' views on mobile etiquette. Retrieved on September 10, 2015 (2015)
2. Sheldon, K.M.: Creativity and selfdetermination in personality (1995)
3. Kore Geetanjali, T., Naik Suman, R., Sanap Kavita, V., Mawale Ashwini, J.: Automobile silent system (2014)
4. <https://stackoverflow.com/questions/38260175/android-onrequestpermissionsresult-grantresulits-size-1>
5. <https://stackoverflow.com/questions/6220354/android-two-imageviews-side-by-side>
6. Genero, M., Fernández-Saez, A.M., Nelson, H.J., Poels, G., Piattini, M.: Research review: a systematic literature review on the quality of UML models. *J. Database Manag. (JDM)* **22**, 46–70 (2011)
7. Verma, P.: Effect of different UML diagrams to evaluate the size metrics for different software projects. *Glob. J. Comput. Sci. Technol. Softw. Eng.* **15**(8). Version 1.0 (2015)
8. <http://www.flaticon.com>

Chapter 25

Traffic Flow Prediction: An Approach for Traffic Management



Utkarsh Jha, Lubesh Kumar Behera, Somnath Mandal, and Pratik Dutta

Abstract Transportation has always been a boon to humankind, and with the advent of motorized vehicles, humans can travel quicker and faster, but with the new solution arose new problems, these problems are usually caused by the bottlenecking of roads. These roads are not able to accommodate a huge number of vehicles; this, in turn, causes the whole transportation system sometimes to collapse. This issue is tried to be avoided, using a method called ‘Time-Series Forecasting’, using historic data and two statistical models namely, ARIMA and regression. These algorithms will be trained according to the data provided. The trained model could now be used, to predict the traffic conditions of a spot at a particular time. Comparing with ARIMA and regression model, ARIMA always becomes cost-effective and sometimes unable to set up in terminal stations. In this paper, the dataset has been restructured in such a way that the linear regression model itself be able to reach satisfactory performance.

25.1 Introduction

The traffic control center (TCC) is the nerve center of Germany’s traffic control system. In addition to being linked to data collection systems, they are also utilized to control variable message sign systems and prepare broadcast messages. In order for road users to receive accurate travel time forecasts from traffic information systems, forecasting road traffic is a critical part of traffic management and traffic management. Increased traffic demand prompted the addition of advanced traffic monitoring, control, and management capabilities to the TCC.

Predicting future behavior based on a time series of collected previous data is a tough job in many scientific applications. A time series is a set of time-ordered data values that may be used to measure a physical process. The most frequent measuring

U. Jha · L. K. Behera · S. Mandal · P. Dutta (✉)

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

e-mail: pratikdutta@soa.ac.in

source in traffic management is loop detectors. They mostly provide traffic information. These parameter predictions are crucial, especially for traffic information systems.

As a convention, ARIMA and SARIMA models are the algorithms to do time series forecasting. ARIMA stands for *autoregressive integrated moving average*, and SARIMA stands for *seasonal autoregressive integrated moving average*. These algorithms are usually used for the prediction of data points in the future. To make the ARIMA algorithm work on the dataset, it is needed to check if the dataset is for seasonality and stationarity. Like most of the models working on time series, the data need to be stationary to keep the bias minimum and data more interpretable for the model. This in turn will keep the accuracy to the maximum. Figuring out the stationarity of the dataset is pretty straightforward, if the statistical approach is used, a need to make a null hypothesis which is going to be ‘Data is not stationary’ and an alternate hypothesis that tells us ‘Data is stationary’. After importing the *adfuller* test from *statsmodel* in Python, finding the P -value of the objective, the ad fuller test gives a value of P which if greater than ‘0.05’, will reject the null hypothesis. This is going to confirm that the data are stationary. After this test, ARIMA and SARIMA models can be used on the datasets to do time series prediction.

Regression on time series data is pretty ‘as is’; regression does not involve moving averages and neither does it involve rolling averages. The data are taken as in the form it is given, thus getting coefficients and plotting a graph that helps in getting the predictions. It is much faster and sometimes less accurate than ARIMA models. Multivariate regression has advantages over normal regression; it also considers the residual data and does make the model more accurate.

25.2 Related Study

Artificial neural networks (ANNs) are the most effective way of forecasting [1]. The findings are heavily reliant on historical data since past traffic flow data are utilized to forecast future traffic flow. One issue with traditional statistical methods is that the forecasts tend to converge around the historical data average. For prediction, the heuristic technique utilized by ANN is superior to traditional statistical methods. This is due to the nonlinear nature of the ANN, which is similar to the human brain system in terms of performance. Because of its easy computation and quick performance, ANN is recommended. Furthermore, it reduces mistakes in a short period, enhancing the system’s accuracy and efficiency. It is long-lasting, and as a result, it outperforms other computational models. There are two methods [1] to deal with the traffic problem: (a) User-centered design. (b) Concentrate on the city long-term and short-term forecasts may both be made using ANN. Long-term forecasts will assist drivers/users in gaining an understanding of traffic conditions on each road, allowing them to pick a less congested route to their destination, decreasing traffic congestion. Consider the case where there are two roads, A1 and B1, each having capacity of 5 and 10. Traffic conditions on each of these routes may be forecasted

using ANN. If the circumstances on B1 are better than on A1, the user will choose B1 instead of A1, decreasing traffic on A1. However, this will not be a comprehensive solution to our situation. For example, if a large number of cars pick road B1, B1's traffic flow will swiftly grow. Focusing on the user is ineffective. As a result, the system's concentration should be on the city. Rather than giving customers other routes, it is preferable to make the traffic light system dynamic. We need short-term predictions for this, not a long-term predictions.

In feed-forward networks, there are no loops; thus, information constantly propagates forward. Other networks, on the other hand, include feedback loops in which the input is dependent on the output. Partially recurrent networks [2] perform better than traditional feed-forward neural networks for this purpose. A partially recurrent network improves temporal data representation. This is due to the extra layer employed in partly recurrent networks: the context layer, which establishes a recurrent data flow. This paper's experimental results include a comparison of outcomes from traditional feed-forward neural networks such as multi-layer. For stock price forecasting, perceptrons with the outputs of a partly recurrent network is used. The Elman-type neural network was utilized as the partly recurrent neural network. This network's prediction is shown to be significantly more accurate and reliable than the widely utilized feed-forward neural network. Using an Elman-type neural network, the mean prediction error is the smallest. Elman-type neural networks outperform other neural networks in terms of correlation coefficient and rise and fall prediction.

As previously indicated, we require short-term prediction to synchronize traffic lights. The traffic volume is the parameter that is utilized to make predictions. Jordan's neural network is the best method for short-term prediction, according to this article [3]. They have a strong generalization ability. From conventional time series forecasting to current and sophisticated neural networks, there are several approaches for prediction. All of these approaches have one difficulty in common: determining the optimum structure for which the prediction is most accurate. Every model's prediction error is determined, and the one with the smallest error is chosen. However, the variation of the parameters used for prediction affects this inaccuracy. The variance between training and test data might be different, impacting the model's accuracy. As a result, variables other than mean square error must be taken into account when choosing the optimal model. Recurrence in partly recurrent networks helps them recall recent past data without complicating the general structure, as we learned in the previous study. The context layer improves the accuracy and efficiency of the network. Jordan's sequential network is an example of a partly recurrent network.

Many research on short-term traffic flow prediction has been published in the literature. Nonlinear methods like neural network models [3–7] and linear models like Kalman filters [8–11] and autoregressive integrated moving average (ARIMA) models [12] are examples of short-term forecasting models. ARIMA models are linear estimators based on the modeled time series' previous values [13]. The modeling approach used for traffic prediction is determined by the nature of the data and the type of application. Schmitt and Jula [14] looked at the limits of widely used linear models and found that 'near-future travel times can be better predicted by a combined predictor'. A linear combination of a historical mean predictor and a

present real-time predictor is known as a combined predictor. Guo et al. [15] evaluated several modeling techniques for short-term traffic prediction and found that utilizing a prediction error feedback strategy increases forecast accuracy in both normal and abnormal situations. Smith et al. [16] examined para-metrically (seasonal ARIMA) and non-parametric (data-driven regression) models and found that ‘traffic condition data are often stochastic, rather than chaotic’.

To increase the feasibility by making it as simple as possible, these three points demonstrate it: (a) Operational feasibility—the solution once established and the dataset trained will not be a resource hog. It will be requiring less computational power and will be able to give results faster even on lower-end processors; (b) Technical feasibility—having used more than one model for the prediction purpose. This is going to help the user get a better picture of what to expect from the output; and (c) Economic feasibility—users will be able to save time, which is not going to be wasted on traffic jams.

25.3 Proposed Framework

This article recommends using a linear regression model to forecast traffic flow, so making static traffic signals dynamic. Several forecasting methods, such as ARIMA and SARIMA, can be used to anticipate traffic flow. However, the problem is that the answer is complex and requires a lot of computing power. Linear regression, on the other hand, is very light and requires little computer power because it only requires a basic mathematical operation to achieve the goal once the coefficient values have been obtained. To do this, the time series dataset was restructured so that it could be fitted into the model.

The suggested work’s class diagram is shown in Figure 25.1. The framework can be used to any traffic flow dataset for a specific intersection. In the actual world, vehicle identification (counting) is accomplished by placing multiple stationary sensors at a given place. Because the goal is to predict, unlike standard forecasting algorithms, the dataset must be reconstructed uniquely, as explained briefly in Sect. 25.3.1. After that linear regression was used to attain the goal.

25.3.1 Dataset

The dataset (PEMS-SF) was downloaded from UCI machine learning repository, which could further be processed and used with statistical models. These models help us predict the flow of traffic. Dataset obtained daily data for 15 months from the California Department of Transportation’s PEMS Web site. The statistic represents the occupancy rate of distinct car lanes on San Francisco Bay region freeways, which ranges from 0 to 1. The measurements are taken every 10 min and cover the period from January 1st, 2008 to March 30th, 2009. Each day in this database is treated as

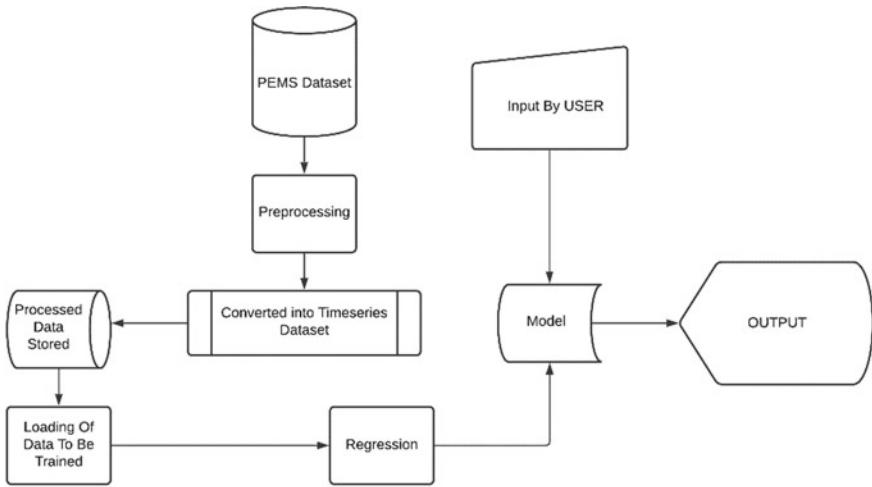


Fig. 25.1 Flow of proposed framework

a separate time series with a dimension of 963 (the number of sensors that worked reliably throughout the study period) and a length of $6 \times 24 = 144$. Along with public holidays, two days with anomalies (March 8th, 2009 and March 9th, 2008) when all sensors were muted between 2:00 and 3:00 a.m. from the dataset have been omitted.

Throughout the time period, Figure 25.2 displays the instance of vehicle occupancy in the abovementioned network. It is clear that during peak hours, occupancy is higher than at other times during the day. The dataset is built and used to categorize each day of the week. Only seven days of data have been shown in Figure 25.3 for a better comprehension of the dataset.

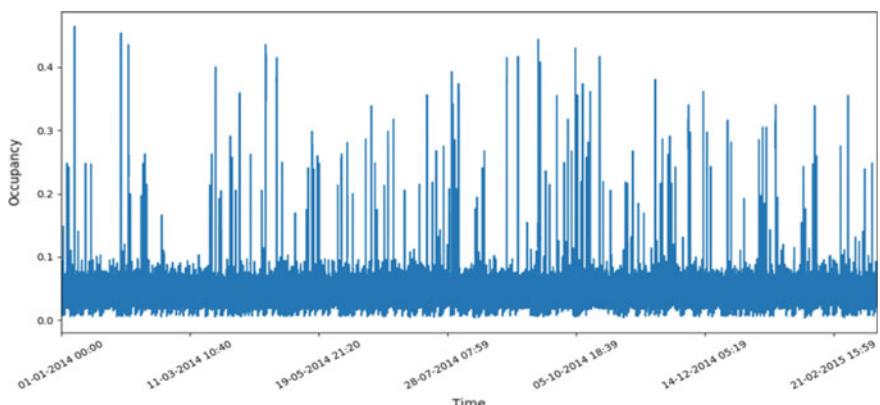


Fig. 25.2 Occupancy of vehicle in one station throughout 440 days

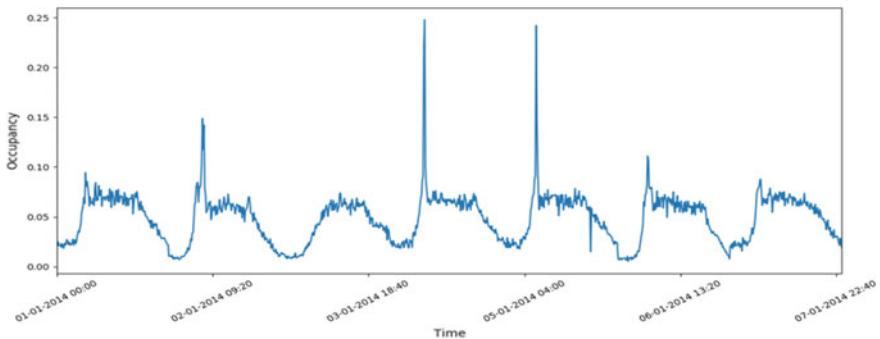


Fig. 25.3 Occupancy of vehicle in one station throughout seven days

Restructuring the Dataset

The dataset has 440 days of data. Each day is having $(963 \times 144) = 138,672$ entries. To fit the dataset into linear regression model, the dataset need to be restructured. And that can be achieved in the following process:

- Separating the data from all 963 sensors, resulting in $440 \times 144 = 63,360$ occupancy entries in each file.
- Using the window slide technique, the first M values are selected as input features, and the $M + 1$ th item is regarded as output. When $M = 6$ for the first sensor, this results in a dataset containing $(63,354 \times 7)$ entries. Table 25.1 depicts the outcome after restructuring the dataset with $M = 10$. However, finding the optimal M is the critical one, and after experimenting with different M , the optimal M has been found as 15. To train and test the linear regression model, the 7:3 ratio has been used.

Table 25.1 Restructured dataset for sensor 1

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	Y
0.026	0.021	0.025	0.022	0.022	0.020	0.024	0.020	0.022	0.019	0.024
0.021	0.025	0.022	0.022	0.020	0.024	0.020	0.022	0.019	0.024	0.022
0.025	0.022	0.022	0.020	0.024	0.020	0.022	0.019	0.024	0.022	0.023
...
...
0.020	0.022	0.019	0.024	0.022	0.023	0.027	0.026	0.022	0.025	0.022
0.022	0.019	0.024	0.022	0.023	0.027	0.026	0.022	0.025	0.022	0.023
0.019	0.024	0.022	0.023	0.027	0.026	0.022	0.025	0.022	0.023	0.022

25.3.2 *Methodologies*

The ARIMA model does not automatically update; thus, if new data points are supplied, the entire model must be rerun. This characteristic allows ARIMA models to perform better in short-term forecasting in general. Because regression is more flexible and parametric, it can be predicted to have better long-term accuracy. However, above a certain threshold, accuracy tends to drop off dramatically.

ARIMA used imperceptible values, and those values are moving averages, and regression calculations normally include a lot of assumptions. This distinguishes their methods to prediction. Regression models are more flexible, whereas ARIMA models are usually less complex and more resilient than regression models. After repeating the tests, it was discovered that multivariate regression was capable of producing very good results, with accuracy reaching as high as 99% in some cases.

The goal of the suggested research is to find a low-cost, simple, and reliable algorithm that can deliver a significant and satisfactory outcome. It is clear from the preceding discussion that ARIMA may generate good results for forecasting traffic vehicles in the road segment; however, the algorithm is quite expensive. The work has been recognized with the knowledge that the system setup in a traffic light is insufficient. This is not an issue with linear regression, and it is also compatible with low-cost systems. The key distinctions between ARIMA and REGRESSION are in Table 25.2.

25.4 Experimental Setup and Result Analysis

The proposed work has experimented under the following system configuration discussed in Sect. 25.4.1 and observed the result discussed in Sect. 25.4.2.

Table 25.2 Understanding difference between ARIMA and LR

ARIMA	Regression
Accuracy is heavily dependent on datasets length. 100+ data points are preferred	Regression usually works with significantly lesser data. As low as four data points
ARIMA models are usually combines the autoregressive with moving averages that usually could be made within a regressive model	Regression is the base model for ARIMA and SARIMA models. This can be tuned as per requirements. This makes it more flexible
It does not usually requires a target a set of predictor variables	Regression is dependent on the set of predictor variables for forecasting
ARIMA, uses unobservable data points as a moving averages	Regression tends to work with visible data points, i.e., data that are clearly visible to user

25.4.1 Experimental Setup

The entire model was thoroughly tested and trained on the dataset to test out the settings. A system with 4 GB RAM and a 2.4 GHz quad-core processor was used for linear regression. However, to cooperate with ARIMA, Google Colab was employed, and the above configuration is incapable of handling the large storage requirements.

25.4.2 Result Analysis

In this section, the thorough observation of the experiment will be discussed. The dataset has been applied to ARIMA and linear regression model as early discussed. The following accuracy metric ($1 - \text{error}$) has been used to justify the findings.

- **Mean Absolute Error (MAE)**—the mean absolute error is the average of the absolute difference between the actual and anticipated values in the dataset. It calculates the average of the dataset's residuals. It can be expressed as

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

- **Mean Squared Error (MSE)**—the mean squared error is the average of the squared difference between the original and projected values in the data collection. It calculates the residuals' variance. It can be expressed as

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

- **R Squared (R^2)**—the R -squared coefficient of determination reflects the fraction of the variation in the dependent variable that the linear regression model can explain. It is a scale-free score, meaning that regardless matter how tiny or huge the values are, R square will be less than one. It can be expressed as

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

where \hat{y} represents predicted value of y .

The accuracy of MAE and MSE can simply be found by subtracting the values from 1 multiplying to 100. Tables 25.3, 25.4, and 25.5 represent the accuracy (%) based on the data that had been taken.

Table 25.3 depicts the accuracy for all the sensors in a road segment. MSE has an accuracy range of 95.88–99.60%, with a variation of 0.211. Similarly, MAE

Table 25.3 Accuracy for all sensors

Sensor	Acc_MSE	Acc_MAE	R^2
Sensor 001	98.84984	99.49091	84.85661
Sensor 002	98.5817	99.40806	88.41685
Sensor 003	98.40915	99.20281	89.47212
Sensor 004	98.46321	99.16235	88.36688
...
Sensor 960	98.47571	99.21906	89.96967
Sensor 961	97.96347	99.20663	87.43588
Sensor 962	98.43422	99.25586	89.15719
Sensor 963	98.53399	99.31092	81.4081

Table 25.4 Comparative study between models

Model	Acc_MSE	Acc_MAE	R^2
ARIMA	99.83349	97.29013	89.0652
Linear regression	98.58569	99.33832	88.90407
Difference	-1.2657	2.06183	-0.18124

Bold values identifies the best accuracy for a methodology compare to other methodologies that has been found from experiment

Table 25.5 Identification of optimal M

M	Acc_MSE	Acc_MAE	R^2	M	Acc_MSE	Acc_MAE	R^2
6	98.559	99.336	88.938	13	98.554	99.336	89.009
7	98.549	99.332	88.874	14	98.561	99.336	89.103
8	98.546	99.331	88.873	15	98.585	99.338	88.904
9	98.560	99.338	89.000	16	98.544	99.336	88.805
10	98.553	99.337	88.921	17	98.540	99.330	88.802
11	98.560	99.337	89.017	18	98.546	99.331	88.873
12	98.559	99.336	88.997	19	98.549	99.332	88.874
				20	98.546	99.331	88.873

Bold values identifies the best accuracy for a methodology compare to other methodologies that has been found from experiment

(accuracy) is found to be within a range of 97.77–99.76% with a variance of 0.034%. Finally, the variation for R^2 is 7.12%, ranging from 88.39 to 98.135%. As a result, it can be stated that linear regression with the rebuilt dataset performs admirably.

Table 25.4 depicts the performance comparison between linear regression and ARIMA model. The results have achieved based on average of all sensor findings. Both the models have been tested on 70% training data with 30% test data. It can be

observed that ARIMA provide better accuracy in terms of Acc_MSE and R^2 , while in R^2 , linear regression is the superior. It can also be identified that the difference in accuracy between two models are negligible as -1.26657% , 2.06183% , and -0.18124% , respectively, with respect to linear regression.

Table 25.5 depicts the identification of optimal value for M . The experiment has been done with the different value of M starting from 6 up to 20. It has been observed that for $M = 15$ which is restructuring the dataset with 15 input features gives the best result among others.

25.5 Conclusion and Future Scope

The contrast between a linear regression and an autoregressive technique is clarified in this paper. The abovementioned algorithms were compared for estimating traffic flow based on historical data. This article makes a contribution by restructuring time series data in a specific way so that it can be used in a light-weight model like linear regression. Extensive testing has revealed that the autoregressive integrated moving average (ARIMA) performs well (with prediction accuracy ranging from 89.06 to 99.83%), while the linear regression model predicts accuracy (88.90–99.33%) based on several accuracy measurement metrics. As a result, if the dataset is restructured, a light-weight prediction model (e.g., linear regression) can be used to substitute a heavy-weight model like ARIMA.

References

1. More, R., Mugal, A., Rajgure, S., Adhao, R.B., Pachghare, V.K.: Road traffic prediction and congestion control using artificial neural networks. In: 2016 International Conference on Computing, Analytics and Security Trends (CAST), pp. 52–57. IEEE (2016). <https://doi.org/10.1109/CAST.2016.7914939>
2. Merkel, D.: Partially recurrent neural networks in stock forecasting. In: Artificial Intelligence in Economics and Management, Tel Aviv, 08–10 Jan 1996
3. van Lint, J.W.C., Hoogendoorn, S.P., van Zuylen, H.J.: Accurate freeway travel time prediction with state-space neural networks under missing data. Transp. Res. Part C Emerg. Technol. **13**(5/6), 347–369 (2005)
4. Yu, J., Chang, G.-L., Ho, H.W., Liu, Y.: Variation based online travel time prediction using clustered neural networks. In: Proceedings of 11th International IEEE Conference on Intelligent Transportation Systems, Oct 2008, pp. 85–90
5. Chan, K.Y., Dillon, T.S., Singh, J., Chang, E.: Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm. IEEE Trans. Intell. Transp. Syst. **13**(2), 644–654 (2012)
6. Park, D., Rilett, L.R., Han, G.: Spectral basis neural networks for realtime travel time forecasting. J. Transp. Eng. **125**(6), 515–523 (1999)
7. Ye, Q.Y.Q., Wong, S.C., Szeto, W.Y.: Short-term traffic speed forecasting based on data recorded at irregular intervals. In: Proceedings of 13th International IEEE Conference ITSC, pp. 1541–1546 (2010)

8. Okutani: The Kalman filtering approaches in some transportation and traffic problems. *Transp. Res. Rec.* **2**(1), 397–416 (1987)
9. Xie, Y., Zhang, Y., Ye, Z.: Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition. *Comput.-Aided Civ. Infrastruct. Eng.* **22**(5), 326–334 (2007)
10. Ji, H.J.H., Xu, A.X.A., Sui, X.S.X., Li, L.L.L.: The applied research of Kalman in the dynamic travel time prediction. In: Proceedings of the 18th International Conference on Geoinformatics, pp. 1–5 (2010)
11. Wang, Y., Papageorgiou, M.: Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transp. Res. Part B Methodol.* **39**(2), 141–167 (2005)
12. Van Der Voort, M., Dougherty, M., Watson, S.: Combining Kohonen maps with ARIMA time series models to forecast traffic flow. *Transp. Res. Part C Emerg. Technol.* **4**(5), 307–318 (1996)
13. Williams, B.M., Durvasula, P.K., Brown, D.E.: Urban freeway traffic flow prediction application of seasonal autoregressive integrated. *Transp. Res. Rec.* **1644**, 132–141 (1998)
14. Schmitt, E.J., Jula, H.: On the limitations of linear models in predicting travel times. In: Proceedings of the IEEE Intelligent Transportation Systems Conference, Sept 2007, pp. 830–835
15. Guo, F., Polak, J.W., Krishnan, R.: Comparison of modeling approaches for short term traffic prediction under normal and abnormal conditions. In: Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, pp. 1209–1214 (2010)
16. Smith, B.L., Williams, B.M., Keith Oswald, R.: Comparison of parametric and nonparametric models for traffic flow forecasting. *Transp. Res. Part C Emerg. Technol.* **10**(4), 303–321 (2002)

Chapter 26

Detection and Classification of Encephalon Tumor Using Extreme Learning Machine Learning Algorithm Based on Deep Learning Method



**Premananda Sahu, Prakash Kumar Sarangi, Srikanta Kumar Mohapatra,
and Bidush Kumar Sahoo**

Abstract The ordinary people cannot have the capability to detect and classify the brain tumor, so the radiologists or the clinical experts are the only person who can detect the encephalon tumor due to their practice and awareness. If the manual detection process will be carried out by the clinical experts, then they will face a lot of problem like time delay, classification accuracy etc., so for this reason, automatic encephalon tumor detection and classification process are required. In this paper, we have detected the encephalon tumor from three schemes, that is, initially, we have collected some raw data and put into the machine learning model; secondly, the classification process is done by extreme learning machine (ELM); and finally, the tumors are extracted from most commonly method gray level co-occurrence method (GLCM). For detecting the tumor region, we have used the one and only common method magnetic resonance image which have also some noises or the doted pictures are available which give a clear image to the clinical experts for detection process. In this work, the classification accuracies have produced as an approximation of 98.09% which has produced the better outcome as compared to the studies that are presented in the related work. From this work, the result is shown as an effective manner for the radiologists as it uses the deep learning method.

P. Sahu (✉)

Department of Computer Science and Engineering, SRMIST, Delhi-NCR, Ghaziabad, India

P. K. Sarangi (✉)

School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

S. K. Mohapatra · B. K. Sahoo

Department of Computer Science and Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

26.1 Introduction

As we know, in human body, there are so many organs present, but if we will discuss regarding the brain, then it is an essential organ. Once the brain is going to be damaged, then that particular human being is valueless in nature. So we want to detect the tumor if the person has any, but in the early stage of a tumor if it is going to be diagnosed, then that person may recover from a big problem. In this case, we have used the magnetic resonance image (MRI) whether the tumor is present in the encephalon or not [1]. The encephalon takes information superficially and represents the concentration of mind, so the human being is the best creature in the world. The total detection of tumor process can be very hectic and difficult task if the clinical experts or radiologists have less knowledge, so only the experienced or knowledgeable person can detect the problem [2]. If the manual inspection of cancer is analyzed by clinical experts, then it comes some error when the numbers of images are going to be increase. To eschew the human mistake, exploration and displacement require an automatic system for medical diagnosis. For this purpose, different segmentation algorithms as well as the optimization techniques worn for encephalon tumor reduction and categorization process from MRI. As the manual detection process for tumor detection method takes more time or it can gives less classification exactness, here we can help to the clinical experts for removing of the above demerits because here we are using a fully automatic method for both classification and segmentation process. In addition to the encephalon tumor, it can divide into two processes that are benign (noncancerous cell) and malignant (cancerous cell) [3]. In this paper, we have used the most commonly used segmentation method; hybrid fuzzy C mean clustering algorithm and automation-based segmentation have been used [4]. When the images are available from MRI, there are a lot of images are present, so we have to extract the images for reduce the duplicity of images. In this regard, we have also used the most common method feature extraction method; gray-level co-occurrence matrix (GLCM) where it increases the speed of learning process as well as the efforts of radiologists. For using the above method, we have to use a dataset, so we have used the ADNI 2020 dataset to appraise the accomplishment of model.

Previously, different researchers have worn for detecting and classification of encephalon tumor by supervised learning process which uses the labeled data to forecast the output, but in this work, we have used the deep learning method which is unsupervised learning method where it uses the unlabeled data. Initially, the classifier who wants to train the model uses the trained data; later, he/she tries to label the unlabeled data which increases the prediction label much higher [5]. Again, we have used a new technique; extreme learning machine (ELM) which belongs to single feed forward network hidden layer that to be modeled for various functions named as threshold value, weight calculation etc., which is used for better classification of encephalon tumor, and it is also very compatible with deep learning method in case of linear data approximation method [6]. Before some years ago, all the detection process was held by computed axial tomography where all the classification process was done by deep convolutional neural network as this is more compatible with

the above scanning process [7]. Now, all the images are scanning through the MRI, so there are some demerits are coming in case of CNN method; (a) if we want to train the model, then it takes more computational time if there are several layers are present in CNN. (b) If the dataset is small, then the processing capacity with the training method to the ANN that cannot predicted [8]. For this reason, extreme learning machine is used for the categorization purpose which reduces some of the demerits of previously used classification method. To increase the speed of learning method, here, we have used the local receptive fields' method which is a spatial term that provides the input to a set of units. This implies that it can give some advantage regarding the complexity issues. If the extreme learning machine can combine with the local receptive fields, then it can give better classification result as compared to the previous one [9].

This paper is synchronized in following manner: Related work has been described in Sect. 26.2, Methodology has entitled in Sect. 26.3, exploratory result has been depicted in Sect. 26.4, and Conclusion and future scope have been referred in Sect. 26.5.

26.2 Related Work

In the past days, various researchers have developed so many image segmentation techniques, feature extraction methods as well as the classification approaches. Some of them has been described in the following manner.

An and Liu [10] describe that if the errors are occurred in minimal set of data points, then the deep learning model can be implemented with combination of adaptive dropout model. Here, the images are segmented by optimized convolutional neural network which is very compatible with adaptive dropout model which gives best result. Minaee et al. [11] surveyed that for the image segmentation, the deep learning algorithms give an immense performance as compared to the other techniques present over there. Here, they have also used the enlarged version of convolutional neural network with the comparison of so many datasets for performance evaluation. Shanker and Bhattacharya [12] have signified the image segmentation process from the combination of K means clustering with fuzzy C means algorithm. Here, they have presented “Hierarchical Centroid Shape Descriptor” which worn to find the abnormal area, and then after, they have validated the results which has given a supreme outcome.

Latif et al. [13] have implemented the content-based image retrieval technique because from MRI, when the pictures are acquired, there are some different types of colors, shapes as well as the textures have produced. Due to the variants of datasets regarding the images, all the above properties cannot be removed properly. So the above extraction method has been used if the bonding of low-level features are also been applied. But as per the authors, if the above technique has been applied in the deep learning method as well as the unlabeled dataset, then it takes more time. Salau and Jain [14] surveyed the recently used existing feature extraction techniques, but

they have used the gray-level difference statistics technique for the feature extraction which has used the properties like entropy, energy, and mean. But the accomplishment and the exactness of these techniques are the key aspects. Jalali and Kaur [15] stated many extracting feature methods that are gray-level co-occurrence matrix, local binary patterns as well as the conventional image–intensity feature extraction technique. Here, the researchers have used the deep learning-based feature extraction which provides the superior classification exactness.

Varuna Shree and Kumar [16] proposed the encephalon segmentation process into cancerous as well as noncancerous. They have used the most commonly extracting feature methods gray-level co-occurrence matrix for the removal about noisy images with smoothening the images, but as they have used the dataset is very small, so if the accuracy required then a large dataset as well as the GLCM technique are required. Wang et al. [17] surveyed that extreme learning machine has given a good response in case of implementation process. As the trained output bias of ELM is always zero, if it will be compatible with deep learning method, then it may produce better classification accuracy. Mishra et al. [18] proposed the detection and classification of brain tumor from MRI by the help of MASCA-PSO-based LLRBFNN model with segmentation process by IFRFCM algorithm. They have indicated that the above technique is fully automatic for detection of cancerous tissues from MRI, but again, they have specified that if the deep learning method with the extreme learning machine classifier are used, then it may give the better classification accuracy.

From the above literature survey, we have observed that for the encephalon tumor detection and classification method not only the supervised learning method are required, but also the non-supervised learning method and the better classifiers are also required.

26.3 Methodology

Here, the research work is mainly focused into the classification and detection of encephalon tumor from MRI which is shown on Fig. 26.1.

The investigation effort has developed from the following processes:

- The encephalon tumor has been segmented by convolutional neural network initially.
- The features have been extracted through the most commonly reduction method, gray-level co-occurrence matrix (GLCM), after segmentation.
- The removed area has been classified by extreme learning machine (ELM) classifier.
- Finally, after classification, the images have detected whether it is malignant or benign by the help of deep learning Method.

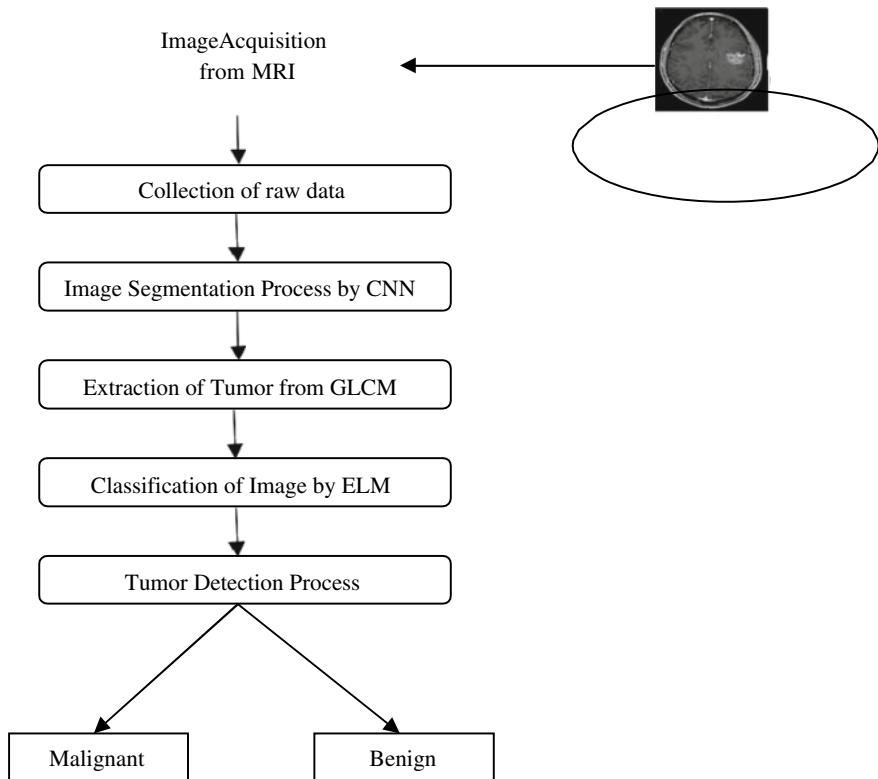


Fig. 26.1 Research workflow block diagram

Proposed Convolutional Neural Network Method for Image Segmentation

As after acquiring from MRI, if some dotted pictures are coming, then it needs to be segmented. We have used the CNN method in this paper for image segmentation process. It is a form of neural network model that allows the user to remove advanced representations for image input. The traditional image detection process distinctly requires defining the image characteristics very clearly. It takes raw data from image, trains the model, and then extracts the features for improved categorization automatically. The basic structure of CNN is illustrated in Fig. 26.2.

We have used CNN for image classification where in the diagram, convolutional layer is used for convolutional operation or combine the input function and pass its result to the next layer; pooling layer extracts important features which is taken advantage of reducing the computational complexity and in fully connected layer, all the neurons are connected to each neuron of previous layer which is predictable to the hidden layer [19].

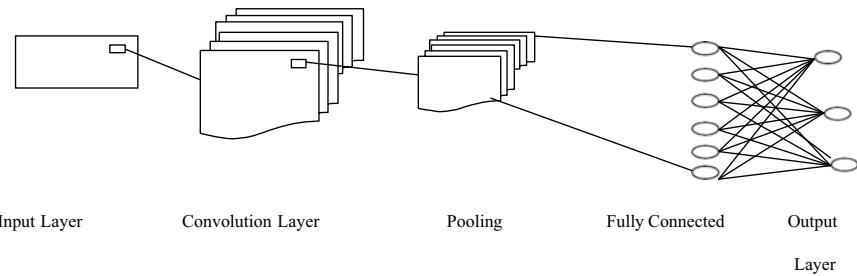


Fig. 26.2 Structure of CNN

GLCM (Gray-Level Co-occurrence Matrix) Technique

As MRI consists of so many pictures that has to be captured, for reduction of large number of pictures to the noisy picture, there is a removal technique used. This is called as feature extraction process. In this work, we have used GLCM technique for feature extraction where it collects superior stage of facts of a picture like structure, size, and color. Now, input MRI pictures will undergo the process of tumor detection as well as the segmentation process.

Extreme Learning Machine

It is a new type of algorithm which uses single-layer feedforward neural network that is mainly used for classification purpose. It has just to build the hidden layer numbers that completely not require to alter the input amounts on a regular basis which has ample of advantages including less human interaction, faster learning speed, and potential for generalization [20]. Presumably, we can say that extreme learning machine is very fast to train the model, but it cannot encode more than one layer at a time. So we have used the above technique based on the deep learning method which may detect and classify the encephalon tumor in a better way. The basic building block of ELM is depicted in Fig. 26.3.

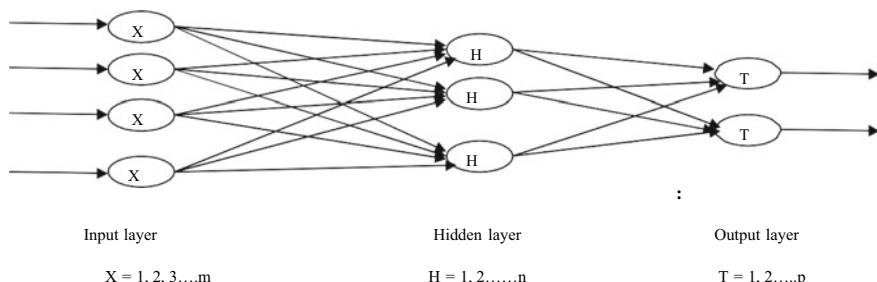


Fig. 26.3 Basic block diagram of ELM

Algorithm

It is a very useful as well as proficient training algorithm. Now, the output of ELM ($f_H(x)$) is determined as

$$f_H(x) = \sum_{i=1}^H v_i b_i(x) = \sum_{i=1}^H v_i b(w_i \times x_j + k_i), \quad j = 1, 2, 3, \dots, N \quad (26.1)$$

where H = number of hidden units, N = number of training samples, w = weight vector between input and hidden layer, v = weight vector between hidden and output layer, b = activation function, k = bias vector, and x = input vector.

Now, if we will assume beta matrix as special matrix, then the output (T) will be considered as

$$T = Z\beta \quad (26.2)$$

where

$$Z = \begin{bmatrix} b(w_1 \times x_1 + k_1) & \dots & b(w_H \times x_1 + k_H) \\ \vdots & & \vdots \\ b(w_1 \times x_1 + k_1) & \dots & b(w_H \times x_N + k_H) \end{bmatrix}_{N \times H}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_H^T \end{bmatrix}_{H \times P} \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times P}$$

where p is number of outputs, and Z is hidden layer output matrix.

26.4 Exploratory Result

We have collected the MRI datasets from ADNI public database, where it has used 156 MRI samples has been taken for training as well as testing [21] and 39 patients record has also recorded where 22 images has been used for training with 17 images has been used for testing. The experimental work has been built by Python on the Google co-lab framework with the Keras deep learning API and a TensorFlow in backend. Here, the ADNI texts have used on behalf of dividing process as well as outcome have presented in Fig. 26.4.

In this work, we have used the ELM method for image classification. If we will use the local receptive field which is much compatible with ELM in which during the convolution process, a neuron inside a convolutional layer is exposed to a defined segmented area inhabited by the content of input data. So we have used local receptive fields-based extreme learning machine for generating filters by Gabor functions

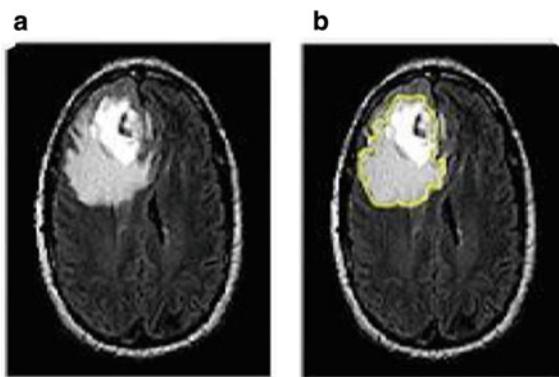


Fig. 26.4 **a** MRI original image. **b** ELM result

which may provide the better classification accuracy [22]. Here, all input images were resized to $32 * 32$ before giving input to the extreme learning machine-based local receptive fields. It has four tunable parameters: convolution filter size, convolution filter number, pooling size, and regulation coefficient. Now, the convolution and pooling action outcome has been signified in Fig. 26.5.

In extreme learning machine-based local receptive fields, all the images are removed by directly but in Gabor functions, all the images have been extracted by the tumor area only. In statistical-based studies, also, same actions have been taken.

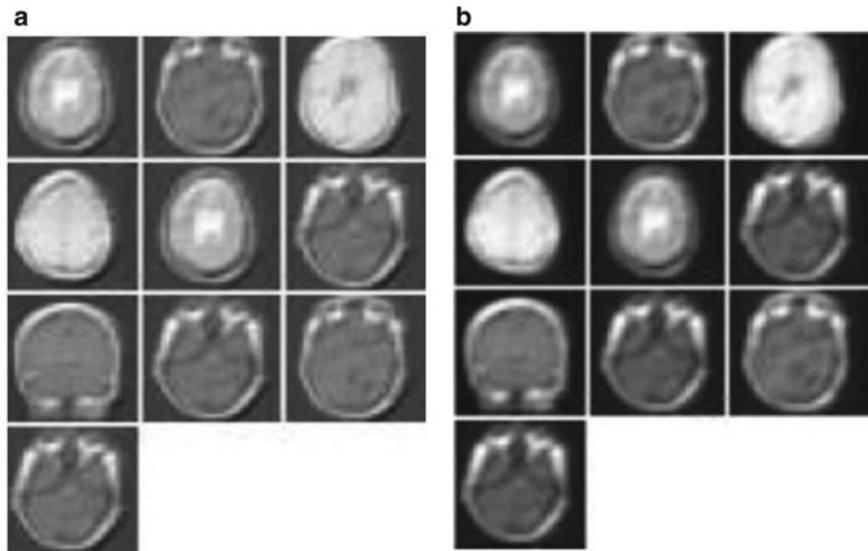


Fig. 26.5 **a** Convolution action. **b** Pooling action

Now, if we will calculate the computational time with the above-noted technique, then the evaluation presentation has come in Table 26.1.

Now, the evaluation results again has to be calculated for better classification accuracy because the accuracy is an assessment of the system's efficiency in executing the entire categorization that determines the whole amount of correctly categorized encephalon magnetic resonance pictures [23]. The terms which have been used for better performance measure or for better classification accuracy has been depicted in Table 26.2.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP = number of true positives, TN = number of true negatives, FN = number of false negatives, and FP = number of false positives.

From the accuracy table, we have seen that ELM-LRF method produces enhanced classification result in contrast to others. As our result is showing good approximation process, but in the new era, the importance of Internet of things is also increased very tremendously if we will classify the encephalon tumor as there have to be a brief amount of data are essential [24].

Table 26.1 Evaluation presentation of different methods

Technique	Classification affluence rate (%)
Gabor features	94.03
Statistical features	96.37
CNN	96.83
ELM-LRF	97.35

Table 26.2 Performance assessment of different classifiers

Technique	Sensitivity	Specificity	Accuracy (%)
Gabor features	0.93	0.95	94.07
Statistical features	0.96	0.92	95.03
CNN	0.96	0.98	97.12
ELM-LRF	0.97	0.99	98.09

26.5 Conclusion

In this work, the MRI process is used for encephalon tumor detection and classification process. From this study, it has been observed that the CNN model has segmented the images in a better way. After segmentation, all the images fed as input for the extraction of images where we have used common and well-known method GLCM has been used. Then, we have used the ELM technique for the classification purpose, but as the convolutional layer has come that supports deep learning method, so local receptive fields have been used for generating the filters which gives better classification accuracy. The outcomes reported in this paper exhibit distinctiveness of this research work as well as comparison outcomes has further presented. The proposed work has revealed excellent potential for categorizing the tumor as benign or malignant which may help to the clinical practitioners.

Particle swarm intelligence-based extreme learning machine may be exploited for encephalon tumor classification process as benign or malignant with feature reduction process may be used to increase the classification accuracy can be used as future work.

References

1. Mohammed Thaha, M., Pradeep Mohan Kumar, K., Murugan, B.S., Dhanasekeran, S., Vijayakarthick, P., Senthil Selvi, A.: Brain tumor segmentation using convolutional neural networks in MRI images. *J. Med. Syst.* (2019). <https://doi.org/10.1007/s10916-019-1416-0>
2. Banday, S.A., Mir, A.H.: Statistical textural feature and deformable model based MR brain tumor segmentation. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 657–663 (2016). <https://doi.org/10.1109/ICACCI.2016.7732121>
3. Rouhi, R., Jafari, M., Kasaei, S., Keshavarzian, P.: Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Syst. Appl.* **42**(3), 990–1002 (2015)
4. Hemalatha, K.L., Prabha, V., Gowda, G.S., Bhavani, S., Chitrashree, K.: RELM: a machine learning technique for brain tumor classification. *Perspect. Commun. Embedded Syst. Signal Process.* **4**(5) (2020). <https://doi.org/10.5281/zenodo.4018991>
5. Ahmad, J., Farman, H., Jan, Z.: Deep learning methods and applications. In: Deep Learning: Convergence to Big Data Analytics. SpringerBriefs in Computer Science. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-3459-7_3
6. Yıldırım, H., Revan Özkal, M.: The performance of ELM based ridge regression via the regularization parameters. *Expert Syst. Appl.* **134**, 225–233
7. Praveen, G.B., Agrawal, A.: Multi stage classification and segmentation of brain tumor. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACoM), pp. 1628–1632 (2016)
8. Ayaz, F., Ari, A., Hanbay, D.: Leaf recognition based on artificial neural network. In: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), pp. 1–5 (2017). <https://doi.org/10.1109/IDAP2017.8090240>
9. Pang, S., Yang, X.: Deep convolutional extreme learning machine and its application in handwritten digit classification. *Comput. Intell. Neurosci.* **2016**, 1–10. Article ID 3049632 (2016). <https://doi.org/10.1155/2016/3049632>

10. An, F.-P., Liu, J.: Medical image segmentation algorithm based on optimized convolutional neural network-adaptive dropout depth calculation. *Complexity* **2020**, 1–13. Article ID 1645479 (2020). <https://doi.org/10.1155/2020/1645479>
11. Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021)
12. Shanker, R., Bhattacharya, M.: Brain tumor segmentation of normal and lesion tissues using hybrid clustering and hierarchical centroid shape descriptor. *Comput. Methods Biomed. Eng. Imaging Vis.* **7**(5–6), 676–689 (2019). <https://doi.org/10.1080/21681163.2019.1579672>
13. Latif, A., Rasheed, A., Sajid, U., Ahmed, J., Ali, N., Ratyal, N.I., Zafar, B., Dar, S.H., Sajid, M., Khalil, T.: Content-based image retrieval and feature extraction: a comprehensive review. *Math. Probl. Eng.* **2019**, 1–21. Article ID 9658350 (2019). <https://doi.org/10.1155/2019/9658350>
14. Salau, A.O., Jain, S.: Feature extraction: a survey of the types, techniques, applications. In: 2019 International Conference on Signal Processing and Communication (ICSC), pp. 158–164 (2019). <https://doi.org/10.1109/ICSC45622.2019.8938371>
15. Jalali, V., Kaur, D.: A study of classification and feature extraction techniques for brain tumor detection. *Int. J. Multimed. Inf. Retr.* **9**, 271–290 (2020)
16. Varuna Shree, N., Kumar, T.N.R.: Identification and classification of brain tumor MRI images with feature extraction using DWT and probabilistic neural network. *Brain Inf.* **5**, 23–30 (2018). <https://doi.org/10.1007/s40708-017-0075-5>
17. Wang, J., Lu, S., Wang, S.-H., Zhang, Y.-D.: A review on extreme learning machine. *Multimed. Tools Appl.* (2021). <https://doi.org/10.1007/s11042-021-11007-7>
18. Mishra, S., Sahu, P., Senapati, M.R.: MASCA-PSO based LLRBFNN model and improved fast and robust FCM algorithm for detection and classification of brain tumor from MR image. *Evol. Intell.* **12**(4), 647–663 (2019)
19. Rachapudi, V., Devi, G.L.: Improved convolutional neural network based histopathological image classification. *Evol. Intell.* 1–7 (2020)
20. Cao, W., Gao, J., Wang, X., Ming, Z., Cai, S.: Random orthogonal projection based enhanced bidirectional extreme learning machine. In: Cao, J., Vong, C., Miche, Y., Lendasse, A. (eds.) *Proceedings of ELM 2018. ELM 2018. Proceedings in Adaptation, Learning and Optimization*, vol. 11. Springer (2020). https://doi.org/10.1007/978-3-030-23307-5_1
21. <https://www.nature.com/articles/s41598-020-79243-9.pdf>
22. He, B., Song, Y., Zhu, Y., et al.: Local receptive fields based extreme learning machine with hybrid filter kernels for image classification. *Multidimens. Syst. Signal Process.* **30**, 1149–1169 (2019). <https://doi.org/10.1007/s11045-018-0598-9>
23. Nayak, D.R., Dash, R., Majhi, B.: Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests. *Neurocomputing* **12**(177), 188–197 (2016)
24. Datta, P., Sharma, B.: A survey on IoT architectures, protocols, security and smart city based applications. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE (2017)

Chapter 27

Comparative Study of Medical Image Segmentation Using Deep Learning Model



Pubali Chatterjee, Simran Sahoo, Subrat Kar, and Pritikrishna Biswal

Abstract In the planning and quantitative assessment of brain tumor treatment, determining the tumor extent is a major challenge. Non-invasive magnetic resonance imaging (MRI) has developed as a diagnostic technique for brain malignancies without the use of ionizing radiation. Gliomas (tumors) are the most common and severe kind of brain tumor due to their infiltrative nature and rapid growth. Identifying tumor boundaries from healthy cells in the clinical environment is still a challenging task. The fluid-attenuated inversion recovery (FLAIR) MRI method can provide clinicians with information on tumor infiltration. Many recommendations include using deep neural networks (DNN) in image segmentation because they perform well in automated brain tumor image segmentation. Due to the complexity of the gradient diffusion problem, training a deeper neural network requires a lot of time and a lot of computing resources. As a result, utilizing fluid-attenuated inversion recovery (FLAIR) MRI data, this project compares two deep learning architectures, U-Net, ResNet, AlexNet, VGG16-Net, and V-Net, for totally automated brain lesion diagnosis and segmentation. In contrast to traditional supervised machine learning techniques, these deep learning-based algorithms do not rely on natural features, and instead, construct a pyramid increasingly complex characteristics are drawn directly from the data.

27.1 Introduction

To diagnose patients, doctors typically employ multimodal three-dimensional (3D) MR scans. Magnetic resonance imaging (MR imaging) is a type of medical imaging that employs techniques such as magnetic resonance signal production and acquisition, as well as spatial encoding and Fourier sampling. 3D MR image technology, which reconstructs a series of 2D images into 3D images, is the most significant imaging technique in brain tumor detection procedures. Doctors may obtain image

P. Chatterjee (✉) · S. Sahoo · S. Kar · P. Biswal

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

e-mail: pubalichatterjee@soa.ac.in

data from 3D MR scans at any point in space, making quantitative and qualitative mathematical and medical investigation easy for staff. MR pictures of various modalities can be scanned in the 3D MR imaging process due to variations in auxiliary circumstances such as contrast agents.

The main objective of this paper is about comparison with various segmentation techniques. Segmentation is a method of separating a picture into smaller sections with common characteristics in gray levels. Additional features include color, texture, and brightness. The following is the goal of medical image segmentation: To explore the inner structure of human anatomy in depth, it aids in the identification of segments such as tumor location.

In this paper, we analyze some segmentation algorithm with some convoluted network based on the U-Net, AlexNet, ResNet design. To increase segmentation accuracy, this project used a comprehensive data augmentation technique. We also utilized a “Soft” dice-based loss algorithm that was created by the researchers. Because some areas below the tumor may represent only a small part of the total tumor volume, the soft dice loss feature has a unique advantage and can accommodate unbalanced samples, which is essential for brain tumor segmentation. In this step, the noise and blur are eliminated. The next step is the segmentation of medical images. Segmented pictures are used for classification and reconstruction. As a result, we are able to acquire a precise medical image, which is then used to illustrate the results.

27.2 Literature Survey

The edge of the pictured item is generally varied in biomedical pictures, and the intricate patterns of the pictured item are common. In order to deal with segmentation for objects with detailed patterns, we are finding a general segmentation algorithm that can be applied to different kind of images [1]. Long et al. [2] in order to create complete segmentation, a skip architecture was designed to merge the high-level graphics from the deep decoding layer with the presence graphics from the shallow encoding layer. This method has proven effective for natural photography and can also be used for biomedical imaging. To overcome the cell tracking problem, Ronneberger et al. [3] offered the U-Net, which uses skipping the architecture. The AlexNet proposed by Krizhevsky et al. [4] was the most successful at image classification of the training set of ImageNet, making convolutional neural networks become the key research object in computer vision, and this research continues to deepen. Milletari et al. [5] suggested the V-Net of the U-Net network structure as a 3D deformation structure. The decoder maps the low-resolution discriminative feature map learned by the encoder to the high-resolution pixel space to realize the category labeling of each pixel. SegNet [6]. Zhao et al. [7] proposed the pyramid scene parsing network (PSPNet). Through the pyramid pool module and the proposed pyramid scene parsing network, it aggregates the ability to mine global context information based on the context information of different regions. Another important segmentation model is mask R-CNN. Faster R-CNN [8] is a popular target detection framework, and mask

R-CNN extends it to an instance segmentation framework. These are used for image segmentation very classic network model. Furthermore, there are other methods of construction, such as those done by recurrent neural network (RNN), and the more meaningful weakly supervised methods. DeepLab [9] solved the difficulty of segmentation caused by differences of the same object scale in the same image. Nyul et al. It is difficult to delineate the boundaries of tumors on magnetic resonance imaging due to the similarity of the tumors [4]. Convolutional neural networks (CNNs) have been used for the segmentation of brain tumors by multimodal magnetic resonance imaging [10]. The CNN model is a progressive strategy that is associated with the extraction and classification of features in a single model. Waiting in Pereira. Lang and Zhao [11], the author projected a deepCNN model with a smaller 3×3 convolutional nucleus to segment tumors on MRI images. These CNN techniques [12] are patch-based methods that can mark pixels and classify the patches placed by pixels. Also, because the gradient disappears throughout the training process, the CNN system is not easy to train. When the depth of the network increases, the problem of deprivation will appear, and the precision will be overwhelmed [13–15]. The ResNets idea is based on the accumulation between the output of layer and its input. This simple change promotes the training of deep networks because some fast contacts correspond to their regular convolutional layers [15, 16].

27.3 Methodology

27.3.1 Deep Convolutional Network Using U-Net

The network model is created in U-Net and contains a downstream sampling path (encoding) and an upstream sampling path (decoding). There are five convolutional blocks in the sampling reduction path. Each block includes two convolutional layers, and each block is composed of a 3×3 filter, a bidirectional length of 1 step, and the rectifier activation, so that the total number of maps of features reaches 1024. The maximum grouping with a step size of 2×2 is useful at the end of each block, but not in the last block used to reduce resolution, and the size of the feature map is minimized from 240×240 to 15×15 . Each block in the up-sampling path starts with the size of the 3×3 filter in the deconvolution layer of 2×2 steps, while reducing the number of feature maps by two, doubling the size of the feature maps in both directions, and it complete the feature map is also provided. The size is from 15×15 to 240×240 . The two convolutional layers minimize the number of feature maps from the combination of the deconvolution feature maps and encoding path feature maps in each up-sampling block. Here, it engage zero padding for all the convolutional layers to retain the output dimension for both the down-sampling and up-sampling paths, unlike the original U-Net design [3]. In the network, no completely linked layer is included (Fig. 27.1).

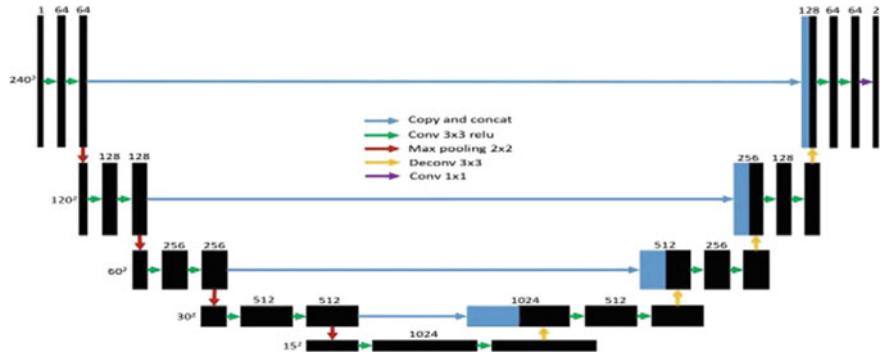


Fig. 27.1 U-Net architecture

The main advantages of U-Net are it can train with a small number of annotated images, making it suitable for image segmentation in biology or medicine, and it doesn't require several executions to accomplish image segmentation.

The limitation of this network is sluggish since each patch requires its own network, and due to overlapping patches, there is a lot of laying-off. There is a trade-off among precision and context utilization in localization.

27.3.2 Deep Convolutional Network Using AlexNet

The model has an eight-layer structure and offers its own set of picture classification benefits [17]. The structure of AlexNet is seen in Figure 27.2.

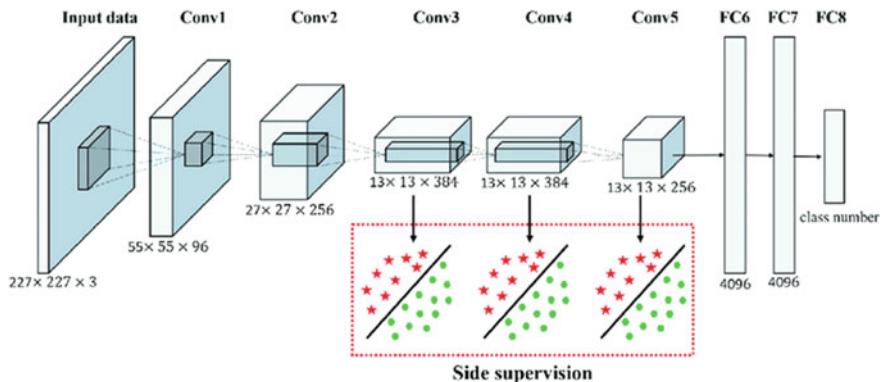


Fig. 27.2 AlexNet network architecture

The data source's input format is $227 \times 227 \times 3$ pixels, where 227 pixels represent the input image's width and height, and three represents the data source's three-channel RGB mode, which chains color pictures in usually used arrangements and removes the need for extra setups for the novel data source composed cropped.

Convolution (Conv), rectified linear unit (ReLU), maximum pooling (max-pooling), and normalization are the top two layers' computing operations (normal). The second layer's output is exposed to convolution for 256 feature maps, of 5 kernel size, 1 for stride, and the rest of the settings the similar as the first layer. Only, convolution and ReLU operations were achieved by the third and fourth layers. The fifth layer is indistinguishable to the first, but it hasn't been regularized. Change the fifth layer's result to a long vector and use a three-layer fully linked model to input it into a standard neural network.

27.3.3 Deep Convolutional Network Using V-Net

Figure 27.3 depicts the V-Net structure. The dice coefficient loss function is used instead of the usual cross-entropy loss function in the V-Net construction. It convolves the picture using a 3D convolution kernel and lowers the channel dimension with a $1 \times 1 \times 1$ convolution kernel. An increasingly compressed path, separated into numerous phases, may be found on the network's left side. There are one to three convolutional layers in each step. To make each step learn a parameter function, the input and output of each stage are combined to achieve residual function learning. Each stage of the convolution procedure uses a convolution kernel that is 5×5

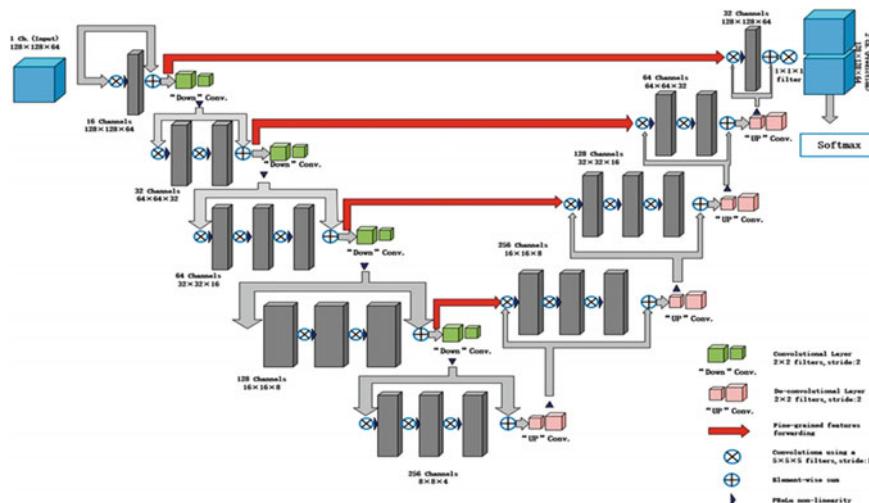


Fig. 27.3 V-Net architecture

× 5. The convolution process is used to extract data features while also reducing the resolution of the data at the end of each “stage” using the proper step size. A gradually decompressed route may be found on the network’s right side. It collects and combines required information to provide dual-channel volume segmentation by extracting features and expanding the spatial support of lower resolution feature maps. The network’s final output size is the same as the initial input size.

27.3.4 Deep Convolutional Network Using ResNet

He et al. [18] developed the ResNet network, which are constructed on deep network that have demonstrated strong convergence tendencies and convincing correctness. ResNet was created using many stacked residual units and a variety of layer counts: 18, 34, 50, 101, 152, and 1202. The amount of operations, on the other hand, might vary depending on the architecture [18]. The residual units are made up of convolutional, pooling, and layering for all of the above. ResNet [7] is comparable to VGGNet [8]; however, ResNet is roughly eight times deeper. Authors presented different ResNet configurations with 18, 34, 50, 101, and 152 levels in the ResNet model [9]. Each ResNet layer is made up of several blocks. ResNet 50 is used for segmentation in our model because it has deeper layers than ResNet 34 and has fewer restrictions than ResNet 101, 152, which shortens training time. Figure 27.4 shows the ResNet 50 network.

Without increasing the training error percentage, the networks with big amount (even thousands) of layers can be trained easily. ResNets help in undertaking the disappearing gradient problematic using uniqueness mapping. The restrictions of ResNet are improved difficulty of network and application of batch regularization layers since ResNet deeply be contingent on it.

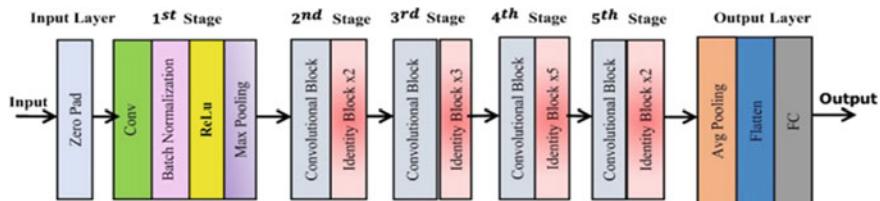


Fig. 27.4 The structure of the ResNet 50

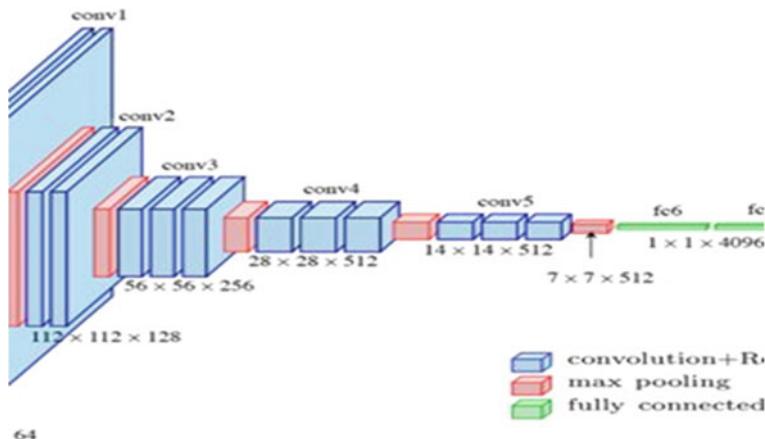


Fig. 27.5 VGG16-Net architecture

27.3.5 Deep Convolutional Network Using VGG16-Net

VGG16 is convolutional neural network [4] model. Modify AlexNet by successively replacing the filters with large kernel sizes (11 and 5 in the first and second convolutional layers) with many filters of 33 kernel sizes.

The input to the cov1 layer is a 224×224 picture with a static size. The picture is processed through a heap of convolutional (conv.) layers with an extremely tiny interested field: 3×3 (the minimum size to detention the notions of left/right, up/down, and center). The 1×1 convolution filter in one of these settings can be thought of as a linear change of the input channel. For 3×3 convolutional space filling, the convolution step size is set to 1, and the convolution layer input is set to 1 pixel. Preserve the layers of spatial resolution for subsequent convolution.

The maximum grouping is done with a step size of 2 in a window of (2×2) pixels. The following is a group of convolutional layers (different in depth in many designs), and three fully connected (FC) layers are additional. In all networks, fully connected layers are arranged similarly (Fig. 27.5).

27.4 Result and Discussion

Executing these five models on a different and not dependent testing dataset might give a more impartial assessment. Second, there are a few parameters in five networks that need to be fine-tuned. Recently, all of the parameters have been set on by pragmatic research. It is not discovering a substantial performance increase after adding L1, L2, or dropout to the network for regularization. The five models were trained with a real picture alteration, and it is impossible to overfit the network with a

Table 27.1 List of deep learning model in ImageNet large scale visual recognition challenge

Architecture	Article	Top 5 error rate (%)	Number of parameters (million)
U-Net	Ronneberger et al. (2015)	14.3	35
AlexNet	Krizhevsky et al. (2012)	16.4	60
VGGNet	Simonyan et al. (2014)	7.3	138
ResNet	Kaiming et al. (2016)	3.57	25.6
V-Net	Milletari et al. (2016)	3.06	35

significant amount of training data. The problem might be resolved by loading all multimodal MRI and doing joint training using some others datasets. Despite these flaws, the se five approach has shown promising segmentation results that are both efficient and effective (Table 27.1).

27.5 Conclusion and Future Scope

In computer-assisted diagnosis, clinical trials, and treatment planning, computer-aided segmentation is a critical step. In recent years, a number of methods for segmenting MR images have been presented, each with its own set of advantages and disadvantages. In this project, we discussed the fundamentals of segmentation methods and their characteristics.

Here, we analyze five fully automated brain tumor identification and segmentation approach based on U-Net, AlexNet, V-Net, ResNet, and VGG16 deep convolutional networks in this study. We have proven that all techniques can offer both well-organized and healthy segmentation when likened to the hand defined ground truth using tests on a well-established dataset that contains brain tumors (or glioma). Furthermore, compared with these techniques, deep convolutional networks can produce equivalent results for the entire tumor area and better results for the center of tumor area. The five-fold cross-validation technique is used for validation in this paper; nevertheless, we can see a simple utilization on not dependent testing datasets as well as claims for different datasets. These suggested techniques allow for the automatic generation of a patient-definite brain tumor segmentation method, which might lead to objective lesion evaluation for clinical activities including analysis, treatment preparation, and patient monitoring.

It is possible that these five algorithms will be tweaked in the future. As a result, we may obtain a quick segmentation method for medical imaging. The study demonstrated that each segmentation approach has its own set of advantages and disadvantages. There is no one algorithm that can divide all of the areas in MRI brain images. As a result, the best feasible methodology, or a combination of approaches, must be found to help in the automated segmentation of all parts. This is due to the fact that the constitution of various organs varies.

References

1. Lateef, F., Ruichek, Y.: Survey on semantic segmentation using deep learning techniques. *Neurocomputing* **338**, 321–348 (2019)
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440 (2015)
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241 (2015)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
5. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision*, IEEE, pp. 565–571 (2016)
6. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Machine Intell* **39**(12), 2481–2495 (2017)
7. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890 (2017)
8. Too, E.C., Yujian, L., Njuki, S., Yingchun, L.: A comparative study of fine-tuning deep learning models for plant disease identification. *Comput. Electron. Agric.* **161**, 272–279 (2018)
9. Sergey Ioffe, C.S.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning ICML*, pp. 448–456 (2015)
10. Pereira, S., Pinto, A.: Brain tumor segmentation using convolution neural networks in MRI images. *IEEE Trans. Med. Imaging* **1240**–1251 (2015)
11. Lang, R., Zhao, L.: Brain tumor image segmentation based on convolution neural network. In: *International Congress on Image and Signal Processing Biomedical Engineering and Informatics*, pp. 1402–1406 (2016)
12. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: Brain tumor segmentation and radiomics survival prediction. In: *Brats 2017 Challenge International MICCAI Brainlesion Workshop*, pp. 287–297. Springer (2017)
13. Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation. *J. Digit. Imaging* **32**, 582–596 (2019)
14. Altaf, F., Islam, S.M.S., Akhtar, N., Namjua, N.K.: Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access* **7**, 99540–99572 (2019)
15. Ma, Z., Tavares, J.M.R.S., Jorge, R.M.N.: A review on the current segmentation algorithms for medical images. In: *International Conference on Imaging Theory and Applications* (2009)
16. Yu-Qian, Z., Wei-Hua, G., Zhen-Cheng, C., Tang, J.-T., Li, L.-Y.: Medical images edge detection based on mathematical morphology. *IEEE Eng. Med. Biol.* **6492**–6495 (2006)

17. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)

Chapter 28

A Review of Challenges and Solution in Peer-to-Peer Energy Trading of Renewable Sources



Ritweek Das, Stuti Snata Ray, Gayatri Mohapatra, and Sanjeeb Kumar Kar

Abstract Peer-to-peer energy trading is the futuristic approach to conserve and trade renewable sources of energy in an economical manner. Peer-to-peer trading leads to a decentralized open free market which benefits both the prosumers who have surplus energy and the consumer with energy deficit. This paper provides a review on the challenges, outcomes, solutions and future research that should be conducted in this area. The various challenges are integrating generation, transmission in a large scale, efficient control of microgrid, developing smart energy meter, complex behavior of prosumers and consumers. The areas of consideration by the previous researchers are game theory, simulation, trading platform, blockchain, optimization and algorithms. We provide a solution by creating a POWERITER cryptocurrency to trade power within local microgrid within blockchain ecosystem for transparency and secured features. It will be a public distributed ledger with proper timestamp consensus. The miners will be rewarded POWERITER tokens to maintain the ledger. This paper will help the researchers for qualitative and quantitative analysis of peer-to-peer energy trading technologies. It is a relatively new topic; there must be a further research to implement this concept in the real-time environment.

Nomenclature

P2P	Peer to peer
Prosumers	People who produce and consume simultaneously
DER	Distributed energy resources
USD	United States Dollar
ESD	Energy storing devices
ICT	Information and communication technology

R. Das · S. S. Ray · G. Mohapatra (✉) · S. K. Kar

Department of Electrical Engineering, Siksha 'O' Anusandhan (Deemed to be University),
Bhubaneswar, Odisha, India

S. K. Kar
e-mail: sanjeebkar@soa.ac.in

28.1 Introduction

Energy sector is the most crucial and important paradigm for infrastructure, economic development and welfare of nations [1]. Traditionally, we depend on the conventional sources of energy such as coal, natural gases, oil, hydro and nuclear power which lead to pollution. The researchers are constantly working on the renewable sources of energy to meet the increasing demand of electricity. According to IBEF report, “by 2022, solar energy is estimated to contribute 114 GW, followed by 67 GW from wind power and 15 GW from biomass and hydropower. The target for renewable energy has been increased to 227 GW by 2022” [2].

The global investment in the power sector as shown in Fig. 28.1 is increasing every year. Since there is a huge investment in the production of electricity, there is monopoly of big corporation. They can manipulate the price according to their connivance which will affect the consumers [3]. There must be a free open market to trade electricity which will lead to competitive pricing [4]. Energy trading benefits the corporation such as increasing the overall efficiency and reducing operating cost.

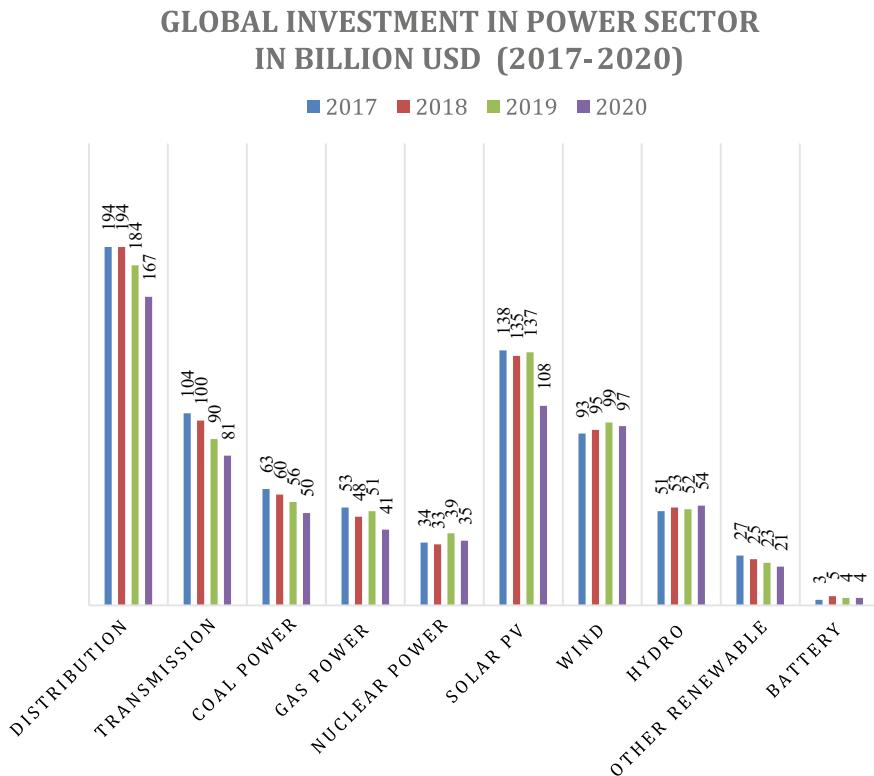
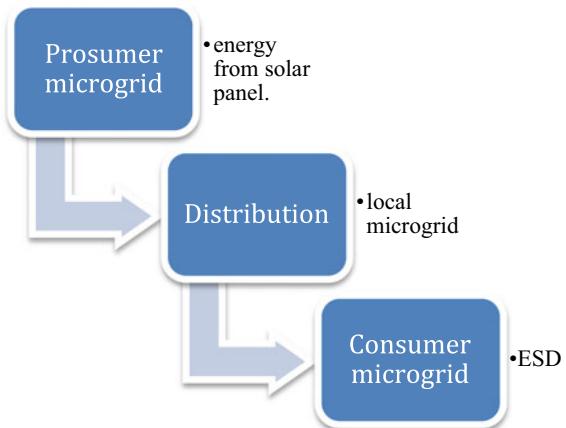


Fig. 28.1 Global investment in power sector

Fig. 28.2 Trading architecture in microgrid



The p2p transfer of electricity is initially possible for prosumers. The generation of electricity is unpredictability and intermittent characteristic of DER [5]. The prosumers have two choices either they can store energy or they can sell it to energy deficit region. The major challenge is to balance the demand and supply effectively [6]. The p2p architecture for transfer to energy has been classified into two types that are trading architecture and optimization model.

28.2 Overview of Peer-to-Peer Architecture

28.2.1 *Trading Architecture*

The energy trading architecture is the decentralized transfer of power among the peers without involving any single entity. In this, the prosumers generate electricity in a small scale from the rooftop solar panel. When prosumers have surplus of energy, they can sell to consumers in the power grid as given in Fig. 28.2. There are four layers for p2p energy trading, namely business layer, control layer, ICT layer and power grid layer.

28.2.2 *Optimization Model*

In power systems, consumer demand varies from time to time. During the peak hours, the load can cause frequency deviation which can lead to system failure. The utility companies always balance the demand and supply by load scheduling and price-based optimization.

Load scheduling is achieved through various methods including interruptible load, direct load control (DLR) which allows the power company to manipulate customer appliances and demand side bidding (DSB) where the customer is allowed to prepare a bid.

There are three types of optimization model.

Centrally controlled model

In this model, there is a central controller which connects two local grids. It can optimize the availability in the peak hours. This model is a non-profitable model because the local cannot compete with the utility company pricing. This model can only be implemented by utility companies, and it can be used in rural areas.

Incentive-driven model

The incentive model helps to reduce the transportation cost. The prosumers get incentive to contribute more in an ecosystem. The grid operator coordinates energy sharing by matching supply and demand between sellers and buyers as well as by providing the infrastructure for energy routing. The incentive can be in form of price, auction and smart contracts.

Cooperative-based model

Cooperative model allows a base station having local renewable energy to perform energy trading with the main grid based on coordinated multi-point communication powered by smart grids. “It minimizes the energy cost of cellular systems by providing an algorithm that allows the BS and the smart grid to manage the two-way energy trading. The demand from BS varies over time due to the fluctuation in power generated from renewable energy sources. The proposed model allows the BS to sell electricity to the grid when the power generated from renewable energy sources is excess and to buy power from the grid when the production is low” [6].

28.3 Comparison with Existing Models

We have done the tabulation of the following data from the proposed models [7–11] as given in Table 28.1.

28.4 Challenges in the Proposed Model

Energy management in smart grid

There are many challenges while depending on only renewable or traditional sources of energy. There is a challenge in optimizing peak demand and load dependency. In traditional system, the operating cost is very high [12]. In the renewable system,

Table 28.1 Comparison of the previous proposed models and their results

Project	Country	Year	Objective	Network	Purpose	Result	Comment
Piclo	UK	2014	It was made to help suppliers	National	Business	P2p trading platform	Not for local markets
Vandeboron	Netherlands	2014	It benefits suppliers	National	Business	P2p trading platform	Not for local market
Peer energy cloud	Germany	2012	Cloud-based trading platform	Microgrid	Energy network	Cloud-based platform	No discussion on control system
Smart watts	Germany	2011	Optimization of energy supply	Regional	Energy network	Smart meter interface	No discussion on control system
Yeloha and mosaic	USA	2015	Solar sharing network	Regional	Business	Terminated due to funding issues	Not for local market
SonnenCommunity	Germany	2015	Energy trading with storage system	National	Energy network	Online trading platform	Not for local market
Lichtblick swarm energy	Germany	2010	IT market for supplier and consumer	National	Energy network	Many services provided by suppliers	Not for local market
Transactive grid	USA	2015	Microgrid blockchain	Microgrid	Energy network	Automatic trading	Less interactive
Electron	UK	2016	Billing and energy metering	Unknown	Energy network	Not started	Not started

the initial cost is high for households. The transmission is bit complicated, and the system is less efficient.

Security, transparency and transaction time

In traditional system, there is less transparency. The transaction speed is very slowly. The security in a transaction is comprised by utility operators. There are many new technologies which have almost solved the problem like blockchain, smart contract, IoT and AI, cryptocurrencies, etc. [13].

28.5 Proposed Solution

In this model, we are using solar energy for energy management for multiple homes and transfer of electricity from one home to other in case of energy deficiency faced by the neighborhood grids. The primary goals are to maximize financial benefit and reduce the peak of the upstream grid [14]. This way, the proposed manages each home energy hub's energy generation and storage, as well as energy purchase and sale, on two levels: lower and upper level. As the solar panel is set to full power point monitoring (MPPT) so we will connect a DC-DC converter to extract maximum power from the solar panel. This will operate the solar panel at maximum power and will step down to supply the HVDC bus. We have used a bidirectional battery which is beneficial in rainy days [15]. Then, we will connect an inverter that will have input as DC and will convert it to 230 V AC which will be supplied to home. A common AC bus bar to the houses will transfer the energy when needed [16].

28.6 Future Outcomes

The p2p architecture will be the integral part of the power system. There will be highly optimized smart grid which will reduce the operating cost and increase the efficiency. There will be complete dependency on renewable sources of energy to reduce carbon footprint in the ecosystem. The new technologies will create a transparent, secured method of payment [17]. The algorithm of various technologies can be applied in the power system which will helpful in generation, transmission and consumption. As we know now government is promoting e vehicle and there is huge opportunity we can use our system to charge electric cars [18].

28.7 Conclusion

We are designing a business model which will be helpful for both consumer and suppliers. Our POWERITER cryptocurrency can be used in energy sector to trade

surplus power among prosumers. We have free market to have competitive pricing which will benefit the consumers. We have given the incentive to suppliers (miners) to generate POWERITER cryptocurrency and make profit while trading it with fiat currency. Everything is transparent because of public distributed ledger blockchain. We promote the use of renewable energy by giving incentive to both buyer and supplier.

References

1. IEA: World Energy Investment 2020. IEA, Paris (2020). <https://www.iea.org/reports/world-energy-investment-2020>
2. India and India, P.: Power Sector in India: Market Size, Industry Analysis, Govt Initiatives | IBEF (Govt. Trust). [Online]. Ibef.org (2021). Available at: <https://www.ibef.org/industry/power-sector-india.aspx>. Accessed 28 Mar 2021
3. Malik, A., Ravishankar, J.: A review of demand response techniques in smart grids. In: Proceedings of the Electrical Power and Energy Conference (EPEC), Ottawa, ON, 12–14 Oct 2016, pp. 1–6. [Google Scholar]
4. Yingdan, F., Xin, A.: The review of load scheduling model research based on demand response method. In: Proceedings of the 2013 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), Kowloon, 8–11 Dec 2013, pp. 1–5. [Google Scholar]
5. Xu, J., Zhang, R.: Cooperative energy trading in comp systems powered by smart grids. IEEE Trans. Veh. Technol. **65**, 2142–2153 (2016). [Google Scholar] [CrossRef]
6. Zhang, C., Wu, J., Cheng, M., Zhou, Y., Long, C.: A bidding system for peer-to-peer energy trading in a grid-connected microgrid
7. SonnenCommunity Website. Available at: <https://www.sonnenbatterie.de/en/sonnenCommunity>
8. Swarm Energy, Lichtblick Website. Available at: <https://www.lichtblick.de/privatkunden/schwarm-energie>
9. TransActive Grid Project Website. Available at: <http://transactivegrid.net>
10. Electron Website. Available at: <http://www.electron.org.uk>
11. Yeloha Website. Available at: <http://www.yeloha.com>
12. Samadi, P., Wong, V.W., Schober, R.: Load scheduling and power trading in systems with high penetration of renewable energy resources. IEEE Trans. Smart Grid **7**, 1802–1812 (2016). [Google Scholar] [CrossRef]
13. Mousa, M., Javadi, M., Pourmousavi, S.A., Lightbody, G.: An advanced retail electricity market for active distribution systems and home microgrid interoperability based on game theory. Electr. Power Syst. Res. **157**, 187–199 (2018). [Google Scholar]
14. Nguyen, P.H., Kling, W.L., Ribeiro, P.F.: Smart power router: a flexible agent-based converter interface in active distribution networks. IEEE Trans. Smart Grid **2**, 487–495 (2011). [Google Scholar] [CrossRef]
15. Zhu, T., Xiao, S., Ping, Y., Towsley, D., Gong, W.: A secure energy routing mechanism for sharing renewable energy in smart microgrid. In: Proceedings of the 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), Brussels, 17–20 Oct 2011, pp. 143–148. [Google Scholar]
16. Brocco, A.: Fully distributed power routing for an ad hoc nanogrid. In: Proceedings of the 2013 IEEE International Workshop on Intelligent Energy Systems (IWIES), Manchester, 5–7 Dec 2013, pp. 113–118. [Google Scholar]

17. Pegueroles-Queralt, J., Cairo-Molins, I.: Power routing strategies for dense electrical grids. In: Proceedings of the 2014 11th International Multi-Conference on Systems, Signals & Devices (SSD), Barcelona, 11–14 Feb 2014, pp. 1–6. [Google Scholar]
18. Zhang, J., Seet, B.-C., Lie, T.-T., Foh, C.H.: Opportunities for software-defined networking in smart grid. In: Proceedings of the 2013 9th International Conference on Information, Communications and Signal Processing (ICICS), Tainan, 10–13 Dec 2013, pp. 1–5. [Google Scholar]

Chapter 29

Data Transfer Using Light Fidelity (Li-Fi) Technology—A Methodological Comparative Study



Lirika Singh, Manish Rout, J. S. Bishal, and Jayashree Ratnam

Abstract Wireless communication serves as the backbone in day-to-day life activities due to its ubiquitous nature. The technology of using light wave carriers for carrying data in a wireless medium is termed as light fidelity (Li-Fi). Li-Fi is a bidirectional wireless communication/network mechanism, typically using the visible spectrum. A Li-Fi system provides enhanced bandwidth, better signal quality, good connectivity as well as secure connection, compared to traditional wireless fidelity (Wi-Fi) system. An optical carrier (in Li-Fi) can be modulated hundred times faster than radio frequency carrier (in Wi-Fi) and cannot be intercepted through electromagnetic induction (EMI) resulting in an inherently secure, high-speed system. In this paper, we discussed some of the data transfer methodologies reported in recent literature, which incorporate Li-Fi-based communication. The brief review presented therein helps in identifying a suitable methodology for any given application as well as in understanding the implications of choosing a particular setup. The potential of each of the implementation is immense and can be further improvised in terms of data accuracy, response time (high speed), and coverage area by adopting state-of-the-art technologies. In this paper, we have examined case studies covering four typical applications of Li-Fi systems/networks and stated our observations.

29.1 Introduction

The role of communication has become very crucial in our day-to-day life, especially the wireless communication. In this paper, we present the salient features of different methodologies employed in implementing an evolving wireless communication technology known as light fidelity (Li-Fi). Li-Fi is a wireless communication technology in which, light-waves or optical frequency ($\sim 10^{14}$ Hz) carriers are utilized to achieve data transfer. Although, optical frequency spectrum covers both visible

L. Singh · M. Rout · J. S. Bishal · J. Ratnam (✉)

Department of Electronics and Communication Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India
e-mail: jayashreeratnam@soa.ac.in

and infrared regions, Li-Fi technology uses carriers from visible light region, especially in the indoor environment. An optical frequency carrier can be modulated hundred times faster than radio frequency carrier (used in Wi-Fi) and also cannot be intercepted through electromagnetic induction (EMI), thereby offering an inherently secure communication system. Since the signal transmission takes place in unguided (free space) medium, it is also referred to as optical wireless communication, especially in the indoor environment. Recent literature reveals that Li-Fi is highly suitable for electromagnetically sensitive scenarios, such as airplanes, thermonuclear energy plants, chemical plants, hospitals, or emergency clinics, and also underwater communication [1, 2].

In near future, Li-Fi will be a better alternative for Wi-Fi, as it is easily accessible, faster, more secure, and cheaper than Wi-Fi. Since Li-Fi transfers data in the form of intensity-modulated light signals, it is also called as visible light communication (VLC). Light emitting diodes (LEDs) are the major sources for VLC technology, which apart from illumination, are also be used for simultaneous data transmission. LED arrays are used in studies conducted at University of Edinburgh to enable parallel transmission of multiple channels. Further, constituent R/G/B colors of white light are used to frequency encode distinct data channels, thereby enhancing the overall data rate [3].

Most of the Li-Fi prototypes have targeted mainly toward transmission of textual information and their improvisations [4–7]. However, there are some studies consider voice/image transfer between computing terminals by making use of microcontrollers [8, 9]. However, recent literature reports studies on examining the potential of Li-Fi technology in realizing Internet of things and vehicle-to-vehicle communication [10, 11]. In this paper, we carried out a comparative study on different Li-Fi system setups reported in recent literature in order to find out which conditions were best suitable for high data accuracy, good response time (high speed), and over larger distances. In this paper, we have examined various aspects of Li-Fi systems under four different case studies in Sect. 29.2, compared them in Sect. 29.3 and stated our observations and remarks in Sect. 29.4.

29.2 Implementation Methodologies

The main purpose of carrying out this study is to compare the different Li-Fi-based system models and examine different parameters such as speed, accuracy, and distance. This helps one to choose the optimum set of parameters suitable for one of the given application scenarios, viz., an indoor network communication, a short-span building-to-building link connectivity, a vehicle-to-vehicle data transfer, a vehicle-to-lamp-post communication etc. We group the systems under study into three case studies and discuss about each as follows.

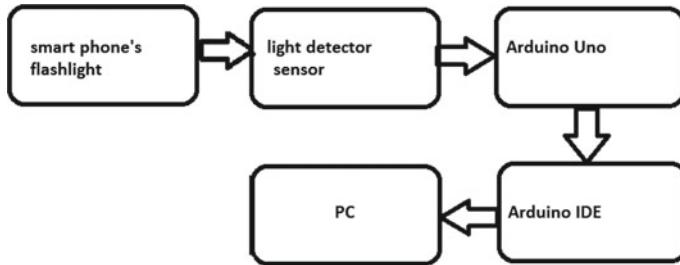


Fig. 29.1 Block diagram of smartphone-to-PC data transfer [4]

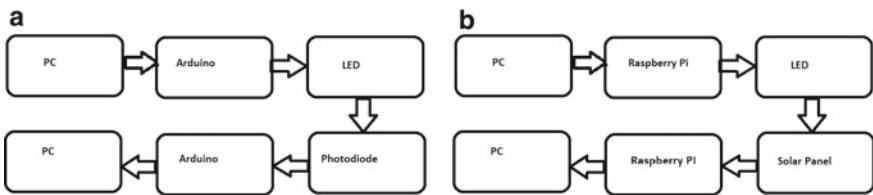


Fig. 29.2 **a** PC-to-PC data transfer (Arduino) [5]. **b** PC-to-PC transfer (Raspberry Pi) [6]

29.2.1 Case Study 1: Smartphone-to-Smartphone Communication

Here, we consider some Li-Fi systems [4–7] which broadly achieve data transfer between any two user terminals, which could be the ubiquitous smartphone/personal computer/laptop. The block diagrams of such Li-Fi systems are as shown in Figs. 29.1 and 29.2. In [4], communication is established between two smartphones or a PC and a smartphone by making use of the in-built flashlight and light sensor in the respective cameras. However, this proposed system achieves only 50 bps data bit rate even with a fast light sensor like BH1750, which has a better response time than a more common place light detecting resistor (LDR). The authors observed that enhancement to the data transmission rate up to 100 bps can be achieved by using a photodiode which has faster response time than BH1750. Another technique suggested was to use an array of LEDs at the transmitter end so that higher signal intensity improves higher received data rates. Further, pulse width modulated (PWM) binary data was found to result in higher data rates over the ON–OFF keyed (OOK) NRZ binary data.

29.2.2 Case Study 2: PC to PC Communication

Here, we consider various setups that have been used in recent times for implementing Li-Fi-based communication among personal computers. In [5], the system utilizes an

array of 8 (surface mounted devices-SMD) white LEDs as the light source, Arduino UNO devices/Arduino IDE software for encode/decode operations from text to digital bits conversion and a light detection sensor BPV10NF for photo detection as shown in Fig. 29.2a. Data transfer rates up to 147 bps with 100% accuracy were reported over a distance of 20 cm using this setup.

In [6], the setup consists of PCs employing Python programming for text to bit conversion and Raspberry Pi microcontroller (as shown in Fig. 29.2b) to interface with the light source and solar panel in the transmitter and receiver sections, respectively. The photo-detected analog signal from the solar panel is A/D converted before feeding it to Raspberry Pi circuit. The laboratory setup in [3] employs virtual system models like Proteus 7.0 to simulate the Arduino-based designs of Li-Fi systems.

In [12], the authors discuss the design procedure to determine the electronic components for optimum operation of the light source and photodetector. The system employs Net Beans IDE to create a Java application for sending and receiving the data. An Arduino board hosts the transceiver section as well as a serves as a medium to transfer software programs between the Arduino microcontroller and the PCs.

29.2.3 Case Study 3: Li-Fi for Internet of Things (Sensors)

The need for data collection from different sensing environments in order to take a comprehensive assessment regarding pollution, traffic control etc., finds useful role for Li-Fi technology in an IoT (Internet of things) platform. As seen in Fig. 29.3, a light dongle delivers optical sensory data from the server to three different user terminals which analyze the received data and update the relevant database [10].

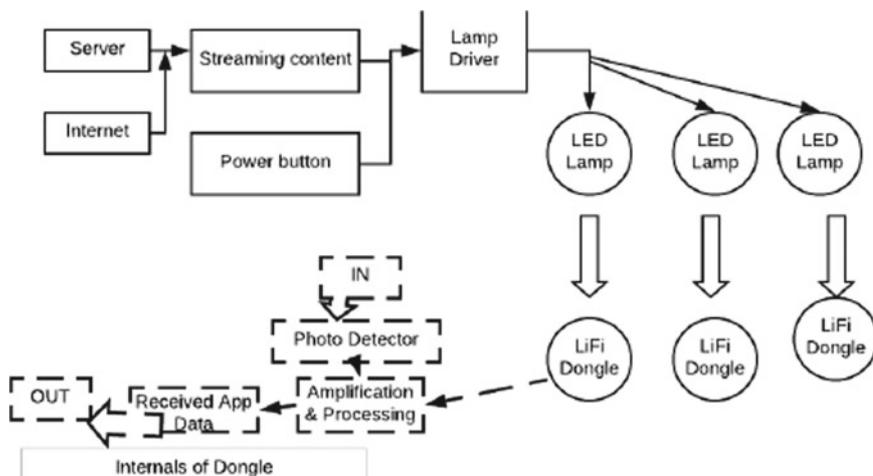


Fig. 29.3 Block diagram of Li-Fi-based communication for wireless sensors [10]

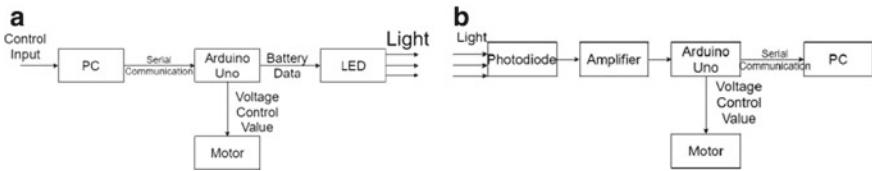


Fig. 29.4 **a** Transmitter side [11]. **b** Receiver side [11]

Each dongle consists of a detector and signal processing electronics. Further, a neural network is trained to help arrive at an appropriate subsequent action.

29.2.4 Case Study 4: Vehicle to Vehicle Communication

A more futuristic application for Li-Fi systems is in autonomous car environment or vehicle-to-vehicle communication. As seen in Fig. 29.4a, b, the vehicle is replicated with a motor in both transmitter and receiver sections. In [11], a full-fledged Li-Fi standard IEEE 802.15.7 is implemented at three layers which is capable of achieving rates up to 100 Mbps. Modulation techniques could be one of SCM or OFDM in order to achieve high spectral efficiency. Here, one DC motor's speed is controlled by the speed of another DC motor through an Arduino Uno-driven Li-Fi link. The data signal controlling the motor attached on transmitter side is transferred as optical signal through LED and converted to electrical signal at the photodiode on the receiver side. The received signal controls the motor on the receiver side. Arduino Uno on both sides provides the PC–motor connection. This serves as a low-end prototype for vehicle-to-vehicle communication.

29.3 Comparative Study

See Table 29.1.

Table 29.1 Main features of Li-Fi systems in different application scenarios [4, 6, 10, 11]

Li-Fi application	Source	Detector	Microcontroller/IDE	Modulation
Smartphone-to-PC	LED	LDR	Arduino/Raspberry Pi	OOK
PC-to-PC	White LED	Photodetector	Arduino/Raspberry Pi	OOK/PWM
IoT platform	LED array	Light dongle	Net Beans IDE	CSK/OFDM
Vehicle-to-vehicle	LED array	Solar panel	VSM Proteus 7.0	OFDM/CSK

29.4 Conclusions

This paper describes the various methodologies used to transmit information on a Li-Fi-based network connection among devices like smartphones, personal computers, sensing devices, or moving vehicles. The various laboratory (prototype) Li-Fi setups reported in recent literature are discussed under four different case studies. In all the cases, the setup makes use of a microcontroller device, viz., Arduino Uno or Raspberry Pi to provide connection between the user device and LED/LDR. A corresponding IDE platform on the PC or smartphone enables relevant software to be loaded into microcontroller devices, needed for generating and interpreting the transmitted data at the transmitter and receiver, respectively. Four different communication scenarios incorporating Li-Fi-based data transfer mechanisms, have been described, viz., smartphone to PC, PC to PC, IoT platform, and vehicle-to-vehicle communications. Some of the techniques that were suggested to improve the data transfer rates include using (i) array of LEDs, (ii) fast response photodetectors/solar panels, (iii) multi-carrier modulation schemes like orthogonal frequency division multiplexing (OFDM), sub-carrier multiplexing (SCM), and color shift keying (CSK).

There are numerous possibilities and several challenges to be solved since Li-Fi has its own problems; for instance, light cannot travel through opaque objects. Li-Fi can be interrupted by sunlight, when visible light spectrum is utilized for carrying data. On the otherhand, bad weather can affect Li-Fi operated on free space optical channels. We will see more Li-Fi devices for data transmission in the years to come on the market, both in the commercial and consumer sectors. A more prudent approach would be to use Li-Fi along with Wi-Fi in an integrated way to get the best of both worlds. For instance, Li-Fi can be deployed to provide an indoor service while Wi-Fi can extend it to outdoor over a short range.

References

1. Tsonev, D., Videv, S., Haas, H.: Light fidelity (Li-Fi): towards all-optical networking. In: Proceedings SPIE OPTO, Volume 9007, Broadband Access Communication Technologies VIII (2014)
2. Sarkar, A., Agarwal, A.S., et al.: Li-Fi technology: data transmission through visible light. *Int. J. Electron. Commun. Eng.* **3**(6), 1–12 (2015)
3. Deka, K., Bora, P., et al.: Design of a Li-Fi based data transmission system. *Int. J. Innov. Res. Sci. Eng. Technol.* **6**(9), 19325–19331 (2017)
4. Aldarkazaly, Z.T., Younus, M.F., et al.: Data transmission using Li-Fi technique. *Int. J. Adv. Sci. Technol.* **29**(3), 7367–7382 (2020)
5. Aldarkazaly, Z.T., Alwan, Z.S.: Transfer data from PC to PC based on Li-Fi communication using Arduino. *Int. J. Adv. Sci. Eng. Inf. Technol.* **11**(2), 433–439 (2021)
6. Neelopant, A., Yavagal, M., et al.: PC to PC data transfer using Li-Fi. *Int. Res. J. Eng. Technol.* **7**(8), 2224–2227 (2020)
7. Bakinde, N.S., Ifada, E., et al.: Li-Fi based technology for PC to PC data transmission. *Int. J. Inf. Process. Commun.* **8**(1), 153–162 (2020)

8. Ifada, E., Bakinde, N.S., et al.: Implementation of data transmission system using LiFi technology. In: 2nd International Conference of the IEEE Nigeria Computer Conference, pp. 1–7 (2019)
9. Rekha, R., Priyadarshini, C., et al.: Li-Fi based data and audio communication. *Int. J. Eng. Res. Technol.* **8**(5), 558–561 (2019)
10. Revathi, G., Sujana, G., et al.: Li-Fi based data transmission and analysis using IoT platform. In: International Conference on Physics and Photonic Processes in Nano Sciences. *J. Phys. Conf. Ser.* **1362**, 012025, 1–12 (2019)
11. George, R., Vaidyanathan, S., et al.: Li-Fi for vehicle to vehicle communication—a review. In: International Conference on Recent Trends in Advanced Computing. *Procedia Comput. Sci.* **165**, 25–31 (2019)
12. Pushkala, S.P., Renuka, M., et al.: Li-Fi based high data rate visible light communication for data and audio transmission. *Int. J. Electron. Commun. Eng.* **10**(2), 83–97 (2017)

Chapter 30

Study on Surveillance of Crop Field Using Smart Technique



Manesh Kumar Behera, Sobhit Panda, Sonali Goel, and Renu Sharma

Abstract The Indian economy is mainly reliant on agriculture and Indian farmers livelihoods are greatly reliant on the Monsoon rains. Agriculture-related issues have long posed a danger to the country's growth. The only way out to this problem is smart agriculture by revolutionizing the existing accustomed techniques of agriculture. This paper presents methods of crop saving from wild animals by using various techniques. Along with this, various sensors are used to monitor temperature, soil moisture and humidity of the agricultural land for controlling the irrigation. The real-time sensed data is stored in Google Sheets for decision making and controlling the irrigation. The system used is ESP8266 Wi-Fi module for simple and faster connectivity. This paper also presents solar-powered repelling units for saving the crops from wild animals. The system uses an advanced object detection technique called YOLOv3 to detect the elephant by making noise through an alarm system. The farmer can access all the information from anywhere in the world.

30.1 Introduction

In recent years, there has been a steep rise in population in India as well all across the world, and this has led many serious environmental problems like increase in pollution, shortage of water, deforestation, uneven rainfalls and erratic weather conditions. Agriculture is the backbone of Indian economy. However, the erratic weather conditions like uneven rainfall, increase in temperature and floods due to deforestation have been causing a great amount of loss to the agriculture. The steep growth in population has led to increase in human settlements in India that has caused a decrease in elephant

M. K. Behera · S. Goel (✉) · R. Sharma

Department of Electrical Engineering, ITER, SOA (Deemed to be University), Bhubaneswar, India

R. Sharma

e-mail: renusharma@soa.ac.in

S. Panda

Instrumentation and Electronics Engineering, CET, Bhubaneswar, India

habitats, so there is an increase in elephant human conflicts. With the increase of such conflicts resulted in destruction of crops and property. The crop damage accounts for major economic losses in India. But we can reduce crop damage and also maximize the agricultural production if we can do proper assessment of the effect of climatic changes on agriculture and built an adaptive farming system. Crop quality is based on the soil moisture, temperature and humidity. So, by using advanced technology like IoT, we can built a system that can monitor on all the essential information that affects the crop and provide a defense system against animals such as elephants and wild boar as well as provide an adaptive irrigation system. In future, smart farming is one of the applications of various advance concepts of IoT. Sensors can be used to collect real-time information about soil moisture, temperature and humidity. With the help of object detection techniques, we can identify the presence of elephants. In real time that can help to reduce potential damage to the crops by building an alarm system. An efficient monitoring and controlling system can be achieved through IoT.

This paper describes a crop field monitoring and protection system. The system uses sensors to monitor temperature, soil moisture and humidity of the agricultural land for controlling the irrigation. The real-time sensed data is stored in Google Sheets for decision making and controlling the irrigation. The system used ESP8266 Wi-Fi module for simple and faster connectivity. The system uses an advanced object detection technique called YOLOv3 to detect an elephant and has an alarm system to reduce human elephant conflicts. The farmer can access all the information from anywhere in the world.

Yuvaraju and Priyanga [1] proposed a method for smart irrigation system using Raspberry Pi-3, resistive soil moisture sensor, water-level sensors and color sensors. This method used the Arduino IDE and Proteus Professional. The main disadvantage of the system is it is not accessible through net and also the Raspberry Pi model is costlier. A PG student Rajalakshmi and Devi Mahalakshmi [2] proposed a model using Arduino Uno, DHT11 sensor, LDR, NRF2 4L01 transmitter and receiver to make smart crop field monitoring system, but the data cannot be accessed globally.

Ogunti, from University of Technology, Nigeria [3], proposed a model for smart crop monitoring system and irrigation using DHT11 sensor, soil moisture sensor and nodeMCU ESP8266 board. The main highlight of this system is that it uses THINGSPEAK to store the data online. The updating of data is relatively slow in things speak. Rao and Sridhar [4] in his paper proposed a model for smart crop monitoring and irrigation with help of Raspberry Pi-3 and soil moisture sensor. It used LM35 temperature sensor instead of DHT11 sensor. Data is stored on a local network, so it can only be accessible for that local network range.

Sukumar et al. [5] proposed a model for crop field monitoring and automatic irrigation using Arduino UNO. The main highlight of this model is use of GSM module for connectivity. But there is no provision for storing the data for future analysis. Pernapati [6] proposed a rather low cost smart irrigation system with the use of nodeMCU, soil moisture sensor, DHT11, relay and water pump. But it is not connected to the Internet and cannot be controlled globally.

Mishra and his group [7] proposed a system for automatic irrigation system with IoT-based approach, and it is an Arduino model. This model is not connected to

Internet as well. Vaishali and his group [8] proposed a model for smart irrigation and monitoring system using IoT. The model is a Raspberry Pi-3 model. The main highlight of this model is that it uses Bluetooth for connectivity using an app named BlueTerm, and this model also is slightly costlier for Raspberry Pi board and also is not connected to Internet.

Fazil and his group [9] proposed a model for elephant detection system based on IoT. The sensor panel used is a geophone used as low frequency underground (Seismic) sensor for detecting ground waves. It has a multiple power source which includes solar panel, battery pack (Li-ion battery), and it has a public alerting unit that alerts the villagers when the elephant is detected via sms. The main disadvantage is that it does not have any provision to protect the crops.

30.2 System Requirement

This system is aided with different hardware and software components. These are mentioned below:

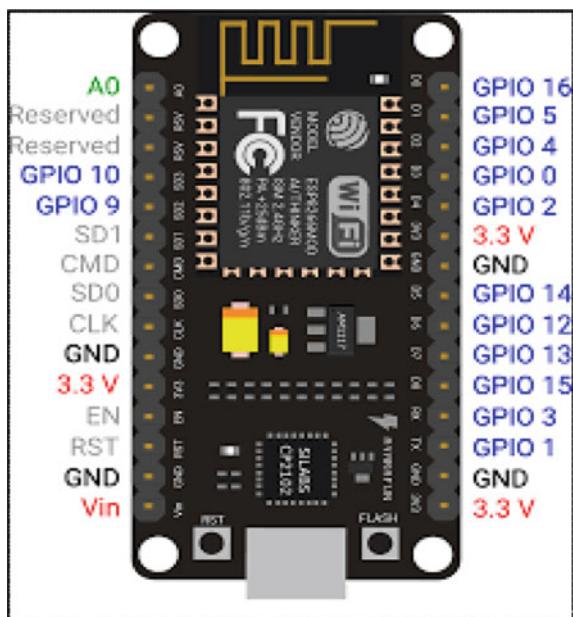
30.2.1 *Hardware Requirement*

The microcontroller mounted into a single printed circuit board is a single board microcontroller. This board offers all the requisite circuits for useful control functions: microprocessor, I/O circuits, clock generator, RAM, memory of the storage program and any required ICs. The aim is that the board is to be automatically beneficial to the application developer, without needing them to expend time and effort improving the controller hardware.

30.2.1.1 NodeMCU

NodeMCU is a low price open source IoT framework [10, 11]. Initially, it contained Espressif Systems' firmware running on ESP8266 Wi-Fi SoC, as well as the hardware-based ESP-12 module [12, 13]. ESP32 32-bit MCU support was introduced later on. NodeMCU is implemented in C and is layered on the NON-OS SDK of Espressif. The programming paradigm of NodeMCU is close to that of Node.js, with Lua only. It is event-driven and asynchronous. Therefore, various functions have parameters for callback functions. NodeMCU has a storage memory of 128 kB and a 4 MB to store data and program (Fig. 30.1).

Fig. 30.1 Pin configuration of NodeMCU



30.2.1.2 DHT11

The DHT11 is a simple and cheap digital temperature and humidity sensor. It employs a capacitive humidity sensor and a thermistor to detect the surrounding air and outputs a digital signal on the data pin. It is simple to use, but it needs careful timing in order to collect data. The only major disadvantage of this sensor is that we can only collect new data from it every one or two seconds (Fig. 30.2).

30.2.1.3 Capacitive Soil Moisture Sensor

The capacitive soil moisture sensor, like other sensors on the market, uses capacitive sensing rather than resistive sensing to detect soil moisture levels. It is composed of rust-resistant material, which means it will last for a long time. It can be inserted into the soil to inspire others with real-time soil moisture data. An ADC converter is required for Raspberry Pi compatibility. This soil moisture sensor pairs well with the 3-pin “Gravity” interface, which may be linked directly to the gravity I/O expansion shield (Fig. 30.3).

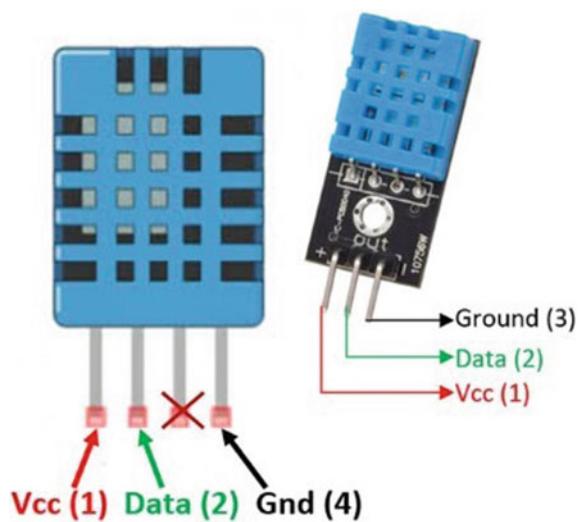
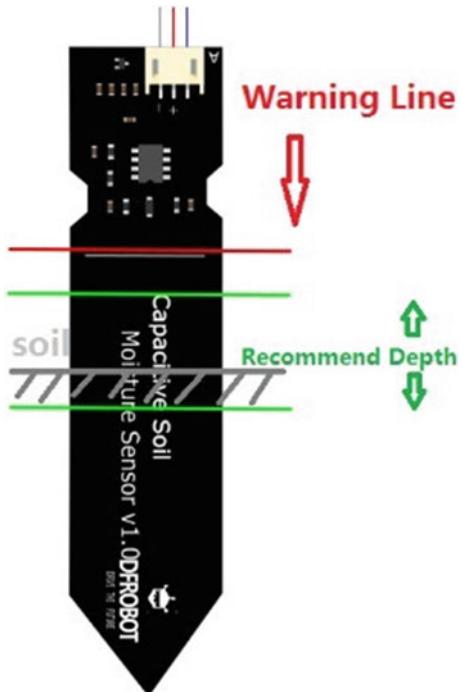
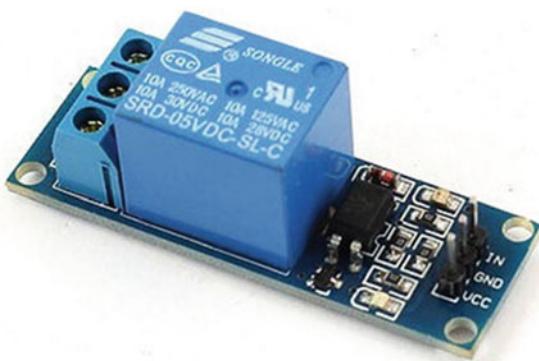
Fig. 30.2 DHT11 sensor**Fig. 30.3** Capacitive soil moisture sensor

Fig. 30.4 Single channel 5 V relay module



30.2.1.4 Single Channel 5 V Relay Module

This is the Arduino PIC AVR DSP ARM 1 Channel 5 V relay board module. It may be controlled by a wide variety of microcontrollers, including Arduino, AVR, PIC and ARM. Each requires a 15–20 mA driving current and is equipped with a high-current relay: DC 5 V/10 A, AC 250 V/10 A (Fig. 30.4).

30.2.2 Software Requirement

30.2.2.1 ARDUINO IDE

This is a C and C++-based cross-platform application (for Windows, macOS and Linux). It is used to create and upload programmes to Arduino-compatible boards, as well as other vendor development boards using third-party cores (Fig. 30.5).



Fig. 30.5 Arduino integrated development environment (IDE)

Fig. 30.6 Darknet 53 architecture

Type	Filters	Size	Output
Convolutional	32	3×3	256×256
Convolutional	64	$3 \times 3 / 2$	128×128
1x	Convolutional	32	1×1
Convolutional	64	3×3	
	Residual		128×128
	Convolutional	128	$3 \times 3 / 2$
	Convolutional	64	1×1
2x	Convolutional	128	3×3
	Residual		64×64
	Convolutional	256	$3 \times 3 / 2$
	Convolutional	128	1×1
8x	Convolutional	256	3×3
	Residual		32×32
	Convolutional	512	$3 \times 3 / 2$
	Convolutional	256	1×1
8x	Convolutional	512	3×3
	Residual		16×16
	Convolutional	1024	$3 \times 3 / 2$
	Convolutional	512	1×1
4x	Convolutional	1024	3×3
	Residual		8×8
	Avgpool		Global
	Connected		1000
	Softmax		

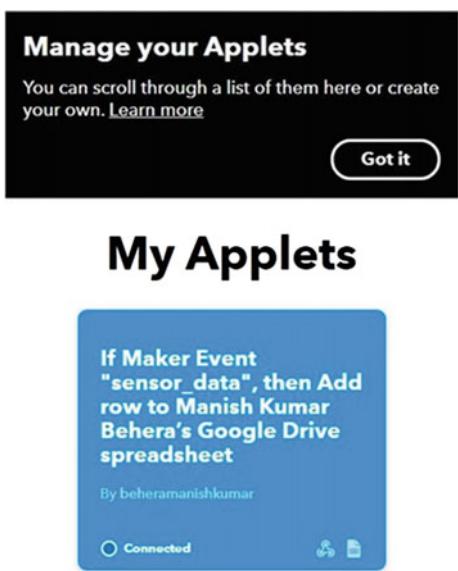
30.2.2.2 YOLOv3

In 2018, Redmon suggested YOLOv3 to use a solitary neural system to process the whole image. It divides the input image into grid cells of the same dimension, and for each grid cell, it predicts bounding boxes and probabilities. DARKNET-53, consisting of 53 convolution layers to capture deep features, is the base of the enhanced YOLO model and has been more popular than Darknet-19, ResNet-101 [14] (Fig. 30.6).

30.2.2.3 IFTTT

IFTTT gets its name from the programming conditional statement “if this, then that.” What the organization gives is a product stage that associates apps, gadgets and services from various developers so as to trigger at least one mechanizations including those applications, gadgets and services. The automations are practiced by means of applets—which are similar to macros that interface numerous applications to run computerized undertakings. You can turn on or off an applet utilizing IFTTT’s

Fig. 30.7 IFTTT applet interface



Web site or mobile applications (as well as the versatile applications' IFTTT gadgets). You can likewise make your own applets or make varieties of existing ones by means of IFTTT's easy to understand, direct interface (Fig. 30.7).

30.3 Proposed Model

The fundamental goal of our framework is to

- (i) Install a solar-powered intermittent high-frequency sound-generating device (sonic electronic repellent) to drive animals away from agriculture fields.
- (ii) Build up a smart protection and irrigation framework that can be accessed by the rancher remotely all through the world by means of IoT stage.

The solar-powered intermittent high-frequency sound-generating device (sonic electronic repellent) can be put in farmers' fields to drive away animals such as elephants, wild boars and monkeys from crop fields. These ultrasonic electronic repellents are incapable of being heard by humans, but it emits high-frequency sound waves, to repel wild animals, thereby saving crops. The repellent units will be put in farmers' fields that are particularly vulnerable to wild animals such as elephants, monkeys and wild pigs. These sound-generating devices may cover an area of up to 4–5 acres. These are powered by solar, battery and commercial electricity. It includes a solar panel, a frame, a battery and a device. The device comes with a battery, which has a backup time of 6–7 h. These devices are weatherproof, portable and self-installable and are ideal for outdoor use. Three sorts of noises are produced

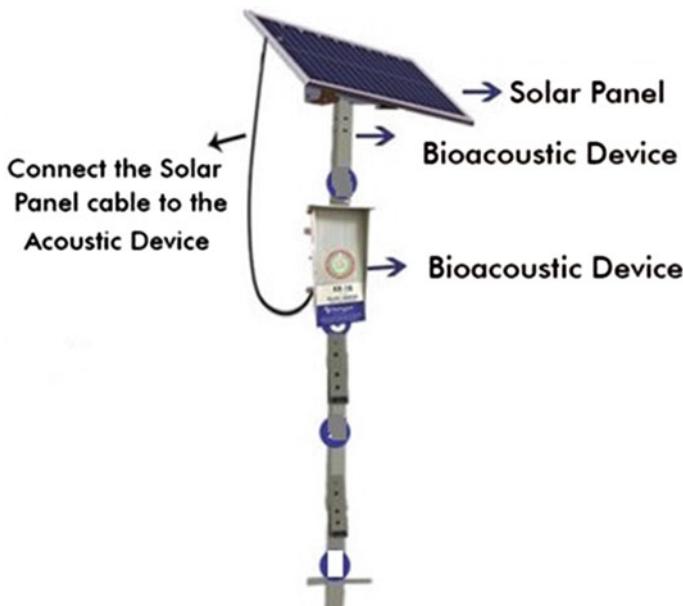


Fig. 30.8 Solar-operated bioacoustic device

by the solar-powered bioacoustics devices: Alarm, distress, and predator calls. The bioacoustics device is shown in Fig. 30.8.

Acoustics coverage

1. Latest call sequences deter various species of animals.
2. The coverage area depends on the ambient sound level.
3. One Q3 covers 5–15 acres and Q5 covers 2.5–3 acres of land.
4. This system is 80–90% effective.

Various authors across the globe have studied the solar-powered smart monitoring system [15–20]. For getting the real-time moisture data, temperature data, humidity data through capacitive soil moisture sensor and DHT11 temperature and humidity sensor, the proposed system is connected to nodeMCU. All these data get stored in Google sheet via IFTTT such that it can be accessible to the farmer from any remote location. This system also provides a smart protection system from the elephant, wild boar, etc. Using object detection technique YOLOv3, the system detects the presence of elephant in an image. If it detects the elephant, wild boar, etc., and it triggers a buzzer or an alarm via IFTTT.

The framework can keep the yields from water logging impact by estimating the dampness of the field by utilizing soil moisture sensor. It can forestall the loss of water that happens due to over irrigation. Temperature sensor can be utilized to quantify the temperature in the field and on the off chance that the harvest is excessively touchy, at that point it very well may be chilled off by sprinkling the water naturally.

The system protection system uses an alarm to alert the farmer and scare the elephant away from the crop field. This system prevents damaging of crop from the elephant.

This technique contains three areas:

1. The initial segment comprises setting up nodeMCU board and interfacing it with all the sensors.
2. The second part comprises building up an IoT stage and connecting it to IFTTT to send the information to the Google sheet.
3. The third part consists of configuring the YOLOv3 model; it includes training of the YOLO model on COCO dataset and enables it to detect elephant in the image and giving a trigger to ON the alarming system.

The block diagram shows that the DHT11 temperature, humidity and the soil moisture sensor is connected to NodeMCU which collects all the sensor data and then it sends to Google sheets via IFTTT using its inbuilt Wi-Fi module. The elephant detection system part consists of camera which captures the images and the YOLOv3 object detection model detects the presence of elephant in the image and sends a request to the IFTTT which triggers the buzzer. The block diagram of the proposed system is shown in Fig. 30.9.

The prototype of the proposed model as a hardware implementation is shown in Fig. 30.10. In this model, the DHT11 sensor and the soil moisture connected to the nodeMCU board which provides the necessary soil data like temperature, humidity and soil moisture and a relay is connected to the model which controls the switching of the water pump for irrigation.

To implement YOLOv3, the system requirements should be: 8 GB RAM and i5 processor. Though it can also be implemented on a Raspberry Pi using OpenCV because it is being quite slow with respect to the complete desktop setup. YOLO is

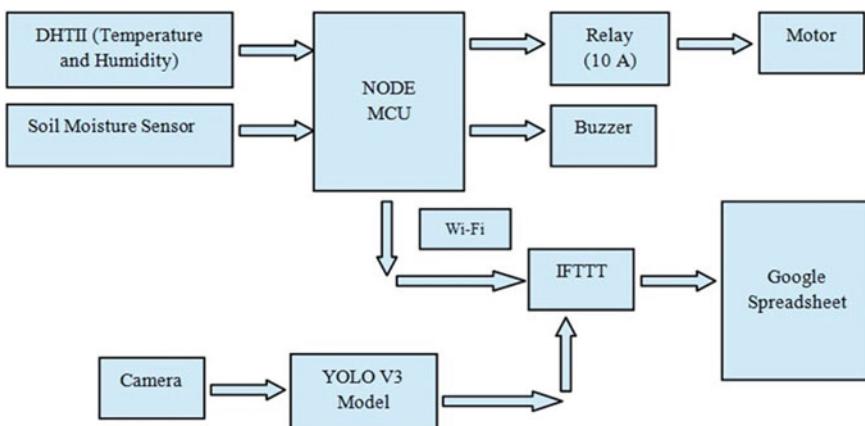


Fig. 30.9 Block diagram of proposed system

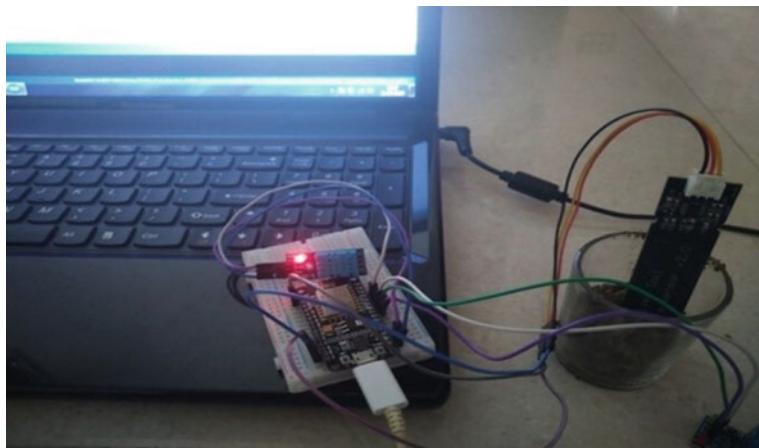


Fig. 30.10 Prototype of the proposed model

still a growing open source community, so the support for Raspberry Pi would come soon.

Basic System Requirement

Processor: Intel Core i5 8250U

Installed Memory (RAM): 8.00 GB

System type: 64-bit

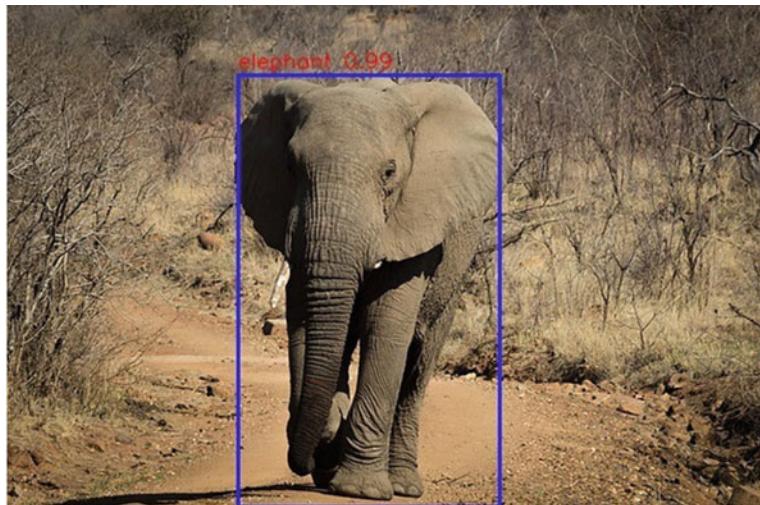
Operating system: Windows 10.

30.4 Multiple Criterions Decision-Making Principle

Many criteria decision assist (MCDS) techniques are decision-making tools that were created to cope with all of the data in order to support complicated decision making with multiple goals. MCDS approaches have already been applied to forestry and other natural resource management planning challenges and is found to be effective one. The MCDM approach is a suitable tool for helping complicated decision making in animal detection. The tool to be used should be chosen to the scenario at hand: In other words, case-by-case consideration is always required in order to create an effective decision support process.

30.5 Results and Discussion

The system protection system detects elephant and uses an alarm to alert the farmer. Figure 30.11 shows the result of YOLOv3 model, the output image rectangular box



```
In [17]: import cv2
import numpy as np
f='my.jpg'
img = cv2.imread('elephant1.jpg')
type(img)
image=detect_image(img,yolo,all_classes)
cv2.imwrite('elephant_12.jpg',image)

time: 36.68s
class: elephant, score: 0.99
box coordinate x,y,w,h: [195.05962372 58.74124628 220.6537056 370.19756234]
```

```
Out[17]: True
```

Fig. 30.11 Output image from the YOLOv3 model

around the elephant image with its type and score printed on the top left corner of the box.

Figure 30.12 shows the real-time soil data that is the temperature, humidity and the soil moisture which is stored in the spreadsheet and it can be accessed globally from all around the world.

30.6 Conclusion

An IoT-based smart crop field monitoring, protection and automatic irrigation system has been developed with low complex circuitry. The manual effort of the farmers of the state Odisha can be reduced by using this smart technique that can also be used as a destiny factor of agriculture. This technique can have monetary benefit to the farmers of Odisha due to non-damage of crops due to wild animals and can also be used for

A	B	C	D	E	F	G	H	I	J
1 DATE&TIME	temperature	humidity	soil moisture						
2 January 28, 2020 at 03:26PM	16	18	100						
3 January 28, 2020 at 03:27PM	21	20	125						
4 January 28, 2020 at 04:59PM	80	60	225						
5 January 28, 2020 at 05:01PM	40	50	125						
6 January 29, 2020 at 02:57PM	12	50	225						
7 January 29, 2020 at 03:25PM	12	50	125						
8 February 3, 2020 at 03:59PM	24.4	56	350						
9 February 3, 2020 at 04:03PM	24.3	55	358						
10 February 3, 2020 at 04:03PM	24.4	53	358						
11 February 3, 2020 at 04:04PM	24.3	53	359						
12 February 3, 2020 at 04:04PM	24.3	53	403						
13 February 3, 2020 at 04:05PM	24.3	54	445						
14 February 3, 2020 at 04:07PM	24.4	56	546						
15 February 3, 2020 at 04:07PM	24.4	53	367						
16 February 3, 2020 at 04:08PM	24.4	52	367						
17 February 3, 2020 at 04:08PM	24.3	53	368						
18 February 3, 2020 at 04:08PM	24.3	53	455						
19 February 3, 2020 at 04:09PM	24.3	53	367						
20 February 3, 2020 at 04:09PM	24.3	54	368						
21 February 3, 2020 at 04:10PM	24.3	54	369						
22 February 3, 2020 at 04:10PM	24.4	55	100						

Fig. 30.12 Real-time soil data stored in the Google sheets

reducing the human animal conflict. The system uses sensors to monitor temperature, soil moisture and humidity of the agricultural land for controlling the irrigation and advanced object detection technique called YOLOv3 to detect an elephant and is effectively interfaced. All perceptions and tests demonstrate that proposed framework is a finished answer for field exercises, irrigation problems and so on. Usage of such a framework in a field will assist with improving the field of yields and the general creation and assurance. This IoT-based smart crop-field monitoring, protection and automatic irrigation system will assist farmers with increasing the harvest production rate. The device specifications should be 8 GB RAM and I5 processor to incorporate YOLOv3. While it can also be used on a Raspberry Pi with OpenCV since it is very slow with regard to the full configuration of the desktop. YOLO is already a growing open source community, so Raspberry Pi support will come soon.

References

1. Yuvaraju, M., Priyanga, K.J.: An IoT based automatic agricultural monitoring and irrigation system. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **4**(5) (2018). ISSN: 2456-3307
2. Rajalakshmi, P., Devi Mahalakshmi, S.: IOT based crop-field monitoring and irrigation automation. In: 2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, pp. 1–6 (2016)
3. Ogunti, E.: IoT based crop field monitoring and irrigation automation system. *IJISSET-Int. J. Innov. Sci. Eng. Technol.* **6**(3) (2019)
4. Rao, R.N., Sridhar, B.: IoT based smart crop-field monitoring and automation irrigation system. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, pp. 478–483 (2018)

5. Sukumar, P., Akshaya, S., Chandraleka, G., Chandrika, D., Dhilip Kumar, R.: IoT based agriculture crop field monitoring system and irrigation automation. *Int. J. Intellect. Adv. Res. Eng. Comput.* **6**(1)
6. Pernapati, K.: IoT based low cost smart irrigation system. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, pp. 1312–1315 (2018)
7. Mishra, D., Khan, A., Tiwari, R., Upadhyay, S.: Automated irrigation system-IoT based approach. In: 2018 3rd International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Bhimtal, pp. 1–4 (2018)
8. Vaishali, S., Suraj, S., Vignesh, G., Dhivya, S., Udhayakumar, S.: Mobile integrated smart irrigation management and monitoring system using IOT. In: 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, pp. 2164–2167 (2017)
9. Fazil, M., Firdhous, M.: IoT-enabled smart elephant detection system for combating human elephant conflict. In: 2018 3rd International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, pp. 1–6 (2018)
10. Zeroday: A LUA based firmware for WiFi-SOC ESP8266. Github. Retrieved 2 Apr 2015
11. Wiguna, H.: NodeMCU LUA firmware. Hackaday. Retrieved 2 Apr 2015
12. Jump up to: a b systems, Espressif. Espressif Systems. Espressif-WikiDevi. Archived from the original on 1 Dec 2017. Retrieved 3 June 2017
13. Benchhoff, B.: A dev board for the ESP LUA interpreter. Hackaday. Retrieved 2 Apr 2015
14. Hu, Y., Wu, X., Zheng, G., Liu, X.: Object detection of UAV for anti-UAV based on improved YOLO v3. In: 2019 Chinese Control Conference (CCC), Guangzhou, China, pp. 8386–8390 (2019)
15. Kumar, N.M., Chopra, S.S., de Oliveira, A.K.V., Ahmed, H., Vaezi, S., Madukanya, U.E., Castañón, J.M.: Solar PV module technologies. In: Photovoltaic Solar Energy Conversion, pp. 51–78. Academic Press (2020)
16. Aghaei, M., Kumar, N.M., Eskandari, A., Ahmed, H., de Oliveira, A.K.V., Chopra, S.S.: Solar PV systems design and monitoring. In: Photovoltaic Solar Energy Conversion, pp. 117–145. Academic Press (2020)
17. Kumar, N.M., Chopra, S.S., Rajput, P.: Life cycle assessment and environmental impacts of solar PV systems. In: Photovoltaic Solar Energy Conversion, pp. 391–411. Academic Press (2020)
18. Kumar, N.M., Dash, A., Singh, N.K.: Internet of Things (IoT): an opportunity for energy-food-water nexus. In: 2018 International Conference on Power Energy, Environment and Intelligent Control (PEEIC), Apr 2018, pp. 68–72. IEEE
19. Kumar, N.M., Mallick, P.K.: The Internet of Things: insights into the building blocks, component interactions, and architecture layers. *Procedia Comput. Sci.* **132**, 109–117 (2018)
20. Kumar, N.M., Atluri, K., Palaparthi, S.: Internet of Things (IoT) in photovoltaic systems. In: 2018 National Power Engineering Conference (NPEC), Mar 2018, pp. 1–4. IEEE

Chapter 31

Smart Student Performance Monitoring System Using Data Mining Techniques



Jay Bijay Arjun Das, Saumendra Kumar Mohapatra,
and Mihir Narayan Mohanty

Abstract In current pandemic situation the education system stands with the support of computer and Internet. Mostly the online teaching, online attendance, and online examinations could perform to some extent. However the activity and performance of the student could not evaluated properly. This article proposes the system that can monitor the student from their activities from distance place. Similarly, the performance can be predicted and monitored with the help of their day to day academic progress. The parameters can be caused as the attendance, the activities with imaging, question answer, assignment, small projects, and weekly assessments. The system can be developed to satisfy those criteria with different modules. The proposed work represents a framework for analyzing the student's academic performance by using the data mining models. Three different types of model are considered for this purpose and from the result it is observed that random forest model is providing better result as compared to others.

31.1 Introduction

Nowadays everything got online even if education. In present situation, we are preferring online study instead of classroom study. And the educational institutions keeping all the information regarding the students like, personal data, attendance, exam scores, etc. in their data bases. The most highly discussed topic in the research in educational system which has been conducted for years is the attendance and the performance in

J. B. A. Das

Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

S. K. Mohapatra

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India

M. N. Mohanty (✉)

Department of Electronics and Communication Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

e-mail: mihirmohanty@soa.ac.in

exams, in college courses of a student. In our education system, i.e., Indian education system, checking the performance of a student is very essential for higher education. Many higher educational institutions consider the performance of a student as one of its crucial part. Schools, colleges, and other educational institutions are running on high swiftness to provide scholars in this competitive world. Most of the institutions use CGPA, internal assessment, external assessment, examination final score, and extra co-curricular activities of the student as prediction criteria. These educational institutions focus on generating graduates with good academic performances [1]. They need to keep track on how the students are performing in particular fields and in what field they need more training for improvement. Day by day the volumes of data is increasing so, to analyze we need to generate algorithms using data mining.

Nowadays, data mining is playing a vital role in educational institutions. The data mining applications are becoming widespread because large amounts of data can be stored in the database and can easily processed. Educational data mining is the application of data mining (is the process of extracting important and useful data from a large dataset) and is helpful to predict useful and important information from educational databases to improve educational performance, better understanding and to have better assessment of the learning process. The educational institutions are getting automated by the advanced technologies. The use of Internet and e-learning in the field of education has facilitated the students [2]. To create student model, the data mining techniques can be applied to the data like, family background, students behavior during the classes, performance in the academics, the data collected from the classroom interaction, and class tests. Prediction of student performance is one of the most important subjects of educational data mining. Prediction of student performance in online learning environments according to students' browsing data is important in terms of increasing the effectiveness of these environments. The education process can be planned more effectively by analyzing the student performance [3]. Video capture technology has become increasingly common in higher education as educators try to provide more flexibility for students and appeal to more learning styles. Total grade point average and internal evaluations (quizzes, laboratory studies, participation, etc.) of variables were frequently used as input variables in estimating student performance. Similarly, tried to predict students 'academic performance by using the students' overall weighted grade point average, letter grades taken from some courses, midterm and final exam scores. Firstly in the study, the data of the students in the database were taken from the database, the data taken were cleared according to certain criteria and the remaining data were normalized and ready to be analyzed. This criteria's and normalization process explained in data gathering and preprocessing section. Then, the data were separated as training and test data, model was created with training data according to specified parameters and accuracy scores of test data and these models were calculated.

31.2 Literature Survey

Now a days in our education system student's performance is depends on many factors. A student can perform better in their education by improving their ability. To improve the education and performance of the student we use data mining technique. Applying data mining in education system is an interesting work now a days. This helps the educational institutions and teachers to improve their educational system. The educational institutes contains lots of academic database of students. In these student databases there are other attributes like family background, family's annual income, etc. it will helps us to identify those students and provide them a chance to pay heed. We can prepare a structure which will analyze the performance of the students by using the concepts of data mining under classification. Classification algorithm like decision tree, naïve Bayes, and support vector machine will help us to predict the performance of students. This technique can help the parents as well as the teachers to keep in track the students and provide them required council ling. These technique may help to provide scholarship and other required training to the aspiring students. Now a day's predicting the performance of the students is very much challenging due to huge amount of databases. It can be improved to lead a better system. For this we present two different approaches. The first one is we can use regression algorithm to predict the performance of the students. In data mining regression is a function which predicts a number. The regression algorithm can estimate the values of the variables which are dependent. In a model the relation between the predictors and target are summarized. The second one is by using the root mean square error method to find the error rate of regression algorithm [4–6]. By this we can identify the students who are weak and can provide them better and additional education to improve their performance.

Online lecture and lecture capture is broadly use in educational system to record the lecture material for the viewing of the students through online. Due to this the attendance of the students in the online classes are decreasing day by day. The online lecture capture makes a negative relation with the attendance of the students online lecture capture has no significant relationship with the attainment of the lecture attendance. In India, we do not have any technique or system by which we can manage or analyze or monitor the progress and performance of the students. Every institutions have their own method and criteria for analyze the performance of the students. This is happens due to lack of study and investigation on existing technique. So to understand the problem, using the data mining technique a detail lecture on predicting students performance is proposed. By this we can improve the performance of the student. Data mining is the analysis of huge dataset to extract the useful pattern and use those pattern to predict the future events. Now a day's data mining is a very important field in educational system [7–9].

31.3 Methodology

Here, we are using the data mining technique to extract the data of a particular student from a student data base of the college. Three data mining algorithm to show the information about a particular student in the student data base. Before predicting the performance a smart platform for storing the academic data of a student is also designed.

31.3.1 Logistic Regression

Logistic regression [10] algorithm inspects the amalgamation that exists between one independent variable and diploid dependent variables. The difference between the logistic regression and linear regression is the continuous dependent variable.

31.3.2 K-Nearest Neighbor

K-NN classifier [11] performs classification by establishing the closest neighbors and utilizes this distance calculation to predict a new class label. The system's main drawback is the computational complexity that current computational-powered systems can overcome.

31.3.3 Support Vector Machine

Support vector machine [12] classify by constructing a hyperplane. If there are more than two classes, then it will perform multi-class SVM. Since our problem is two classes, a single SVM is enough to classify the data points into either patients having heart disease or not.

31.3.4 Random Forest

The decision tree [13] builds a tree-like classifier to traverse the tree to reach the final leaf outcomes. The class label will be present in the leaf outcomes. A random forest classifier [6] is an ensemble classifier that constructs a weak classifier's combination to produce final classification results.

31.4 Result

A smart student performance record storing system is designed in the initial stage of the proposed work. First we have to enter the name of the student. Then it shows the information of the particular student. It shows the information contains the name, regd no., branch, sec, etc. then we have the option to find the performance of the student. The performance are the semester result, class performance, quiz test. Figure 31.1 shows the snapshot of the proposed smart system. Then three different data mining model is considered for predicting the performance of the student. The prediction result is presented in Table 31.1.

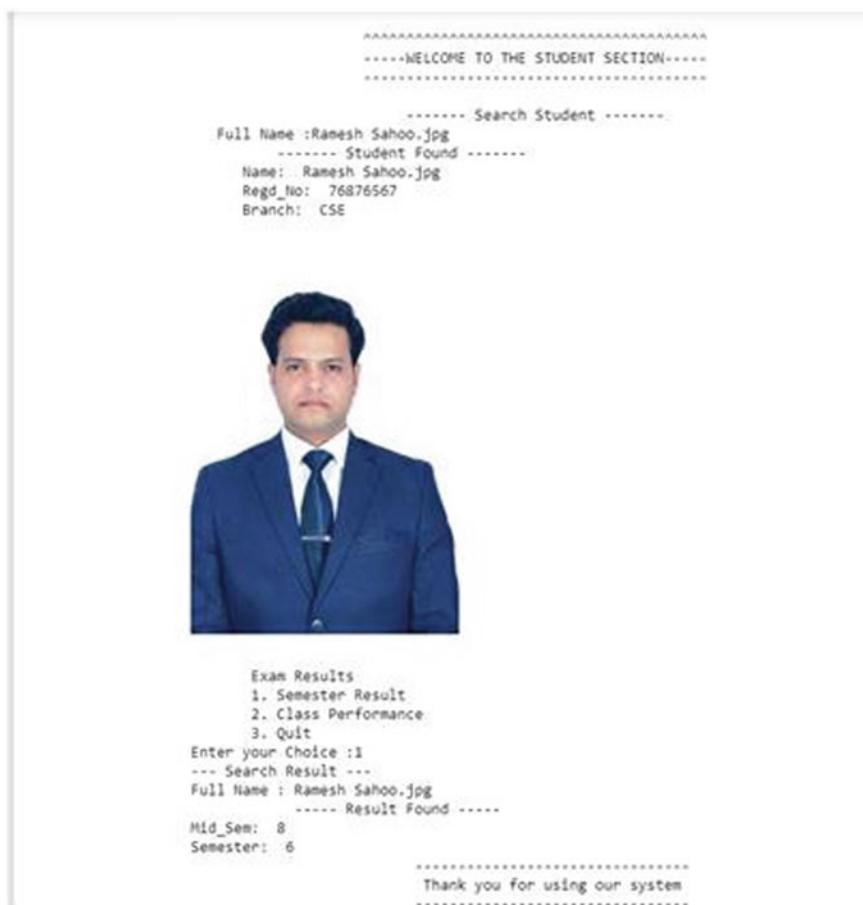


Fig. 31.1 Smart student academic performance system

Table 31.1 Prediction performance

Method	Accuracy (%)
Logistic regression	79
K-nearest neighbor	82
Random forest	89

31.5 Conclusion

The proposed smart student academic performance prediction system will be a supportive tool for both teachers and student during this pandemic situation. It will help to easily predicting the students performance as well as their academic record. Three different data mining models are considered for this purpose and from the result it is observed that the random forest ensemble model is providing better result as compared to other. Further different ensemble and deep learning models can be considered for this purpose to increase the performance.

References

1. Bhattacharya, S., Nainala, G.S., Das, P., Routray, A.: Smart attendance monitoring system (SAMS): a face recognition based attendance system for classroom environment. In: 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), pp. 358–360. IEEE, July 2018
2. Ananta, A.Y., Rohadi, E., Ekijono, E., Wijayaningrum, V.N., Ariyanto, R., Noprianto, N., Syulistyo, A.R.: Smart monitoring system for teaching and learning process at the university. In: IOP Conference Series: Materials Science and Engineering, vol. 732, no. 1, p. 012042. IOP Publishing (2020)
3. Abdullah, A.Z., Isa, M., Rohani, M.N.K.H., Jamalil, S.A.S., Abdullah, A.N.N., Azizan, N.: Development of smart online partial discharge monitoring system for medium voltage power cable. Int. J. Power Electron. Drive Syst. **10**(4), 2190 (2019)
4. Patel, R., Patel, N., Gajjar, M.: Online students' attendance monitoring system in classroom using radio frequency identification technology: a proposed system framework. Int. J. Emerg. Technol. Adv. Eng. **2**(2), 61–66 (2012)
5. Gagare, P.S., Sathe, P.A., Pawaskar, V.T., Bhave, S.S.: Smart attendance system. Int. J. Recent Innov. Trends Comput. Commun. **2**(1), 124–127 (2014)
6. Hu, Y., Huang, R.: Development of weather monitoring system based on Raspberry Pi for technology rich classroom. In: Emerging Issues in Smart Learning, pp. 123–129. Springer, Berlin (2015)
7. Supovitz, J.A., Klein, V.: Mapping a course for improved student learning: how innovative schools systematically use student performance data to guide improvement (2003)
8. Krishnapillai, L., Veluppillai, S., Akilan, A., Naomi Saumika, V., Dhammadika, K.P., De Silva, H., Gamage, M.P.A.W.: Smart attendance and progress management system. In: Innovations in Electrical and Electronic Engineering, pp. 771–785. Springer, Singapore (2021)
9. Jo, J., Park, K., Lee, D., Lim, H.: An integrated teaching and learning assistance system meeting requirements for smart education. Wireless Pers. Commun. **79**(4), 2453–2467 (2014)
10. Wright, R.E.: Logistic regression (1995)

11. Sarangi, L., Mohanty, M.N., Patnaik, S.: Design of ANFIS based e-health care system for cardio vascular disease detection. In: International Conference on Intelligent and Interactive Systems and Applications, pp. 445–453. Springer, Cham (2016)
12. Mohapatra, S.K., Behera, S., Mohanty, M.N.: A Comparative analysis of cardiac data classification using support vector machine with various Kernels. In: 2020 International Conference on Communication and Signal Processing (ICCSP), pp. 515–519. IEEE
13. Mohapatra, S.K., Mohanty, M.N.: Analysis of resampling method for arrhythmia classification using random forest classifier with selected features. In: 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), pp. 495–499. IEEE

Chapter 32

BertOdia: BERT Pre-training for Low Resource Odia Language



Shantipriya Parida, Satya Prakash Biswal, Biranchi Narayan Nayak, Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Satya Ranjan Dash

Abstract Odia language is one of the 30 most spoken languages in the world. It is spoken in the Indian state called Odisha. Odia language lacks online content and resources for natural language processing (NLP) research. There is a great need for a better language model for the low resource Odia language, which can be used for many downstream NLP tasks. In this paper, we introduce a Bert-based language model, pre-trained on 430,000 Odia sentences. We also evaluate the model on the well-known Kaggle Odia news classification dataset (BertOdia: 96%, RoBERTaOdia: 92%, and ULMFit: 91.9% classification accuracy), and perform a comparison study with multilingual Bidirectional Encoder Representations from Transformers (BERT) supporting Odia. The model will be released publicly for the researchers to explore other NLP tasks.

S. Parida (✉) · M. Fabien · E. Villatoro-Tello · P. Motlicek
Idiap Research Institute, Martigny, Switzerland
e-mail: shantipriya.parida@idiap.ch

M. Fabien
e-mail: mael.fabien@epfl.com

E. Villatoro-Tello
e-mail: evillatoro@cua.uam.mx

S. P. Biswal
The University of Chicago, Chicago, USA
e-mail: sbiswal@chicagobooth.edu

B. N. Nayak
Capgemini Technology Services India Limited, Bangalore, India
e-mail: biranchi125@gmail.com

M. Fabien
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

E. Villatoro-Tello
Universidad Autónoma Metropolitana, Mexico City, Mexico

S. R. Dash
KIIT University, Bhubaneswar, India
e-mail: sdashfca@kiit.ac.in

32.1 Introduction

Odia¹ is one of the oldest languages in India as it dates back to the tenth century CE. It originated from Ardhmagadhi Prakrit. This is the mother tongue of around 31 million people living in Odisha state. The term Odia comes from the ancient Sanskrit Odra. The Odrakas are described in Mahabharata as great warriors who fought in that battle. Odia is one of the six classical languages (Sanskrit, Tamil, Telugu, Kannada, and Malayalam) identified by the Indian Government. Odia language has more than 2 lakh manuscripts documented, which makes it the second-highest manuscript holder among Indian languages. Odia is agglutinative.² It differentiates between plural and singular number; male and female gender; second, first, and third persons. But it does not have gender biasing in verbs, nouns, or pronouns like other languages, which reduces the complexity. Odia language allows compounding but does not allow elision. It has 6 vowels 28 consonants, 9 diphthongs, 0 ending with consonants, and 4 semivowels. Odia's vocabulary is influenced by Sanskrit, and is also a little influenced by Arabic, Persian, Austronesian languages as the Kalinga Empire (Odisha's ancient name) was connected to other kingdoms.³ Odia script is an Abugida that is written from left to right.

In recent years there is a growing interest in the NLP community using pre-trained language models for various NLP tasks, where the models are trained in a semi-supervised fashion to learn a general language model [5]. A better language model is the key component of the automatic speech recognition system (ASR) [19]. Building a language model is a challenging task in the case of low resource languages where the availability of contents is limited [1]. Researchers proposed many techniques for the low resource NLP tasks such as feature engineering, and knowledge transfer across domain [9, 10]. However, these approaches do not use a pre-trained general language model, rather they perform pre-training for each task individually. We focus on building a general language model using the limited resources available in the low resource language which can be useful for many language and speech processing tasks.

Our key contribution includes building a language-specific BERT model for this low resource Odia language and as per our best knowledge, this is the first work in this direction. The overall architecture of the proposed model is shown in Fig. 32.1.

32.2 Related Work

Low resource languages have been drawing the attention of several recent works in language model pre-training [3]. Although Multilingual Bidirectional Encoder

¹ https://en.wikipedia.org/wiki/Odia_language.

² <https://www.mustgo.com/worldlanguages/oriya/>.

³ <https://www.nriol.com/indian-languages/oriya-page.asp>.

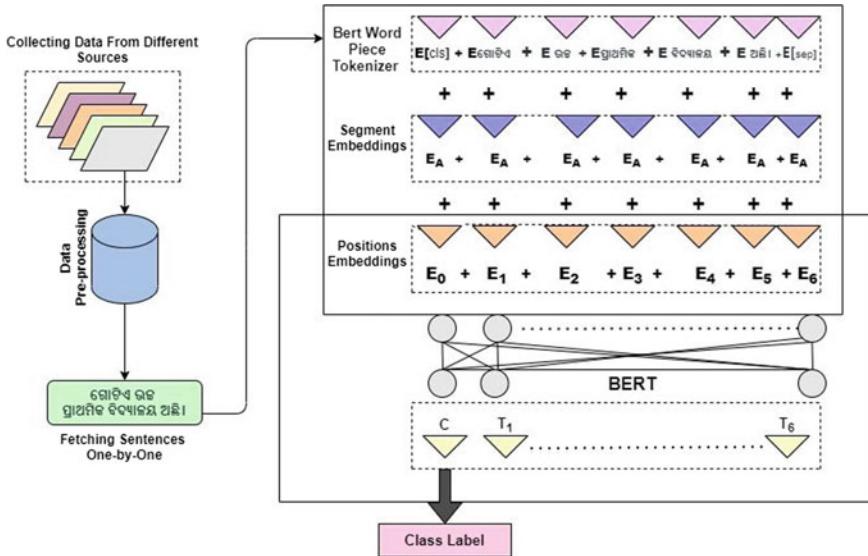


Fig. 32.1 The proposed model: visualization of our experimental model used for the single sentence classification task with BERT embedding layers

Representations from Transformers (M-BERT)⁴ successfully covers 104 languages and is being further extended, a large number of languages are still not covered. It is one of the research trends to extend M-BERT for low resource languages [24]. Some researchers found that language-specific BERT models perform better compared to multilingual BERT models [14].

The sufficient availability of online contents remains one of the major challenges for many low resource languages including Odia [16]. Recently, a few research projects were initiated to build a language model for many low resource Indian languages including Odia. In particular, the Natural Language Toolkit for Indic Languages (iNLTK)⁵ released different language models for the Indian languages including Odia using 17 K Odia Wikipedia articles, and the model is tested on the classification task using IndicNLP News Article Classification Dataset-Oriya [2]. There is also the multilingual IndicBERT⁶ model based on BERT that supports 12 Indian languages including Odia available in Huggingface transformers library. It also has IndicGLUE, a natural language understanding benchmark for the evaluation of a few tasks for Odia [8].

⁴ <https://github.com/google-research/bert/blob/master/multilingual.md>.

⁵ <https://github.com/goryu001/inltk>.

⁶ <https://github.com/AI4Bharat/indic-bert>

Table 32.1 Dataset statistics

Source	Sentences	Unique Odia tokens
OdiEnCorp2.0	97,233	174,045
CVIT PIB	58,461	66,844
CVIT MKB	769	3944
OSCAR	192,014	642,446
Wikipedia	82,255	236,377
Total deduped	430,732	1,123,656

32.3 Data Source

We have collected monolingual Odia text from the recently released OdiEnCorp 2.0 [17].⁷ In addition to this, we have used Odia corpus from OSCAR [15].⁸ We also included in our dataset the parallel corpus sources by the Center for Visual Information Technology (CVIT).⁹ This contains both CVIT PIB [v0.2] (Sentence aligned parallel corpus between 11 Indian Languages, crawled and extracted from the press information bureau website) and CVIT MKB [v0.0] (The Prime Minister’s speeches—Mann Ki Baat, on All India Radio, translated into many languages). Finally, we added a Kaggle dataset scraped from Wikipedia.¹⁰

All the aforementioned datasets were merged into a single training file and then deduped to remove any duplicate sentences. The statistics of the dataset are shown in Table 32.1. This dataset also covers different domains such as the Bible, Wikipedia articles, literature websites, government portals, literature, and dictionaries.

32.4 Training and Evaluation

There are different variants of BERT [4] available for the language model such as “A Lite BERT” (ALBERT) allowing to fit the model into memory and to increase the speed of BERT [12], StructBERT considering word level and sentence level ordering [6], and RoBERTa [13] that iterates on BERT’s pre-training procedure, including training the model longer, with bigger batches over more data, removing the next sentence prediction objective, training on longer sequences and dynamically changing the masking pattern applied to the training data.

In our experiment, we built the language model based on BERT and RoBERTa. We used a single GPU (NVIDIA Tesla V100-SXM2-16GB) for training our models. We train for 30 epochs with a training time of around 12 h. During training, we did not

⁷ <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3211>.

⁸ <https://oscar-corpus.com/>.

⁹ <http://preon.iiit.ac.in/~jerin/bhasha/>.

¹⁰ <https://www.kaggle.com/disisbig/odia-wikipedia-articles>.

Table 32.2 Training configurations

Parameter	BERT RoB	ERTa
Learning rate	5e-5	5e-5
Training epochs	5	10
Dropout Prob	0.1	0.1
MLM Prob	0.1	0.2
Self-attention layer	6	6
Attention head	12	12
Hidden layer size	768	768
Hidden layer Activation	gelu	gelu
Total parameters	84 M	125 M

consider upper and lower case letters, since the Odia language does not distinguish between them.¹¹ The configuration parameters are shown in Table 32.2.

32.4.1 BERT/RoBERTa Model Training

We explored training both BERT and RoBERTa models on the same dataset. RoBERTa is built on BERT’s language masking strategy, wherein the system learns to predict intentionally hidden sections of text within otherwise unannotated language examples.

We used Huggingface’s interface to train both BERT and RoBERTa models. For the RoBERTa model, we chose to train a byte-level Byte-Pair Encoding (BPE) tokenizer (the same as GPT-2), with the same special tokens as RoBERTa. We arbitrarily picked its vocabulary size to be 52,000. The advantage of using a byte-level BPE (rather than a WordPiece tokenizer that was used in BERT) is that it will start building its vocabulary from an alphabet of single bytes [21]. Hence, all words will be decomposed into tokens which were reflected in the results we obtained. For the BERT model, we chose to train a WordPiece tokenizer. Again we arbitrarily picked vocabulary size to be 52,000.

The hyper-parameters are shown in Table 32.2. The learning curve for BertOdia training is shown in Fig. 32.2.

32.4.2 Model Fine-Tuning

To evaluate the models, we fine-tuned the BERT/RoBERTa models on a downstream task, i.e., classification task. We used the Kaggle Odia news classification dataset¹²

¹¹ <https://fontmeme.com/odia/>.

¹² <https://www.kaggle.com/disisbig/odia-news-dataset>.

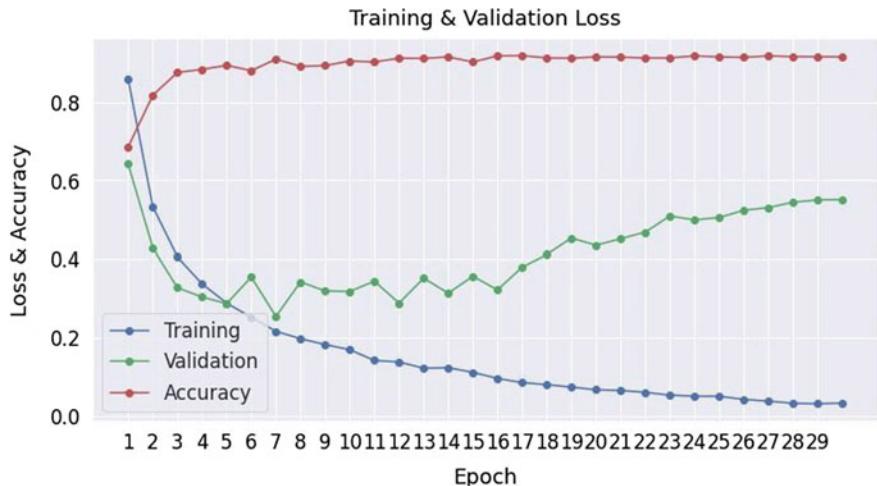


Fig. 32.2 Training: performance versus epochs

for this task. This dataset is scraped from Odia daily news papers. It has headlines and the section of the newspaper from which they are scrapped. The dataset contains 3 classification labels (sports, business, entertainment) for the headlines and is divided into train/test sets. We used the same tokenizer that was trained and saved earlier. After tokenization, special tokens were added. Then the sentences were padded to 512 blocks. We trained the final layer using the classification dataset.

32.4.3 ULMFiT Model Training

We also used a pre-trained ULMFiT model to benchmark against our BERT-based models [7]. This model uses AWD_LSTM architecture. It also uses the default parameters and a drop_mult of 0.3. The model is trained on Odia Wikipedia articles¹³ and available on GitHub.¹⁴ We fine-tuned the model for the downstream classification task with the same news classification data used for the BERT and RoBERTa models.

The evaluation results of our models on the news classification are shown in Table 32.3. The BertOdia model performance outperformed RoBERTaOdia by up to 4.5% (relative) and RoBERTaOdia reached performances on this task compared to the performance of ULMFiT.

¹³ <https://www.kaggle.com/disisbig/odia-wikipedia-articles>.

¹⁴ <https://github.com/goru001/nlp-for-odia>.

Table 32.3 BertOdia performance

Model	Task	Accuracy
BertOdia	Text classification	96.0
RoBERTaOdia	Text classification	92.0
ULMFiT	Text classification	91.9

The best performance is highlighted in bold

32.4.4 Language Model Evaluation

To evaluate the BERT model, we have used the perplexity (PPL) score using BERT masking approach [20, 22]. It is important to notice that although the PPL metric applies specifically to classical language models, it is not well defined for masked language models like BERT. When dealing with masked languages, PPL can be thought of as an evaluation of the model’s ability to predict uniformly among the set of specified tokens in a corpus, meaning that the tokenization procedure has a direct impact on a model’s perplexity. Additionally, when working with approximate models (e.g., BERT), we typically have a constraint on the number of tokens the model can process, e.g., 512 tokens. Thus, when computing the PPL of BERT, the input sequence is typically broken into subsequences equal to the model’s maximum input size, considering a sliding-window strategy.

This involves repeatedly sliding the context window so that the model has more context when making each prediction.

Accordingly, we evaluated our model in a small set of unseen data (500 sentences) extracted from a news website.¹⁵ For all the considered sentences, the length is equal to or less than 512 tokens. Hence, considering the approach described above, we obtained a mean score of $PPL = 134.97$ with a standard deviation of 196.98. The perplexity score differs (low-high) based on the sentences as depicted in Fig. 32.3.

In [18] authors describe the training process performed on Embeddings from Language Models (ELMo) for many Indian languages. Among other languages, they report a perplexity of ($PPL = 975$) for the Odia language model. Although we cannot make a direct comparison, our trained model obtains a better performance for this morphologically rich language [11].

32.4.5 IndicGLUE Tasks

We also benchmarked our model against a few IndicGLUE tasks [8]. Despite our model being trained on 6% of data used for training IndicBert, we got comparable results as shown in Table 32.4.

For the Cloze-style Multiple-choice QA task, we feed the masked text segment as input to the model and at the output, we have a softmax layer that predicts a

¹⁵ <http://news.odialanguage.com/>.

ID	Odia Sentences and its English Translation	Perplexity (Odia Sentence)
1	ସେଇମ ରେଣ୍ଟିହୁଯାର ଅବନିକାଣ୍ଡତେ ପ୍ରତି ମୃତ୍ୟୁକ୍ଷେତ୍ର ପରିବାରକୁ ମିଳିବ ୨.୫ ଲକ୍ଷ ଟଙ୍କା, ଯୋଗତା କଲେ SII ଅଧ୍ୟୟକ୍ଷୀ। Serum Institute fire: Rs 2.5 million per family of deceased, SII chairman announced.	6.51
2	ନଯାଗର୍ଥ କିଲାଲ ରଣ୍ଧ୍ରୀର ବଳକ୍ଷର ନହିଁପୂର ଶୀଃପର ପାଇଁ ନିଆଁ ଖୋଲିବା ବେଳେ ମାତ୍ର ତଢ଼ୁ ବାହାରିଲ ୧୦ ଖୁଦ୍ଦୁରୁ ଅଧିକ କରାଯାଇଛି। Nachipur village in Ranpur block of Nayagarh district: More than 10 bushels of cowries came out of the ground while digging for a house.	171.40
3	୧୮ ଜାନିଙ୍ଗରେ ଶାଖାତ୍ତକ ଚପଣଟିଆ ଲେଲଗେଇ ଥାଏଗୋଲନ । Farmers' six-hour rail strike on the 18th.	615.10
4	୧୫ ଜାନ୍ମ ଦେବାରୀକ ହେବ ସବୁ ଅଧାରତ, ନିର୍ଦ୍ଦିଷ୍ଟ ଅଧାରତ, ନିର୍ଦ୍ଦିଷ୍ଟ ନିର୍ଦ୍ଦିଷ୍ଟ ପାଇଁ SOP ରାଖି । The SOP will continue for all courts, lower courts and tribunals from the 15th.	808.86
5	ଉତ୍କଳ ବିଶ୍ୱବିଦ୍ୟାଳୟ ଛାତ୍ରବାଚକ ଦ୍ୱାରା ଉତ୍କଳ କାର୍ଯ୍ୟକ୍ରମରେ ମଧ୍ୟ ଲାଗନ୍ତିରୁ ଅଧ୍ୟୟକ୍ଷ, ୧୭୪୦ ଟଙ୍କା ଫେଁ ଛାତ୍ର । PG Council president agreed to Utkal university student's demand, Rs 1740 fee waive off.	15.62

Fig. 32.3 Perplexity score for sample Odia sentences**Table 32.4** Comparison of BertOdia with IndicBERT. BertOdia was trained on 6% of the data of IndicBERT

Model	Article genre classification	Cloze-style multiple-choice QA
XLM-R	97.07	35.98
M-BERT	69.33	26.37
IndicBERT base	97.33	39.32
IndicBERT large	97.60	33.81
BertOdia	96.90	23.00

The best performance is highlighted in bold

probability distribution over the given candidates. We fine-tune the model using cross-entropy loss with the target label as 1 for the correct candidate and 0 for the incorrect candidates [8]. The code for the fine-tuning model is given in the IndicGLUE GitHub.¹⁶

For the Article Genre Classification task we used the IndicGLUE dataset for news classification.¹⁷ It is generated from an Odia daily newspaper source except it has one more genre called crime. More details of the IndicGLUE datasets could be found on its website.¹⁸

¹⁶ https://github.com/AI4Bharat/indic-bert/tree/master/fine_tune.

¹⁷ <https://storage.googleapis.com/ai4bharat-public-indic-nlp-corpora/evaluations/inltk-headlines.tar.gz>.

¹⁸ <https://indicnlp.ai4bharat.org/indic-glue/>.

32.5 Conclusion

In this paper, we presented BertOdia, a pre-trained Odia language model which can be useful for many language and speech processing tasks for this low resource language. BertOdia will be the first language-specific BERT model in Odia which can be used by researchers for many language and speech processing tasks. Our studies will help researchers working on low resource languages. The code and dataset are available at:

https://colab.research.google.com/gist/satyapb2002/aeb7bf9a686a9c7294ec5725ff53fa49/odiabert_languagemodel.ipynb

Future work will include:

- Incorporating OdiEnCorp1.0¹⁹ [16] and IndicCorp data sets²⁰ which have around 8.2 M sentences combined. We want to include a large volume of Odia text covered in a variety of domains to build a robust language model for better performance. Even the IndicBERT large has poor performance on Assamese and Odia—the two languages due to the smallest corpora sizes as compared to other Indian languages [8].
- Developing NLP datasets (natural language inference, question answering, and next sentence prediction) for the Odia language.
- Preparing a dashboard similar to GLUE [23]²¹ and IndicGLUE,²² called “OdiaGLUE” for evaluating the model on various natural language understanding (NLU) tasks specific for Odia language.
- Enriching OdiaGLUE with more NLU tasks as the recent IndicGLUE supports limited NLU tasks for Odia.

Acknowledgements The authors Shantipriya Parida and Petr Motlcek were supported by the European Union’s Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

The authors do not see any significant ethical or privacy concerns that would prevent the processing of the data used in the study. The datasets do contain personal data, and these are processed in compliance with the GDPR and national law.

Esaú Villatoro-Tello, was supported partially by Idiap Research Institute, SNI CONACyT, and UAM-Cuajimalpa Mexico.

¹⁹ <https://indat.mff.cuni.cz/repository/xmlui/handle/11234/1-2879>.

²⁰ <https://indicnlp.ai4bharat.org/corpora/>.

²¹ <https://gluebenchmark.com/>.

²² <https://indicnlp.ai4bharat.org/indic-glue/>.

References

1. Adams, O., Makarucha, A., Neubig, G., Bird, S., Cohn, T.: Cross-lingual word embeddings for low-resource language modeling. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 937–947 (2017)
2. Arora, G.: inltk: Natural language toolkit for indic languages. In: Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pp. 66–71 (2020)
3. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pretrained bert model and evaluation data. In: Practical ML for Developing Countries Workshop@ ICLR 2020 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4171–4186 (2019)
5. Grießhaber, D., Maucher, J., Vu, N.T.: Fine-tuning bert for low-resource natural language understanding via active learning. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 1158–1171 (2020)
6. Hazem, A., Bouhandi, M., Boudin, F., Daille, B.: Termeval 2020: Taln-ls2n system for automatic term extraction. In: Proceedings of the 6th International Workshop on Computational Terminology, pp. 95–100 (2020)
7. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 328–339 (2018)
8. Kakwani, D., Kunchukuttan, A., Golla, S., N.C.G., Bhattacharyya, A., Khapra, M.M., Kumar, P.: IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In: Findings of EMNLP (2020)
9. Kocmi, T., Parida, S., Bojar, O.: CUNI NMT system for WAT 2018 translation tasks. In: Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation. Association for Computational Linguistics, Hong Kong, 1–3 Dec 2018, <https://www.aclweb.org/anthology/Y18-3002>
10. Korzeniowski, R., Rolczynski, R., Sadownik, P., Korbak, T., Mozejko, M.: Exploiting unsupervised pre-training and automated feature engineering for low-resource hate speech detection in polish. Proceedings of the PolEval2019 Workshop, p. 141 (2019)
11. Kumar, S., Kumar, S., Kanodia, D., Bhattacharyya, P.: “A passage to India”: pre-trained word embeddings for Indian languages. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pp. 352–357 (2020)
12. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. In: International Conference on Learning Representations (2019)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint [arXiv: 1907.11692](https://arxiv.org/abs/1907.11692) (2019)
14. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty French language model. In: ACL 2020-58th Annual Meeting of the Association for Computational Linguistics (2020)
15. Ortiz Suárez, P.J., Romary, L., Sagot, B.: A monolingual approach to contextualized word embeddings for mid-resource languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1703–1714. Association for Computational Linguistics, Online, July 2020, <https://www.aclweb.org/anthology/2020.acl-main.156>

16. Parida, S., Bojar, O., Dash, S.R.: Odiencorp: Odia–English and Odia-only corpus for machine translation. In: Smart Intelligent Computing and Applications, pp. 495–504. Springer, Berlin (2020)
17. Parida, S., Dash, S.R., Bojar, O., Motlcek, P., Pattnaik, P., Mallick, D.K.: Odiencorp 2.0: Odia–english parallel corpus for machine translation. In: LREC 2020 Workshop Language Resources and Evaluation Conference, 11–16 May 2020, p. 14
18. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long Papers), pp. 2227–2237 (2018)
19. Raju, A., Filimonov, D., Tiwari, G., Lan, G., Rastrow, A.: Scalable multi corpora neural language models for asr. Proc. Interspeech **2019**, 3910–3914 (2019)
20. Salazar, J., Liang, D., Nguyen, T.Q., Kirchhoff, K.: Masked language model scoring. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2699–2712 (2020)
21. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL (2016)
22. Wang, A., Cho, K.: Bert has a mouth, and it must speak: Bert as a Markov random field language model. In: Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation, pp. 30–36 (2019)
23. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A multitask benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 353–355. Association for Computational Linguistics, Brussels, Belgium, Nov 2018. <https://doi.org/10.18653/v1/W18-5446>, <https://www.aclweb.org/anthology/W18-5446>
24. Wang, Z., Mayhew, S., Roth, D., et al.: Extending multilingual bert to low-resource languages. arXiv preprint [arXiv:2004.13640](https://arxiv.org/abs/2004.13640) (2020)

Chapter 33

A Machine Learning Approach to Analyze Mental Health from Reddit Posts



**Smriti Nayak, Debolina Mahapatra, Riddhi Chatterjee, Shantipriya Parida,
and Satya Ranjan Dash**

Abstract Reddit is a platform with a heavy focus on its community forums and hence is comparatively unique from other social media platforms. It is divided into sub-Reddits, resulting in distinct topic-specific communities. The convenience of expressing thoughts, a flexibility of describing emotions, inter-operability of using jargon, the security of user identity makes Reddit forums replete with mental health-relevant data. Timely diagnosis and detection of early symptoms are one of the main challenges of several mental health conditions for which they have been affecting millions of people across the globe. In this paper, we use a dataset collected from Reddit, containing posts from different sub-Reddits, to extract and interpret meaningful insights using natural language processing techniques followed by supervised machine learning algorithms to build a predictive model to analyze different states of mental health. The paper aims to discover how a user's psychology is evident from the language used, which can be instrumental in identifying early symptoms in vulnerable groups. This work presents a comparative analysis of two popular feature engineering techniques along with commonly used classification algorithms.

S. Nayak
Silicon Institute of Technology, Bhubaneshwar, India

D. Mahapatra
National Institute of Technology Patna, Patna, India

R. Chatterjee
Heritage Institute of Technology, Kolkata, India

S. Parida
Idiap Research Institute, Martigny, Switzerland
e-mail: shantipriya.parida@idiap.ch

S. R. Dash (✉)
KIIT University, Bhubaneswar, India
e-mail: sdashfca@kiit.ac.in

33.1 Introduction

The evolution in social networking has augmented opportunities for people to communicate on the Internet. The number of social media users increased from 2.31 billion users in 2016 to 4.33 billion users in 2021. The global social penetration rate has surged by 87% in the last five years. Social media allows us to connect with people, share and discover ideas, find and review businesses, trade goods, services, shop online and even converse anonymously. Social media is widely perceived to have provided humans a safe space to express themselves. Social media platforms furnish a peer support network, facilitate social interaction and promote engagement and retention in treatment and therapy. Reddit [1] is a network of communities based on people's interests. It has been graded as the 18th-most-visited Web site in the world. It is one of the most popular social networks with over 100,000 active communities and 52 million daily users. Unlike Facebook and Twitter, Reddit does not restrict its users on the length of their posts. Reddit segregates its contents via sub-Reddits that are dedicated to a specific topic, including various communities, thus making it a very valuable information pool.

Mental health ailments are broadly defined as health conditions that alter an individual's thoughts, perceptions, emotional states and behaviors causing profound distress, thus disrupting the effective functioning of their personal, professional and social lives. According to a World Health Organization report, 450 million people around the world experience a mental illness in their lifetime. The report also estimates that roughly 1 in 4 people is likely to have a psychiatric disorder or a neurological disorder [2]. Unfortunately, mental health conditions continue to be under-reported and under-diagnosed [3]. This increases the risk of early symptoms being unacknowledged and is further exacerbated due to lack of sufficient awareness, negative attitudes and prevalent social stigma which act as barriers to care. Approximately, two-thirds of the individuals do not take any professional assistance. Certain studies have found that a significant number of people use social media as a medium to share personal thoughts and experiences, vent out their negative emotions, search for information about mental health disorders and treatment/therapy options. These people give and receive support from others, often anonymously, facing similar challenges.

Now considered as a new medical research paradigm, artificial intelligence and natural language processing have been useful in detecting and predicting mental health conditions from social media data. This work emphasizes the role of social media as a potentially feasible intervention platform for providing support to people with any mental ailment, boosting engagement and retention in care and augmenting existing mental health services, all while keeping safety in mind [4]. In this paper, we have summarized the current research on the use of social media data for the mental health analysis of individuals. Various machine learning models have been implemented to analyze and classify Reddit posts into related categories of psychiatric conditions.

The paper aims to serve as a preliminary study in this domain which can be further expanded by using more sophisticated algorithms. It is organized as follows: Sect. 2

provides the literature review, and Sect. 3 states the methodology along with the experimental results and analysis. Finally, the paper ends with the conclusion, and future work is given in Sect. 4.

33.2 Literature Review

Natural language processing (NLP), an interdisciplinary branch born out of the trinity of computational linguistics, computer science and artificial intelligence, is committed to the comprehension, interpretation and analysis of human language. The overarching objectives of mental health applications are to aid in understanding mental health, act as a channel of support and promote the well-being of individuals. These are realized through textual mining, sentiment detection, sentiment analysis, emotion classification and building technological interventions [5]. As NLP techniques attempt to bridge and facilitate human–computer interactions, it, therefore, becomes a veritable tool for achieving the said objectives.

Seal et al. [6] constructed an emotion detection method that is comprised of text preprocessing, keyword extraction from sentences using POS-tagging and keyword analysis to categorize the revealed emotional affinity. Herzog et al. [7] introduced an ensemble methodology of bag of words (BOW) and word embedding-based classifier, to perform emotion detection. In the latter classifier, document representations that were exercised are continuous bag of words (CBOW) and term frequency-inverse document frequency (TF-IDF). The field of machine learning (ML) has considerably augmented the understanding and scope of research in the vast domain of mental health. The usage of several computational and statistical methods to build better systems for diagnosis, prognosis and treatment of mental ailment symptoms has been accelerated by ML techniques [8]. These techniques are employed to extract psychological insights from predominantly four arenas of data pools, namely data, sensors, structured data and multi-modal technology interactions [9].

A significant portion of the literature is focused on the early detection and monitoring of indicators of depression. Zhou et al. [10] built a model using computer vision and data mining techniques to project a continuous, multifaceted view of one's mental health. Support vector machine (SVM) classifier and logistic regression have been utilized for emotion inference in their work. Fatima et al. [11] identified depressive posts and the degree of depression from user-generated content by investigating the linguistic style and associated sentiment of the posts. Random forest (RF) classifier was utilized for the aforementioned purpose. In the context of social media analysis, a novel framework to detect users prone to depression via conducting a temporal analysis of eight basic emotions exhibited by Twitter posts was performed by Chen et al. [12]. Their methodology focused on examining the textual contents of tweets by employing SVM and RF classifiers with optimized parameters. The classification accuracy of detecting emotions was augmented by 8% for the SVM classifier and 3% for the RF classifier when temporal measures of emotions were included in the prediction task.

Similarly, Suhasini et al. [13] employed Naive Bayes (NB) and k-nearest neighbor algorithm (KNN) to execute textual mining of emotions of Twitter posts and label them into four distinct emotional categories, by exploring the characteristics of the tweets. The NB algorithm performed better than KNN, showing an accuracy of 72.6%. Gaind et al. [14] devised a composition scheme of NLP techniques like emotion-words set (EWS) and ML classification algorithms like sequential minimal optimization (SMO) algorithm and J48 algorithm to detect, classify and quantify Twitter posts. They presented six classification categories of emotions, namely happiness, sadness, surprise, anger, disgust and fear.

Saha et al. [15] attempted to infer expressions of stress from Reddit posts of college communities using a stress classifier. They adopted a transfer learning approach to build a binary SVM classifier that scrutinized and determined posts as either “high stress” or “low stress”. The classifier achieved an accuracy of 82%. Further, suicidal ideation detection (SID) methods, from a computational approach, can greatly benefit from engaging ML classifiers. This is illustrated by a study conducted by Pestian et al. [16]. Their robust ML classifier, based on unweighted SVM, could compute the risk of suicide for individuals before them showing acute suicidal tendencies. By extracting structural, lexical and semantic information from manually labeled suicide notes, the authors in [17] built an automatic emotion detection system to determine 15 classes of emotions.

A noteworthy amount of feature engineering is requisite for conventional ML models to show optimal performance. This step of preprocessing impedes their efficiency, owing to being a tedious and resource-consuming process. State-of-the-art deep learning (DL) algorithms aim to directly map the input data features through a multi-layer network structure [18]. This inevitably leads to models showing superior performance.

Ragheb et al. [19] applied deep transfer learning for sentiment analysis and detection in textual conversations. Self-attention mechanisms and turn-based conversation modeling were used. Gkotsis et al. [17] used a convolutional neural network (CNN) that demonstrated a good accuracy of 91.08% to determine Reddit posts in the domain of mental health discussions and further classified those posts, according to the type of symptoms exhibited, with an accuracy of 71.37%. Sekulic et al. [20] used a hierarchical attention network (HAN) model [21] to predict if a social media user has a specific mental condition, out of nine others that are prevalent. Dheeraj et al. [22] presented a mechanism that inspects emotions from psychiatric texts to discern negative emotions. Multi-head attention with bidirectional long short-term memory and convolutional neural network (MHA-BCNN) was utilized for the said purpose. Their choice of model was influenced due to the selection of long text sequences. Kim et al. [23] collated sub-Reddit posts from the mental health community and developed a CNN model to accurately ascertain a user’s potential mental state and the kind of mental ailment one could be facing, out of the category of depression, anxiety, bipolar disorder, borderline personality disorder (BPD), schizophrenia and autism.

Table 33.1 Sample of the dataset

Title	Text	Sub-Reddit
Exposure does not work!	“I have struggled with social anxiety from childhood and the main.”	Anxiety
Paranoia	“Does anyone here deal with paranoia on a daily basis?....”	BPD
Depression coming back?	“So the thing is that for some years I’ve been on and off with self-harm”	Depression
I’m a sociopath...	“My therapist said I have sociopathic tendencies so does that mean I”	Mental health
Unwanted obsessive and upsetting thoughts?	“I’ve been stressing out so much lately over things that haven’t happened”	Bipolar
Auditory hallucinations (non-verbal)	“I’ve realized lately that I have auditory hallucinations”	Schizophrenia
Am I autistic?	“I get super anxiety and am uncomfortable even going into the store....”	Autism

33.3 Methodology

33.3.1 Data Collection

We have used the dataset collected by Kim et al. [11] in their work. The dataset comprises posts under the following six sub-Reddits associated with the subject of mental health: r/depression (258,495 posts), r/Anxiety (86,243 posts), r/bipolar (41,493 posts), r/BPD (38,216 posts), r/schizophrenia (17,506 posts), r/autism (7143 posts) as well as r/mental health (39,373 posts).

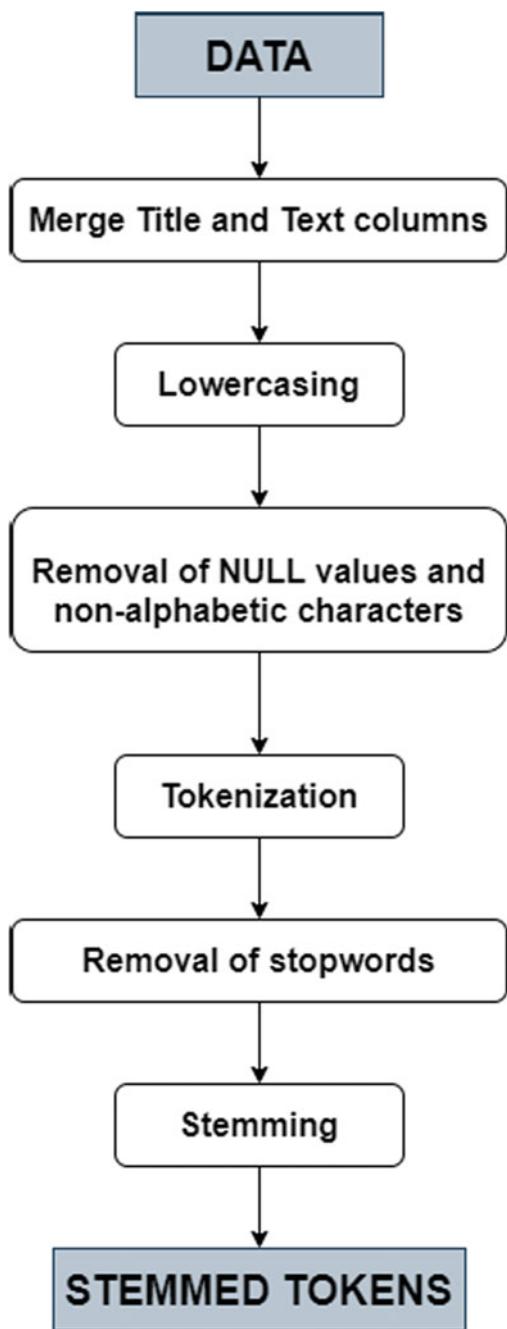
A brief snapshot of the dataset has been provided in Table 33.1.

33.3.2 Preprocessing

The dataset originally contains three columns: title of the post, text content of the post and the sub-Reddit category associated with it. The raw data collected must be preprocessed before training, evaluating and using machine learning models. We have used the following preprocessing pipeline as given in Fig. 33.1.

Initially, the title and the text columns were concatenated, followed by lowercasing to normalize the text in each of the Reddit posts. The unnecessary punctuation marks, white spaces, null values and other non-alphabetic characters were removed from each post. Tokenization separates words in a sentence via space characters. Certain words like *a*, *an*, *the*, *is*, *are*, etc., are commonly used words that do not carry any significant importance in the text. Hence, all the stop-words present are removed from

Fig. 33.1 Text preprocessing pipeline



the post. This step is followed by stemming which is a process to remove morphological affixes from words. It converts the words to their root form and decreases the number of word corpus. For example, the stemmed form of eating, eats and eaten is “eat”. The natural language toolkit (NLTK) [24] is used to perform various preprocessing steps.

33.3.3 *Feature Extraction*

For classifying our posts into their respective mental health category, we first need to extract features from them. Text data cannot be directly given to a machine learning model as they accept numeric feature vectors as input. It is important to convert textual data into vector representation before feeding them into a machine learning model. The features are numerical attributes that exhibit abstraction of the raw data at the level of whether (or to what extent) a particular characteristic is present for a given post. This entails using some NLP techniques to process the raw data, convert text to numbers to generate useful features in the form of vectors. In this work, we have used bag of words (BoW) and term frequency-inverse document frequency (TF-IDF) for feature engineering.

Bag of Words

The simplest encoding of text into vectors is achieved by using bag of words. In this approach, a single feature vector is built using all of the terms in the vocabulary obtained by tokenizing the sentences in the documents. Each text document (or post) present in the dataset is represented uniquely by converting it into a feature vector representation. Each word is treated as a distinct property. As a result, the number of features in the vocabulary is equal to the number of unique words.

Every Reddit post in the dataset is considered a sample or record. If the word appears in the sample, it stores the “frequency” of the word, which is regarded as the feature, and if the word does not appear, it is zero. A word is represented by each column of a vector. In this process of word encoding, the representation of the word takes precedence over the order of the words. This approach returns an encoded vector with a total vocabulary length and an integer count of how many times each word appears in the document.

Term Frequency-Inverse Document Frequency

TF-IDF is an acronym for *term frequency-inverse document frequency* and is used to determine the relevance of a term in a particular corpus by calculating the weight of each term. Term frequency is used to calculate the occurrence of a particular term in the entire corpus. The frequency value of a term is normalized by dividing it by the total number of words in the corpus. Document frequency is a metric to assess the significance of a document in the context of the entire corpus.

33.3.4 Experimental Results

Machine learning approaches can open new avenues for learning human behavior patterns, recognizing early symptoms of mental health disorders and risk factors, making illness progression predictions and customizing and improving therapies. Each Reddit post in the dataset has a label associated with it which determines the related mental health condition of the individual. By using machine learning algorithms on this dataset, the predictive model learns different associations between the post and its corresponding label.

The dataset underwent two types of feature extraction—bag of words and term frequency-inverse document frequency. Following this stage, feature vectors of size 2000 were obtained, respectively. This data was split into 80% for training and 20% for testing. Machine learning algorithms like MultinomialNB, decision tree, random forest classifier, logistic regression and ensemble techniques like AdaBoost and XGBoost were applied to both the feature sets.

The accuracy of the machine learning models is used to evaluate the performance of each of the feature engineering techniques used, as shown in Fig. 33.2 and stated below in Table 33.2.

It can be inferred that multinomial logistic regression, when implemented using TF-IDF feature vector, provides the highest accuracy of 77% and performs better than the other machine learning models. Among the ensemble techniques used, XGBoost exhibits a good performance with nearly 76% on both feature sets.

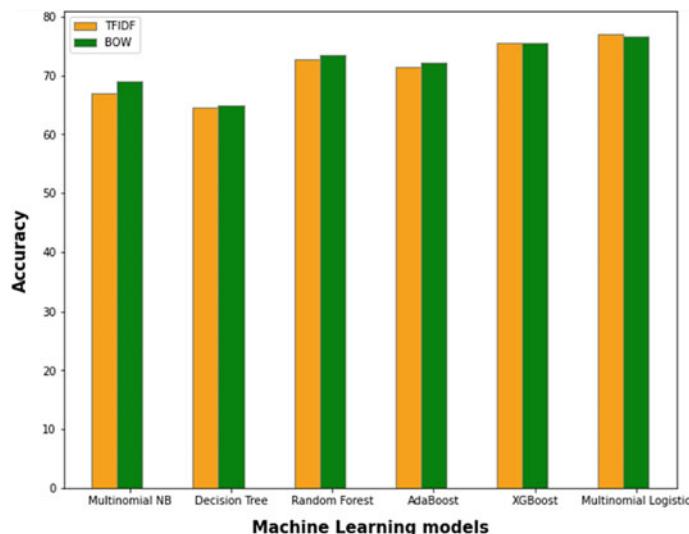


Fig. 33.2 Graph for performance analysis of machine learning models

Table 33.2 Performance analysis of machine learning models

Machine learning model	TF-IDF features	BoW features
Multinomial Naïve Bayes	0.67	0.69
Decision tree	0.646	0.65
Random forest classifier	0.728	0.735
AdaBoost classifier	0.714	0.7225
XGBoost classifier	0.7547	0.7553
Multinomial logistic regression	0.7710	0.7666

33.4 Conclusions and Future Work

Reddit has several concentrated and structured mental health forums which provides its users an opportunity to anonymously share their experiences and engage in peer-to-peer support groups. The main objective of this paper is to use these posts and apply NLP techniques to build a predictive machine learning model targeted to identify a possible mental health condition. Previous works based on this dataset have focused on binary classification task for detection of each mental health category individually. In this work, we have treated it as a multi-class classification problem and have reported the performance of machine learning models with feature engineering which will serve as a preliminary study on using Reddit data for analyzing mental health.

Future work shall be done on building more complex and efficient predictive models that can help us to resolve the imbalanced data problem. Deep neural network-based techniques can be applied to obtain refined results using automated feature extraction supported by these models. The insights gained from our work will help researchers build better predictive models on this dataset using some more sophisticated approach.

References

1. Reddit. <https://www.reddit.com/>
2. The World health report: 2001: Mental health: new understanding, new hope (2001). World Health Organization: Institutional Repository for Information Security. <https://apps.who.int/iris/handle/10665/42390>
3. Ritchie, H.: Global mental health: five key insights which emerge from the data. Our World in Data (2018). <https://ourworldindata.org/global-mental-health>
4. Naslund, J.A., Bondre, A., Torous, J., Aschbrenner, K.A.: Social media and mental health: benefits, risks, and opportunities for research and practice. *J. Technol. Behav. Sci.* **5**, 245–257 (2020)
5. Calvo, R.A., Milne, D.N., Hussain, M.S., Christensen, H.: Natural language processing in mental health applications using non-clinical texts. *Nat. Lang. Eng.* **23**(5), 649–685 (2017)

6. Seal, D., Roy, U.K., Basak, R.: Sentence-level emotion detection from text based on semantic rules. In: Information and Communication Technology for Sustainable Development, pp. 423–430. Springer, Singapore (2020)
7. Herzig, J., Shmueli-Scheuer, M., Konopnicki, D.: Emotion detection from text via ensemble classification using word embeddings. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, pp. 269–272 (2017)
8. Ryan, S., Doherty, G.: Fairness definitions for digital mental health applications
9. Thieme, A., Belgrave, D., Doherty, G.: Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans. Comput.-Hum. Interact. (TOCHI)* **27**(5), 1–53 (2020)
10. Zhou, D., Luo, J., Silenzio, V.M., Zhou, Y., Hu, J., Currier, G., Kautz, H.: Tackling mental health by integrating unobtrusive multimodal sensing. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
11. Fatima, I., Mukhtar, H., Ahmad, H.F., Rajpoot, K.: Analysis of user-generated content from online social communities to characterise and predict depression degree. *J. Inf. Sci.* **44**(5), 683–695 (2018)
12. Chen, X., Sykora, M.D., Jackson, T.W., Elayan, S.: What about mood swings: identifying depression on twitter with temporal measures of emotions. In: Companion Proceedings of the Web Conference 2018, pp. 1653–1660 (2018)
13. Suhasini, M., Srinivasu, B.: Emotion detection framework for twitter data using supervised classifiers. In: Data Engineering and Communication Technology, pp. 565–576. Springer, Singapore (2020)
14. Gaind, B., Syal, V., & Padgalwar, S.: Emotion detection and analysis on social media (2019). arXiv preprint [arXiv:1901.08458](https://arxiv.org/abs/1901.08458)
15. Saha, K., De Choudhury, M.: Modeling stress with social media around incidents of gun violence on college campuses. *Proc. ACM Hum.-Comput. Interact.* **1**(CSCW), 1–27 (2017)
16. Pestian, J., Santel, D., Sorter, M., Bayram, U., Connolly, B., Glauser, T., DelBello, M., Tamang, S., Cohen, K.: A machine learning approach to identifying changes in suicidal language. *Suicide Life-Threat. Behav.* **50**(5), 939–947 (2020)
17. Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T.J., Dobson, R.J., Dutta, R.: Characterisation of mental health conditions in social media using Informed Deep Learning. *Sci. Rep.* **7**(1), 1–11 (2017)
18. Su, C., Xu, Z., Pathak, J., Wang, F.: Deep learning in mental health outcome research: a scoping review. *Transl. Psychiatry* **10**(1), 1–26 (2020)
19. Ragheb, W., Azé, J., Bringay, S., Servajean, M.: Attention-based modeling for emotion detection and classification in textual conversations (2019). arXiv preprint [arXiv:1906.07020](https://arxiv.org/abs/1906.07020)
20. Sekulić, I., Strube, M.: Adapting deep learning methods for mental health prediction on social media (2020). arXiv preprint [arXiv:2003.07634](https://arxiv.org/abs/2003.07634)
21. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)
22. Dheeraj, K., Ramakrishnudu, T.: Negative emotions detection on online mental-health related patients texts using the deep learning with mha-bcnn model. *Expert Syst. Appl.* **182**, 115265 (2021)
23. Kim, J., Lee, J., Park, E., Han, J.: A deep learning model for detecting mental illness from user content on social media. *Sci. Rep.* **10**(1), 1–6 (2020)
24. Natural Language Toolkit. NLTK 3.6.2 documentation. <https://www.nltk.org/>.

Chapter 34

A Computational Approach to Assess Breast Cancer Risk in Relation with Lifestyle Factors



Oindrila Das, Itismita Pradhan, Riddhi Chatterjee, and Satya Ranjan Dash

Abstract Lifestyle changes alter behavioral and psychological makeup in a human being. Many people in our society adopt harmful lifestyles due to psychosocial pressure without thinking about their physical health which turns to be a risk factor, expanding the possibility of getting affected by deadly diseases. One of the major diseases, a plausible result of unhealthy lifestyle choices like poor diet, less physical activity, stress, obesity, consumption of a high amount of alcohol, frequent smoking, is breast cancer. The harmful effects of the lifestyle choices listed above impact the therapies related to breast cancer. Among the risk factors, we considered some of the ones which are scientifically established and some which are yet to be properly established. Modification of lifestyle choices offers hope for breast cancer survivors to improve their overall quality of life. In this paper, we focused to evaluate, by using computational models, the chances of getting breast cancer, based on the rise in estrogen level and BRCA1 gene mutation, as a result of four risky health behaviors—stress, smoking, alcohol, and obesity.

34.1 Introduction

Cancer ranks second in the leading cause of death and constitutes a major health problem for people all over the world. Breast cancer is noted as most common in women and very rare in men, due to lifestyle changes. At the age of puberty due to

O. Das

School of Biotechnology, KIIT University, Bhubaneswar, Odisha, India

I. Pradhan

Trident Academy of Creative Technology, Bhubaneswar, Odisha, India

R. Chatterjee

Heritage Institute of Technology, Maulana Abul Kalam Azad University of Technology, Kolkata, West Bengal, India

S. R. Dash (✉)

School of Computer Applications, KIIT University, Bhubaneswar, Odisha, India

e-mail: sdashfca@kiit.ac.in

the association of estrogen with growth hormone, there is a development of breast in female humans. Subcutaneous fat that surrounds a network of ducts, merge on the nipple and form a cellular organization that grants the breast its perfect size and shape [1]. Generally, breast cancer occurs in either ducts or lobules of the breast. There is a remarkable association between the growth of tumors in breast cells and estrogen receptors (ER) [2]. Encouragement to the growth of the MCF-7 cells, a breast cell line of a human containing progesterone receptors and ERs, is dependent on both low level and high levels of estrogen which may differ from case to case and articulation of different proteins in human breast cancer cells is based on the level of estrogen [3]. In most cases, MCF-7 cells are accustomed to studying action mechanisms related to cell proliferation and estrogen level, and thus in the in-vitro studies, consideration of MCF-7 cells occurs due to the preparation of estrogen through the ER in the form of estradiol [4]. Lifestyle changes have a greater effect not only on our physical health but also on our mental health and the primary effect of which is stress, one of the fundamental factors leading its way to breast cancer. Innumerable factors influence breast cancer but, in our work, we have focused on four particular lifestyle changes—stress, smoking, alcohol intake, obesity, and their carcinogenic effects on our health. The detailed discussion is as follows:

1. Stress—Unpleasant life situations trigger mental cycles that influence health behaviors and neuroendocrine modulation through four basic ways:

- Stressors can prompt manifestations of distress and increment risks of psychological problems.
- The negative feelings associated with mental pressure affect psychological well-being unfavorably [5].
- The intense negative reactions lead to biological dysregulation that can add to pointers of subclinical sickness.
- Endeavors to adapt to negative enthusiastic reactions can prompt an expansion in harmful practices (tobacco, substance abuse, over or under-eating, etc.).

Studies from molecular biology uncovered the physiological impacts of the chemical cortisol on the mammary gland. They are also related to organ improvement, estrogen movement, and other intracellular pathways associated with breast cancer. Over time, logical proof has fortified concerning the natural believability that upsetting life occasions lead to an expanded shot at creating bosom malignancy in females. Women with high levels of caregiving stress had considerably lower levels of estradiol than women who did not provide care, according to the Nurses' Health Study [6], implying that greater stress may protect against breast cancer by lowering endogenous estrogen levels.

2. Smoking—The most harmful carcinogenic material used in cigarettes is polycyclic aromatic hydrocarbons. Particularly tobacco nitrosamines are metabolically effective and help in the restraining of DNA in epithelial cells (the cells present on the surface) of the breast. They permit permanent mutation through

replication by the breakaway of repair mechanisms by cells and it is correlated with p53 and various oncogenes [7]. Mainly the ER expressions are correlated with the components of cigarettes which lead to the recurrent tumor in mammary glands. Particular identification of expressions interconnected to ER is very much important for therapies related to breast cancer. Smoking increases the proliferation activity of breast cancer cells and is also a reason for the accumulation of fat in the abdominal area and weight gain [8]. The immune system incapacitates due to heavy smoking and also makes it difficult to fight against cancerous cells, as it is a cause of unregulated cell growth. Smoking not only damages the cell but also makes the cell resistant to repair itself [9]. Blood circulation is the way to harm the mammary gland tissues through the damage of DNA by the effect of tobacco and its harmful components. Mainly the estrogen receptor expressions are correlated with the bad components of cigarettes which leads to a recurrent tumor in mammary glands [10].

3. Alcohol Consumption—The intake of alcohol beyond tolerable limits was established to be linked with increased estrogen levels in the blood and also has a high impact on the function of ER [2]. ER is not present in all cancer cells but in taking higher levels of alcohol can be crucial for the receptor-containing cancer cells and more crucial for postmenopausal women. Ethanol is related to have a carcinogenic effect not only on the stromal cells but also on epithelial cells [11]. Alcohol intake activates the estrogen-sensitive cells to behave as cancerous cells. It can turn androgens to estrogen to some extent by escalating plasma estrogen levels. The breast cells have ERs that get affected by a higher intake of alcohol and thus bind to DNA and control various gene activities. At the time of breakdown of alcohol in the body, there is the formation of acetaldehyde which plays a key role in DNA mutagenesis through reactive oxygen species [12]. ER- α -independent genotoxic pathway is a track that gives rise to breast cancer related to estrogen [11].
4. Obesity—Obesity is generally described as a weight record of Body Mass Index (BMI). A BMI of 30 or more serves as an outstanding ally for mortality and in most cases an unequivocal cause of death from cardio metabolic diseases like diabetes and coronary course disorder [13]. A high BMI has a large association with an extended risk of inflammatory breast cancer (IBC), which is the most destructive kind of breast cancer in both premenopausal and postmenopausal women. As the number of young, overweight individuals continue rising, an associated speed expansion in the overall infection inconvenience is presumably going to follow. In robust women, observational examinations exhibit 2.12-cross-over development in the overall peril of death from breast cancer and 6.25-overlay extension in the general peril of death from uterine cancer [14]. The crucial wellspring of estrogen is from the difference in androgen precursor, androstenedione in the periphery adipocytes to estrogen.

The diagram below (Fig. 34.1) gives an overview of the four lifestyle factors considered for breast cancer risk assessment.

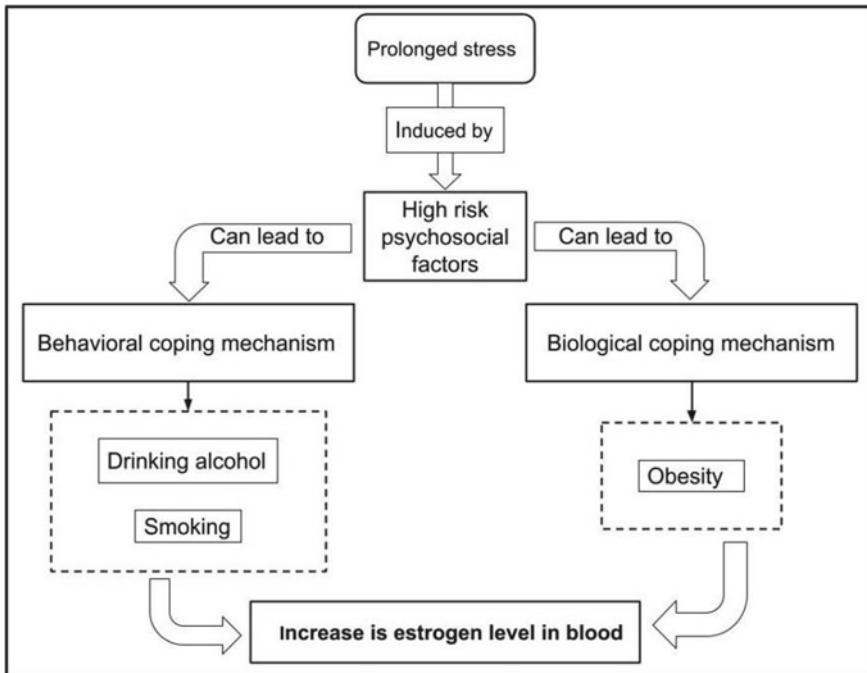


Fig. 34.1 Pictorial representation showing the evolution and connection of the four lifestyle factors with an increase in estrogen level concerning breast cancer

34.2 Our Approach

We propose two computational models (Figs. 34.2 and 34.3) to elucidate the various kinds of the probability of developing breast cancer in individuals. Our first model is a risk assessment model that demonstrates the influence of the four lifestyle factors, and their severity. The second model takes on a pathophysiological approach, using artificial intelligence (AI) on estrogen receptor (ER), to exemplify the chances of getting breast cancer.

Two potential components are suggested to aid in the understanding of how these psychosocial risk factors increase the chance of breast cancer. The first one is a social component such that the cooperation between high danger psychosocial states and prompted unhealthy practices like liquor consumption, tobacco intake inevitably leads these variables to fill in as mediators for breast cancer risk. The second component is a biological system-high danger psychosocial factors lead straightforwardly to physiological states related to the development of breast tumor malignancy through endocrine or immunological variables. Hormones assume a significant part in the advancement of breast cancer. In this manner, different upsetting life occasions without social help may impact malignant growth. They are damaging through the initiation of the autonomic sensory system and the hypothalamic-pituitary-adrenal

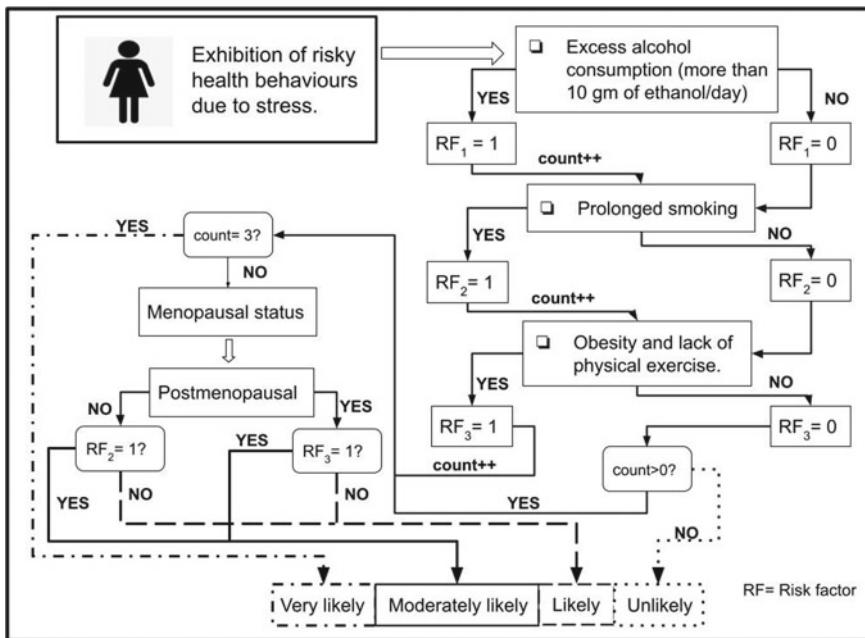


Fig. 34.2 Computational model showing the four chosen risk factors augmenting breast cancer risk through a four-part probability spectrum

axis by bringing an increment in endogenous estrogens. MCF-7 is greatly affected by an increase in the amount of nicotine in our bodies [4]. As the MCF-7 has estrogen and progesterone receptors, heavy smoking is dangerous as unrestrained hormonal stimulation is a cause of tumor growth. Mutation of gene BRCA1/2 due to heavy smoking is a common endangered factor. As alcohol provokes a rise in estrogen level, it works as a tumor inducer. However, it also acts as a primary cause for metastasis [15]. The high amount of alcohol consumption can decrease the BRCA1 expression [16]. Studies clearly show that interaction molecules such as E-cadherin and catenins expressions are noticeably low due to the cause of higher exposure to alcohol. They are specific for the cell to maintain tissue integrity. Destruction of DNA in the MCF-7 cells and p53 DNA repair pathways are affected by alcohol intake. Extended levels of ability of estrogen assemble cell development and angiogenesis through various mechanisms, including confining to the ER and stimulating the IGF1 hailing pathway in chest cancer. The main cell type responsible for estrogen biosynthesis in human chest preadipocytes or fat stromal cells. These cells drive the release of aromatase enzyme which converts androgens to estrogens. Leptin-induced MCF-7 cell augmentation is connected with an extended explanation of ER (α).

Figure 34.2 illustrates a computational model that determines the risk probability of individual developing breast cancer, taking into consideration four recognized lifestyle causal factors—stress, alcohol, smoking, and obesity. Current literature has

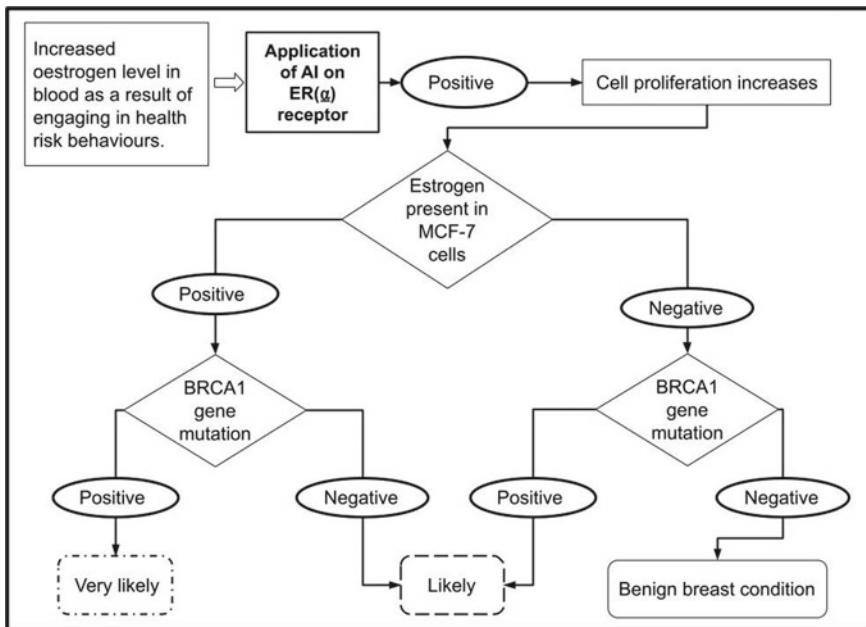


Fig. 34.3 Model showing the rise in estrogen level (a consequence of the four mentioned lifestyle factors), its effect, and the influence of the BRC1 gene on breast cancer risk

shown convincing evidence of these four factors attributing to breast cancer; hence our model is built on them. The model employs customized information and user responses, from a binary set of choices, to compute the likelihood of a person risking the disease. The quantitative definition of what constitutes “excess”, in the case of the first determinant, is based on the recommendation of The World Cancer Research Fund, which suggests a safe drinking limit equivalent to “10 gm of ethanol per day.” The menopausal status of a woman has been taken into account as various studies demonstrate dependency between it and the second and third risk factors. The findings of a cohort study [17] reveal the association of long-term smoking with an increased risk of breast cancer, especially among young women. Similarly, Botoroff et al. [18] assert that young smokers have a higher chance of developing premenopausal breast cancer. A multitude of observational studies denotes a strong linkage between obesity and elevated risk of breast cancer in postmenopausal women [19]. A survey of 1.2 million women who are aged between 50 and 64 years was conducted by Reeves et al. [20]. The survey illustrated that those women who are obese are at a 30% higher risk of being affected by breast cancer.

Based on the user input, a risk assessment is performed. A neural network model can be utilized for prediction. Machine learning techniques [21] can be used to analyze the collected user responses for the mining of beneficial insights thus aiding further research in the domain of breast cancer. Frequently updating the model, making use of feedback and expert suggestions, will enable a better user experience.

Figure 34.3 illustrates a model showing the process in how the increase in estrogen level, due to the four lifestyle factors, impacts breast cancer risk. The estrogen receptor ER (α) targets its presence in the MCF-7 cell line present in the breast adipose tissues and cell proliferation increases due to the rise in estrogen level. Artificial intelligence (AI) is applied to ER (α), which acts as a checkpoint to estimate the level of estrogen hormone. If there is proved the presence of estrogen higher than normal level in MCF-7 cells, and along with that BRCA1 gene mutation has also occurred, then the chances of breast cancer become very likely. However, if BRCA1 gene mutation has not occurred, then the probability of breast cancer risk is slightly reduced than the aforementioned likelihood. Likewise, if no unusual rise in estrogen level in the MCF-7 cells is detected, but there is a mutation in the BRCA1 gene, the possibility of breast cancer risk remains the same as before. We consider it to be a benign breast condition when neither unusual rise in estrogen level is detected in the MCF-7 cells nor BRC1 gene mutation is observed. This condition, in very rare cases, may lead to breast cancer.

34.3 Discussion

Modification in lifestyle choices is essential for keeping distance from breast cancer occurrence and for treatment and survival from the disease. Physical activity and a healthy diet are the two main keys to maintain a healthy body, stress-free, and decreasing the risk of breast cancer. Substance use disorders escalate the path to different types of cancer and other diseases. Many studies and models show the impact of lifestyle changes but show no specific correlation with one another. In this paper, we attempted to show the four most common lifestyle factors that are stress, alcohol consumption, smoking, obesity, and their link with breast cancer. We proposed two computational models showing the correlation of the four lifestyle factors for assessing breast cancer risk probability. We further demonstrated how these four causal factors lead to an increase in estrogen level in the MCF-7 cells present in breast adipocyte tissues. The rise in estrogen levels leads to uncontrolled cell proliferation which in turn leads to malignant breast tumors and thus breast cancer. These factors also have varied effects on premenopausal and postmenopausal women. Smoking has a greater effect on premenopausal women, whereas alcohol consumption has a greater effect on postmenopausal women. Obesity acts as a major risk factor due to the deposition of adipose tissues in the breast and due to the overall increase in BMI. Although in the case of obesity, other hormones like leptin produced by fat cells have an effect on the rise in estrogen level causing obesity. Our second model also showed the effect caused by the mutation of the BRCA1 gene. Although there is not much of a connection between mutation of the gene and rise in estrogen level leading to breast cancer but much works of literature have also suggested it as a valid point and can be progressed for future work. If both estrogen and BRCA1 gene alteration occurs due to lifestyle change factors, then there is a high risk of breast cancer but if either of the two takes place then there is

a moderate chance of breast cancer. Although there is less connection between the effect of stress and breast cancer, as it is a psychosocial matter, it may lead to mental pressure and while adapting to this imbalanced state, people can consume junk food and succumb to unhealthy lifestyle habits which eventually cause deadly diseases. Thus, consideration of a healthy and active lifestyle is the ultimate way to stay away not only from cancer but from many diseases.

34.4 Conclusion and Future Prospect

Breast cancer has gained immense public attention due to being a life-threatening disease. Extensive research is conducted on it and its cure. However, the impact of lifestyle choices on the disease remains unclear. Though lifestyle factors have unrealized and prospective effects, mingling with biological factors provides a route to breast cancer. The growth of breast cancer is distinguished by hormonal control and in most cases, the notable amount of estrogen receptors are there. Though there is no proper significant evidence of the direct connection of intake of alcohol with breast cancer many works of literature already showed that there is a little higher chance of getting breast cancer by consuming alcohol post menopause [22]. There is also a chance of breast cancer due to alcohol intake in the pre-menopause state but the chances are less than post menopause [12]. There exist many contradictory statements regarding the relationship between alcohol and breast cancer. The amount of alcohol intake in day-to-day life is a variable too [22]. Carcinogens present in tobacco smoke may induce the risk of breast cancer [23]. Many studies have been done on the duration of smoking, starting age of smoking, the regular quantity of smoking, and their corresponding relationship towards breast cancer, but it is not established yet with appropriate same evidence in all cases [24]. Women who smoke regularly have a higher chance of breast cancer. It is also related to the number of cigarettes and alcohol consumed in daily life but in some cases taking to the habit of smoking after menarche or early age, especially before the first pregnancy exacerbates chances of breast cancer [25]. The adverse conditions in our life manifest excessive stress which not only affects our mental stability but also our hormonal levels. Due to over fluctuation of the hormonal levels some women put on excess weight. Thus, the condition of stress is indirectly related to breast cancer via weight gain. Predictable, free, and positive affiliations have been found between obesity and breast cancer in postmenopausal women. In most of the studies on obesity-related with helpless anticipation of breast tumor, malignant growth is observed in both premenopausal and postmenopausal women. The direct relation to stress and breast cancer is not yet known, and hence marks as a great work in the future. Women who do not breastfeed after pregnancy also develop some chances of breast cancer.

Accurate datasets related to the combination of lifestyle changes, estrogen level, cell proliferation, and occurrence of breast cancer are needed in future studies to analyze the chances of breast cancer by evaluating estrogen level varies depending on the lifestyle factors. Neural network models can be used for predicting and analyzing

the affecting factors to interpret the chances of breast cancer clinically. Cellular, physical and genomic changes due to unhealthy lifestyles should be observed to distinguish between a normal cell and a tumor cell. This is also required to predict the chances of breast cancer in the early stage by considering estrogen levels [11]. It is expected that obesity in women will show its impact on the expanded rate of breast tumor malignant growth in postmenopausal women in the coming years [26]. Ergo, a general well-being strategy, arranging for well-being training efforts are direly needed to resolve the rampant and ubiquitous complications of breast cancer.

References

1. American Joint Committee on Cancer: Breast. AJCC Cancer Staging Manual, pp. 223–240. Springer, New York (2002)
2. Saceda, M., et al.: Regulation of the estrogen receptor in MCF-7 cells by estradiol. *Molekul. Endocrinol.* **2**(12), 1157–1162 (1988)
3. Hagini, Y., et al.: Estrogen inhibits the growth of MCF-7 cell variants resistant to transforming growth factor-beta. *Japan. J. Cancer Res.* **79**(1), 74–81 (1988)
4. Hsieh, C.-Y., et al.: Estrogenic effects of genistein on the growth of estrogen receptor-positive human breast cancer (MCF-7) cells in vitro and in vivo. *Cancer Res.* **58**(17), 3833–3838 (1998)
5. Bowen, D.J., et al.: The role of stress in breast cancer incidence: risk factors, interventions, and directions for the future. *Int. J. Environ. Res. Public Health* **18**(4), 1871 (2021)
6. Michael, Y.L., et al.: Influence of stressors on breast cancer incidence in the Women's Health Initiative. *Health Psychol.* **28**(2), 137 (2009)
7. Kispert, S., McHowat, J.: Recent insights into cigarette smoking as a lifestyle risk factor for breast cancer. *Breast Cancer: Targets Therapy* **9**, 127 (2017)
8. Samaras, K., et al.: Tobacco smoking and estrogen replacement are associated with lower total and central fat in monozygotic twins. *Int. J. Obes.* **22**(2), 149–156 (1998)
9. Brown, K.F., et al.: The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. *Br. J. Cancer* **118**(8), 1130–1141 (2018)
10. Takada, K., et al.: The effect of smoking on biological change of recurrent breast cancer. *J. Transl. Med.* **18**(1), 1–12 (2020)
11. Liu, Y., Nguyen, N., Colditz, G.A.: Links between alcohol consumption and breast cancer: a look at the evidence. *Womens Health* **11**(1), 65–77 (2015)
12. Seitz, H.K., Stickel, F.: Molecular mechanisms of alcohol-mediated carcinogenesis. *Nat. Rev. Cancer* **7**(8), 599–612 (2007)
13. Carmichael, A.R., Bates, T.: Obesity and breast cancer: a review of the literature. *The Breast* **13**(2), 85–92 (2004)
14. Rubinstein, M.M., Brown, K.A., Iyengar, N.M.: Targeting obesity-related dysfunction in hormonally driven cancers. *Br. J. Cancer* 1–15 (2021)
15. Al-Sader, H., et al.: Alcohol and breast cancer: the mechanisms explained. *J. Clin. Med. Res.* **1**(3), 125 (2009)
16. Russo, J., Russo, I.H.: The role of estrogen in the initiation of breast cancer. *J. Steroid Biochem. Mol. Biol.* **102**(1–5), 89–96 (2006)
17. Jones, M.E., et al.: Smoking and risk of breast cancer in the generations study cohort. *Breast Cancer Res.* **19**(1), 1–14 (2017)
18. Bottorff, J.L., et al.: Young women's responses to smoking and breast cancer risk information. *Health Educ. Res.* **25**(4), 668–677 (2010)
19. Picon-Ruiz, M., et al.: Obesity and adverse breast cancer risk and outcome: mechanistic insights and strategies for intervention. *CA: Cancer J. Clin.* **67**(5), 378–397 (2017)

20. Reeves, G.K., et al.: Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. *BMJ* **335**(7630), 1134 (2007)
21. Behravan, H., et al.: Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Sci. Rep.* **10**(1), 1–16 (2020)
22. Zhao, M., et al.: p53 pathway determines the cellular response to alcohol-induced DNA damage in MCF-7 breast cancer cells. *PLoS One* **12**(4), e0175121 (2017)
23. Reynolds, P.: Smoking and breast cancer. *J. Mammary Gland Biol. Neoplasia* **18**(1), 15–23 (2013)
24. Cui, Y., Miller, A.B., Rohan, T.E.: Cigarette smoking and breast cancer risk: update of a prospective cohort study. *Breast Cancer Res. Treat.* **100**(3), 293–299 (2006)
25. Clemons, M., Goss, P.: Estrogen and the risk of breast cancer. *N. Engl. J. Med.* **344**(4), 276–285 (2001)
26. Kelsey, J.L., Bernstein, L.: Epidemiology and prevention of breast cancer. *Annu. Rev. Public Health* **17**(1), 47–67 (1996)

Chapter 35

Digitalization of Education: Rural India's Potential to Adapt to the Digital Transformation as New Normality



Ankita Sahu and Swati Samantaray

Abstract Education at its core is a social endeavour and the use of technology in education has a decisive role in providing innovative forms of support to teachers, students and the learning process more copiously. As a source of empowerment in education, technology is enhancing teaching and learning in a blended learning environment. The digital education in India is strengthening the education level and developing the nation to take a stand in the worldwide competition. With an increase in the active Internet subscribers, there is a rise in the use of diverse technologies as well as the web tools for the teaching and learning process. In order to enhance the educational system, the government is trying its best to provide all possible technological facilities for the students as well as the teachers alike so that they may equip themselves with the advanced technologies and web tools. This paper not only assesses the digital shift in education, but also estimates the capability of rural India to embrace the forthcoming digital changes. In the rural areas, adapting to newer technologies in education is of course vital, yet still a challenging task. Hence, this paper makes a modest attempt in spreading awareness regarding the digitalization needs in education and the implementation of various technological schemes launched by the government for digital transformation of education system amongst the learners in the rural areas in order to meet the future technological challenges.

35.1 Introduction

'Digitalization' and 'technology' are the buzz words ruling the world's development currently, being used in almost every aspect of our lives, including the education sector as well. Learners' inclination towards advanced technologies are replacing the printed book-based learning. Using Internet is more convenient, time saving and provides infotainment, which is a vital reason for digital transformation in education system in India. The use of diverse technologies in education has not

A. Sahu · S. Samantaray (✉)

School of Humanities, KIIT Deemed To Be University, Bhubaneswar, India

only enhanced the performance of students and teachers, but its use has magnificent impact upon the educational institutions as a whole. Essentially, educational institutions are compelled to adopt digital platform for blended teaching-learning process. The consequences of the changes at a global level have raised the demand for digital technologies in education sectors. E-learning opportunities and the use of web tools in the education system have journeyed beyond traditional educational system; moreover, they are more informative, convenient, interesting as well as interactive, entertaining at the same time and thus largely adopted by the organizations. The availability of texts, animated videos and practical assessments online, is a boon for the differently-able learners. In order to have an effective management in teaching and learning, applications like Internet of things (IOT) have separate study materials, according to the requirements for both students and teachers. Other technologies like virtual reality and gamification, artificial intelligence, machine learning are also used for education purposes. Use of these technologies is extremely flexible for both students and teachers anywhere and anytime. They help learners to have an individualized educational experience in a productive way. India is moving towards a holistic and multidisciplinary form of education while leveraging technology to strengthen teaching-learning from school to higher level. In this digital era it is important for the individuals, communities and organizations to have a proper access to the Internet and its facilities [1].

35.2 Literature Review

A survey was conducted (2021) in a blended learning setting by Md. Kabirul for his work *Promoting Student-Centred Blended Learning in Higher Education: A Model* [2]. The study aimed at examining the course design strategies adopted by the specific course teachers in a blended learning setting so as to enhance the students' online communication in a collaborative manner. The findings of the survey suggest a model that would strengthen the connection between a student and the teacher in both synchronous and asynchronous modes of education system. Kurt D. Squire (2021) in the paper *From Virtual to Participatory Learning with Technology During COVID-19* [3], has reflected that the utilization of technology by the youth in both virtual schooling and outside the school within United States mark a vast difference in their performances. Students' interaction and participation during the virtual school remains grounded and limited within the walls of the school while learning. But they are actively involved, extremely anxious and even put an extra effort in order to find information, while using technology for their personal entertainment or learning at home. Their work focuses on the need to re-think about the role of technology for school learning if the situation compels to continue with the virtual mode by introducing better aid for communication, and emphasising upon creative teaching learning with better technical infrastructure for conducting better virtual schooling.

Yet another research work in 2021 by Elizabeth Wargo entitled *On the Digital Frontier: Stakeholders in Rural Areas Take on Educational Technology and Schooling*

[4], insists upon the perception of a stakeholder on educational technology concerning six district schools of the state of Idaho (United States). According to the stakeholder, the district schools equally understand the importance and accept inculcating technologies in education system. Due to lack of resources and improper management of educational facilities hinders the educational growth in these areas. Therefore, this study exemplifies how by paying particular attention to the real experts' (stakeholder) perceptions in rural places can help prepare students and communities for future, by leveraging the best what technology has to offer. Similarly a case study on *Supporting New Online Instructors and Engaging Remote Learners During COVID-19: A Distributed Team Teaching Approach* [5] by Vanessa P. (2021), explores upon assisting pre-service teachers/instructors through distributed team strategy. The findings suggests that this collaborative process was extremely helpful in teaching and learning by reducing other academic and technical related tasks and focusing more upon the students individual requirements related to the curriculum during such adverse condition.

Lokanath Mishra's work (2020) on *Online Teaching-Learning in Higher Education During Lockdown Period of COVID-19 Pandemic* [6], delineates a key to overcome obstacles in teaching- learning process during pandemic by making small changes in management procedure and online education. The study is basically about a productive transformation of the educational institutions from physical to online classes using virtual mode and digital tools amid COVID-19 wave. In 2020, Shivangi Dhawan's *Online Learning: A Panacea in the Time of COVID-19 Crisis* [7], mentions about the Indian education which followed traditional set up before COVID-19, while the blended learning was already introduced in the education system but was not prioritised as traditional system of education. The onset of pandemic challenged the education system across world overnight this drastic shift needs an online transfer. Thus, the article shows the importance of online learning and strengths, weaknesses, opportunities and obstacles during the period of crisis. The paper also illustrates on the rise in educational technological companies during pandemic and concurrently suggests educational institutions to deal with the issues related to online education. In another paper by Hye Jeong Kim (2019) titled *The Roles of Academic Engagement and Digital Readiness in Students' Achievements in University E-Learning Environment* [8] scrutinizes the perceptions of the university students regarding online education according to their personal experiences and the availability of digital tools and institutional management inside the University for Educational Developments. While the students respond positively about the online education experience within the campus. But in order to have complete involvement in digital environment they essentially need to have a command over digital skills. The study suggests various effective results to amplify accepting online educational environments by the academic institutions.

An editorial by Shailendra Palvia (2018) *Online Education: Worldwide Status, Challenges, Trends and Implications* [9] emphasises upon the reasons for an increase in the online education globally, like the introduction of the advanced technologies rise in the use of Internet worldwide with an increase in the demand for digital skills in order to fulfil the needs of digital economy. Simultaneously, the paper brings up

various obstacles blocking the smooth functioning of online education within the five territories of the world. The reason being variety of people with different cultures, no proper management, shortage of required resources and lack of training facilities and skill developments are some of the common reasons discovered as the barriers for the smooth functioning of online education in most of the regions of the world. The chapter entitled *Adopting New Digital Technologies in Education: Professional Learning* by Wan Ng (2015) talks about the educators who supposed to have good content of knowledge as they are the key determinant of students' achievement. In order to have an upgraded version of teaching quality, it is essential to include current technologies into their education system [10].

35.3 Objectives

Though plenty of researches are done on the benefits of using digital technology in education, convenience of blended learning, teachers' role while choosing the online curriculum, transformation of education during COVID-19 pandemic, yet there are some unexplored areas. This paper makes a modest attempt to show how the introduction of new technologies in the contemporary era like AI, IoT, cloud computing and the like are capable of transforming the entire education system and the youth's reaction towards the online education. Further, the paper also discusses on the potential of rural India in embracing the technological changes related to education, focusing upon various roadblocks as well as the possible solutions for overcoming the pitfalls.

35.4 An Interplay of Technology and Education

Digital technologies are used in various educational institutions for skill development, testing operating system and other similar tasks. In the upcoming days new technologies may efficiently work for streamlining admin tasks, where teachers would have ample time and freedom to look after the requirements of the students. Thus, machine will take over maximum charge of the teacher's extra work pressure and both teachers and machine may work for the best outcome of students. The advanced technologies will help students to access classrooms at a global level. Multi-lingual learners, differently-able students, and the students who missed classes for some unavoidable reasons can use presentation translator which is a free plug-in for PowerPoint that creates subtitles in real time about the lecture. The sophistication of the machine will lead to various possibilities like, it could detect the confused expressions of the student who is struggling to understand a topic; the machine might modify a lesson for the student according to his flexibility and understanding. Certainly in the upcoming days technologies like Internet of things, artificial intelligence, virtual and augmented realities, robotics will inculcate many novel features

that will make teaching and learning extremely flexible and convenient. Furthermore, educational institutions may not have any specific physical location, education will not be limited to students' age and class, and there will be a wide platform open for all to explore on any topic beyond their curriculum.

Clearly the unrestricted availability of Internet helps students to access the experts and the content beyond the limited existing material within the classrooms. The vast availability of the materials related to the school curriculums is well explained and presented by the experts online with minimum or no charges. Considering the above mentioned factors the changes in future education is obvious. EdTech is booming—the improvements are a shift in the paradigm as it has evolved to encompass a plethora of new technological pedagogies. According to the Organization for Economic Co-operation and Development (OECD) reports, the global expenditure in the EdTech industry is predicted to grow from USD 163 billion in 2019 to USD 404 billion in 2025 [11]. The pandemic has resulted in leveraging online education as a new normality and augmented in broadening our knowledge and idea about smart technologies. Employment of technologies in education is not only time saving, but it also amplifies the rate of learning, simultaneously increasing the educational productivity with affordable costs.

Technology is regarded as a superb equalizer and a mode of getting in touch with people that may not have been possible ever before [12]. Virtual reality creates an artificial setting of real experience on specific topics for the students that will help them in diverse learning needs. It also reduces the gaps in the learning process by connecting them with other sources and international students. VR in the education field also takes into count cultural competence, the ability to understand different cultures and values, which is again an important skill in the present interconnected global society. Augmented reality possess the feature of creating a magnified real physical world using various features like digital visual elements, sound, or other sensory stimuli delivered through technology. The introduction of these features in the contemporary education system has definitely transformed the pattern of teaching and learning process to make it more informative and time saving. Augmented reality obviously helps learners to understand geometric concepts easily through the 3D geometric forms from different perspectives. Mixed reality is the combination of both the real world with the virtual world in order to create an absolutely new environment and visualization, where physical and real objects co-exist and interact in real time. Mixed reality can be considered as the hybrid form of reality and virtual reality. The new technologies helps learners to understand data set, complex formulae and abstract concepts which seem to be difficult to grasp in the verbal instructions of the teachers by presenting artificial images in the real forms. It goes without saying that technology works wonders in education sectors because visual activities and explanations are more fascinating and easy to grasp than only listening to lectures. Further, laboratory testing and experimenting are easier with the required equipment.

Adaptive learning, as a part of the new norm uses algorithms as well as artificial intelligence to coordinate proper interaction with the learners and delivers customized resources and learning activities to address the unique needs of each learner (with different learning outcomes). Learning analytics is expected to dive into the education

sector in the coming future in the educational research, learning and assessment science, educational technology. It will help in tracking the students' performances, and also will help in designing assessments that can help in the improving, and filling the gaps in both teaching and learning for a long run. 'Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs' [13].

35.5 Rural India's Potential and Challenges for Digital Transformation in Education

World's development is taking a technological turn in every sector today. In the terms of development India too have relentless potential to become successful economically and technologically because half of the population in India comes under the working group and at present students are keenly inclined towards adapting the advanced technologies which are essential for future developments. This present pandemic has resulted a shift from physical access of classroom education to digital or online form of education and Indian education system have whole heartedly tried accepting digital form of education. Though digital education was practiced in India before pandemic, yet traditional form was much more prioritized. However, recent status shows a rise in the Internet consumers in the country. The following graph from Statista shows the Internet penetration rate of the country in the past few years.

Figure 35.1 clearly portrays the rise in the penetration rate of Internet in India from 2009 to 2020 [14]. With the increase in the use of the Internet there has been rise in the demand for digital use in almost every sector of the society. The utilization of digital technology has made the country progress in myriad ways, specifically the education system utilizing the virtual mode of education. Looking at the inclination of the scholars towards the technologies, schools and other educational institutions

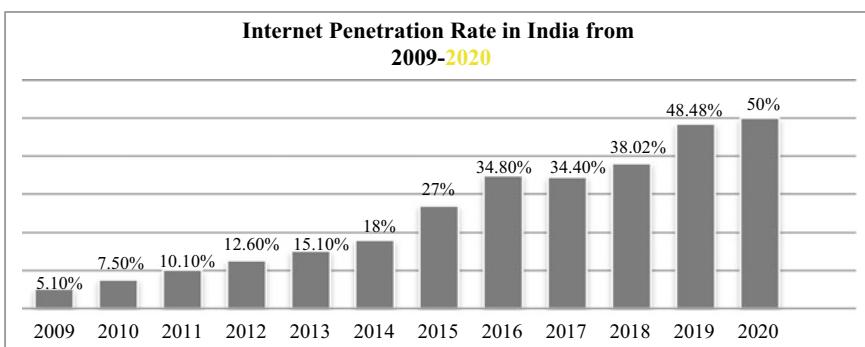


Fig. 35.1 Internet penetration rate in India from 2009 to 2020

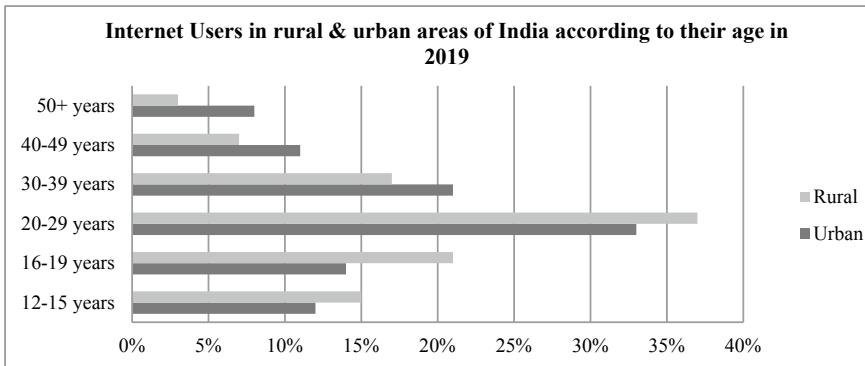


Fig. 35.2 Internet users in rural and urban areas of India by age group in 2019

are including various web tools in the classroom to make learning more fun, interactive and informative. These digitalized educational tools turn out to be encouraging and collaborative facilitating a keen communication bonding between educators and learners. This is a vital reason why it is fascinating yet surprising to find this inclination and increase in the rate of penetration of Internet is more among the rural population compared to the urban areas.

Figure 35.2 is a Survey of Research Department 2021 reflects a verified survey of the Internet consumers more in rural than in urban area which is 37% compared to 33% in urban areas, consumers are between 20 and 29 years of age [15].

35.5.1 Reasons for Digital Learning and the Escalation of Internet

Government initiatives of low cost of Internet, easy-access, e-commerce activities, many schemes to encourage rural digital education like Disksha Portal, PM eVidya, Vidya-Daan, Samagra Shikha and the like, and increase in income stream among the rural population are some of the basic components determined in the upward movement of Internet consumption. Entertainment and communication sectors, is another section which is adding on a huge amount of customers to its lists. Looking back, over time rural Indians' are potentially has taken an upswing in purchasing power and a better economic progress digitally, due to more number of youth population. It is a known fact that Millennial (Defined as those currently in age-band between 15 and 35 years) are the central reason for the digital growth of every country economically. Thus, India remains no exception to it. The Millennial population of India is around 450 Million out of which around 300 Million Millennial (67% of total population) live in rural India [16], and are responsible for an increase in number of Internet users recently.

Germinating digital footprint has witnessed a greater democratization of knowledge and access to skills and services. The weaker sections of the society concerning to gender or community or geography is entitled to incorporate digital revolution today. The present pandemic situation is one of the vital reasons for enormous consumption by the Internet users. The survey shows that there has been a visible increase in the enrolment rate of the students in government schools in order to continue with their studies in online mode irrespective of problem of inequality and digital divide in rural India. According to 15th Annual Status of Education Report (ASER 2020) Rural, reflected the resilience shown towards learning as 11% of all rural families bought a new phone since the lock down began and 80% of those were smart phones [17]. There is a deeper penetration of mobile learning seen among rural areas expanding the scope of online education. All the lectures are being conducted with the help of video conferencing, video recording and distributions of projects, assignments and other study related notes can be transferred through WhatsApp or Google classroom. The present situation is an opportunity for the students residing in rural areas to understand advanced technologies and expand their knowledge through digital platform. Students are able to enrol themselves into many certificate courses and internship programmes online. Thus, students residing in rural areas are considering it as an advantage and are trying to get acquainted with these opportunities and also taking initiative to learn and understand the advanced digital technologies more. With the advent of time students, teachers and parents in the rural areas of India have understood the value of education, as well as the benefits of inculcating digital technologies in education to enhance the teaching learning process. For instance, in remote village of Maharashtra teachers keep a note about the students who have access to Internet and digital facilities and accordingly merge the children residing nearby in groups so that it becomes convenient for the students without digital facilities to take notes and other required study materials. Similarly, a farmer residing in the village of Uttar Pradesh devotes most of his time helping his son with the online education. He travels down to the Panchayat office from his area for a better Internet access, download pdfs and other study materials for different subjects, understand them and then teaches his son. The authorities of New Adarsh Public School in Karnal, Haryana try to conduct virtual classes contacting the parents of the students detailing them about the class timings, assignments, study materials and videos through online mode in WhatsApp groups. Hence, they have received a positive response from both students and parents. The school students have also created their own home study group, where they share their devices with the students who do not possess digital devices or Internet access. Besides there are Non-Governmental Organizations (NGOs) working for the digital developments of rural section in India. Mumbai-based Think Sharp Foundation is a Non-Governmental Organization working for the improvement in the quality of education in rural schools by providing them with digital facilities and infrastructure access. They have successfully provided digital set up to many rural schools across Maharashtra both for online and offline mode of education. Though great efforts are initiated in order to strengthen the education in rural India, yet there are lacunas. The facilities need to travel and

reach every nook and corner of rural India in order to achieve an overall digital development.

Education is known to be the mainstream of digital landscape where digital learning found its fertile ground to germinate into a fruitful plant. The estimating calculation says that increase in the demand for the Internet and a rapid adoption of advanced technologies among the students residing in rural areas will largely put a huge impact upon the economic status of the country. According to the report 2.8 million jobs in rural India would be generated by AI and IOT Apps by 2029 [18]. Thus introducing to the world of digital learning has generated opportunities for the students residing in rural India to get familiar with the digital learning tool kits. The teachers residing in rural areas are also trying hard to make an optimum utilization of digital platform so as to make sure that students do not miss out in the pursuits of basic education. Vijaylakshmi Akki a teacher teaching in a village school of Magadi in Karnataka says that she do not have much knowledge about technology nor did she received any kind of training related to digital skills. Yet she says, 'But I am trying to leverage all sorts of platforms like WhatsApp, YouTube, and even SMS to keep the learning going for children' [19]. A lot many organizations and individuals are coming forward in order to help in overcoming the tribulation encountered by the rural education today, besides majority of population residing in rural areas are striving to stay in line with the developments in the world of digital era, but the lack of proper access and knowledge about the digital platform, is pulling India backward in digital development.

35.5.2 *The Challenges*

Need of proper management

Government's dream of transforming Indian society to digitally empowered society faces many challenges where improper management leads to an imbalance in the required facilities for digitalizing rural education. Every individual fails to avail the benefits provided by the government due to mismanagement. This is the reason why there is a digital divide among rural and urban areas. Non availability of digital facilities, proper training and education to the rural educationists results in lack of digital knowledge and skills required to understand and use the digital terminologies and devices. Hence, rural population remains backward even after so many schemes and facilities provided by the government.

Language barrier and reluctant attitude of teachers

Many teachers living in rural areas are above 40 or 50 years of age. They may be considered as misoneists, since they are least concerned about adopting new technologies and inculcating them into education. Despite being aware about its benefits, self-learning becomes quite difficult and time consuming for them. They prefer staying in their comfort zones and make use of the traditional methods of teaching pedagogies. Above all most of the people staying in rural areas do not know

the proper use of English language. Therefore, unavailability of quality content in the local regional languages result in the slow growth of adopting digital methods in education.

Lack of information and awareness

A large part of rural population is still unaware about the importance and benefits of using digital technology in education. Apart from this there are no proper circulated information regarding the digital schemes introduced by the government in the recent years for rural educational development. Therefore, it is extremely essential for the rural population to have accurate knowledge benefits of advanced technologies in education.

35.6 Approaches to Overcome the Challenges

Proper planning needs proper management, in order to build a digitalize nation, Government should not only focus upon the planning and launching of various developmental schemes, but keeping a check upon the implementation and failure is equally important. The basic needs like infrastructure, electricity, high speed Internet, proper Internet connectivity and the digital equipments should have a balance in their distribution according to the population and the geographical area. Once the planning and the resources are properly allocated. It is essential to raise awareness regarding the digital use in education, and its requirement in the upcoming digital environment. Every individual should be consciously aware about the digital facilities and schemes available for the rural educational developments. This can be possible if the information is spread 24 h through all the wireless channels with or without Internet, in order to increase an awakening among the people about the needs and importance of digital use in education. The use of regional language while delivering information will bring more clarity in understanding the requirements and benefits of digital learning. Once they understand the essence of inculcating digital use in education system, it is essential to upgrade the educationists residing in rural areas with digital skill through trainings. As they are the instructors they need to have an exact knowledge and idea about the manner of executing the digital devices. Language being one of the vital reasons for teachers' disinterests towards learning new technology, providing training in regional language will develop a deep understanding of the technology with interest, without fearing or showing reluctant attitude towards changes.

Planning and solution will only work effectively if the proper execution of the plans takes place. All the government schemes available for the digital development in rural areas will be beneficial if used wisely. Therefore, it is the responsibility of the central and as well as the state government to keep a check on the developments, furthers requirements and the drawbacks. Tracking down the progress of the rural population and benefits attainable by the students from the facilities provided by the government will help recording the developments and the pitfalls at ground level.

This will help mending the gaps and continue with the smooth functioning of the developmental plans. Even a small progress will act as a catalyst to work for further development and help government to accomplish its dream to build a 'Digital India'.

35.7 Conclusion

The current digital atmosphere might cause challenges that can be exhilarating. However with consistent efforts to adapt the current changes can help harmonizing with the situation. It is an age of information, where computers are been instructed to work accordingly. The development and advancement in the technologies has risen too much extent that these technologies are making way for huge opportunities to build an intelligent smart future generation. The upcoming curriculums will be based upon child's interests towards specific topics. Children will be the centre of new education system.

The future new paradigm the curriculum will be one that follows the child. It begins with the children: what they are interested in, what excites them, what they are capable of, and how they learn. This paradigm does not assume children are the same; therefore, it does not impose artificial standards or age-based grade-level expectations. It helps children move forward from where they are. Furthermore, it does not believe children are simply empty vessels ready to be filled with knowledge, but rather it assumes that each child is a purpose agent who actively interacts with the outside world [20].

In fine, for availing the opportunities in future, the rural learners of India should be prepared to inculcate and implement the diverse technological advancements, overcoming all the present obstacles and become capable enough to adapt to the changes with regards to the digitalization of education.

References

1. Ankita, S., Swati, S.: Cyberspace: a contemporary path for religious education. In: Sata-pathy, Chandra, S., Bhateja, V., Mohanty, J.R., Üdgata, K.S. (eds.) Smart Intelligent Computing and Applications. SIST, vol. 160, pp. 155–163. Springer Nature, Singapore (2019). <https://doi.org/https://doi.org/10.1007/978-981-32-9690-9>
2. Islam Kabirul, M., Sarkar Foud Hossain, M., Islam Saiful, M.: Promoting student-centred blended learning in higher education: a model, E-learning and digital media: Sage J. (2021). <https://doi.org/10.1177/20427530211027721>
3. Kurt, S.D.: From virtual to participatory learning with technology during Covid-19. E-Learn. Digit. Media: Sage J. (2021). <https://doi.org/10.1177/20427530211022926>
4. Wargo, E., Chellam Carr, D., Budge, K., Davis Canfield, K.: On the digital frontier: stakeholders in rural areas take on educational technology and schooling. J. Res. Technol. Educ. **53**, 140–158 (2021). <https://doi.org/10.1080/15391523.2020.1760753>
5. Dennen, P.V., Bagdy, M.L., Arslan, O., Choi, H., Liu, Z.: Supporting new online instructors and engaging remote learners during COVID-19: a distributed team teaching approach. J. Res. Technol. Educ. (2021). <https://doi.org/10.1080/15391523.2021.1924093>

6. Mishra, L., Gupta, T., Shree, A.: Online teaching-learning in higher education during lockdown period of COVID-19 pandemic. *Int. J. Educ. Res. Open* **1** (2020). <https://doi.org/10.1016/j.ijer.2020.100012>
7. Dhawan, S.: Online learning: a panacea in the time of COVID-19 crisis. *J. Educ. Technol. Syst.* **49**, 5–22 (2020). <https://doi.org/10.1177/0047239520934018>
8. Kim, J.H., Hong, J.A., Song, D.H.: The roles of academic engagement and digital readiness in students' achievements in university E-learning environment. *Int. J. Educ. Technol. Higher Educ.* **16** (2019). <https://doi.org/10.1186/s41239-019-0152-3>
9. Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., Sindhi, S.: Online education: worldwide status, challenges, trends and implications. *J. Glob. Inf. Technol. Manage.* **21**, 233–241 (2018). <https://doi.org/10.1080/1097198X.2018.1542262>
10. Wan, N.: Adopting new digital technologies in education: professional learning. In: *New Digital Technology in Education*, pp. 25–48, Springer, Switzerland (2015). <https://doi.org/10.1007/978-3-319-05822-1>
11. Vlies, V.R.: Coronavirus Took School on line, but how digital is the future education? <https://oecdudedtoday.com/coronavirus-school-on-line-how-digital-future-education/>
12. Mohanty, J.R., Samantaray, S.: Cyber feminism: unleashing women power through technology. *Rupkatha J. Interdiscip. Stud. Hum.* **IX**(2), 328–336 (2017). <https://doi.org/10.21659/rupkatha.v9n2.33>
13. Kamath, A.: Industry Study: EdTech in India and How OnlineTyari Uses Data to Succeed, <https://www.moengage.com/blog/industry-study-edtech-in-india-and-how-online-tyari-uses-data-to-succeed>
14. Keelery, S.: Internet Penetration Rate in India 2007–2021, <https://www.statista.com/statistics/792074/india-internet-penetration-rate/>
15. Keelery, S.: Distribution of Urban and Rural Internet Users in India 2019 by Age Group, <https://www.statista.com/statistics/1115242/india-share-of-urban-and-rural-internet-user-by-age-group/>
16. Anura.: Kantar IMBR the Rural Millennial Market (2019). <https://www.scribd.com/document/440087168/The-Rural-Millennial-Kantar-IMRB-Dialogue-Factory>
17. Gohain, P.M.: How Rural India Took to Online Education in Lockdown. <https://m.timeofindia.com/how-rural-india-took-to-online-education-in-lockdown/articleshow/78925273>
18. Mishra, P.: 2.8 Million Jobs In Rural India Would Be Generated By AI and IOT Apps By 2029. <https://dazeinfo.com.2019/07/04/2-8-million-jobs-in-rural-india-ai-iot-apps-2029/>
19. Balaji, R.: Teacher's Day: Here's How Teachers in Rural India are Striving to Impart Education Online. <https://yourstory.com/socialstory/2020/09/teachers-day-rural-india-digital-education/amp>
20. Zhao, Y.: The Future of Education: 2030. In: Yu, S., Niemi, H., Mason, J. (eds.) *Shaping Future Schools with Digital Technology: An International Handbook*, pp. 1–24. Springer, Singapore (2019). <https://doi.org/10.1007/978-981-13-9439-3>

Chapter 36

Methodologies and Tools of Sentiment Analysis: A Review



Bijayalaxmi Panda, Chhabi Rani Panigrahi, and Bibudhendu Pati

Abstract Sentiment analysis has become an interesting research field in the current era. In recent years, social media users, online portal users are increasing exponentially. People are using blogs, forums, and question answer portals for their basic needs in day to day life. Sentiment analysis mainly focuses on online reviews regarding product, movie, health care, and in many more areas. Applications used for various purposes in day to day life are beneficial for mankind and increase their level of satisfaction. In this work, we have performed a detail analysis of various steps involved along with tools and techniques used in sentiment analysis. We performed a brief comparison among the techniques to analyze which technique offers better performance. This study covers the survey included in the research articles of movie review, product review, and health care. In this work, we have provided a descriptive analysis of preprocessing and noise reduction techniques used in text mining. The machine learning algorithms used in different domain are also compared and future direction of this particular area is identified.

36.1 Introduction

In the world of technology, people like to share their issues online. By 2020, data generated is expected to be very large amount such as it is around 45ZB and most of the data out of them will be textual [1]. Sources available for data communication are blogs, forums, social media websites like twitter, Facebook, mobile instant messaging apps, Instagram, WhatsApp and so on. Sentiment analysis (SA) recognizes the text expressions and analyzes it. So the focus area of SA is to identify sentiments, and classify them on the basis of their polarity. SA can be regarded as a classification process. Different categories of classification present in sentiment analysis such as document level, sentence level and aspect level. The whole document is taken as input in document level [2] and states the expression as positive or negative statement [3]. Sentence level takes each sentence as input then it defines the sentence as subjective

B. Panda (✉) · C. R. Panigrahi · B. Pati

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

or objective [4]. Sentence level SA will identify the sentence as positive or negative opinions in subjective sentence. In aspect level, sentiment analysis the properties of a product or service defined as positive, negative or neutral [5].

Sentiment analysis regarded as a vast topic under which there are three main areas of application such as movie review [6], product review [7], and health care [8]. SA uses automated processes to perform the related issues without user support. This paper is to emphasize on the three areas. In movie review people share their opinion on a particular movie through comments on twitter, face book, etc. Likewise in product review users share their views, likes, dislikes, and ratings for specific products those belongs to popular online shopping sites like Amazon, Flipkart, etc. Mostly health care deals with patient reviews, drug reviews related blogs and forums to share their health problems and need suggestions for their medication; treatment, etc. SA uses natural language processing (NLP) that faces challenges in multiple directions. Some challenges depend on the type of data and others are related to textual analysis [9].

The paper organization is given below: A brief description of sentiment analysis is given in Sect. 36.2. Section 36.3 includes data sources and their comparison. Section 36.4 includes the techniques used for SA. Section 36.5 includes tools used for SA and their uses. Finally, Sect. 36.6 enclosed conclusion and future direction.

36.2 Overview of Sentiment Analysis

This section describes different levels of SA, feature/aspect level, and text pre-processing techniques used in SA.

36.2.1 *Levels of Sentiment Analysis*

Document level mining: The input is taken as the whole document and the output obtained represented as positive, negative or neutral regarded as polarity of the document. A document contains a particular subject or topic, so sentiment representation is about that topic only. This is not applicable for blog data because it contains different kind of sentences that may not be related to a particular topic and the blogs are having comparative sentences which may results in inaccurate classification [6].

Sentence level mining: The sentence in this level is divided into two classes: subjective and objective. The aim of subjective sentence is recognition of polarity of a sentence according to information content of the sentence. It includes feelings like and dislikes of the speaker [6]. An objective statement does not contain the author's feelings or sharing's.

Entity/Phrase level mining: This is also called phrase level SA. In this level, we can take into consideration the like or unlike aspect of an object. It is better evaluation of opinion performed directly. This generates sentiment for complete sentence. The operations present here are rearranging and extracting the properties of an object, produce summary as a whole based on features of the entity [6].

36.2.2 Feature/aspect Level

It is one of the most important tasks for SA [10]. It is mainly used in product review. In this level the piece of text is identified as a feature of some product. When people are buying or selling product online they need feedback about that product. They follow the ratings given for the product which is helpful for choosing any product. The tasks that need to be done for feature-based SA are [11]:

1. Preparation of dataset with reviews
2. Using part of speech tagging
3. Extraction of features/attributes
4. Opinion expressing word extraction
5. Word polarity identification
6. Sentence polarity identification
7. Generation of summary.

36.2.3 Data Preprocessing

The preprocessing is the process of cleaning up the dataset which contains data, results in complexity reduction of a document.

Lowercasing: It is applicable in some of the problems such as NLP, when data size is small, searching, etc. For this purpose text data must be converted into lowercase that helps in maintaining in output. It is also helpful in search indexing. It converts text data into simpler form.

Tokenization: It is the process of dividing the sentence into different parts called tokens [12]. The tokens include words, keywords, phrases, symbols, etc. A sentence can be a token in a paragraph.

Part of Speech tagging (POS): It is done after tokenization where the same word having separate meanings. Let us take an example suppose two sentences are there [13] [“the orange is very sweet” and “the building is orange in color”] it is necessary to identify “orange” and “orange” accurately identify the relationship between sentences. Another thing regarding POS tagging is that it should be done before any words are replaced to preserve structure of the sentence.

Removing unnecessary punctuation, tags: This includes the following:

Removal of stop words: These are frequently used words in a language. In English, frequently used words are “a”, “the”, “is”, “are”, etc. Since these words are giving less information these can be removed from the text, so that attention can be given to important words [12] which give more information. Stop words are mainly used in search systems, applications of text classification, topic modeling, topic extraction, etc.

Stemming: Stemming is the process where prefixes or suffixes cut from the beginning or end of the word that reduces inflection in words which form their root. For example, “playing” becomes “play”. So in this process words are transformed into their original form. Sometimes cutting process becomes successful and sometimes it fails [12].

Lemmatization: This is the way of translating the words into its original form and the surface is very similar to stemming. In this process, morphological analysis of words is done to replace inflections and map a word to its root forms. It uses dictionary based approaches such as WordNet, rule-based approach, etc. Dictionaries are available for every language to provide this kind of analysis [12]. Lemmatization follows the proper way to map the word into its root form. For example, “studies” converted to “study”.

Normalization: Text normalization is implemented for noisy texts that belong to social media comments, blog posts, WhatsApp messages, where lots of words are present with misspelling, out-of-vocabulary words (ooh), abbreviations, etc. In text normalization, the text is converted into its canonical form. It also maps the words into their nearly identical form. For example, “b4” can be transformed into “before”.

NLP is also having some preprocessing techniques. Among them syntactic parsing and shallow parsing are vastly used.

Syntactic Parsing or Dependency Parsing: This gives syntactic structure to the sentence. It converts flat input sentence to hierarchical structure. Multiple parse trees are used on part of speech tags [14] for resolving ambiguities. Parse trees are applied on semantic analysis, grammar checking, etc.

Shallow parsing: Generally, grammar shows how a sentence is derived from a set of rules. A parse tree with a set of grammar rules is known as parsing. A parse tree generates a series of words to form phrases and relationship between these phrases and also gives POS tags. Shallow parsing is used to get part of the parse tree [14]. POS tagging is used to get last layer of the parse tree which contains tags like verb/noun/adjective associated with individual words.

Noise reduction techniques: A noise reduction technique is another preprocessing technique that replaces the digits, special characters, symbols, etc. from the text. This is a process of filtering inflected text. The different types of noise reduction methods include:

Tweet Cleansing: Usernames, hash tags, punctuations marks, URLs, digits, and special characters are replaced from tweets to make ready the text for analysis. Emoticons and slangs are not included in special characters.

Tokenization: In tokenization, tweets are converted to tiny parts of text. The tokenized words include: “Beautiful”, “:)” and “Yoyo”, etc. There are many tools available for tokenization and one of the best tools used for this purpose is Python-based NLTK tokenizer.

Slang Filtering: Sometimes tweets are associated with slangs which can be identified using slang dictionaries such as “noslang.com”, “onlineslangdictionary.com”, “netlingo.com”, etc. If a slang term, detected out of the tweet is replaced. For example, a tweet may be “wish u good luk guys. Have a gr8 day”, here “gr8” regarded as slang, replaced and saved in a text file.

Emoticon Filtering: If the word is defined as an emoticon in any of the repositories available, online is replaced from the tweet [7] and stored in a text file for future processing along with the tweet.

Case Conversion: All uppercase expressions are converted to lower case in a tweet [7]. There is a popular online tool available that is *textfixer.com* for this purpose.

Lemmatisation: Lemmatisation converts the terms to its original form, e.g., “car”, and “caring” to “car” and “care”. The WordNet Lemmatiser tool in python is used to perform lemmatisation [15].

Spelling Correction: Sometimes a term cannot be considered as either valid English, slang, or an emoticon and that is regarded as a word with incorrect spelling and forwarded for correction. To improve classification accuracy spelling correction is required, if it cannot be solved is dropped. The Python-based library named as Aspell [16] is used for this purpose.

Negation Handling: Negative words deal with the polarity which converts positive to negative and vice versa. Available set of negative terms, like “no”, “not”, “never” is applied to check the presence or absence of negation in tweets. Presence of negation is converted based on the polarity of a nearby sentiment word by multiplying the score of the sentiment word by -1 [7].

36.3 Data Sources Used for Sentiment Analysis

In this study, we focus on three different aspects of SA that is product review, movie review, health care. They need different data sources to perform analysis. Movie review and product review are mainly based on twitter data set and other social media blogs. Sometimes product review extracts data source from Amazon, flipkart, etc. The remaining area of SA is health care which includes patient reviews and drug reviews. Some specific source of data used in health care domain. A list of health

Table 36.1 Different datasets used for sentiment analysis

Authors and Year	Data source
Denecke et al. [19], 2009	Mayo clinic, Yedda and NHS. WebMD, AskDrWiki and MedlinePlus
Yang et al. [20], 2011	Medline, pubmed
Cambria et al. [21], 2012	patientopinion.org.uk
Huh et al. [18], 2013	WebMD
Sharif et al. [22], 2014	AskaPatient and Pharma
Noferesti et al. [14], 2015	askapatient.com druglib.com webmd DailyStrength dailymed.nlm.nih.gov
Denecke et al. [23], 2015	DrugRatingz.com WebMD
Na et al. [24], 2015	WebMD
Lycett et al. [25], 2016	NHS choice
Stuart et al. [26], 2016	NHS choice www.patientopinion.org.uk
Gopalakrishnan et al. [27], 2017	Aska Patient
Chen et al. [28], 2018	Twitter Amazon WeChat circle of friends druglib.com drugs.com
Malberg et al. [4], 2018	Drugs.com druglib.com
Kaur et al. [29], 2020	twitter dataset

care data set is given, some of them are very popular and used frequently in health care analysis. A comparison of frequently used dataset is presented in this work. In the given study, “webmd.com” is most frequently used data source where we can find patient reviews as well as drug reviews. Some other important aspect of this data source is related to telemedicine [17] and moderator’s help to the patients in emergency situation [18] (Tables 36.1 and 36.2).

36.4 Techniques Used for Sentiment Analysis

Extraction of features involves generating the efficient way for machine learning that contains text views ($T_1; T_2; T_3; \dots; T_n$) are converted into valuable word features ($wd_1; wd_2; wd_3; \dots; wd_n$) by the help of feature engineering approaches. Feature extraction is one of the methods for constructing effective classifiers. If the features extracted are accompanied with the factors of polarity such as positivity and negativity, then classification will be effective. Bag of Words (BoW), Bag of Phrases (BoP), n-gram, and Bag of Concepts (BoC) are some common techniques used for feature extraction [1]. Sentiment, negation, and opinion words are another set of features. SentiWordNet is a lexical resource that represents sentiment scores to each synset in WordNet. Sometimes emotions and moods present in directional text. For this AffectWordNet is the lexical resource which includes a subset of WordNet

Table 36.2 Different datasets used for sentiment analysis over year

Year of use	Twitter	National health service	Webmd.com	Medline	Patient opinion.org.uk	Askapatient.com	Druglib.com	Drugs.com
2009 [19]		Y	Y	Y				
2011 [20]			Y					
2012 [21]				Y				
2013 [18]			Y					
2014 [22]				Y				
2015 [14]			Y		Y			
2015 [24]		Y						
2015 [23]				Y				
2015 [17]		Y						
2016 [26]	Y							
2016 [25]	Y			Y				
2017 [27]					Y			
2018 [28]	Y					Y		
2018 [4]						Y		
2020 [30]	Y							

synsets that represent affective concepts. Based on the data source feature extraction varies such as the content of face book, twitter, etc. usually captures hash tags, abbreviations, emojis, etc. Machine learning approaches play a great role in large number of application areas. It has an important effect on SA. It is populated with many algorithms that handle large volume of data. We have discussed some commonly used machine learning algorithms for SA along with their performance.

Naïve Bayes: Bayes' theorem is used in this classifier. It acts as a probabilistic model for retrieving data. It is essential for classification because it is fast and easy to understand. It uses minimal dataset for training and predicts parameters needed for classification [1]. According to this classifier the specific feature value is independent of each other. It mainly used for large dataset to reduce complexity [6].

Support vector machines: It is a supervised technique on the basis of learning algorithms. A support vector machine is a universal learner that learns linear threshold functions [1]. It may be used to learn in accordance with radial basis function and sigmoid neural networks trained with polynomial classifiers. SVM is used for both binary and multiclass classification.

Logistic regression model: This supervised technique is used in the form of identifying probability of a target variable and used in the field of spam detection, disease prediction, etc. [30]. It is mainly used for binary classification. Another kind of logistic regression [4] is multinomial logistic regression but it has no quantitative significance.

Random Forests: Random forest (RF) algorithms depends on decision trees and act as a classification and regression method based on the ensemble learning that can generate many kinds of classifiers that generate results as a whole. Two commonly proposed technique for classification are boosting and bagging. The operating procedure of this algorithm adds randomness to bagging [30].

Ensemble classifier: An ensemble of classifier is the combination of more than one classifier based on voting mechanism to find out result. The classification depends on voted classifier [1, 27] where classification achieved through the majority vote of the prediction.

Incremental learning: In a semi-supervised learning the user can use a system in a specific domain and assign the system to a use-case. For instance, a railway could use this system to identify opinions from passengers and consider negative opinions fast. It could also find out the positive opinions to analyze the happiness of the passengers and increase resources on that area that is working well. In a cross domain solution or a non-supervised system we can ignore the focus in brief that a domain specific model can reach. All words considered as sentiment words used in the use-case to add new entries to the dictionary. Random-walk algorithm is used to adapt the addition of each and every word to maximize the number of matches. The addition of the word depends on the domain and completed according to the domain. According to random-walk we will make a change in one random word of the dictionary as a result

Table 36.3 Comparison of machine learning algorithms used for sentiment analysis

Machine learning algorithm	Accuracy
Naive Bayes [31]	89.5
Support vector machine [30]	94.16
Logistic regression [30]	91.52
Random forest [30]	90.13
Ensemble approach [27]	78.2%
Incremental model [27]	78%
Strength of association [25]	67%
Maximum entropy [7]	83.8
Semantic analysis (Word Net) [7]	89.9

of which the number of matches increases then it will be preserved, otherwise we try a different change [27].

Strength of the association: This algorithm states the way in which independent variable affects the dependent variable. Negative values for ordinal and numeric data occur when larger values of one variable correspond to smaller values of the other variable. Strength of the association is the observed associations in very large sample size. This algorithm refers to an extent of an association rather than to its statistical significance [25].

Maximum entropy: When there is no availability of information distribution should have maximal entropy. It is the classification on the basis of probability distributions of the data [7]. Training data which has limitations on distribution and find out minimum non-uniformity.

Semantic Analysis: In sentiment analysis all the terms are related to each other, it is derived from WordNet database which contains linking English words. This database is of English words which are linked together. Words with same meaning are semantically similar. It is supporting to find out the polarity of user's sentiment [7]. The above explained machine learning algorithm plays great role in different sectors of SA. Their performance comparison established on the basis of different aspect of health care. In the following table we have given a comparison of accuracy of different algorithms which is best proved in different research articles. According to the comparison the best proved algorithm is support vector machine which is most popular algorithm used in all the fields of SA (Table 36.3).

36.5 Tools Used for Sentiment Analysis

Several SA tools are available for conducting operations on several data sets. We are discussing some of them in this paper and comparing their accuracy using some datasets.

Table 36.4 Different tools used in sentiment analysis

Author and Year	Tools used
Noferesti et al. [14], 2015	Factnet, Stanford coreNLP, SentiwordnNet
Huh et al. [18], 2013	LIWC
Sharif et al. [22], 2014	SentiStrength, Sentiment 140, Opinion fnder, FSH, SentiWordNet, AffectWordNet, FRRF
Denecke et al. [23], 2015	SentiWordNet, AffectWordNet
Mumtaz et al. [6], 2016	Review Seer tool, Opinion observer, StanfordCoreNLP, Google Analytics
Asghar et al. [7], 2017	SentiWordNet
Gopalakrishnan et al. [27], 2017	StanfordCoreNLP

FRRF: This feature representational richness framework (FRRF) for sentiments analysis used on Health 2.0 data. It extracts sentiments, semantics, aspects, and domain specific features [22]. These feature set parameters are combined which is able to transfer information experienced by people.

SentiStrength: This is an important stand-alone SA tool uses a sentiment lexicon for assuming scores to negative and positive phrases in text. Phrase level scores are combined to identify sentence level polarities [22]. It is directly applied for testing data.

OpinionFinder: It runs in two modes, to process documents such as batch and interactive [32]. It takes a number of documents as input. This tool identifies subjectivity within sentences.

Many other tools are also available like FSH, Word baseline, Sentiment140, n-gram, etc. [22] which is useful in analyzing sentiment gives result for polarity, subjectivity, etc. In this paper, we have discussed some of the tools applied on different papers (Table 36.4).

36.6 Challenges for Sentiment Analysis

The development and application areas of SA are in demand for computer science and its subfields. So, it has to focus a lot of challenges as follows:

36.6.1 *Developing SA Methods Require Handling Sparse, Uncertain*

- Public opinion regarding any product, service or person is not conclusive. SA approaches are required to maintain the quality and truth regarding public opinion.

- Integrated SA required measuring the polarity of the post and the method to identify consumers' views on a large scale basis.
- Misinterpretation may cause due to the sentences having unclear meanings and feelings that is difficult to understand.
- Since the reorganization of user is not clear, concept of opinion faking arises. So some companies get benefit or some are losing their popularity.
- For a particular topic, a product, etc. customer's opinion changes over time. So, it is difficult to analyze the fact.
- Spelling mismatch, text length, sense of case, and abbreviations cause difficulty in handling methods of SA.
- Users always prefer their local language to post comments on a particular topic which makes the process of SA complicated.

36.7 Conclusion and Future Direction

This article presents a survey enclosing concept of the research articles from 2009 to 2020. In our work, we have discussed a brief idea of SA working with different fields. SA at different levels is discussed. Data preprocessing is needed for using a better quality data so almost all the preprocessing activities as well as noise reduction techniques are discussed. Mainly three areas of SA are reflected such as movie review, product review, and health care. We have focused on health care domain and discussed several data sources available related to health care and their frequency of use. Then some feature extraction techniques are discussed. Some of the machine learning approaches frequently used in health care and their performance comparison are established. Now-a-days several tools are available for finding out subjectivity, polarity and many more aspects of SA. Finally, we have discussed specific challenges of SA. Future direction of this article is to study in details about the healthcare sector and its sub fields such as patient reviews, drug reviews, etc. and how these can be analyzed using different approaches.

References

1. Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P.S., Chung, Y.W., Al-Garadi, M.A.: Sentiment analysis of big data: methods, applications, and open challenges. *IEEE Access* **6**, 37807–37827 (2018)
2. Balaji, P., Nagaraju, O., Haritha, D.: Levels of sentiment analysis and its challenges: a literature review. In: 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), pp. 436–439. IEEE (2017)
3. Lighthart, A., Catal, C., Tekinerdogan, B.: Systematic reviews in sentiment analysis: a tertiary study. *Artif. Intell. Rev.* 1–57 (2021)
4. Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S.: Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In: Proceedings of the 2018 International Conference on Digital Health, pp. 121–125 (2018)

5. Nanli, Z., Ping, Z., Weiguo, L., Meng, C.: Sentiment analysis: a literature review. In: 2012 International Symposium on Management of Technology (ISMOT), pp. 572–576. IEEE (2012)
6. Mumtaz, D., Ahuja, B.: Sentiment analysis of movie review data using Senti-lexicon algorithm. In: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (ICATccT), pp. 592–597. IEEE, 2016
7. Asghar, M.Z., Khan, A., Khan, F., Kundi, F.M.: RIFT: a rule induction framework for Twitter sentiment analysis. *Arab. J. Sci. Eng.* **43**(2), 857–877 (2018)
8. Gautam, G., Yadav, D.: Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In: 2014 Seventh International Conference on Contemporary Computing (IC3), pp. 437–442. IEEE, 2014
9. Khan, M.T., Khalid, S.: Sentiment analysis for health care. In: Big Data: Concepts, Methodologies, Tools, and Applications, pp. 676–689. IGI Global (2016)
10. Chandrakala, S., Sindhu, C.: Opinion mining and sentiment classification a survey. *ICTACT J. Soft Comput.* **3**(1), 420–425 (2012)
11. Tribhuvan, P.P., Bhirud, S.G., Tribhuvan, A.P.: A peer review of feature based opinion mining and summarization. *Int. J. Comput. Sci. Inf. Technol. (IJCSIT)* **5**(1), 247–250 (2014)
12. Khan, M.T., Khalid, S.: Sentiment analysis for health care. In: Big Data: Concepts, Methodologies, Tools, and Applications (pp. 676–689). IGI Global (2016)
13. <https://towardsdatascience.com/machine-learning-text-processing>. Accessed 12 Aug 2021
14. Noferesti, S., Shamsfard, M.: Using linked data for polarity classification of patients' experiences. *J. Biomed. Inform.* **57**, 6–19 (2015)
15. <https://towardsdatascience.com/a-game-of-words-vectorization-tagging-and-sentiment-analysis>. Accessed 12 Aug 2021
16. http://www.nltk.org/_modules/nltk/stem/wordnet.html. Accessed 12 Aug 2021
17. Kamsu-Foguem, B., Tiako, P.F., Fotso, L.P., Foguem, C.: Modeling for effective collaboration in telemedicine. *Telematics Inform.* **32**(4), 776–786 (2015)
18. Huh, J., Yetisgen-Yildiz, M., Pratt, W.: Text classification for assisting moderators in online health communities. *J. Biomed. Inform.* **46**(6), 998–1005 (2013)
19. Denecke, K., Nejdl, W.: How valuable is medical social media data? Content analysis of the medical web. *Inf. Sci.* **179**(12), 1870–1880 (2009)
20. Yang, H., Swaminathan, R., Sharma, A., Ketkar, V., Jason, D.S.: Mining biomedical text towards building a quantitative food-disease-gene network. In: Learning Structure and Schemas from Documents, pp. 205–225. Springer, Berlin (2011)
21. Cambria, E., Benson, T., Eckl, C., Hussain, A.: Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Syst. Appl.* **39**(12), 10533–10543 (2012)
22. Sharif, H., Zaffar, F., Abbasi, A., Zimbra, D.: Detecting adverse drug reactions using a sentiment classification framework. *Socialcom* 1–10 (2014)
23. Denecke, K., Deng, Y.: Sentiment analysis in medical settings: new opportunities and challenges. *Artif. Intell. Med.* **64**(1), 17–27 (2015)
24. Na, J.C., Kyaing, W.Y.M.: Sentiment analysis of user-generated content on drug review websites. *J. Inf. Sci. Theory Pract.* **3**(1), 6–23 (2015)
25. Bahja, M., Lyett, M.: Identifying patient experience from online resources via sentiment analysis and topic modelling. In: Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, pp. 94–99 (2016)
26. Stuart, K., Botella, A., Ferri, I.: A corpus-driven approach to sentiment analysis of patient narratives. In: 8th International Conference on Corpus Linguistics, vol. 1, pp. 381–395, 2016
27. Gopalakrishnan, V., Ramaswamy, C.: Patient opinion mining to analyze drugs satisfaction using supervised learning. *J. Appl. Res. Technol.* **15**(4), 311–319 (2017)
28. Chen, Y., Zhou, B., Zhang, W., Gong, W., Sun, G.: Sentiment analysis based on deep learning and its application in screening for perinatal depression. In: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), pp. 451–456. IEEE (2018)
29. Kaur, C., Sharma, A.: COVID-19 sentimental analysis using machine learning techniques. In: Progress in Advanced Computing and Intelligent Engineering, pp. 153–162. Springer, Singapore (2021)

30. <https://pypi.python.org/pypi/aspell-python-py2/1.13>. Accessed 12 Aug 2021
31. Neethu, M.S., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–5. IEEE (2013)
32. <https://mpqa.cs.pitt.edu/opinionfinder/>. Accessed 12 Aug 2021

Chapter 37

A Review: Convolutional Neural Network Application for MRI Dissection and Detection of Brain Tumor



Dillip Ranjan Nayak, Neelamadhab Padhy, Pradeep Kumar Mallick, and Dilip Kumar Bagal

Abstract From the year 2000 onwards, deep learning methodology has gained widespread acceptance in the area in medical image processing, medical image analysis, and bioinformatics. The result of this transformation is that deep learning has significantly improved the methods of identification, estimation, and diagnosis in the application of medical fields, including neuroscience, brain tumors, lung cancer, abdominal cavity, heart, retina, and others. Deep learning is also being used to improve the accuracy of medical imaging. The idea of this article is to examine key deep learning issues that are relevant to brain tumor research, in the applications for deep learning like segmentation, classification, prediction, and evaluation. This article provides an overview of a large number of scientific study in deep learning for brain tumor analysis.

37.1 Introduction

The brain is the major component and complex part of the body. Due to appearance of the skull around the brain, it is impossible to investigate and the process of identifying illnesses. The mind, unlike other parts of the body, is not predisposed to a particular disease; nevertheless, it may be actuated by the uncontrolled growth of

D. R. Nayak (✉) · N. Padhy

School of Engineering and Technology (CSE), GIET University, Gunupur, Odisha 765022, India
e-mail: dilipranjan.nayak@giit.edu

N. Padhy

e-mail: dr.neelamadhab@giit.edu

P. K. Mallick

School of Computer Engineering, Kalinga Institute of Industrial Technology, Deemed to be University, Bhubaneswar 751024, India
e-mail: pradeep.mallickfcs@kiit.ac.in

D. K. Bagal

Department of Mechanical Engineering, Government College of Engineering, Kalahandi, Bhawanipatna, Odisha 766002, India

brain cells, which results in an alteration in its conduct and design [1]. The location of such tumors may be identified using magnetic resonance imaging (MRI) [2]. Brain tumors are mainly classified into two types, one is cancerous that is called malignant tumors, and other is non-cancerous that is called benign tumors which is shown in Fig. 37.1. Malignant tumors again classified into 04 ranks as I to IV by World Health Organization (WHO). Grade-I tumors and Grade-II tumors are less aggressiveness semi-malignant tumors while Grade-III and Grade-IV are malignant tumors are highly operative on the brain of the patient.

Nowadays, computers play an important role in human life. Each task is usually done automatically without the intervention of human activities. Hence, the same has been also applied heavily in the field of medical science. In medical science, technology plays a very important role as it performs task with ease and least tolerance. As the medical domain is the most sensitive domain; hence, there should be as least margin of error as possible. One of the applications of technology in medical science is the case of automatic disease identification just by providing few information [4]. MRI is a popular imaging technique for analysis the human brain. It gives adequate

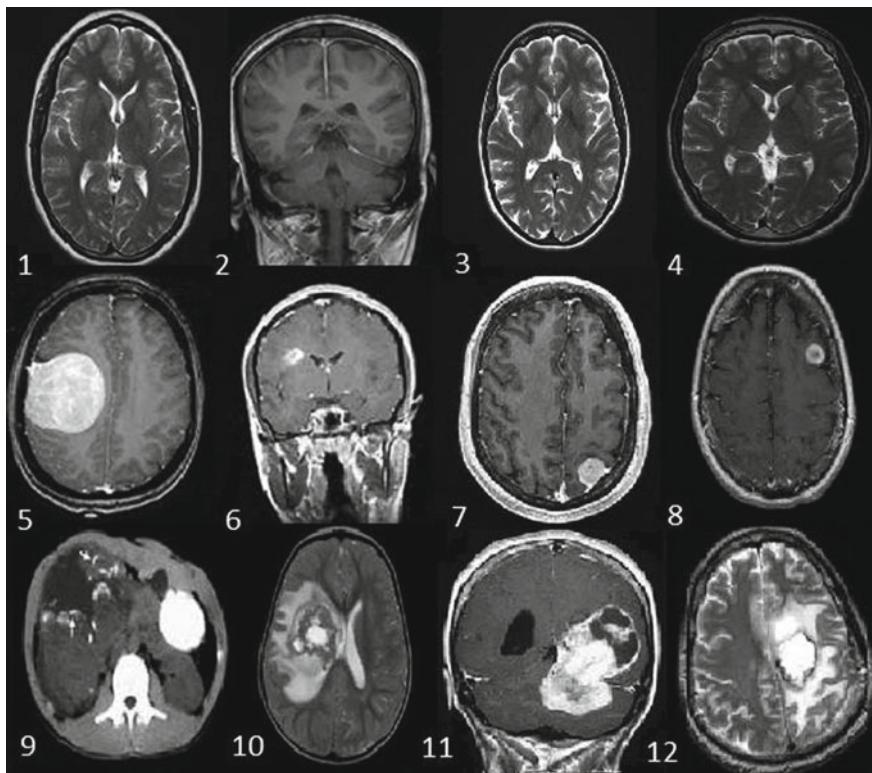


Fig. 37.1 Sample datasets of brain tumor MRI images normal brain MRI (1–4) benign tumor MRI (5–8) malignant tumor MRI (9–12) [3]

information about brain nerves for medical diagnosis and biomedical research. But still, there is a need of image preprocessing for enhancement, noise reduction, and feature extraction for further experimental works. Lot of research has been performed for image preprocessing such as application of fuzzy Gaussian filter, Gabor filter, and fuzzy entropy [5]. Similarly, preprocessing can be performed by the help of 2D or 3D wavelet transform [6]. In recent days, deep learning is one of the efficient ways of designing classifier and segmentation models by using CNN [7, 8].

Automatic detection of the functional morphological structure of the brain with a deep learning network is the best tool for accurate classification [9, 10]. A hybrid method that consists of deep autoencoder along with Bayesian fuzzy clustering is a new approach for brain tumor segmentation [11]. This paper pivoted on review of diverse deep learning methods for automatic detection of brain tumors and highlights of their pros and cons. Comparative analysis of different technologies used for classification of tumors can be fascinated as new area of research in the medical field.

37.2 General Approach for Brain Tumor Classification

Deep learning algorithms for classifying brain tumors into different grades include four main steps, namely preprocessing, segmentation, feature extraction, and classification.

37.2.1 Preprocessing

It is the first step to resize the brain images for accurate observations of tumor. It includes different methods like intensity normalization, fuzzy image preprocessing, morphological filtering, adaptive contrast enhancement, and skull stripping for better classification.

37.2.2 Segmentation

Segmentation is a crucial step to separate out tumor portion from the brain MRI. Different supervised and unsupervised learning approaches like fuzzy thresholding, fuzzy clustering, neural network, water shed, edge-based, and fuzzy C-means are used for segmentation. The detailed structure is shown in Fig. 37.2.

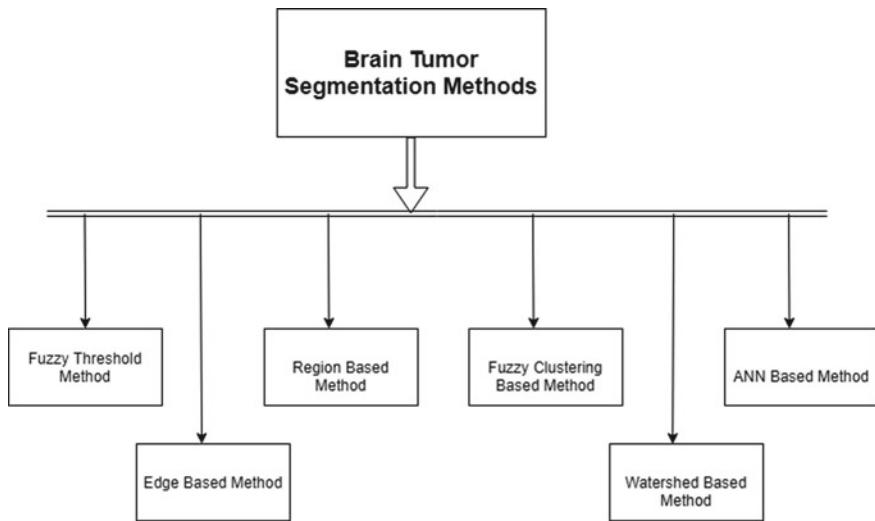


Fig. 37.2 Different brain tumor segmentation methods

37.2.3 Feature Extraction

Feature extraction means the meaningful representation of features like shape, size, texture, and increase the contrast also [5]. Different feature extraction methods are used like CNN, autoencoder, PCA, and regression tree are used for brain tumor cases.

37.2.4 Classification

Brain tumors are classified using different CNN or deep CNN methods. Mainly, tumors are classified as benign and malignant tumors. Malignant tumors are again decoupled into types glioma, meningioma, and pituitary which is shown in Fig. 37.3.

37.3 Reviews

Elamri and Planque proposed an advanced method based on 3D CNN for the separation of the glioblastoma (GBM) tumor [12]. Menze et al. [13] utilized a dataset of 65 multi-contrast MRI results from a large number of glioma patients to conduct about twenty division computations in order to detect tumors. According to the researchers, various estimations were suitable for distinct tumor districts, and no one computation was found to be dominant for all sub regions.

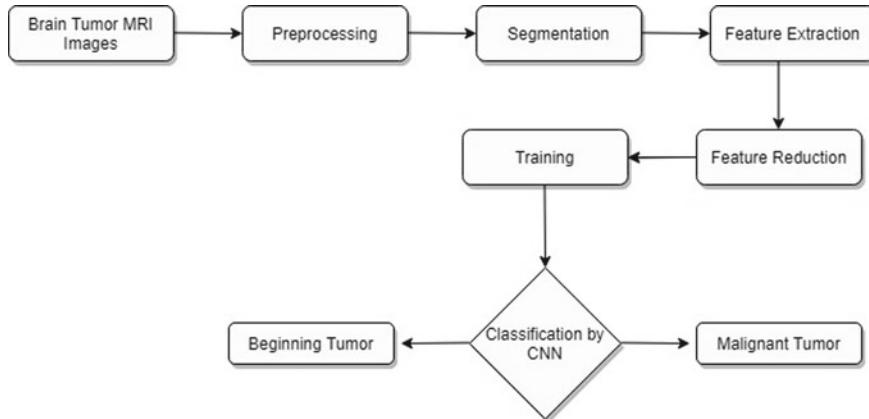


Fig. 37.3 Detailed structure of brain tumor classification

The convolutional deep network is a powerful tool for brain image identification and prediction [14, 15]. On the other hand, Havaei and colleagues [16] presented a programmed approach for brain tumor segmentation that relied on flexible, high-capacity deep neural networks to segment tumors. The difficulties associated with the irregularity in the naming of the tumor are addressed via the use of a preparatory approach that is divided into two phases.

To properly analyze and treat cancer patients, as well as to evaluate treatment outcomes, it is essential to divide the tumor district using the basis of semi-automatic and automated procedures and approaches. When it comes to pattern categorization problem in tumor segmentation, support vector machines (SVMs) [17] and random forest (RF) [18–20] are frequently utilized. As a result of their superior performance in medical image processing areas such as target identification [21], picture classification [22], object recognition, and semantic segmentation [23–25], deep learning methodologies are gaining momentum in brain tumor segmentation.

Zhao et al. [14] suggested mechanism by integrating fully conventional neural networks (FCNNs) and conditional random fields (CRFs) for the separation in brain tumors using 2D image patches and slices. They utilized image data from BRATS 2013, 2015, and 2016 for their experiments. They indicated that the segmentation robustness factors, such as picture pitch size and number of training images, may be improved by using the above-mentioned method, and that they had also obtained substantial performance of tumor segmentation models based on flare, T1CE, T1, and T2 scans. The various types of brain detect tumors are shown in Fig. 37.4 together with MRI pictures.

With the use of discrete wavelet decomposition on convolutional neural networks (CNNs), Sarham et al. [27] investigated three different kinds of brain tumors: meningioma, glioma, and pituitary tumors and discovered an accuracy rate of 99.3% in his experiments. He said that the proposed wavelet-based CNN (WCNN) methodology outperformed the traditional support vector method (SVM) technique in terms of

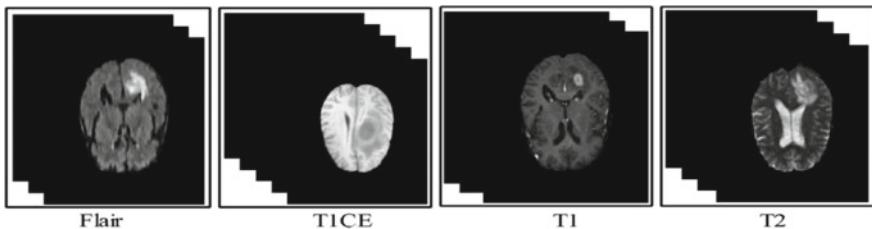


Fig. 37.4 Different types of brain tumor [26]

success rate, and that this was due to the greater success rate. In their trial with 153 patients, Siar and Teshnehlab [28] analyzed the application of feature extraction with convolutional neural network (CNN) for the diagnosis of brain tumors and discovered an accuracy rate of 99.12%.

In the clinical sector, the doctors divided the tumor images for use in applications such as irregularity detection, meticulous arrangement, and post-medical operation assessment. Many methods have been suggested for dividing pediatric cerebrum cancers, with the majority of them centered on the sufficiency of two distinct surface features, as well as the utilization of multimodal MRI images for sectioning and grouping juvenile cerebrum tumors. The piecewise-triangular-prism-surface-area (PTPSA) algorithm is used to infer the fractal spotlight of one of the surfaces, while the fractional Brownian motion algorithm is used to infer the texture of the other. The algorithm used to infer the fractal spotlight of one of the surfaces is the piecewise-triangular-prism-surface-area (PTPSA) algorithm. Roy and Bandyopadhyay [29] used symmetric research to examine and dignify a tumor on a brain MRI scan and were successful.

There have been a variety of brain tumor image preparation methods and tactics used to diagnose and treat cerebrum tumors over the years. This division is the most significant achievement in medical imaging techniques [30], and it is used to separate the polluted regions of cerebrum tissue from MRIs. A hybrid deep learning method which is combined the autoencoder with a Bayesian fuzzy clustering segmentation method was used in the BRATS 2015 database by Raja and Rani [31] to investigate brain tumor classification accuracy. They showed 98.5% the classification accuracy. They also used the Bayesian fuzzy clustering (BFC) with hybrid strategy that combined deep autoencoder (DAE) which based on Jaya optimization algorithm (JOA) with the Softmax regression method to classify the pictures.

Kadal et al. [32] did their experiment using a differential deep CNN classifier to classify different types of brain tumor. They got accuracy 99.25% on Tianjin Universal Center of Medical Imaging and Diagnostic (TUCMD) data. Sajja et al. [33] classified the brain images into malignant and benign tumors using VGG16 network with Fuzzy C-means algorithm. They got the 96.70 in accuracy. Sajjad et al. [34] proposed multigame brain classification system using VGG19 CNN by using extensive data augmentation techniques like rotation, flip, emboss. They assessed 94.58% accuracy on CE-MRI dataset. Jia and Chen [35] introduced a fully automatic

Table 37.1 An overview of techniques for brain tumor classification on MRI using different data set

Author	Classification method	Data source	Accuracy (%)
Sarham et al. [27]	WCNN	MRI	99.3
Siar and Teshnehlab [28]	CNN	MRI	99.12
Raja and Rani [31]	Bayesian Fuzzy	BRATs 2015	98.5
Kadal et al. [32]	Deep CNN	TUCMD	99.25
Sajja et al. [33]	VGG16	BRATS	96.70
Jia and Chen [35]	FAHS-SVM	MRI	96.70

heterogeneous segmentation using support vector machine (FAHS-SVM) approach for the identification and segmentation of MRI images of human brains and achieved 98.51% accuracy in detection of abnormal and normal tissue in their experiments. Mahalakshmi and Velmurugan [36] built up a calculation to recognize cerebrum tumor utilizing molecule swarm advancement. The detailed achieving better accuracy method shown in Table 37.1, and performance is analyzed in Fig. 37.5.

In the areas of brain tumor image segmentation, detection, and prediction, CNNs surpass all other deep learning methods and techniques combined. Two-dimensional CNNs (2D-CNNs) [37–40] and three-dimensional CNNs [41, 42] have been used to develop techniques for brain tumor segmentation, grouping, and prediction. It is

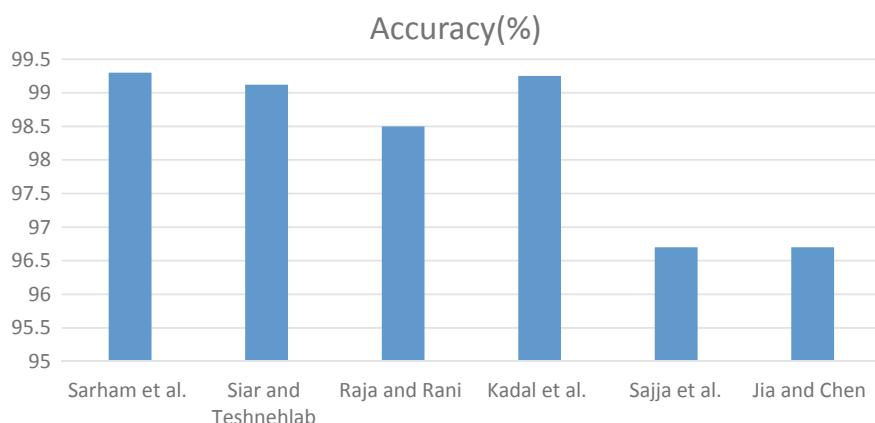


Fig. 37.5 Performance analysis of techniques for brain tumor classification on MRI using different data set

split into different categories based on the segmentation procedures that have been used, such as necrosis (death), stable tissues (survival), edema (death), enhancing the heart (heart enhancement), and non-enhancing the core.

El-Dahshan et al. [43] investigated several characterization methods and division strategies and conceived that PC assisted discovery is a major problem in the MRI of the human cerebrum, which they concluded is a serious issue. According to Dong et al. [44], a completely automated technique for segmenting brain tumors using deep convolutional networks based on the U-Net was created. Padole and Chaudhari [45] developed a method for identifying brain cancers from MRI images via part inspection, in which the normalized cut (Ncut) and mean shift algorithms were coupled to naturally identify the cerebrum tumor surface zone.

According to Pereira et al. [46], a programmed segmentation method based on CNN was suggested, and it was tested using minor 3×3 kernels. The use of force standardization as a preprocessing endeavor, in conjunction with the increase in information, demonstrated the suitability of the division for MRI photographs. This method has been authorized based on the BRATS 2013 data collection. The basic question of improvement in accuracy is really addressed in the study that was previously stated. This may be addressed by implementing an improved plan that incorporates WCA. Abdel-Maksoud et al. developed a technique for picture segmentation that merged the K-means clustering algorithm with the fuzzy C-means algorithm [47]. Nowadays, different hybrid deep neural network methods are used for brain tumor classification which are explained in Table 37.2, and their performance are analyzed in Fig. 37.6.

From the above state of arts, the authors conclude that deep CNNs have better accuracy than normal CNN methods.

Table 37.2 Different hybrid CNN approaches for brain tumor segmentation

Sl. No.	Reference	Methods	Data	Performance (%)
1	Deb et al. [48]	Fuzzy deep neural network	Brain MRI	99.6
2	Cinar et al. [49]	Deep ResNet	Brain MRI	97.2
3	Francisco et al. [50]	Multi-pathway CNN	3064 MRI	97.3
4	Ullah et al. [51]	WDNN	MRI	98.5
5	Merhotra et al. [52]	AI-based DNN	T1 MRI	99.04
6	XinyuZhou et al. [53]	3D Residual neural network	BRATS 2018	91.21
7	Mohammed et al. [54]	Deep CNN	MRI	96
8	Gumaei et al. [55]	Tenfold CNN	MRI	92.61

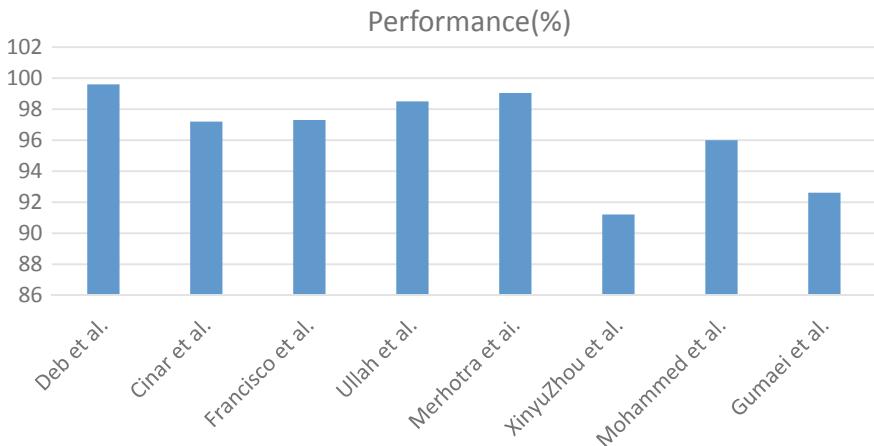


Fig. 37.6 Performance analysis of different Hybrid CNN approaches for brain tumor segmentation

37.4 Conclusion

The implementation of deep learning methods to detect brain tumors have numerous unique challenges like the lack of large training datasets, slice-by-slice annotations of 3D brain tumor segmentation which is time-consuming task, and efficient learning to analysis of limited data. So various authors have trained their 3D learning models using only 2D learning. Another problem is the data class-imbalance which is occur due to augmentation. Another demerit is patch classification, which may solve the feeding entire image into the deep network.

Artificial general intelligence (AGI) is a term used to describe the ability to attain human-level results in a variety of activities. The use of standardized labeling reports in the health sector, particularly in the research of brain tumors, is anticipated to become increasingly widespread in the future. Text-free and structured reports for network training are anticipated to become more popular in the future, particularly in the area of brain tumor research. According to this article, several academics have done outstanding work on deep learning for the study of brain tumors using publically available datasets. Potential academics and aspirants will benefit from this study since it will provide them with a better knowledge of research initiatives such as the segmentation, diagnosis, and classification of brain tumors in humans and other animals.

References:

1. Gondal, A.H., Khan, M.N.A.: A review of fully automated techniques for brain tumor detection from MR images. *Int. J. Mod. Educ. Comput. Sci. Rev.* **5**(2), 55–61 (2013)

2. Iftekharuddin, K.M.: Techniques in fractal analysis and their applications in brain MRI. In: Medical Imaging Systems Technology. Analysis and Computational Methods, vol. 1. World Scientific, pp. 63–86 (2005)
3. Alam, S., et al.: An efficient image processing technique for brain tumor detection from MRI images. In: IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (2019)
4. Siar H, Teshnehlab M: Diagnosing and classification tumors and MS simultaneous of magnetic resonance images using convolution neural network. In: 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (2019)
5. Pham, T.X., Siarry, P., Ouladji, H.: Integrating fuzzy entropy clustering with an improved PSO for MRI brain image segmentation. *Appl. Soft Comput.* **65**, 230–242 (2018)
6. Chervyakov, N., Lyakhov, P., Nagornov, N.: Analysis of the quantization noise in discrete wavelet transform filters for 3D medical imaging. *Appl. Sci.* **10**(4) (2020)
7. Pereira S., et al.: Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* **35**(5), 1240–1251 (2016)
8. Badža, M.M., Barjaktarovic, M.C.: Classification of brain tumors from MRI images using a convolutional neural network. *Appl. Sci.* **10**(6) (2020)
9. Liu, Z., Jin, L., Chen, I., Fang, Q., Ablameyko, S., Yin, Z., Xu, Y.: A survey on applications of deep learning in microscopy image analysis. *Comput. Biol. Med.* **134**, 12–24 (2021)
10. Díaz-Pernas, F.J., Martínez-Zarzuela, M., Antón-Rodríguez, M., González-Ortega, D.: A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. *Healthcare* **9**(2) (2021)
11. Eser Sert, F., Ozyort, A., Doğantekin, A.: A new approach for brain tumor diagnosis system: Single image super resolution based maximum fuzzy entropy segmentation and convolutional neural network hypotheses (2019)
12. Elamri, C., Planque, T.: A new algorithm for fully automatic brain tumor segmentation with 3-D convolutional neural networks. Stanford University Report, vol. 322 (2016)
13. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **34**(10), 1993–2024 (2014)
14. Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., Fan, Y.J.M.I.A.: A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Med. Image Anal.* **43**, 98–111 (2018)
15. Wang, G., et al.: Slic-Seg: a minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views. *Med. Image Anal.* **34**, pp. 137–147 (2016)
16. Havaei, M., et al.: Brain tumor segmentation with deep neural networks. *Med. Image Anal.* **35**, 18–31 (2017)
17. Li H., Fan, Y.: Label propagation with robust initialization for brain tumor segmentation. In: 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1715–1718. IEEE (2012)
18. Goetz, M., Weber, C., Bloecher, J., Stieljes, B., Meinzer, H.P., Maier-Hein, K.: Extremely randomized trees based brain tumor segmentation. In: Proceeding of BRATS challenge-MICCAI, pp. 006–011 (2014)
19. Kleesiek, J., Biller, A., Urban, G., Kothe, U., Bendszus, M., Hamprecht, F.: Ilastik for multi-modal brain tumor segmentation. In: Proceedings MICCAI BraTS, pp. 12–17 (2014)
20. Meier, R., Bauer, S., Slotboom, J., Wiest R., Reyes M.: Appearance-and context-sensitive features for brain tumor segmentation. In: Proceedings of MICCAI BRATS Challenge, pp. 020–026 (2014)
21. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE conference on computer vision and pattern recognition, pp. 580–587 (2014)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Adv. Neural. Inf. Process. Syst.* **25**, 1097–1105 (2012)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
24. Zheng, S., et al.: Conditional random fields as recurrent neural networks. In: IEEE International Conference on Computer Vision, pp. 1529–1537 (2015)

25. Liu, Z., Li, X., Luo, P., Loy, C., Tang, X.: Semantic image segmentation via deep parsing network. IEEE International Conference on Computer Vision, pp. 1377–1385 (2015)
26. Rehman, A., Khan, M.A., Saba, T., Mehmood, Z., Tariq, U., Ayesha, N.: Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture. *Microsc. Res. Tech.* **84**(1), 133–149 (2021)
27. Sarhan, A.M.: Brain tumor classification in magnetic resonance images using deep learning and wavelet transform. *J. Biomed. Sci. Eng. Appl. Artif. Intell.* **13**(06), 102 (2020)
28. Siar, M., Teshnehab, M.: Brain tumor detection using deep neural network and machine learning algorithm. In: 9th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 363–368: IEEE (2019)
29. Roy, S., Bandyopadhyay, S.K.: Detection and quantification of brain tumor from MRI of brain and its symmetric analysis. *Int. J. Inf. Commun. Technol. Res.* **2**(6) (2012)
30. Mittal, M., Goyal, L.M., Kaur, S., Kaur, I., Verma, A., Hemanth, D.J.: Deep learning based enhanced tumor segmentation approach for MR brain images. *Appl. Soft Comput.* **78**, 346–354 (2019)
31. Raja, P.S., Rani, A.V.: Brain tumor classification using a hybrid deep autoencoder with Bayesian fuzzy clustering-based segmentation approach. *Biocybern. Biomed. Eng.* **40**(1), 440–453 (2020)
32. Kader, I.A., Xu, G., Shuai, Z., Saminu, S., Javaid, I., Ahmad, I.I.S.: Differential deep convolutional neural network model for brain tumor classification. *Brain Sci.* **11**(3) (2021)
33. Sajja, V.R., Kalluri, H.R.: Classification of brain tumors using fuzzy C-means and VGG16. *Turkish J. Comput. Math. Educ.* **12**(9), 2103–2113 (2021)
34. Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., Baik, S.W.: Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J. Comput. Sci.* pp. 174–182 (2018)
35. Jia, Z., Chen, D.: Brain tumor identification and classification of mri images using deep learning techniques. *IEEE Access* (2020)
36. Mahalakshmi, S., Velmurugan, T.: Detection of brain tumor by particle swarm optimization using image segmentation. *Indian J. Sci. Technol.* **8**(22), 1 (2015)
37. Zikic, D., Ioannou, Y., Brown, M., Criminisi, A.: Segmentation of brain tumor tissues with convolutional neural networks. *MICCAI-BRATS* **36**, 36–39 (2014)
38. Dvořák, P., Menze, B.: Local structure prediction with convolutional neural networks for multi-modal brain tumor segmentation. In: International MICCAI Workshop on Medical Computer Vision, pp. 59–71. Springer (2015)
39. Havaei, M., Dutil, F., Pal, C., Larochelle, H., Jodoin, M.J.: A convolutional neural network approach to brain tumor segmentation. *BrainLes*, pp. 195–208. Springer (2015)
40. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI. *BrainLes*, pp. 131–143. Springer (2015)
41. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **36**, 61–78 (2017)
42. Yi, D., Zhou, M., Chen, Z., Gevaert, O.: 3-D convolutional neural networks for glioblastoma segmentation. *arXiv preprint* (2016)
43. El-Dahshan, E.S.A., Mohsen, H.M., Revett, K., Salem, A.B.M.: Computer-aided diagnosis of human brain tumor through MRI: a survey and a new algorithm. *Expert Syst. Appl.* **41**(11), 5526–5545 (2014)
44. Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y.: Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In: Annual Conference on medical Image Understanding and Analysis, pp. 506–517. Springer (2017)
45. Padole, V.B., Chaudhari, D.: Detection of brain tumor in MRI images using mean shift algorithm and normalized cut method. *Int. J. Eng. Adv. Technol.* **1**(5), 53–56 (2012)
46. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* **35**(5), 1240–1251 (2016)
47. Abdel-Maksoud, E., Elmogy, M., Al-Awadi, R.: Brain tumor segmentation based on a hybrid clustering technique. *Egypt. Inf. J.* **16**(1), 71–81 (2015)

48. Deb, D., Roy, S.: Brain tumor detection based on hybrid deep neural network in MRI by adaptive squirrel search optimization. *Multi Media Tools Appl.* **80**, 2621–2645 (2021)
49. Cinar, A., Yildirim, M.: Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture. *Med. Hypotheses* **139** (2020)
50. Francisco, J.P., Mario, Z.M., Miriam, R.A.: a deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. *Healthcare* **9**(2) (2021)
51. Ullah, Z., Farooq, M.U., Lee, S.H.A.: Hybrid image enhancement based brain MRI images classification technique. *Med. Hypotheses*. **143** (2020)
52. Mehrotra, R., Ansari, M.A., Agrawal, R., Anand, R.S.: A transfer learning approach for AI-based classification of brain tumors. *Mach. Learn. Appl.* **2**, 10–19 (2020)
53. Zhou, X., Li, X., Hu, K., Zhang, Y., Chen, Z., Gao, X.: ERV-Net: an efficient 3D residual neural network for brain tumor segmentation. *Expert Syst. Appl.* **170** (2021)
54. Mohammed, B.A., Shaban, M.: An efficient approach to diagnose brain tumors through deep CNN. *MBE* **18**(1), 851–867 (2020)
55. Gumaei, A., Hassan, M.M., Hassan, R., Alelaiwi, A., Fortino, G.: A Hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. *IEEE Access* **7**, 36266–36273 (2019)

Chapter 38

Neutrosophic Logic and Its Scientific Applications



Sitikantha Mallik, Suneeta Mohanty, and Bhabani Shankar Mishra

Abstract The scientific term neutrosophy was first coined by Florentin Smarandache a few years ago. The origins, attribute, extent of neutralities and their interactions with other ideational spectra, and indeterminacy are all investigated in this discipline of study. Neutrosophic logic, a group of many-valued systems which can be regarded as an extension of fuzzy logic, is one of the new theories based on the fundamental principles of neutrosophy. Neutrosophy logic is a new branch of logic that addresses the shortcomings of fuzzy and classical logic. Some of the disadvantages of fuzzy relations are failures to handle inconsistent information and the high processing cost of completing a non-linear program. In neutrosophic sets, truthfulness and falsity are independent, whereas in intuitionistic fuzzy sets, it is dependent. The neutrosophic logic has the ability to manipulate both incomplete and inconsistent data. So, there is a need for research into the use of neutrosophic logic in different domains from medical treatment to the role of a recommender system using new advanced computational intelligent techniques. In this study, we are discussing about basic concepts of neutrosophic logic, fuzzy logic's drawbacks and advantages of using neutrosophic logic, and also the comparison between neutrosophic logic, intuitionistic and interval-valued fuzzy systems, and classical logic on different factors like uncertainty and vagueness.

38.1 Introduction

One of A.I.'s prominent issues and challenges is simulating uncertainty for addressing realistic situations. Managing uncertainties, particularly indeterminate circumstances where it is not true or false, is the utmost goal for decision-makers. As a result, new approaches to attribute interpretation are emerging, like fuzzy logic, intuitionistic,

S. Mallik (✉) · S. Mohanty (✉) · B. S. Mishra
KIIT Deemed to be University, Bhubaneswar, India
e-mail: smohantyfcs@kiit.ac.in

B. S. Mishra
e-mail: bsmishrafcs@kiit.ac.in

interval-valued fuzzy, and neutrosophic models. The fuzzy logic system was first introduced by Prof. L.A. Zadeh in 1965. It is used to deal with the concept of partial truth, where the truth value can be somewhere between true and false. It is one of the soft computing methods for simulating real-world uncertainties. In classical logic, on the other hand, the truth values of variables are either 0 or 1. Models that are fuzzy, intuitionistic fuzzy, or imprecise are constrained because they cannot express indeterminacy and contradiction, which is a very important feature of human thought. Florentin Smarandache proposed the theory of neutrosophic logic since fuzzy logic is unable to exhibit indeterminacy on its own [1]. Neutrosophic logic (NL) is a set of logic that generalizes fuzzy logic, paraconsistent logic, intuitionistic logic, etc. The first part of neutrosophic logic is the degree of membership (T), the middle part is indeterminacy (I), and the third part is the degree of non-membership of each set element.

The rest of the chapter is organized as follows: Sect. 38.2 describes about the overview of the neutrosophic logic system and its basic concepts. Section 38.3 presents the relationship between neutrosophic logic, classical logic, and fuzzy logic, and Sect. 38.4 describes about differences between the two logic systems. Section 38.5 tells about the advantages of neutrosophic logic over fuzzy logic. Section 38.6 differentiates between different expert systems. Section 38.7 describes the different inference systems of different expert systems. Section 38.8 shows the applications of neutrosophic logic and Sect. 38.9 comprises the conclusion and future work.

38.2 Background

38.2.1 Crisp Logic (Classical Logic)

It is similar to Boolean logic (either 0 or 1). True (1) or false (0) is the outcome of a statement. Fuzzy logic, on the other hand, captures the degree to which something is true.

38.2.2 Fuzzy Logic

Fuzzy logic is a type of many-valued logic with membership degrees ranging from 1 to 0. The fuzzy set theory proposed by fuzzy logic states that a fuzzy set is a collection of ordered pairs represented by

$$A = \{(y, \mu_A(y)) | y \in U\}$$

where $\mu_A(y)$ = membership function in fuzzy set A ,

U = universe of discourse.

There is partial membership in the situation of fuzzy sets. The membership function, with a value ranging from 0 to 1, is used to calculate the degree of uncertainty of the fuzzy set's elements. The truth value of variables in fuzzy logic can be any real number between 0 and 1, both inclusive [2].

38.2.3 *Intuitionistic and Paraconsistent Logic*

Paraconsistent logic refers to a branch of logic that studies and develops “inconsistency-tolerant” logic systems that reject the idea of explosion, whereas intuitionistic logic includes the general principles of logical reasoning which have been derived by logicians from intuitionistic mathematics.

38.2.4 *Neutrosophic Logic*

The term “neutrosophic logic” refers to a logic in which the proposition’s parameters are defined as follows:

1. Truth (T), percentage of truth
2. Falsehood (F), percentage of falsity
3. Indeterminacy (I), the condition or percentage of being indeterminate

where truth, indeterminacy, and falsehood are standard or non-standard real subsets of $]0,1 + [$, that is not inherently connected.

The total value of the elements in single-valued neutrosophic logic is:

1. $0 \leq t + i + f \leq 3$, when each of these elements is self-contained.
2. $0 \leq t + i + f \leq 2$, when two elements are interdependent, but the final part is unaffected.
3. $0 \leq t + i + f \leq 1$, when all three elements are interconnected.

There is scope for inadequate information (total < 1), paraconsistent and conflicting information (total > 1), or complete information (total = 1) when three or two of the T, I, F components are independent. If all three aspects of truth, indeterminacy, and falsity are interconnected, one can offer partial (total < 1) or entire (total = 1) information in the same way.

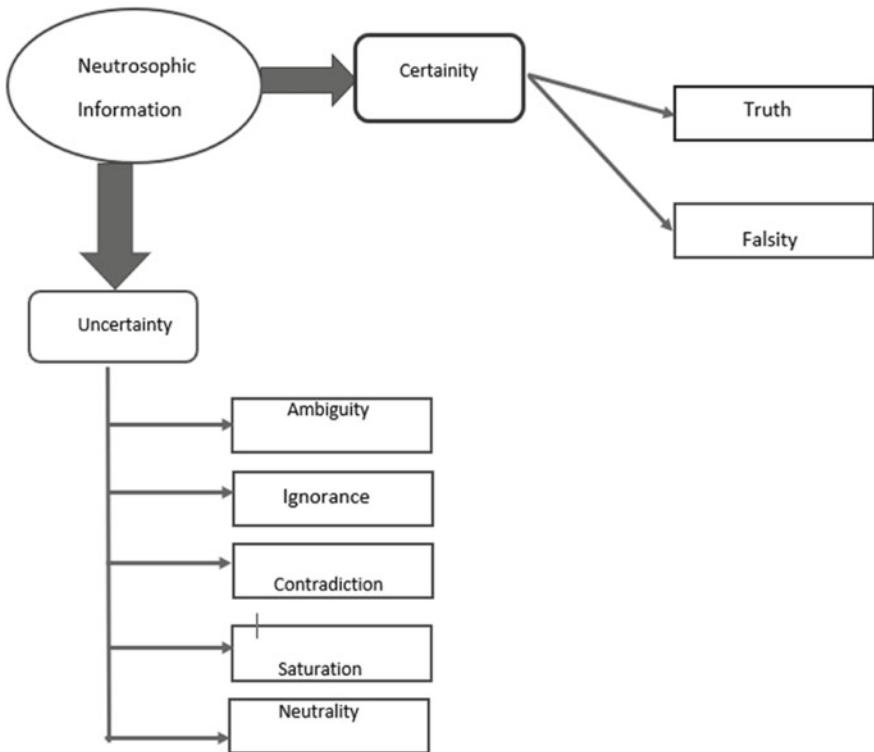


Fig. 38.1 Neutrosophic data

38.2.5 *Neutrosophic Data*

Neutrosophic data consists of certainty and uncertainty data. Certainty data is further subdivided into truth and falsity. In the case of uncertainty data, it consists of ambiguity, ignorance, contradiction, saturation, and neutralities (Fig. 38.1).

38.3 Relationship Between Neutrosophic and Fuzzy Logic

38.3.1 *Neutrosophic System Equation*

$$X \circ R = Y$$

where

$X \Rightarrow$ Neutrosophic data.

$Y \Rightarrow$ Neutrosophic output/information.



Fig. 38.2 Neutrosophic system equation

$R \Rightarrow$ Neutrosophic rules.

$\circ \Rightarrow$ Inference Mechanism.

Below is a block diagram of a neutrosophic system along with various I/O neutrosophic processors. The neutrosophic system is at the core and can interact with all of the processors (Fig. 38.2).

The neutrosophic system works on the principle of the neutrosophic system equation ' $X \circ R = Y$ '. The neutrosophic data ' X ' is input to the neutrosophic system, where R is the collection of neutrosophic rules that the system utilizes to obtain the output. The output obtained is neutrosophic information ' Y '. The neutrosophic knowledge is obtained by extracting knowledge from neutrosophic data. As a result, after implementing the decision support system procedure on neutrosophic data, the neutrosophic decision is the endpoint. The decision support system processes the data required for solving the computational tasks. The task of launching the I/O program is assigned by the decision support system. It collects neutrosophic data and displays neutrosophic outputs. The interface between the neutrosophic system and the devices is based on the decision support system. It entails a series of events that execute I/O operations before storing the results as neutrosophic output [3].

38.3.2 *Neutrosophic Logic Data Processing*

Neutrosophic data consists of certainty and uncertainty data. The neutrosophic system processes neutrosophic data into neutrosophic information (based on the neutrosophic system equation). Neutrosophic information is converted into neutrosophic knowledge. This knowledge helps in decision-making which is known as neutrosophic decision. The whole is known as neutrosophic data processing (Fig. 38.3).

38.3.3 *Representation of Neutrosophic Set*

See Fig. 38.4.

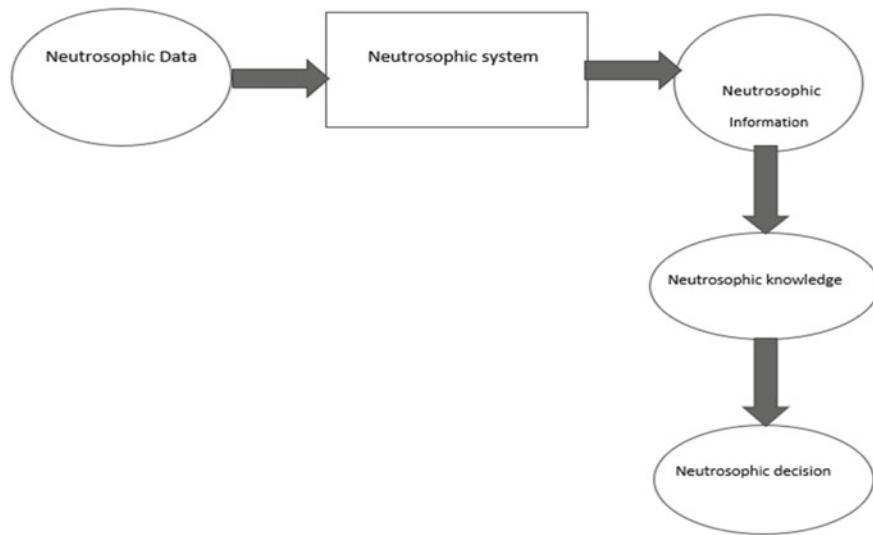


Fig. 38.3 Neutrosophic data processing

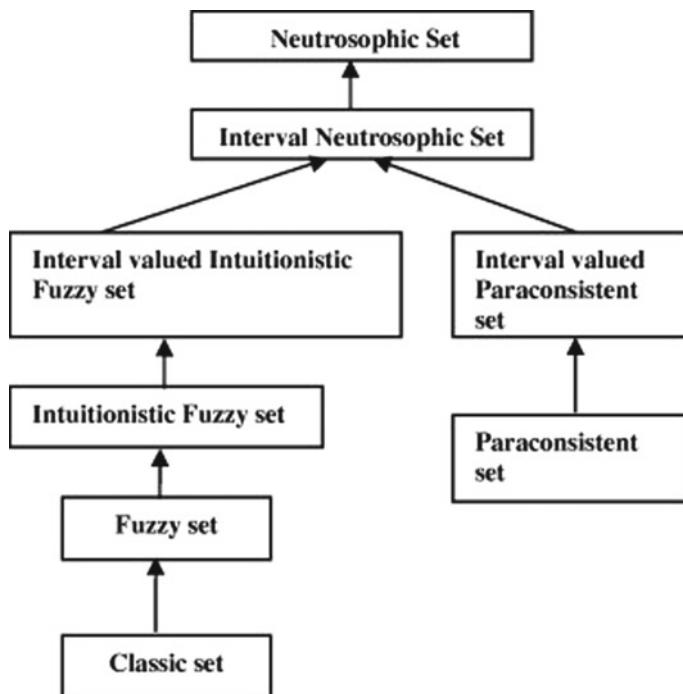


Fig. 38.4 Relation between neutrosophic set and other sets

38.3.3.1 Classical Set

These sets are sets with crisp boundaries. These sets are a collection of distinct objects.

38.3.3.2 Neutrosophic Set

It is a collection that contains triplets with distinct membership values for true, false, and indeterminacy (T, I, F).

38.3.3.3 Interval Neutrosophic Set

It is a neutrosophic set (A) that satisfies the following condition:

For each point z in Z , $T(z), I(z), F(z) \subseteq [0, 1]$.

where degree of membership T , indeterminacy membership function I , and degree of non-membership F are parameters that describe an interval neutrosophic set A in Z .

38.3.3.4 Intuitionistic Fuzzy Set

Atanassov presented a fuzzy set that more correctly quantifies uncertainty and allows for detailed modelling of the problem based on existing knowledge and observations which is known as an intuitionistic fuzzy set. An intuitionistic fuzzy set's extend version or generalization is the neutrosophic set [4].

38.3.3.5 Paraconsistent Set

If a logical consequence relation is not explosive, it is considered para consistent. If any arbitrary conclusion is implied by any arbitrary contradiction, a logical consequence connection is explosive [5].

Because the fuzzy set concentrates solely on the membership degree of members of the fuzzy set, it ignores unpredictability and indeterminacy that characterize the real world. The neutrosophic set is a subset or generalization of the intuitionistic fuzzy set. It successfully and efficiently illustrates real-world problems by taking into account all components of the situation (i.e. falsity, indeterminacy, truthiness) [6], as depicted in Fig. 38.5.

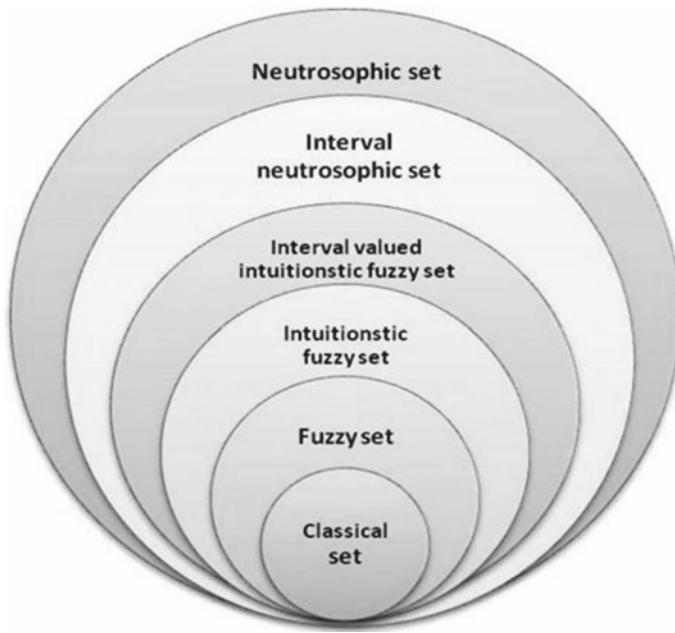


Fig. 38.5 From crisp sets to neutrosophic sets

38.4 Difference Between Fuzzy and Neutrosophic Logic

All three measurements (truth, falsehood, and indeterminacy) are independent in neutrosophic sets. We have inadequate information as a result of how one influences the other in decision-making, and their sum is less than 1. The total of the membership and non-membership components should be equal to 1 because they are based on the intuitionistic fuzzy set. If membership increases in the case of intuitionistic fuzzy, then certainly, the sum of the other two measures will decrease [7].

The neutrosophic logic has the ability to manipulate both incomplete and inconsistent data. Only imperfect information can be handled by fuzzy relations or intuitionistic fuzzy relations. The neutrosophic theory is more versatile and valuable than all other logic. It is more flexible.

The combination of qualifying probabilities can result in fuzzy probabilities that are either insufficiently exact or insufficiently informative. But the factor of indeterminacy plays a vital role in favouring neutrosophic logic systems over fuzzy logic systems [8].

In general, fuzzy logic inference reduces to the answers for a non-linear program; hence, finding strategies to solve such programs can be computationally expensive. To get inference from fuzzy if–then rules which have characteristics of fuzzy probability, there are currently no inexact, low-cost techniques. In addition, we do not have a good

Table 38.1 Relation between fuzzy, intuitionistic fuzzy, and neutrosophic logic

Uncertainty models	Uncertainty types			
	Vagueness	Imprecision	Ambiguity	Inconsistency
Fuzzy	Yes			
Intuitionistic fuzzy	Yes	Yes		
Neutrosophic	Yes	Yes	Yes	Yes

way to infer from possibility-qualified rules within a branch. This is a significant flaw in fuzzy logic (Table 38.1).

38.5 Advantages of Neutrosophic Logic Over Fuzzy Logic

- The fuzzy logic deals with inconsistencies, whereas neutrosophic logic deals with both inconsistencies and incompleteness.
- Although fuzzy logic ensures a particular element's multiple belongingness to multiple classes to varying degrees, it cannot capture neutralities due to indeterminacy. Furthermore, data representation employing fuzzy logic is constrained by the requirement that an element's membership and non-membership values sum to 1 [9].
- According to Leibniz's theory, the following definitions of absolute truth (true in all possible universes) and relative truth (truth in at least one world) are stated. In contrast to neutrosophic logic, the terms of absolute truth and relative truth are defined as NL (absolute truth) = 1^+ and NL (relative truth) = 1. Neutrosophic theory, on the other end, offers a means for dealing with data-driven indeterminacy [10].
- Due to its capacity to overcome indeterminacy difficulties, neutrosophic logic unlike other logic is a better indicator of the real information.
- Similarly, other allied logic like Lukasiewicz logic considered three values (1, 1/2, 0). All values are constrained between 0 and 1 only. A value less than 0 or an extension beyond 1 is not allowed. But in the case of neutrosophic logic, all values are possible.
- Neutrosophic theory's ability to deal with all aspects of a problem, including conflicts, can be combined with other types of sets such as rough sets, soft sets, and bipolar sets due to its hybrid behaviour. Therefore, we are seeing various uses of the neutrosophic theory which is increasingly being applied in various fields including medical diagnostics, data analysis, analysis of images, pattern recognition, aggregation, and cancer treatment.

Table 38.2 Difference between neural network, fuzzy, and neutrosophic logic

	Neural network	Fuzzy logic	Neutrosophic logic
Definition	It is a system that is based on the biological neurons of the human brain to perform computations	It is a style of reasoning and decision-making that is similar to human reasoning	It is a philosophical discipline concerned with the origin, nature, and scope of neutralities
Flexibility	This system cannot easily be modified	This system can easily be modified	It is more flexible
Training	It trains itself by learning from the data set	Everything must be defined explicitly	Everything must be defined explicitly
Inconsistencies, uncertainness	It deals with uncertainness	It deals with inconsistencies	Neutrosophic logic deals with both inconsistencies and incompleteness
Usage	It helps to perform predictions	It helps to perform pattern recognition	It helps to solve indeterminacy problems
Complexity	It is complex than fuzzy logic	It is simpler than a neural network	It is simpler than fuzzy logic
Learning	It is based on learning	It is not based on learning	It is not based on learning
Knowledge	Difficult to extract knowledge	Knowledge can easily be extracted	Knowledge can easily be extracted

38.6 Comparison Among Different Expert Systems

See Table 38.2.

38.7 Inference Systems of Different Expert Systems

38.7.1 Fuzzy Inference System

Fuzzy inference systems are widely used computing frameworks responsible for transforming the mapping from an input (in the case of fuzzy categorization, features are inputs) to an output (in the case of fuzzy categorization, classes are outputs), as shown in Fig. 38.6. Its concept is based on fuzzy if–then rules, fuzzy set theory, and fuzzy reasoning. It is made up of three parts: input fuzzification, knowledge-based system, and output defuzzification. The fuzzy knowledge base contains a collection of fuzzy production rules as well as membership functions defined in fuzzy sets. During fuzzification, the crisp input is converted to a fuzzy output using membership functions from the fuzzy knowledge base. There are some common defuzzification

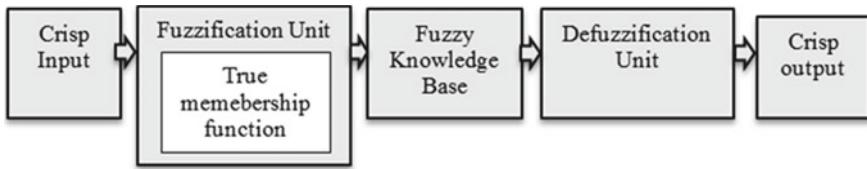


Fig. 38.6 Fuzzy inference system

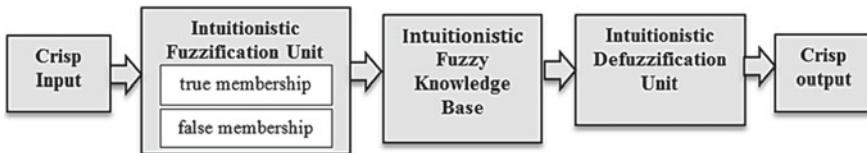


Fig. 38.7 Intuitionistic fuzzy inference system

techniques like centroid, bisector, and maximum methods in which crisp output is generated from the fuzzy output [7].

38.7.2 *Intuitionistic Fuzzy Inference System*

A single value between zero and one represents an element's membership in a fuzzy set. However, because there is a hesitating degree, the degree of non-membership of an element is not equal to a difference between 1 and the degree of membership. For imitating human imprecise decision-making, an intuitionistic fuzzy set is appropriate. Figure 38.7 depicts the intuitionistic fuzzy inference system. The fuzzy knowledge base stores the true and false membership functions of intuitionistic fuzzy sets, as well as a set of intuitionistic fuzzy production rules.

38.7.3 *Neutrosophic Inference System*

The neutrosophic inference system is made up of three parts: a neutrosophication unit that receives crisp input and distributes suitable membership functions, a neutrosophic knowledge base which is present in the central portion of the inference system that connects the input to output variable, and a deneutrosophication unit that takes neutrosophic values and transforms them to crisp values [11] (Fig. 38.8).

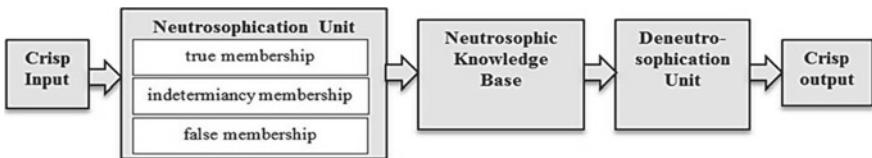


Fig. 38.8 Neutrosophic inference system

38.8 Applications of Neutrosophic System

Neutrosophic applications in different sectors have been rapidly developing in all directions, particularly since the publication of numerous scientific journals and studies on neutrosophic logic. In neutrosophic logic, new theories, methodologies, and algorithms have been rapidly developed. Due to the neutrosophic set's hybridization nature, it can combine with other possibilities of sets is one of the most noticeable developments in neutrosophic theory. Examples of such sets are bipolar set, rough set, soft set, and hesitant fuzzy set. Different hybrid structures introduced in recent works, such as bipolar neutrosophic set, single-valued neutrosophic rough set, single-valued neutrosophic hesitant fuzzy set, rough neutrosophic set, and so on, have found wide application in different sectors. Data analytics, decision-making, cancer treatment, automobiles, medical, social science, and other fields have all benefited from the use of the neutrosophic set [8]. One of the most promising uses of the neutrosophic logic framework could be assisting doctors in better classification of Alzheimer's disease, potentially leading to earlier treatment.

In Table 38.3, we provided details of neutrosophic techniques and their applications in different sectors like healthcare and the medical industry through some prominent scientific journals and articles.

In Paper 1, it was discovered that the role of the neutrosophic logic tool was critical. It was able to properly forecast breast cancer detection in patients in the majority of cases, and more than 80% of the time when predicting the onset of breast cancer [12].

Another example of how neutrosophic logic could help individuals with neurological diseases is Alzheimer's disease. In paper 2, the authors have released a study describing a new A.I. system based on neutrosophic systems that could identify Alzheimer's disease up to six years sooner than current diagnostic procedures. It designed the method by using thousands of brain images to train a machine learning algorithm, making it one of the clearer Alzheimer's detection techniques now available. The AI system based on neutrosophic logic outperformed human clinicians in recognizing patterns of scans of the brain, an indicator that can suggest brain sickness [13].

According to the findings in Paper 3, the algorithm could assist in diagnosing breast cancer by earlier detection than the traditional method in many situations. The findings point to the potential use of neutrosophic logic in mammography to

Table 38.3 Applications of neutrosophic logic in different domains

S. No.	Sources	Year of publication	Selected applications	Methodology
1	Thermogram breast cancer prediction approach based on Neutrosophic Sets and Fuzzy C-Means Algorithm, Tarek Gaber, Gehad Ismail	2015	Breast cancer treatment	Using neutrosophic set, FFCM, and morphological operators, the system first extracted ROI. The SVM was then used to detect normal and abnormal breasts using many variables (statistical, texture, and energy)
2	A recommender system for Alzheimer Patients in Sultanate of Oman using Neutrosophic Logic	2020	Alzheimer's disease early detection	Patients and clinicians can use the recommendations to determine the seriousness of Alzheimer's disease at the different stages of therapy
3	Breast cancer detection using neutrosophic logic, S. Sivaranjani, K. Kaarthik	2019	Breast cancer treatment	Neutrosophic logic designed to better model and predict breast cancer was able to outperform radiologists, in some cases with an ability to detect tiny malignant tissues that would otherwise go unnoticed
4	Analysing age group and time of the day using interval-valued neutrosophic sets, S. Broumi, M. Lathamaheswari	2020	Analysing and forecasting patterns	This paper uses an interval-valued neutrosophic set with thorough description and pictorial depiction to analyse age groups and time (day or night)
5	Hong-yu Zhang, Jian-qiang Wang, Xiao-hong Chen, "Interval neutrosophic sets and their application in multi-criteria decision making problems"	2014	Multi-criteria decision-making problems	Presented a multiple attribute group decision-making approach

(continued)

Table 38.3 (continued)

S. No.	Sources	Year of publication	Selected applications	Methodology
6	Smarandache, Florentin; Leyva-Vázquez, Maikel (2018): Fundamentals of neutrosophic logic and sets and their role in artificial intelligence	2018	Artificial intelligence	Neutrosophic logic is used in A.I. to get better results with greater precision and automation results
7	Challenges and future directions in neutrosophic set-based medical image analysis Deepika Koundal, Bhisham Sharma	2019	Medical imaging	Neutrosophic logic models outperformed traditional machine learning techniques in detecting patterns and discriminative features in brain imaging
8	Applications of neutrosophic logic to robotics: An introduction, F. Smarandache, and L. Vlăduțeanu	2011	Robotics	The paper begins with a brief discussion of a robot's mathematical dynamics, followed by a method for bringing neutrosophic science to robotics
9	Comparison of neutrosophic approach to various deep learning models for sentiment analysis, Mayukh Sharma, Ilanthenral Kandasamy, W.B. Vasantha	2021	Sentiment analysis	Deep learning has been applied in a variety of real industrial applications as well as classification difficulties. Neutrosophic theory depicts neutrality and indeterminacy parameters clearly
10	COVID-19 Vaccine: A neutrosophic MCDM approach for determining the priority groups, Ibrahim M. Hezam, Moddassir Khan Nayeem, Abdelaziz Foul	2021	COVID-19 vaccines	The process is involved in identifying priority groups for receiving COVID vaccines using age and job criteria

diagnose breast cancer earlier, as well as the ability to build AI for other clinical fields with better results [14].

According to paper 4, neutrosophic logic is appropriate for this research since it parallels human behaviour in terms of forecasting age and time (or day-night). Membership values of truth, indeterminacy, and falsehood may be exact numbers or interval numbers, based on human intelligence. This paper uses an interval-valued neutrosophic set with a thorough description and pictorial depiction to analyse age groups and time (day or night). Another relevant area described in the paper is fuzzy logic and the representation of uncertainty, as well as its application to complex systems. These causal models are tools that can assist you to make better decisions regarding uncertainty [15].

In real-world scientific and technical applications, interval neutrosophic sets are useful for dealing with data that is unclear, imprecise, incomplete, or inconsistent. However, because there is not enough research on interval neutrosophic sets, a relatively young branch of neutrosophic sets, there is not enough to go on. The available literature, in particular, does not mention aggregation operators or a multi-criteria decision-making strategy for interval neutrosophic sets. Based on the related research achievements in interval-valued intuitionistic fuzzy sets, the operations of interval neutrosophic sets are defined in paper 5. And the approach to compare interval neutrosophic sets was proposed in the paper. As a result of the proposed operators, a multi-criteria decision-making approach is constructed. The process discussed in the paper includes the ranking of all options that may be calculated using the comparison approach, and the best one can be simply found [16].

For applying the machine learning model, traditional machine learning approaches require a centralized database with direct access to patient data. Such methods are affected by various factors such as rules and regulations, the privacy of patient's data, information security, data ownership, and the overhead to hospitals of developing and maintaining these centralized databases. The application of neutrosophic theory in A.I. is growing in popularity since it is thought to provide the best outcomes. Robotics, autonomous decisions, satellite image classification, healthcare problems, neutrosophic cognitive maps, linear and non-linear programming, and neutrosophic cognitive maps have all used neutrosophic logic, sets, probability, and statistics in the expansion of A.I. (artificial intelligence) tools and methodologies. This wide range of applications of neutrosophic theory in various fields such as A.I. (artificial intelligence) has raised new concerns and brought creative solutions to pressing issues [17].

Medical technicians face a difficult problem in effectively predicting medical imaging scans since the data is extremely complicated and the links between different types of data are poorly understood. Neutrosophic logic models outperformed traditional machine learning techniques in detecting patterns and discriminative features in brain imaging. According to paper 7, neutrosophic logic takes medical imaging to the next level. Thanks to a potent mix of artificial intelligence, neutrosophic logic, and 3D medical imaging, substantial advances in imaging have been made in the last five years [18].

Scaling has become impossible without robots, and solutions can now be deployed with remarkable speed and low downtime. Advances in robotic autonomy, which use various machine learning approaches to improve the robot's ability to recognize, diagnose, and respond to unanticipated conditions, determine the pace and speed of scientific advancement in artificial intelligence. The application of neutrosophic logic in robotics is one of the strategies outlined in paper 8. The paper begins with a brief discussion of a robot's mathematical dynamics, followed by a method for bringing neutrosophic science to robotics. Robotics is the study and development of devices that perform physical activities in an automated manner. The integration of neutrosophic systems has resulted in a new generation of robots that can collaborate with humans and execute a variety of jobs in challenging conditions. Drones, disaster-response robots, and robot aides in home health care are just a few examples [19].

According to paper 9, deep learning has been applied in a variety of real industrial applications as well as classification difficulties. Neutrosophic theory depicts neutrality and indeterminacy parameters. Due to the presence of these factors, it is used for sentiment analysis [20].

Authorities must identify priority groups for COVID-19 vaccine dose allocation through different deep learning techniques, according to the findings in Paper 10. The four primary criteria based on age, health status, women's status, and the type of job are defined in this paper. All of the findings show that healthcare workers, persons with pre-existing medical conditions, elderly, essential employees, and pregnant and nursing moms should be the first to receive the vaccine dosage [21].

38.9 Conclusion and Future Work

As big data analytics, artificial intelligence, machine learning (ML), deep learning, and other technologies are being increasingly widely used in fields such as medical, robotics, smart homes, and automobiles, researchers are actively seeking innovative approaches to training different algorithms which will be effective and give accurate and precise results across different sectors. The neutrosophic system is among the most effective methods which are widely accepted in several research fields. Because of greater accuracy and precise results, neutrosophic techniques are preferred over other deep learning techniques, but there is a need for research in this field. Furthermore, the neutrosophic system can solve more complex issues in decision-making because due to the presence of three truth-membership, indeterminacy, and falsity membership components. As a result, the neutrosophic system, unlike the fuzzy logic system, is more generalized and indeterminacy tolerant in its functioning. So, neutrosophic logic is a generalized form of fuzzy logic and intuitionistic logic. In this paper, we discussed different possible scenarios of uses of neutrosophic logic in the field of medical treatment, breast cancer, artificial intelligence, and robotics. We also discussed about neutrosophic system equation ($X \circ R = Y$). After the study, we have identified some improvements which are left for future work. Different performance factors and criteria discussed here can also be used in future models to rank

and evaluate the performance of models. In conclusion, the study presents that the role of neutrosophic logic in decision-making is very important and vital. Further researches are recommended in the above directions.

References

1. Rivieccio, U.: Neutrosophic logics: prospects and problems. *Fuzzy Sets Syst.* **159**(14) (2008). <https://doi.org/10.1016/j.fss.2007.11.011>
2. Ansari, A.Q., Biswas, R., Aggarwal, S.: Neutrosophic classifier: an extension of fuzzy classifier. *Appl. Soft Comput.* **13**(1), 563–573 (2013). <https://doi.org/10.1016/j.asoc.2012.08.002>
3. Gafar, M.G., Elhoseny, M., Gunasekaran, M.: Modelingneutrosophic variables based on particle swarm optimization and information theory measures for forest fires. *J. Supercomput.* **76**, 2339–2356 (2020). <https://doi.org/10.1007/s11227-018-2512-5>
4. Atanassov, K.T.: Intuitionistic fuzzy sets. Physica-Verlag, Heidelberg (1999)
5. Priest, G., Tanaka, K., Weber, Z.: Para consistent logic. In: Zalta, E.N. (ed) The Stanford Encyclopedia of Philosophy (Summer 2018 Edition). <https://plato.stanford.edu/archives/sum2018/entries/logic-paraconsistent/>
6. Abdel-Basset, M., Manogaran, G., Gamal, A., Smarandache, F.: A hybrid approach of neutrosophic sets and DEMATEL method for developing supplier selection criteria. *Des. Autom. Embedded Syst.* **22**(3), 257–278 (2018)
7. Smarandache, F.: A unifying field in logics: neutrosophic logic. In: Neutrosophy, Neutrosophic Set, Neutrosophic Probability: Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability. Infinite Study. American Research Press, Santa Fe (2005)
8. Smarandache, F. (ed.), Proceedings of the First International Conference on Neutrosophy, Neutrosophic logic, Neutrosophic Set, Neutrosophic Probability and Statistics. University of New Mexico, Gallup Campus, Xiquan, Phoenix, p. 147 (2002)
9. Smarandache, F., Leyva-Vázquez, M.: Fundamentals of neutrosophic logic and sets and their role in artificial intelligence. Journal contribution (2018). <https://doi.org/10.6084/m9.figshare.7048106.v1>
10. Kavitha, B., Karthikeyan, S., Sheeba Maybell, P.: An ensemble design of intrusion detection system for handling uncertainty using neutrosophic logic classifier. *Know.-Based Syst.* **28**, 88–96 (2012). <https://doi.org/10.1016/j.knosys.2011.12.004>
11. Radwan, N., BadrSerousy, M., Riad, A.E.D.M.: Neutrosophic logic approach for evaluating learning management systems. *Neutrosophic Sets Syst.* **11**, 3–7 (2016)
12. Broumi, S., Lathamaheswari, M., Bakali, A., Talea, M., Smarandache, F., Nagarajan, D., Kavikumar, K., Asmae, G.: Analyzing age group and time of the day using interval valued neutrosophic sets. *Neutrosophic Sets Syst.* **32**, 1 (2020)
13. Alzadjali, N., Jereesa, M.S., Savarimuthu, C., Divyajyothi, M.G.: A recommender system for Alzheimer patients in sultanate of Oman using neutrosophic logic. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1–5 (2020)
14. Sivaranjani, S., et al.: Breast cancer detection using neutrosophic logic. *Int. J. Fut. Gener. Commun. Netw.* **12**(5) (2019)
15. Gaber, T., et al.: Thermogram breast cancer prediction approach based on neutrosophic sets and fuzzy c-means algorithm. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4254–4257 (2015). <https://doi.org/10.1109/EMBC.2015.7319334>
16. Zhang, H., Wang, J., Chen, X.: Interval neutrosophic sets and their application in multicriteria decision making problems. *Sci. World J.* **2014**. Article ID 645953, 15 p (2014)

17. Lupiáñez, F.G.: On neutrosophic sets and topology. *Procedia Comput. Sci.* **120**, 975–982 (2017). <https://doi.org/10.1016/j.procs.2018.01.090>
18. Koundal, D., Sharma, B.: 15-challenges and future directions in neutrosophic set-based medical image analysis. In: *Neutrosophic Set in Medical Image Analysis*. Academic Press, pp. 313–343 (2019)
19. Smarandache, F., Vlăduceanu, L.: Applications of neutrosophic logic to robotics: An introduction. *IEEE International Conference on Granular Computing* **2011**, 607–612 (2011). <https://doi.org/10.1109/GRC.2011.6122666>
20. Sharma, M., Kandasamy, I., Vasantha, W.B.: Comparison of neutrosophic approach to various deep learning models for sentiment analysis. *Knowl. Based Syst.* **223**, 107058 (2021). ISSN 0950-7051. <https://doi.org/10.1016/j.knosys.2021.107058>
21. Hezam, I.M., Nayeem, M.K., Foul, A., Alrasheedi, A.F.: COVID-19 vaccine: a neutrosophic MCDM approach for determining the priority groups. *Results Phys.* **20**, 103654 (2021). ISSN 2211-3797. <https://doi.org/10.1016/j.rinp.2020.103654>

Chapter 39

Application of Expectation–Maximization Algorithm to Solve Lexical Divergence in Bangla–Odia Machine Translation



Bishwa Ranjan Das, Hima Bindu Maringanti, and Niladri Sekhar Dash

Abstract This paper shows the word alignment between Odia–Bangla languages using the expectation–maximization (EM) algorithm with high accuracy output. The entire mathematical calculation is worked out and shown here by taking some Bangla–Odia sentences as a set of examples. The EM algorithm helps to find out the maximum likelihood probability value with the collaboration of the ‘argmax function’ that follows the mapping between two or more words of source and target language sentences. The lexical relationship among the words between two parallel sentences is known after calculating some mathematical values, and those values indicate which word of the target language is aligned with which word of the source language. As the EM algorithm is an iterative or looping process, the word relationship between source and target languages is easily found out by calculating some probability values in terms of maximum likelihood estimation (MLE) in an iterative way. To find the MLE or maximum a posterior (MAP) of parameters in the probability model, the model depends on unobserved latent variable(s). For years, it has been one of the toughest challenges because the process of lexical alignment for translation involves several machine learning algorithms and mathematical modeling. Keeping all these issues in mind, we have attempted to describe the nature of lexical problems that arise at the time of analyzing bilingual translated texts between Bangla (as source language) and Odia (as the target language). In word alignment, handling the ‘word divergence’ or ‘lexical divergence’ problem is the main issue and a challenging task, though it is not solved by EM algorithm, it is only possible through a bilingual dictionary or called as a lexical database that is experimentally examined and tested only mathematically. Problems of word divergence are normally addressed at the phrase level using bilingual dictionaries or lexical databases. The basic challenge lies in the identification of the single word units of the source text which are converted into multiword units in the target text.

B. R. Das (✉) · H. B. Maringanti

Department of Computer Application, Maharaja Sriram Chandra Bhanja Deo University, Baripada, India

N. S. Dash

Linguistic Research Unit, Indian Statistical Institute, Kolkata, India

39.1 Introduction

Word alignment is the process of identifying or mapping the exact and corresponding word between two parallel corpora. It is one of the translation relationships of the words between two or more parallel sentences. In some cases, a word is translated by a single word or multiple words, and this is termed as word divergence. If parallel sentences are given, then finding the corresponding relationship among words that may be one-to-one, one-to-many and many-to-many of source and target sentences remains the main task of word alignment. Alignment of source language phrases with corresponding target language phrases or groups of words is the solution of phrase-based translation. If the words of the source sentence are unable to find their appropriate translation in the target language, simply they are assigned null. The movement of translated words in the source sentence to their appropriate position in the target sentence is also done in word alignment. In case of bilingual machine translation, the word reordering may be a necessity, and word alignment helps in achieving it. There are multiple factors for word alignment, i.e., named entities, transliteration similarities, local word grouping, nearest aligned neighbors and dictionary lookup. The various challenges of achieving word alignment include ambiguity, word order, word sense, idioms and pronoun resolution.

39.2 Expectation and Maximization (EM) Algorithm

An EM algorithm is an iterative process, classed as an unsupervised method to find the maximum likelihood estimation (MLE) or maximum a posterior (MAP) estimates of parameters in a statistical model, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E*-step. The main objective of this algorithm is to find the best result or value.

39.3 Word Alignment with Methodology

This paper focuses to learn a conditional probability model of an Odia sentence ‘*O*’ given a Bangla sentence ‘*B*’, which is denoted as $P_\theta(O|B)$. The subscript θ refers to the set of parameters in the model having a dataset D of ‘*n*’ sentence pairs that are known to be the translation of each other, $D = \{(B_1, O_1), (B_2, O_2), (B_3, O_3), \dots, (B_n, O_n)\}$, where each subscript ‘*n*’ indicates a different pair. The model is designed in such a way that there will be no chance of uncovering the hidden word-to-word correspondences in these translation pairs. The model is fully trained from data to

predict the existence of the missing word alignment. These are many ways to define $P(O|B)$. Suppose an Odia sentence O be represented by an array of indices I words, $(O_1, O_2, O_3, \dots, O_I)$ and Bangla sentence B be represented by an array of J words, $(B_1, B_2, B_3, \dots, B_J)$. Now the assumption can be made that each Odia word aligned to exactly one or more Bangla word, and this can be represented as an array a of length I , denoted as $(a_1, a_2, a_3, \dots, a_i)$ where $a_1, a_2, a_3, \dots, a_i$ are one-one alignment variables. An alignment variable a_i takes a value in the range $[0, J]$, and the index of the Bangla word to which Odia word O_i is aligned. If $a_i = 0$, this means that O_i is not aligned to any word in the Bangla sentence, called as null alignment. Consider the sentence pair Bangla–Odia below.

Bangla sentence: নিজেদের দাবি নিয়ে নির্মাণকার্য বন্ধ করার জন্য কৃষকদের সংগঠনের মধ্যে আলোড়ন সৃষ্টি হয়েছে।

Transliteration: ‘Nijeder dAbi niye nirmAnkAryya bandha karAr janna krishakder sangaThener madhye aloRon srisTi hayechhe’.

Odia sentence: ନିଜର ଅଧିକାରକୁ ନେଇ ନିର୍ମାଣ କାର୍ଯ୍ୟ ବନ୍ଦକରିବାକୁ କୃଷକସଂଗଠନଗୁଡ଼ିକରେ ହତେମଟ ସୃଷ୍ଟି ହୋଇ ଯାଇଛି।

Transliteration: ‘Nijara adhikaraku nei nirmanakarjya banda karibaku krushaka sangathana gudikare hatachamata shrusti haijachhi’.

The length of the Bangla sentence is 14, and the length of the Odia sentence is 12. Here I indicates the length of the Odia sentence, and J indicates the length of the Bangla sentence. The words of both the sentences are indexed like $B_1, B_2, B_3, \dots, B_J$, and $O_1, O_2, O_3, \dots, O_I$, respectively, for Bangla and Odia. The value of an alignment array ‘ a ’ will be $\{1, 2, 3, 4, 5, 6, 7-8, 9, 10-11, 12\}$. So the probabilistic model generates the Odia sentence from Bangla using a simple procedure. First the length I is chosen according to a distribution $P(I|J)$, in this case, $P(12|14)$. Then each Odia word position aligns to a Bangla word (or null) according to the valid sentence alignment of the standard corpus (ILCI) is $P(a_i = j|J)$. Finally, each Odia word O_i is translated according to the probability distribution function on the aligned Bangla word, $P(O_i|B_{a_i})$. So for this alignment, all probability values are multiplied likewise $P(\text{NijaraNijeder}), P(\text{adhikarakuldabi}), P(\text{neilnie})$ and so on. The joint probability of the Odia sentence and its alignment conditioned on Bangla is simply the product of all these probabilities.

$$P(O, a|B) = P(I|J) \prod_{i=1}^I P(a_i|J) \cdot P(O_i|B_{a_i}) \quad (39.1)$$

It is simply two tables of numbers: $P(I|J)$, for all pairs of sentence lengths I and J and $P(O|B)$ for all pairs of co-occurring Odia and Bangla words O and B . Since these numbers represent probabilities, the set of valid assignments of numbers to these tables must follow basic rules of probability.

$$\forall_{B,O} P(O|B) \in [0, 1] \quad (39.2)$$

$$\forall_B \sum_O P(O|B) = 1 \quad (39.3)$$

To observe the alignment, just taking care of the $P(O|B)$ and estimating the approximate value through maximum likelihood estimation (MLE) is the process to be adopted. At first, the alignment of the sentences needs to be done properly before starting word alignment between Bangla and Odia. But there is no such type of situation occurring in Bangla–Odia as shown in English–French [19] translation. For example, most of the words of French are aligned with the English words many times, but this never occurs in Bangla–Odia. To bring in clarity further, the MLE function is introduced here to calculate the probability of the given parameters.

$$\prod_{n=1}^N P_\theta(O^{(n)}, a^{(n)}|B^{(n)}) = \prod_{n=1}^N P(I^{(n)}|J^{(n)}) \prod_{i=1}^{I^{(n)}} P(a_i^{(n)}|J^{(n)} \cdot P(O_i^{(n)}|B_{a_i}^{(n)}) \quad (39.4)$$

Now data is observed, and the parameters are estimated, finally needing a probability function to find the highest value as our data (value) is highly probable under this model.

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^N P_\theta(O^{(n)}, a^{(n)}|B^{(n)}) \quad (39.5)$$

In Eq. (39.5), where $\hat{\theta}$ it searches the highest probability value of word alignment by argmax function for each and every word in a sentence. It is basically a searching problem from an infinite number of possible sentences in the case of machine translation. Only one sentence is selected from different possible sentences after translation is in agreement with the corpus. For this case, though the search problem is trivial, there is a closed-form solution for $\hat{\theta}$ when the data described by our model is fully observed. An algorithm is developed to learn θ from our hypothetical aligned data actually, initiating the strategy or model which is described here. It is very simple: The data is scanned and observing the alignments and counting them (means aligned data, Bangla–Odia pairs) for each Bangla–Odia word pair. To obtain probability values (aligned word pair Bangla–Odia), all counts (means probability values) are normalized by the number of times they are observed corresponding to Bangla word participating in any alignment. This yields a very simple algorithm described here.

39.4 Algorithm

- Step 1. Initialize all counts to 0.
- Step 2. For each n value between 1 and N .
- Step 3. For each i value between 1 and I .

- Step 4. For each j value between 1 and J .
- Step 5. Compare $a_i = j$ upto n , i.e., i value.
- Step 6. Count $[(O_i, B_j)]++$
- Step 7. Count $[B_j]++$
- Step 8. For each (O_i, B_j) value in count.
- Step 9. $P(O|B) = \text{Count}(O, B)/\text{Count}(B)$.

This algorithm loops over all pairs of the word in each in order to collect count, a computation that is quadratic in sentence length. This is not strictly necessary: It could have just looped over the alignment variable to collect the counts, which is linear. However, thinking about the algorithm as one that examines all pairs of a word will be useful when it moves to the case of unobserved alignments, which turns out to be an extension of this algorithm. Here two formulae are used to calculate the alignment probabilities after some iterations.

$$\begin{aligned} C(B_i \leftrightarrow O_j; B^s \leftrightarrow O^s) \\ = \frac{P(O_j|B_i)}{\sum_x P(x|B_i)} * (\#B_i \in B^s) * (\#O_j \in O^s) \end{aligned} \quad (39.6)$$

where C = expected count of $B_i \leftrightarrow O_j$ mapping in the context of the parallel corpus $B^s \leftrightarrow O^s$. $\#B_i \in B^s$ = no. of time B_i occurs in B^s . $\#O_j \in O^s$ = no. of time O_j occurs in O^s . ‘ s ’ references to a parallel sentence pair.

$$P(O_j|B_i) = \frac{\sum_s C(B_i \leftrightarrow O_j; B^s \leftrightarrow O^s)}{\sum_s \sum_x C(B_i \leftrightarrow x; B^s \leftrightarrow O^s)} M - \text{step} \quad (39.7)$$

where $P(O_j|B_i)$ is calculated from the ratio of counts of $B_i \leftrightarrow O_j$ mapping in all parallel sentence pairs and the count of mapping of $B_i \leftrightarrow x$, where x is any word in all the parallel sentences. It can be proved that after every iteration of E -step and M -step, the likelihood of the data, which in this case, the parallel corpora, increases monotonically. An equivalent way of describing improvement is the progressive decrease in entropy. So the iterative procedure is greedy. It could have got stuck in a local minimum, but for the fact that the data likelihood expression is a convex one and guarantees global minimum.

A Bangla sentence $B = b_1, b_2, b_3 \dots b_j$ and translated into an Odia sentence $O = o_1, o_2, o_3 \dots o_i$. Among all possible Odia sentences, one is looked for the highest probability $P(O|B)$. Using Bayes rule, it can be written as

$$P(O|B) = P(O)P(B|O)/P(B) \quad (39.8)$$

As the denominator is independent of O , finding the most probable translation e^* will lead to the noisy channel model for statistical machine translation.

$$e^* = \text{argmax} P(O|B) \quad (39.9)$$

$$= \operatorname{argmax} P(O)P(B|O) \quad (39.10)$$

where $P(O)$ is called the language model and $P(B|O)$ is the translation model. In most of the cases, many-to-one and one-to-many word alignment are purely based on phrase-based translation, and there is no other way to do translation when word divergence is seen in word alignment. A bilingual Bangla–Odia lexicon is developed as per the corpus based on the agriculture domain for mapping the words and translated very smoothly by one-to-one correspondence.

39.5 Result and Discussion

In the bilingual dictionary based on the agriculture domain, a small handful of sentences (approximately five thousand) around fifty thousand words are stored in a well-formatted and scientific manner for easy access with observed alignments. All observed alignments are trained, and it produces a good estimate of θ as mentioned in Eq. (39.5). If we think about, to take a very large corpus then we get very good estimation or accuracy, it contains a one-to-one word, many-to-one and many-to-one word correspondence. First of all connections (as one-to-one mapping) are equally likely. After one iteration, the model learns that the connection is made between most similar words from two parallel sentences by finding the probability value. After another iteration, it becomes clear that a connection between previous similar words is more likely as the probability value of the current word. So bigram and trigram are the best methods to find the probability of the sentence and alignment among the words. All probability values are calculated using a bigram in the form of a table/matrix. Expected count, revised expected count and revised alignment probabilities values are calculated among words in each parallel sentence. Revised alignment probabilities give more approximation value between the words in the parallel sentence. The average entropy value with EM alignment 1.4 is already less than the average entropy value of 1.53 with heuristic alignment, signaling that there is a progress toward a better probability distribution.

$$1.4 < 1.53$$

This percentage value can be further enhanced by using other mathematical algorithms (Fig. 39.1).

39.6 Conclusion and Future Work

When translation happens from one language to another, first of all, if a parallel corpus is properly aligned at sentence level, then word-by-word translation is easily done by a

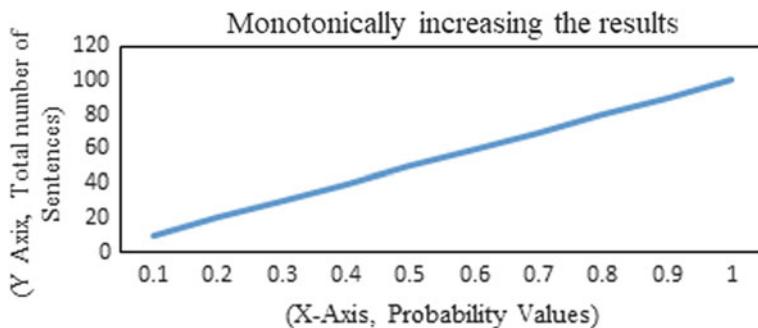


Fig. 39.1 Performance of EM algorithm

machine. With proper estimation, most of the problem converges to one-to-many and many-to-one alignment which are solved by bilingual dictionary and phrase-based translation. A bilingual dictionary including one-to-one, one-to-many and many-to-one correspondence (Bangla–Odia) between two languages is created. Sometimes phrase-based translation is a more appropriate solution for word divergence occurrences. The expectation–maximization algorithm is used for finding the most suitable word pair between two languages (Bangla–Odia) from where the highest probability value is taken. It also helps to translate word by word, phrase wise and finding the appropriate position of the word of the target language with good accuracy. Time complexity is one of the major factors when data is huge for machine translation. So care should be taken to obtain a better result; to optimize this is a challenging task. Space complexity may not be considered, as our data or corpus is huge, space should be increased for this as far as memory is concerned; otherwise, any research work based on NLP or data science will only be superficial.

References

1. Aswani, N., Gaizauskas, R.: Aligning words in English-Hindi parallel corpora. In: Association for Computational Linguistics, vol. 19, pp. 115–118 (2005)
2. Das, B.R., Maringanti, H.B., Dash, N.S.: Word alignment in bilingual text for Bangla to Odia machine translation. Presented in the International Conference on Languaging and Translating: Within and Beyond on 21–23 Feb 2020, IIT Patna, India
3. Das, B.R., Maringanti, H.B., Dash, N.S.: Challenges faced in machine learning-based Bangla–Odia word alignment for machine translation. Presented in the 42nd International Conference of Linguistic Society of India (ICOLSI-42) on 10–12 Dec 2020, GLA University, Mathura, UP, India
4. Das, B.R., Maringanti, H.B., Dash, N.S.: Bangla-Odia word alignment using EM algorithm for machine translation. J. Sci. Technol (Special issue), Maharaja Sriram Chandra Bhanja Deo (erstwhile North Orissa) University, Baripada, India
5. Dubey, S., Diwan, T.D.: Supporting large English-Hindi parallel corpus using word alignment. Int. J. Comput. Appl. **49**(16–19) (2012)

6. Jindal, K., et al.: Automatic word aligning algorithm for Hindi-Punjabi parallel text. In: Conference on Information Systems for Indian languages, pp. 180–184 (2011)
7. Koehn, P., Knight, K.: Empirical methods for compounding splitting. In: EACL ‘03 Association for Computational Linguistics, vol. 1, pp. 187–193, 12–17 Apr (2003)
8. Mansouri, A.B., et. al.: Joint prediction of word alignment with alignment types. *Trans. Assoc. Comput. Linguist.* **5**, 501–514 (2017)
9. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* **19**(2), 263–311 (1993)
10. Koehn, P.: Statistical machine translation (2010)
11. Songyot, T., Songyot, D.C.: Improving word alignment using word similarity. In: Empirical methods in Natural Language Processing, pp. 1840–1845 (2014)
12. Tidemann, J.: Word alignment step by step. In: Proceedings of the 12th Nordic Conference on Computational Linguistics, pp. 216–227. University of Trondheim, Norway (1999)
13. Tidemann, J.: Combining clues for word alignment. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 339–346, Budapest, Hungary, Apr 2003
14. Tidemann, J.: Word to word alignment strategies. In: International Conference on Computational Linguistics (2004)
15. Bhattacharyya, P.: Machine Translation. CRC Press (2017)
16. Jurafsky, D., Martin, J.H.: Speech and Language Processing, 4th edn. Pearson (2011)
17. https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm.
18. https://www.cs.sfu.ca/~anoop/students/annahita_mansouri/annahita-depth-report.pdf.
19. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.421.5497&rep=rep1&type=pdf>.

Chapter 40

Impact of Odisha Gramya Bank on Socio-economic Development of Beneficiaries: A Case Study of Balasore and Bhadrak District of Odisha



Artta Bandhu Jena and Debadutta Nayak

The Future of India lies in its Village

—Mahatma Gandhi

Abstract Indian economy is based on rural in character. After economic reform process, rural India has witnessed the increasing trend in purchasing power of people, income level, literacy, irrigation, infrastructure, technology and telecommunication facility and standard of living also. RBI and Govt. of India have jointly taken the economic bold step for formation of Regional Rural Banks (RRBs), RRBs were set to provide the basic banking and financial services to needy and poor rural people in India. The main objective of RRBs is to mobile the funds from customer and allocate these mobilizations in the form of loan and advance mainly to marginal farmers, agricultural labourers, rural artisans, etc. to meet their financial requirements to develop the rural sectors and also to have the socio-economic development of needy and poor rural people. The rural people face the problem of indebtedness, getting credit and many more. Needy and poor rural people need the credit/loan for agriculture and allied sectors. Rural people need to access the financial institutions that can provide them with credit/loan at lower rate of interest with reasonable terms and conditions. Hence, RRBs can fulfil the credit gap and other gaps also of needy and poor rural people. As we know, 2/3rd people live in rural India. RRBs are the rural banks for the rural people. Odisha Gramya Bank (OGB) is a RRB. Here, an attempt has been made to measure the role played by OGB towards socio-economic development of beneficiaries of Balasore and Bhadrak District of Odisha. The present study has taken both primary and secondary data.

A. B. Jena (✉) · D. Nayak

ICSSR Sponsored Major Research Project (MRP), Department of Business Management, Fakir Mohan University, Vyasa Vihar, Balasore, Odisha, India

40.1 Introduction

The banking facility was the main problem in pre-independence period of India in rural areas. Needy and rural people were depending upon exploitative terms and conditions of credit/loan mainly from private moneylenders, landlords, big farmers, trade merchants, etc. They were usually charging high rate of interest on credit/loan. After independence, Govt. of India noted it seriously the rural banking for needy and poor rural people to fill the gap of micro-finance. As per the report of Banking Commission (1972) and Narsimham Committee Report suggested the establishment of RRBs in 1975 as the new rural banks which will give the regional feelings and solve the finance problems of needy and rural poor people in India. According to RRBs Act, 1976, RRBs provide the loan and advances to farmers, agricultural labourers, artisans and small entrepreneurs to develop the agriculture, trade, commerce, industry and other productive activities in rural areas. With joint shareholdings by Govt. of India, various State Govt. and sponsored banks, the steps were taken to integrate the banking activities within the policy thrust towards social banking. Their equity capital holding is in the proportion of 50:15:35. The growth in branch network of RRBs has facilitated to increase the banking activities in unbanked areas and mobilize the savings and investments for socio-economic development of needy and poor rural people.

40.2 Odisha Gramya Bank (OGB)

OGB is a RRB and has been set upon 7th January, 2013 by amalgamating of Neelachala Gramya Bank, Kalinga Gramya Bank and Baitarani Gramya Bank as per the provisions of RRBs Act, 1976. Indian Overseas Bank (IOB). OGB is jointly owned by Govt. of India, Govt. of Odisha and IOB. The shareholders of OGB are Govt. of India (50%), IOB (35%) and Govt. of Odisha (15%). OGB operates its business in 13 districts and head office located at Bhubaneswar. OGB has 9 regional offices and also 549 branches out of which 446 branches locating in remote rural area. OGB operates its business in 13 districts in by covering 37% geographical area and 52% total population in Odisha.

40.3 Statement of Problem

Credit is the main of a bank function and is also related with deposit. Higher amount of deposits by customers, higher will be the fund to deploy credit to customers. To get more deposit, RRBs need to bring rural people into banking fold. Similarly, deployment of credit depends upon availability of fund, financial performance and

recovery rate also. The deposit mobilization is an important indicator of bank performance. The present study has measured the socio-economic performance of OGB of Balasore and Bhadrak districts of Odisha. These two districts have many socio-economic aspects with regard to population size, male and female ratio, male and female literacy rate, children ratio, occupation status, etc. Thus, primary data has been collected for existing study. The present study would help to Governments, policymakers, regulators and personnel of OGB to know the ground reality in study area towards socio-economic development of needy and poor rural people through its redesigned credit policies of OGB.

40.3.1 Review of Literature

The present study has measured the socio-economic performance of OGB towards beneficiaries/customers Balasore and Bhadrak District of Odisha. Here, an attempt has been made to review the published research articles/reports in the field of socio-economic performance and other aspects of RRBs for the purpose of study. Hence, the research works done by researchers in past have been reviewed and presented in the following paragraph.

Dhaliwal [1] studied to know the problems faced by rural poor people dealing with the banks in the state of Punjab. The study observed that RRBs have been performing well in Punjab state as compared to all India level with regard to recovery performance, reduction of NPAs, enhancing the profitability and also productivity. The study recommended that RRBs should enhance the branch network and equip them with adequate infrastructural facilities and also to employ more staff to provide the better banking services to customers. Ibrahim [2] evaluated the performance of RRBs whether the amalgamation process has improved their financial performance or not. The study highlighted that RRBs have achieved the remarkable business performance in post-merger period as compared to pre-merger period. Ishwara [3] measured the financial performance of RRBs especially after amalgamation process of RRBs. The study concluded that net profit of RRBs has been increased by 200% and business performance by 100% after amalgamation process of RRBs. Mohanty et al. [4] examined to know the financial performance of RRBs. The study observed that rural banking has been transforming gradually with incorporation of technology, launching the customized banking product and providing the better customer service to have the better efficiency of RRBs. The study concluded that parameters like investment, credit and other income of respective sponsoring banks have the insignificant impact on profitability of rural sector banks. Reddy and Prasad [5] studied the financial performance of selected RRBs in post-reorganization period. Their study concluded that Andhra Pragati Grameen Bank has excelled over Sapthagiri Grameena Bank to protect the interest of creditors and quality of earnings. Subbarayudu and Reddy [6] studied the role of RRBs towards rural development. It was observed that beneficiaries face the problems like delay in sanctioning the loan amount, rigid terms and conditions of credit, lack of cooperation from bank personnel, insufficient technical

and legal facilities, etc. The study suggested that RRBs should focus more on the upliftment of the poor people of the society. Soni and Kapre [7] studied on current status of RRBs to know the financial performance. The study concluded that RRBs are the economic vehicles of disbursal of credit in rural area to provide credit to marginal farmers and socio-economically weaker section of society to develop the agriculture and allied sectors. Kemeswari [8] measured the provision of credit facility for rural people. The study highlighted that micro-finance is the main provision, credit and other financial services to needy and poor people who live in rural, semi-urban and urban areas to increase their income level as well as living standard of living. Gupta [9] studied the role of RRBs towards rural financing. His study highlighted that rural finance is provided to protect and also to safeguard the economic interest of rural people from informal source of finance. Rural people always face many problems i.e. the risk of unpredictable production, high dependency on monsoon, finance, seeds, fertilizers, water supply debts, etc. His study concluded that RRBs are the good medium of credit disbursement to agriculture and non-agriculture sectors in backward and rural areas in India. RRBs should not confine the business not only in agriculture sector but also provide the loan and advances to small entrepreneurs, village and cottage industries to boost the rural sectors. Murthy et al. (2012) studied on agricultural credit of RRBs. The objective of the study was to know the agricultural credit provided by RRBs to farmers. The study highlighted that RRBs perform a meagre role as compared to commercial banks for agriculture sector. The loan recovery mechanism of RRBs was also not satisfactory. The study suggested that Govt. of India should provide some schemes to enhance share of RRBs for agriculture credit through institutional agencies to avoid NPAs.

Sanap (2012) measured role and responsibilities of RRBs for rural development in India. The study concluded that RRBs should develop the agricultural and allied sectors by providing the larger amount of term loans to customers to meet their financial requirements. Without development of rural sectors, the objectives of economic planning will remain as dream not as reality. Hence, banks and other financial institutions have the important role for development of rural economy. Navi (2013) measured the impact of rural banking on farmers in Belgaum district. The study concluded that higher rate of interest on loan amount should be less on loan borrowing ability of farmers. Further, the high interest rate may affect the development and growth of farming and business activities in Belgaum district of Karnataka. Rafique and Manswani (2013) critically discussed the role of RRBs for economic development in India. The study concluded that rural economy must be developed to ensure a balanced economic growth in India. The problems faced by rural sectors must be identified, solved and addressed in priority basis by providing the adequate credit facilities to poor and needy people towards their economic upliftment. Kher [10] highlighted the role of rural banks for development of rural economy. The study concluded that RRBs have the tremendous achievements by providing the credit to agriculture and non-agriculture sectors for their development as well as to have the socio-economic development of beneficiaries. Bisoyi and Nilesh (2013) studied on Rushikulya Gramya Bank (RGB) which is one of the RRB sponsored by Andhra Bank. The objective of the study was to assess the financial performance of RGB

towards development of agriculture sector. The study concluded that performance of RGB is satisfactory towards development of agricultural sector. The RGB has provided the sufficient amount of credit for the development of agricultural sector. Jangale (2013) assessed the financial performance of 7 no. of RRBs in Maharashtra state. His study concluded that the financial performance of these RRBs is satisfactory in Maharashtra state. These RRBs play a major role to provide the banking services in remote areas in Maharashtra state for the socio-economic development of the beneficiaries.

Sharma (2013) made an empirical study on credit policies of RRBs for the sustainable development in state of Himachal Pradesh. The study concluded that majority of beneficiaries are satisfied with planning and implementation strategies taken by Himachal Grameena Bank. Swami et al. (2014) measured the trends of post-reform period in deposit mobilization and credit deployment of RRBs in Karnataka state. The study concluded that RRBs have achieved the wonderful progress during the post-reform period with regard to deposit mobilization from customer and credit disbursement to customers. Further, RRBs have also given the credit priority for personal loan to rural artisan, industries and other small-scale industries. Jindal et al. (2014) studied that RRBs are the inseparable part of the rural credit structure and rural economy also. The study concluded, policymakers and regulators of RRBs should take the corrective measures as RRBs are the banks of the poor and common people and are responsible for the development of rural economy. Devi (2014) identified the problems and prospects of RRBs. The study concluded that RRBs should not concentrate their business only in agriculture sector but should also provide the loans and advances to small entrepreneurs and cottage industries. RRBs should remove the bottlenecks and problems to have more business which lead towards good relationship with customers. To get more business, RRBs should provide the speedy and customized banking services to retain existing customers and to attract also the potential customers from rural areas.

Earlier, a few researchers have conducted the limited studies the role of RRBs towards socio-economic development of beneficiaries. But, no researchers have carried out in past on the topic of socio-economic aspects of beneficiaries and role played by OGB in these two sampling districts of Odisha. The present study would fill the following main gaps.

1. There is no study to know the socio-economic performance of OGB
2. There is the gap towards banking service and other aspects of OGB by taking the views of beneficiaries/customers.
3. There is also no study to know the main purpose of taking the credits of OGB by beneficiaries/customers.
4. There is also no study to know the impact of OGB credit to create the employment opportunity.
5. There is also no study to find out the main hindrances of OGB for the growth of micro-finance sectors
6. There is no study to know the impact of OGB loans and advances to develop the agriculture and allied sectors.

40.4 Objective of the Study

The present study has the following objectives.

- (i) To examine the role of OGB for the socio-economic development of beneficiaries i.e. farmers, owners of micro-business units, retailers, unemployed, the marginal workers, etc. of Balasore and Bhadrak districts of Odisha.
- (ii) To identify the main hindrances of OGB for the growth of micro-finance sectors.
- (iii) To suggest the recommendations for OGB to provide the finance to needy and poor rural people to meet their financial requirements.

40.5 Hypothesis of the Study

The followings are the setting hypotheses are the followings for the study.

1. **H-1:** The socio-economic developmental factors positively influence the amount of credit taken from OGB.
2. **H-2:** The underlying dimensions explain that the correlation among set of variables define the hindrances of OGB for growth of micro-finance sectors to provide credit to needy and poor rural people and views of the beneficiaries are correlated.
3. **H-3:** The underlying dimensions explain that the correlation among the set of variables define the recommendations for OGB in providing the finance to needy and poor rural people, and the views of beneficiaries are correlated.

40.6 Research Methodology Adopted

The primary data has been collected from 1200 beneficiaries/customers from both the districts on random stratified basis through a well-structured questionnaire. The secondary data has also been collected for the study. The collected primary data has been analysed and interpreted through SPSS-23 version to get the inference. Further, reliability and validity tests have also been done.

40.7 Analysis and Interpretation of Data

The following pages deal with major findings of the study with regard to demographic profile of beneficiaries and testing of hypothesis.

40.7.1 Demographic Profile of the Beneficiaries/Customers

58.33% beneficiaries are in the age group (25–40 years). 20% of beneficiaries belong to 40–50 years, 14.17% of beneficiaries are in age group more than 50 years. 972 no. of beneficiaries are in male category and 850 no. of beneficiaries are in married category. 59.16% of beneficiaries are in general category and followed by 20% of beneficiaries are in OBC category. 12.5% of beneficiaries are in SC and only 8.34% of beneficiaries are in ST category. Majority beneficiaries (53.33%) depend upon agriculture, and agriculture is the main profession to them, 20.83% of beneficiaries have the business as the occupation to them and 8.34% of beneficiaries ($n = 100$) are still unemployed. 59.16% of beneficiaries are of graduate level and 20% of beneficiaries are of post-graduate level. Still, 10.84% of beneficiaries are of illiterate and remaining beneficiaries are under metric. 61.66% of beneficiaries are living in rural area as compared semi-urban and urban location. 80.83% of beneficiaries have below 5 acre cultivated land and 9.17% of respondents have between 5 and 10 acres cultivated land. 76.66% of beneficiaries have less than 2 acres of irrigated land, whereas 13.34% of beneficiaries have less than 4 acres of irrigated land for agriculture purpose. Majority beneficiaries have the annual income between Rs. 100,000/- and 11.66% of beneficiaries have the annual income between Rs. 1 lakh and Rs. 3 lakhs. 9.16% of beneficiaries between Rs. 3 lakhs and Rs. 5 lakh per annum and remaining of beneficiaries ($n = 70$) have the annual income above Rs. 5 lakhs. Availing the agricultural facilities and getting the govt. subsidy by beneficiaries are ranked as 1st and 2nd, respectively, as the main purposes to open an account in RRBs as compared to other purposes. 816 no. of beneficiaries have taken the credit for various purposes from OGB to meet their financial requirements. Out of total 816 beneficiaries, 558 no. of beneficiaries have viewed that the loan amount is adequate to them, and 142 no. of beneficiaries have expressed that the loan amount is inadequate to them. Rest beneficiaries are with the views of otherwise. Out of total 816 respondents, 56.25% of beneficiaries have opined that the procedures of opening the loan account are simple, and 37.75% of beneficiaries have opined, the procedures of opening the loan account are not easy in OGB. 62.5% of beneficiaries pay the loan amount regularly, whereas rest beneficiaries do not pay the loan amount regularly due to some problems i.e. failure of crops, poor financial condition, drought, flood, etc.

40.8 Views of Beneficiaries on OGB Credit for Socio-economic Development

The socio-economic development of needy and poor rural people is very important so far as economy and sustainable development aspects are concerned in India. OGB play in very vital role towards socio-economic development of needy and poor rural people in Odisha. The beneficiaries have given their views on a five-point rating scale

Table 40.1 Views on OGB credit for socio-economic development ($n = 816$)

Statements		Rating						
		5	4	3	2	1	Rank	Total
A	Helpful in increasing your income level	1825	624	390	220	55	3114	3
B	Helpful in generating employment	950	840	333	288	161	2572	7
C	Helpful in meeting family obligation	1885	656	342	196	63	3142	2
D	Helpful in eliminating the poverty	1950	588	360	180	69	3147	1
E	Helpful in improving your life style	1600	720	480	140	86	3026	5
F	Helpful in increasing your social status	1775	708	369	178	72	3102	4
G	Helpful for your children's higher education	1595	748	333	210	94	2980	6
H	Helpful for OGB to develop the priority sector and SSI	725	880	411	354	137	2507	8

Rating scale 5 to 1: 5—strongly agree, 4—agree, 3—undecided, 2—disagree and 1—strongly disagree

Source Primary data

on OGB credit towards socio-economic development. The data has been collected, tabulated, analysed, ranked and presented the same in Table 40.1.

Table 40.1 measures the view of beneficiaries towards OGB credit for socio-economic development by taking into consideration of 6 statements. It has been noticed from Table 40.1 that OGB credit helps in eliminating the poverty, meeting family obligation and increasing the income level of beneficiaries for socio-economic development of needy and rural poor people which are ranked as 1st, 2nd and 3rd, respectively. Further, it has also been observed that enhancement of social status; better life style and higher education of their children are other possible for socio-economic development of beneficiaries by OGB credit. It has been concluded that socio-economic development of needy and poor rural people can be possible by OGB credit. No doubt, Government of India is doing a lot through OGB credit for socio-economic development of needy and rural poor people. Still huge scope is there, jointly Govt. of India and concerned State Govt. may march ahead through

more social welfare schemes with beating pace for socio-economic development of needy and rural poor people through OGB in rural Odisha.

40.9 Views of Beneficiaries on Main Hindrances of OGB for the Growth of Micro-finance Sectors

The contribution of OGB is praiseworthy towards rural development. OGB faces many problems and challenges so far as their growth and development are concerned. The main hindrances of OGB are herewith mentioned in the Table 40.2 for the effective management of the problems. Some statements are given below relating to growth of OGB. Similarly, the beneficiaries have been appealed to give their views on a five-point rating scale on this aspect. The data so collected from them are analysed and presented the same in Table 40.2.

OGB, in many cases, has the mixed record of successes and failures in its banking business and achievement of goals. Table 40.2 shows the views of beneficiaries on main hindrances for the growth of OGB by taking into considerations of 8 statements. Lack of trained and skilled OGB personnel, lack of awareness programme on products of OGB among customers to avail the same and friendly behaviour of RRBs personnel with customers are found the main hindrances of OGB for the growth of micro-finance sectors and are also ranked as 1st, 2nd and 3rd, respectively, as compared to other hindrances. Further, it is also observed that the helpful attitude of OGB personnel, timely and quick service and easy regulations in applying loan/credit are the other hindrances of OGB for the growth of micro-finance sectors. It is concluded that lack of trained and skilled of OGB personnel, lack of awareness programme on banking products of OGB and the friendly behaviour of OGB personnel with costumers are the main hindrance of OGB for the growth of micro-finance sectors. It may be suggested that regulators and policymakers of OGB should provide the training programme in regular basis to RRBs personnel in order to provide the good customized services to customers in due time as well as to handle the quarries of customers.

40.10 Views of Beneficiaries on Main Recommendations to OGB in Providing Finance to Needy and Poor Rural People

As we know, needy and poor rural people are the prime stakeholders of OGB. OGB is basically a rural bank for rural people in Odisha. OGB usually provides the finance in the form of loan to meet the financial requirements of needy and poor rural people to boost the rural sectors. Through the suitable credit policy of OGB, the informal credit/loan can be checked for the greater economic interest of needy and poor rural people. Some statements are given below relating to few recommendations to OGB

Table 40.2 Views of beneficiaries on main hindrances of OGB for the growth of micro-finance sectors ($n = 1200$)

Statements		Rating						
		5	4	3	2	1	Total	Rank
A	Implementation of easy regulations in applying loan/credit	2295	1400	465	234	119	4513	6
B	There should be the awareness programme by OGB personnel about the loan/credit and its benefits	2275	1480	495	222	99	4571	2
C	Timely and quick service	2205	1456	555	210	105	4531	5
D	OGB personnel should be helpful to uneducated loans in filling up the forms	2180	1524	531	218	97	4550	4
E	OGB personnel should be trained and skilled enough to handle all types of queries of customers	2105	1588	582	240	68	4583	1
F	OGB personnel should have the friendly behaviour with customers	2185	1644	381	232	109	4551	3
G	There should be the easy procedure of sanctioning loan/credit amount	2055	1636	327	230	156	4404	8
H	There should be the easy instalment provision on loan/credit amount	2060	1632	363	234	142	4431	7

Rank from 5 to 1: 5—most important, 4—important, 3—undecided, 2—less important, 1—least important

Source Primary data

in providing the finance to needy and poor rural people. Similarly, the beneficiaries have been appealed to give their views on a five-point rating scale on this aspect. The data, thus, so collected from them are analysed and presented the same in Table 40.3.

OGB is the growth engine and growth partner of rural economy in Odisha. OGB is the rural economy builder towards socio-economic development of needy and poor

Table 40.3 Opinion of beneficiaries on main recommendations to RRBs in providing finance to needy and poor rural people ($n = 1200$)

Statements		Rating						
		5	4	3	2	1	Total	Rank
A	Skilful and trained OGB personnel enough to solve the problems of customers	2205	1240	837	222	59	4563	6
B	Conducting the awareness programme by OGB personnel about loan/credit and its benefits	2185	1232	792	204	89	4502	9
C	Prompt and quick service	2105	1416	771	202	67	4561	7
D	Sanctioning the loan amount within a short period of time	2435	1236	792	136	72	4671	1
E	The fair regulations for availing loan/credit	2340	1300	672	228	69	4609	2
F	Friendly and empathy attitude of OGB personnel towards customers	2295	1324	681	218	74	4592	3
G	Easy procedure of sanctioning loan amount	2195	1308	672	210	105	4490	10
H	Easy instalments on loan/credit amount	2220	1336	684	240	74	4554	8
I	Subsidized rate of interest on loan/credit amount	2245	1316	672	312	42	4587	4
J	Helpful attitude of OGB personnel to uneducated loans in filling up the loan application form	2060	1424	684	234	87	4489	11

(continued)

Table 40.3 (continued)

Statements		Rating						
		5	4	3	2	1	Total	Rank
K	Prompt action against customer complain	2045	1472	774	232	49	4572	5

Rating scale 5 to 1: 5—very important, 4—important, 3—undecided, 2—less important and 1—least important

Source Primary data

rural people in India. Table 40.3 shows the views of beneficiaries on few recommendations to OGB to provide the finance to needy and poor rural people by taking into considerations of 11 statements. Sanctioning the loan amount in time, fair regulations to avail the credit, friendly and empathy attitude of OGB personnel and subsidized rate of interest on loan amount are found the few recommendations of beneficiaries to OGB to provide the finance to customers and also are ranked as 1st, 2nd, 3rd and 4th, respectively, as compared to other recommendations of beneficiaries. Further, it is also observed that prompt action against complains, skilful and trained OGB personnel enough to solve the problems of costumers, prompt and quick services and easy instalments on loan amount are also other recommendations for the growth of OGB as well as for the socio-economic development of needy and poor rural people through the suitable credit mechanism. It is concluded that sanctioning the loan amount in time, fair regulations to avail the credit, friendly and empathy attitude of OGB personnel and subsidized rate of interest on loan amount are the major recommendations of beneficiaries to have the win-win strategy not only to OGB but also to needy and poor rural people in Odisha. It may be suggested that RRBs should sanction the loan amount to customers in a shorter period and should also adopt the easy and fair regulation with regard to sanctioning the loan amount to poor and needy people. So that the needy and poor rural people will get the economic and financial fruits from OGB.

40.10.1 Testing of Hypothesis

As per the setting Objective-1 and Hypothesis-1, the analysis and interrelation are given below.

The multiple correlation coefficient $R = 0.633$ shows the strong correlation between depended variables and variables predicted through regression model. The adjusted R -square value explains, this model shows 39.4% of variance (Table 40.4).

Table 40.5 shows the result of F -test for hypothesis that no predictor variable is associated on amount of loan taken from OGB. Hence, null hypothesis has been rejected ($F = 67.129$, $P < 0.05$). It has been observed that at least one of the independent variables is associated with dependent variable.

Table 40.4 Model summary

Model	<i>R</i>	<i>R</i> -square	Adjusted <i>R</i> -square	Std. error of estimate
1	0.633 ^a	0.400	0.394	1.37898

^aPredictors: (Constant), VAR8, VAR7, VAR6, VAR5, VAR4, VAR3, VAR2, VAR1

Source SPSS output

Table 40.5 ANOVA

Model		Sum of squares	Df	Mean square	<i>F</i>	Sig.
1	Regression	1021.208	8	127.651	67.129	0.000 ^a
	Residual	1528.872	804	1.902		
	Total	2550.080	812			

^aPredictors: (Constant), VAR8, VAR7, VAR6, VAR5, VAR4, VAR3, VAR2, VAR1

^bDependent variable: VAR8

Source SPSS output

Table 40.6 indicates the impact of two variables such as '*Helpful for your children's higher education*' ($B = 0.514, t = 11.815, p < 0.05$), '*RRBs help to develop the priority sector and SSI*' ($B = 0.502, t = 11.950, p < 0.05$), are statistically significant with the dependent variable (amount of loan taken from OGB) at 0.05 significant levels. Therefore, '*Helpful for your children's higher education*' and '*OGB helps to develop the priority sector and SSIs*' has the greater positive influence on the dependent variable and all other variables have insignificant influence on the dependent variable.

As per the setting Objective-2 and Hypothesis-2, the analysis and interrelation are given below.

The results of FA by using principal component method are given in Table 40.9. Hence, FA has been considered to be an ideal tool for correlation matrix. The principal component method has been used to calculate the minimum number of factors which will show the maximum variance.

1. Eigen values more than 1.00 results in 4 no. of factors are extracted. The factors along with Eigen values more than 1.00 are included and remaining factors are excluded.
2. By comparing the varimax rotated factor matrix with unrotated factor matrix has been observed as component matrix. The rotation has been seen the simplicity and also has been found more interpretability. 4 no. factors are also extracted in rotation factor matrix.

Following listed variables are included in analysis. The analysis has undertaken the following variables and shown in Table 40.7.

KMO measures of sampling adequacy is 0.667 which signifies the accuracy of FA. 1st item from output of communalities given in Table 40.10 which has shown how much of variance in variables shown in extracted factors (Tables 40.8 and 40.9).

Table 40.6 Coefficients

Model		Unstandardized coefficient		Standardized coefficients Beta	T	Sig.
		B	Std. error			
1	(Constant)	-1.078	0.422		-2.556	0.011
	Helpful in increasing your income level	0.004	0.045	0.002	0.079	0.937
	Helpful in generating employment	-0.025	0.086	-0.009	-0.295	0.768
	Helpful in meeting family obligation	-0.032	0.063	-0.018	-0.516	0.606
	Helpful in eliminating the poverty	0.023	0.057	0.013	0.399	0.690
	Helpful in improving your life style	-0.001	0.064	0.000	-0.017	0.987
	Helpful in increasing your social status	-0.018	0.058	-0.009	-0.311	0.756
	Helpful for your children's higher education	0.514	0.043	0.367	11.815	0.000
	Helpful for OGB to develop the priority sector and SSI	0.502	0.042	0.372	11.950	0.000

^aDependent variable: VAR00008

Source SPSS output

40.11 Principal Component Analysis

Table 40.10 presents the whole factors extractable and Eigen values from analysis. The percentage of variance attribute is with each factor, cumulative variance of factors and previous factors (Table 40.11).

Table 40.10 shows that 1st factor shows 19.952% of variance, 2nd one is 18.597% and so on.

Table 40.7 Rotated factor matrix

V1	Skilful and trained OGB personnel enough to solve the problems of customers
V2	Conducting the awareness programme by OGB personnel about loan/credit and its benefits
V3	Prompt and quick services
V4	Sanctioning the loan amount within a short period of time
V5	The fair regulations for availing loan/credit
V6	Friendly and empathy attitude of OGB personnel towards customer
V7	Easy procedure of sanctioning loan amount
V8	Easy instalments on loan/credit amount
V9	Subsidized rate of interest on loan/credit amount
V10	Helpful attitude of OGB personnel to uneducated loans in filling up the loan application form
V11	Prompt action against complains

Source SPSS output

Table 40.8 KMO and Bartlett's test

KMO measure of sampling adequacy	0.667	
Bartlett's test of sphericity	Approx. Chi-square	1833.941
	Df	55
	Sig.	0.000

Source SPSS output

Table 40.9 Communities

Variables	Initial	Extraction
VAR1	1.000	0.489
VAR2	1.000	0.747
VAR3	1.000	0.722
VAR4	1.000	0.564
VAR5	1.000	0.641
VAR6	1.000	0.521
VAR7	1.000	0.734
VAR8	1.000	0.522
VAR9	1.000	0.638
VAR10	1.000	0.544
VAR11	1.000	0.639

Extraction method Principal component analysis

Source SPSS output

Table 40.10 Total variance explained

Component	Initial Eigen values			Extraction sums of squared loading			Rotation sums of squared loading		
	Total	% of variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.195	19.952	19.952	2.195	19.952	19.952	2.148	19.524	19.524
2	2.046	18.597	38.549	2.046	18.597	38.549	1.840	16.723	36.247
3	1.181	10.737	49.286	1.181	10.737	49.286	1.248	11.349	47.596
4	1.018	9.253	58.540	1.018	9.253	58.540	1.204	10.943	58.540
5	0.896	8.142	66.681						
6	0.770	7.004	73.685						
7	0.716	6.513	80.198						
8	0.692	6.289	86.487						
9	0.592	5.379	91.865						
10	0.501	4.559	96.424						
11	0.393	3.576	100.000						

Extraction method Principal component analysis

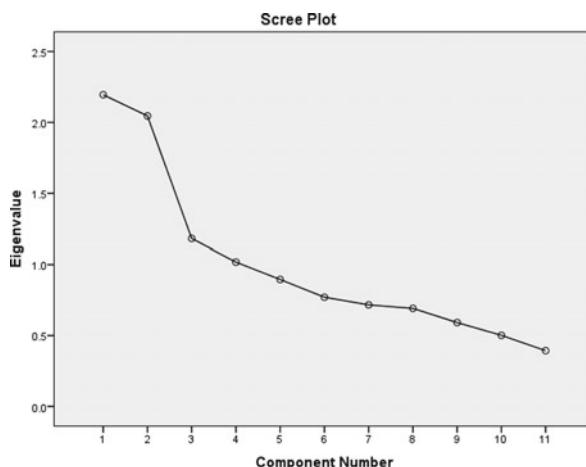
Source SPSS output

Table 40.11 Component matrix

Variables	Component			
	1	2	3	4
VAR1	0.371	-0.213	0.553	-0.022
VAR2	0.021	-0.214	0.821	0.161
VAR3	0.773	-0.132	-0.323	-0.048
VAR4	0.728	-0.147	0.065	-0.092
VAR5	0.781	-0.102	-0.116	-0.082
VAR6	0.485	0.012	0.039	0.251
VAR7	0.153	0.354	-0.004	0.765
VAR8	0.123	0.659	-0.031	0.269
VAR9	0.118	0.719	0.161	-0.285
VAR10	0.068	0.662	0.021	0.022
VAR11	0.162	0.626	0.223	-0.414

Extraction method Principal component analysis

^a4 components extracted



Scree plot also explored four factors

40.12 Rotated Component Matrix

The aim of rotation is to minimize the no. of factors on which the variables under investigation have high loadings. Rotation makes the interrelation in analysis (Table 40.12).

Table 40.12 Rotated component matrix

Variables	Component			
	1	2	3	4
VAR1	0.287	-0.001	0.634	-0.069
VAR2	-0.116	-0.069	0.852	0.045
VAR3	0.833	-0.051	-0.160	0.012
VAR4	0.721	0.032	0.203	-0.049
VAR5	0.799	0.035	0.030	-0.011
VAR6	0.440	-0.010	0.537	0.295
VAR7	0.036	0.023	0.018	0.855
VAR8	0.004	0.480	-0.129	0.525
VAR9	-0.005	0.797	-0.014	0.042
VAR10	-0.041	0.585	-0.113	0.295
VAR11	0.050	0.788	0.060	-0.108

Extraction method Principal component analysis

Rotation method Varimax with Kaiser Normalization

^aRotation converged in 5 iterations

FA has established the relationship between the variables and subsequently the acceptance of hypothesis and explored four important factors such as '*Skillful and trained OGB personnel enough to solve the problems of customers.*', '*Conducting the awareness programme by OGB personnel about loan/credit and its benefits*', '*Prompt and quick services*' and '*Sanctioning the loan amount within a short period of time*'. 1st, 2nd and 3rd factors have three loadings, and 4th factor has only two loadings. 1st factor and 2nd factor are accounted for 19.952% 18.597% of variance, respectively. Further, 3rd factor and 4th factor are accounted for 10.737% and 9.253% of variance, respectively (Table 40.13).

As per the setting Objective-3 and Hypothesis-3, the analysis and interrelation are given below.

Table 40.13 Extracted factors

Factor	Factor interpretation	Variables included in factor
F1	Skilful and trained OGB personnel enough to solve the customers' problems	X3, X4, X5
F2	Conducting the awareness programme by OGB personnel about loan/credit and its benefits	X9, X10, X11
F3	Prompt and quick services	X1, X2, X6
F4	Sanctioning the loan amount within a short period of time	X7, X8

Source SPSS output

The analysis has been undertaken the following variables and shown the same in given below in Table 40.14.

1st item from output of communalities given below highlights that how much of variance in variables has been observed in extracted factors (Tables 40.15 and 40.16).

Table 40.14 Rotated factor matrix

V1	Implementation of easy regulations in applying loan/credit
V2	There should be awareness programme by OGB personnel about the loan/credit and its benefits
V3	Timely and quick services
V4	OGB personnel should help to the uneducated loans in filling up the forms
V5	OGB personnel should be trained and skilled enough to handle all types of queries of customer
V6	OGBs personnel should have the friendly behaviour with customers
V7	There should be easy procedure of sanctioning loan/credit
V8	There should be the easy instalment provision on loan/credit amount

Table 40.15 KMO and Bartlett's test

KMO measure of sampling Adequacy	0.769	
Bartlett's test of sphericity	Approx. Chi-square	788.922
	Df	28
	Sig	0.000

KMO measure of sampling adequacy—0.769 which signifies the accuracy of FA

Source SPSS output

Table 40.16 Communalites

Variables	Initial	Extraction
VAR1	1.000	0.550
VAR2	1.000	0.546
VAR3	1.000	0.630
VAR4	1.000	0.441
VAR5	1.000	0.633
VAR6	1.000	0.404
VAR7	1.000	0.607
VAR8	1.000	0.505

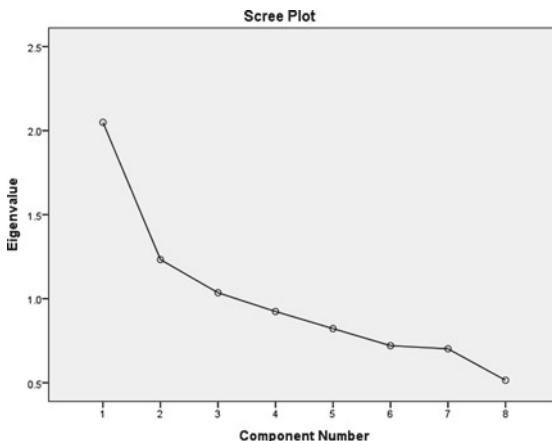
Extraction method Principal component analysis

Source SPSS output

40.13 Principal Component Analysis

Table 40.17 given below presents the factors extractable and Eigen value from analysis. The percentages of variance attributable are found to each factor, cumulative variance of factors as well as previous factors.

It has been observed from Table 40.17 that 1st factor is found 25.614% of variance, 2nd factor is also found for 15.407% and so on (Tables 40.18, 40.19, 40.20 and 40.21).



Scree plot also explored three factors.

FA has established the relationship between the variables and subsequently the acceptance of hypothesis and explored three important factors such as '*Implementation of easy regulations in applying loan/credit*', '*There should be awareness programme by OGB personnel about the loan/credit and its benefits*' and '*Timely and quick services*'. 1st factor has four loadings. But, 2nd and 3rd factors have two loadings. 1st factor is found for 25.614% of variance. 2nd and 3rd factor are found for 15.407% and 12.943% of variance, respectively.

40.14 Concluding Remarks

To conclude, rapid growth of RRBs has helped a lot to minimize the regional disparities towards banking services in Odisha. The steps taken for branch extension, deposit mobilization, rural development and sanction of credit by OGB in rural are noticeable and praiseworthy. OGB has successfully achieved its objective like to provide the services to door steps of rural households in banking deprived rural area, to give credit at low rate to weaker rural area, to encourage rural saving and investment, to generate more employment in rural area and to also reduce the credit cost in rural area. OGB plays in very important role as an economic vehicle of credit

Table 40.17 Total variance explained

Component	Initial Eigenvalue			Extraction sums of squared loading			Rotation sums of squared Loading		
	Total	% of variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of variance	Cumulative %
1	2.049	25.614	25.614	2.049	25.614	25.614	1.879	23.486	23.486
2	1.233	15.407	41.021	1.233	15.407	41.021	1.253	15.667	39.152
3	1.035	12.943	53.964	1.035	12.943	53.964	1.185	14.812	53.964
4	0.924	11.547	65.512						
5	0.822	10.275	75.786						
6	0.721	9.007	84.793						
7	0.702	8.774	93.566						
8	0.515	6.434	100.000						

Extraction method Principal component analysis

Source SPSS output

Table 40.18 Component matrix

Variables	Component		
	1	2	3
VAR1	-0.136	0.710	-0.166
VAR2	-0.203	0.693	-0.156
VAR3	0.729	0.074	-0.306
VAR04	0.664	-0.029	0.006
VAR5	0.645	0.126	-0.449
VAR6	0.104	0.449	0.438
VAR7	0.377	0.155	0.664
VAR8	0.670	0.024	0.237

Extraction method Principal component analysis

^a3 Components extracted

Source SPSS output

Table 40.19 Rotated component matrix

Variables	Component		
	1	2	3
VAR1	0.004	0.741	0.029
VAR2	-0.062	0.736	0.010
VAR3	0.794	0.000	0.015
VAR4	0.600	-0.163	0.234
VAR5	0.779	0.104	-0.124
VAR6	-0.042	0.291	0.564
VAR7	0.090	-0.101	0.767
VAR8	0.517	-0.175	0.455

Extraction method Principal component analysis

Rotation method Varimax with Kaiser normalization

^aRotation converged in 5 iterations

Table 40.20 Extracted factors

Factor	Factors interpretation	Variables included in factor
F1	Implementation of easy regulations in applying loan/credit	X3, X4, X5, X8
F2	There should be awareness programme by OGB personnel about the loan/credit and its benefits	X1, X2
F3	Timely and quick services	X6, X7

Source SPSS output

Table 40.21 Results of testing of hypothesis

S. No.	Objective wise results as per setting hypothesis
1	<p><i>Objective-1:</i> To examine the role of OGB for the socio-economic development of beneficiaries i.e. farmers, owners of micro-business units, retailers, unemployed, the marginal workers, etc. of Balasore and Bhadrak districts of Odisha</p> <p><i>Hypothesis-1:</i> The socio-economic developmental factors positively influence the amount of credit taken from OGB</p> <p>Statistical Test: Multiple regression analysis</p> <p>Decision/Result: Accepted for ‘Development of Education’ and ‘OGB help to develop the priority sector and SSI’ have greater positive influence on the dependent variable</p>
2	<p><i>Objective-2:</i> To identify the main hindrances of OGB for the growth of micro-finance sectors</p> <p><i>Hypothesis-2:</i> The underlying dimensions explain that the correlation among set of variables define the hindrances of OGB for growth of micro-finance sectors to provide credit to needy and poor rural people and views of the beneficiaries are correlated</p> <p>Statistical test: Factor analysis</p> <p>Decision/Result: Accepted</p> <p>FA has established the relationship between the variables and subsequently the acceptance of hypothesis and explored four important factors such as ‘Skillful and trained OGB personnel enough to solve the problems of customers.’, ‘Conducting the awareness programme by OGB personnel about loan/credit and its benefits.’, ‘Prompt and quick services’ and ‘Subsidized interest rate with easy instalment’</p>
3	<p><i>Hypothesis-3:</i> The underlying dimensions explain that the correlation among the set of variables define the recommendations for OGB in providing the finance to needy and poor rural people and the views of beneficiaries are correlated</p> <p><i>Objective-3:</i> To suggest the recommendations for OGB to provide the finance, the needy and poor rural people to meet their financial requirements</p> <p>Statistical test: Factor analysis</p> <p>Decision/Result: Accepted</p> <p>FA has established the relationship between the variables and subsequently the acceptance of hypothesis and explored three important factors—‘Implementation of easy regulations in applying loan/credit’, ‘There should be awareness programme by OGB personnel about the loan/credit and its benefits.’ And ‘Timely and quick services’</p>

delivery in rural areas with objective of credit dispersion among marginal farmers and socio-economically weaker section of people to develop of agriculture and allied sectors. OGB has to remove these hindrances to have the better economic viability. OGB should provide the speedy, qualitative and superior services not only to retain the existing customers but also to attract potential customers. Further, Governments should take the adequate steps for provision of adequate irrigation facility to farmers to enhance the agricultural production. The agricultural production of farmers will have the positive impact on their socio-economic development as well as boosting the Indian economy. Governments should take the bold action against defaulters and should not announce the waiving of loans. OGB should give the due preference into micro-credit scheme and motivate to SHGs and other non-priority sectors. Socio-economic development of needy and poor rural people can be possible by OGB credit. No doubt, Govt. of India and State Govts. are doing a lot through OGB credit for socio-economic development of needy and rural poor people. Still huge scope

is there, Govt. of India and Govt. of Odisha may march ahead through more social welfare schemes with beating pace for socio-economic development of needy and rural poor people through OGB in Odisha.

References

1. Dhaliwal, M.L.: Regional rural banks—a clarification. *Econ. Polit. View Wkly* **XIII**, 23–34 (2010)
2. Ibrahim, M.S.: Performance evaluation of RRBs in India. *Int. Bus. Res.* **3**(4), 203–211 (2010)
3. Ishwara, B.: Impact of Grameena Bank finance on non-farm sector: an empirical study. *Fin. Agricult.* **3**(1), 77–82 (2011)
4. Mohanti, B.: Regional rural banks': economic and political weekly, Feb 12, pp. 45–56 (2011)
5. Reddy, B., Prasad, M.: Rural Banking: misdirected policy changes. *Econ. Polit. Wkly* **XXVIII**(48), 62–78 (2011)
6. Subbarayudu, D.V., Reddy, R.K.: Regional Rural Banks improve performance in rural sector. *J. Indian Inst. Bankers* 76–86 (2011)
7. Soni, A.K., Kapre, A.: A study on current status of RRBs in India. *Natl. Monthly Refereed J. Res. Commerce Manag.* **2**(2), 1–16 (2011)
8. Kemeswari, R.G.: Impact of Grameena Bank finance on non-farm sector: an empirical study. *Fin. Agricult.* **3**(1), 64–78 (2011)
9. Gupta, T.C.: Institutional credit for dairying in Haryana State. *Indian J. Agricult. Econ.* **43**(3), 45–56 (2011)
10. Kher, M.: Performance of a Regional Rural Bank in Uttar Bhaiyan: a study on Keonjhar District. *Indian J. Agricult. Econ.* **47**(3), 424–425 (2003)
11. Annual Report of OGP for the financial year 2018 and 2019
12. Ahmed, J.U.: The efficacy and working of RRBs: An implication in Indian context. *Int. J. Bank. Risk Insur.* **2**(1), 18–29 (2014)
13. Ahmed, J.U.: Performance evaluation of RRBs: Evidence from Indian Rural Banks. *Global J. Manag. Bus. Res. Fin.* **13**(10), 67–76 (2013)
14. Ahmed, J.U., Bhandari, G.P.: Profitability of Regional Rural Banks in India: an empirical analysis in Meghalaya Rural Bank. *J. Rural Ind. Dev.* **1**(1), 1–15 (2013)
15. Ahmed J.U.: Productivity analysis of Rural Banks in India: a case of Meghalaya Rural Bank. *NEHU J.* **XII**(1), 53–76 (2014)
16. Allen, L., Rai, A.: *Operational efficiency in banking: an international comparison*. *J. Bank. Fin.* **20**, 655–672 (1996)
17. Almazar, A.A.: Capital adequacy, cost income ratio and the performance of Saudi Banks. *Int. J. Acad. Res. Account. Fin. Manag. Sci.* **3**(4), 284–293 (2013)
18. Avkiran, N.K.: An application reference for data envelopment analysis: helping the novice researcher. *Int. J. Bank Mark.* **17**(5), 206–220 (1999)
19. Banker, R.D., Charnes, A., Cooper, W.W.: Some models for estimating technical and scale inefficiencies in DEA. *Manag. Sci.* **30**(9), 1078–1082 (1984)
20. Bhandarnayake, S., Ashtha, J.P.: Factors influencing the efficiency of commercial banks in Sri Lanka. *Sri Lankan J. Manag.* **18**(1, 2), 21–50 (2013)
21. Chavan, P., Pallavi, R.: Regional rural banking: challenges ahead. *Indian J. Agricult. Econ.* **18**(3), 21–33 (2004)
22. Geetha, R.S.: Performance evaluation of RRBs with reference to Krishna PragathiGramin Bank, Shimogga District. *IOSR J. Bus. Manag.* **18**(I), 42–56 (2016)
23. Hussain, S.: The assessment of operational efficiency of Commercial Banks in India using cost to income ratio approach. *Int. J. Manag. Bus. Res.* **4**(3), 225–234 (2014)
24. Hadi, A., Bagchi, K.K.: *Performance of RRBs in West Bengal: an evaluation*. Serials Publications, New Delhi

25. Ibrahim, U.R.: Performances and prospects of RRBs', Banking Finance. Indian J. Agricult. Econ. 43(3), 34–41 (2010)
26. Government of India: Report of the Working Group on Rural Banks, New Delhi (1975)
27. Government of India: RRBs Act, Ministry of Law, Justice and Company Affairs, New Delhi (1976)
28. Government of India (July 2001): Chalpathy Rao Committee on RRBs Act Amendments. Ministry of Finance, New Delhi
29. Kalkundrikar, A.B.: Regional Rural Banks and Economic Development. Daya Publishing House, New Delhi (1990)
30. Kaye, L., Tasi, R.: Functioning of RRBs: an overview. Indian J. Agricult. Econ. **39**(10), 101–112 (2006)
31. Kaye, T.: Role of RRBs in economic development. Abhijeet Publications, Delhi (2006)
32. NABARD: Review of the performance of RRBs, as on 31st March 2008, 2009 and 2010, Mumbai
33. Mishra, M.: Grameen Bank of Odisha: a model for financing rural development. Int. J. Dev. Bank. **13**, 15–27 (2006)
34. Patel, T., Shah, N.: A study on performance evaluation of RRBs of India. Pezzottaite J. **5**(2), 128–134 (2016)
35. Raees, S., Pathak, S.: Financial inclusion through RRBs—dream or reality. Chronicle Neville Wadia Inst. Manag. Stud. Res. (2014). ISSN: 2230-9667, pp. 208–213
36. Reserve Bank of India: Report of Kelkar Committee on RRBs. Bombay (1986)
37. Reserve Bank of India: Report of the Committee on Restructuring of RRBs, Bombay (1994)
38. Soni, A.K., Kapre, A.: A study on current status of RRBs in India. Natl. Monthly Refereed J. Res. Commerce Manag. **2**(2), 1–16 (2011)
39. Suresh, R.: Regional rural banks and their performance—an empirical study. J. Rural. Dev. **34**(3), 285–303 (2015)
40. Thakur, P., Gupta, A.: Regional Rural Banks: a way to become a developed economy. In: Proceedings of the International Symposium on Emerging Trends in Social Science Research (2015). ISBN: 978-1-941505-23-6, pp 1–13
41. Shettar, M.: Analysis of socio-economic impact of advances on beneficiaries of Union Bank of India. IOSR J. Econ. Finance **3**(3), 43–48 (2014)
42. Singh, A.K.: Role of RRBs in the promotion of self-help groups in India (an analytical study). Int. J. Sci. Res. Publ. **5**(9), 1–7 (2015)
43. Subbarayudu, D.V., Reddy, R.K.: RRBs improve performance in rural sector. J. Indian Inst. Bankers **43**, 76–86 (2011)

Chapter 41

Performance of Machine Learning Techniques Before and After COVID-19 on Indian Foreign Exchange Rate



Trilok Nath Pandey, Rashmi Ranjan Mahakud, Bichitrana Patra,
Parimal Kumar Giri, and Satchidananda Dehuri

Abstract The FOREX marketplace has seen sudden boom during last couple of decades. The changes carry out a critical function in balancing the dynamics of the marketplace. As a result, the correct prediction of the change price is a critical aspect for the fulfillment of many companies and fund managers. In-spite of the reality, the marketplace is famous for its flightiness and volatility; there exists groups like agencies, banks, and pandemic for awaiting change by several techniques. The goal of this article is to locate and advocate a neural community version to forecast exchange rate to the United States dollar against Indian rupees. In this article, we have analyzed the performance of different machine learning techniques during COVID-19 pandemic situation. This is further extended to find the best model to our purpose. In this paper, we implemented three different types of techniques to predict the foreign exchange rate of US dollar against the Indian rupees with high accuracy rate, before and after the COVID-19 pandemic. The three types of neural network models implemented in this article are artificial neural network (ANN), long short-term memory network (LSTM), and gated recurring units (GRU). The results from the above three models are compared so as to find out which model performs the best as compared to other models, before and after the COVID-19 pandemic. From the empirical analysis of all the models, we concluded that GRU outperformed both ANN and LSTM. We have five sections in this article. Section 41.1 briefly describes

T. N. Pandey (✉) · R. R. Mahakud · B. Patra

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India
e-mail: trilokpandey@soa.ac.in

R. R. Mahakud

e-mail: rashmiranjanmahakud@soa.ac.in

B. Patra

e-mail: bichitranaapatra@soa.ac.in

P. K. Giri

Department of Computer Science, College of Engineering Bhubaneswar, Bhubaneswar, Odisha, India

S. Dehuri

Department of Information and Communication, Fakir Mohan University, Balasore, Odisha, India

about prediction of foreign exchange rate. In Sect. 41.2, we have discussed the methods used in this article for the prediction of foreign exchange rate. Data collection and experimental results have been discussed in Sects. 41.3 and 41.4. Finally, in Sect. 41.5, we have given the conclusion and future scope of this experimental article.

41.1 Introduction

The market place has seen tremendous growth since last few years. Exchange rate has an important role in the exchange market dynamics. So, accurate forecasting of the exchange rate is required to be successful in business investments [1–5]. To some extent, exchange rates are predictable now-a-days. The major input in the forecasting process is the historical data. The conventional model or technique used to forecast the exchange rates is the time series modeling. But, the problem with this technique is the series of exchange rate is noisy and non-stationary. Forecasting of exchange rates is one of the most demanding applications of modern time. Exchange rates are unpredictably noisy, chaotic, and non-stationary [6–8]. These attributes hint toward the incomplete information.

Moreover, this information can be taken from the credentials of such markets to apprehend its holding between coming rates and that in the past. The query is how good are these forecasts? The intent of this paper is to probe and contrast the prediction of USD against the Indian rupee (in Indian exchange rate). The aim of this research is to predict and propose a neural network model to forecast exchange rate of the US dollar against Indian rupees before and after the pandemic COVID-19. This is further extended to find the best model to fit our purpose. In our analysis, we have used foreign exchange rate of twenty-three economically sound countries to predict the exchange rate of Indian rupees against US dollar. The three different models used in this research to predict the exchange rate are described in the next section.

41.1.1 Methods

There are numerous distinct neural network algorithms base in the studies. But, till now, there is no study found to methodically regulate the rationalize performance of each algorithm. Through this paper, we implemented three different neural network models [9–14], namely Artificial Neural Network (ANN), Long Short Term Memory Network (LSTM) and Gated Recurring Units (GRU), inorder to analyse which model forecasts the rate of exchange of Indian Rupee more accurately and how their prediction capability is affected before and after the pandemic COVID-19. Backing, we represent these three algorithms in a nutshell.

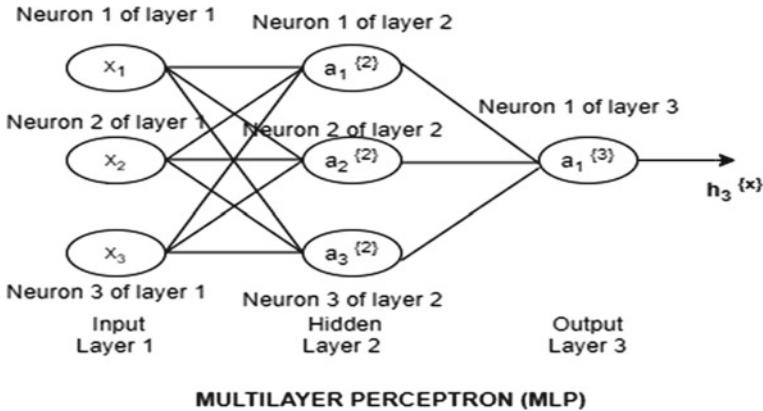


Fig. 41.1 Artificial neural network

41.1.2 Artificial Neural Network (ANN)

Artificial neural network is the simulation of human brain. In the human brain, there are millions of neuron present. These are nothing but the basic unit of brain. All the sensory organs send the input to the brain. The neurons receive it, process it, and generate appropriate output to take an action. When we try to achieve the same thing artificially, then it is called artificial neural network [15–17]. The ANN is trained first using data; then, it is implemented using testing data to produce results. The ANN has three layers, as shown in Fig. 41.1. They are input layer, hidden layer, and output layer [18–20].

$$n_j^{et} = \sum_i w_{ij} * x_i \quad (41.1)$$

$$\varphi = \frac{1}{1 + e^{-\text{net}}} \quad (41.2)$$

41.1.3 Recurrent Neural Network (RNN)

RNN is the advanced version of ANN. The term “Recurrent” itself means “Repetitive.” What happens in RNN is that not only the current input is considered but also the previous output is considered as input repeatedly as shown in Fig. 41.2. This helps the network to yield better outputs than the ANN. The major difference is there are neurons present in the ANN, but here in RNN, they are replaced by the memory blocks which stores the previous output [20–23]. After the discovery of RNN, the understanding and listening ability of the machines has improved tremendously.

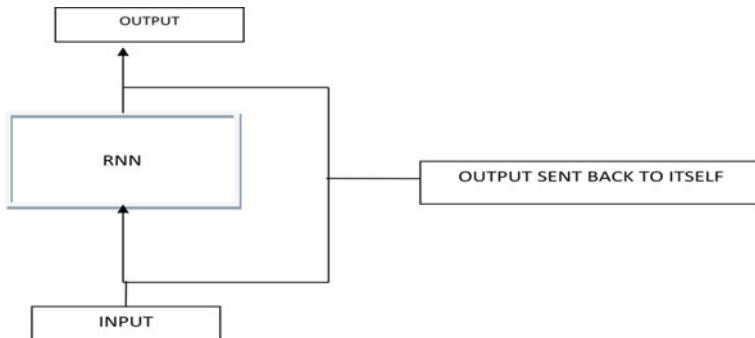


Fig. 41.2 Recurrent neural network

Now-a-days, it has become a very important part of our day-to-day life. Starting from Google translation to auto-completion of messages, RNN is used.

Formula for calculating current state:

$$h_t = f(h_{t-1}, x_t) \quad (41.3)$$

where

- h_t = current state,
- h_{t-1} = previous state,
- x_t = input state.

Formula for activation function (tanh):

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \quad (41.4)$$

where

- W_{hh} = recurrent neuron weight,
- W_{xh} = input neuron weight.

Formula for calculating output:

$$y_t = W_{hy}h_t \quad (41.5)$$

where

- y_t = output,
- W_{hy} = output layer weight.

Advantages of RNN

As RNN remembers previous inputs, it becomes useful for time series prediction. Convolution layer with recurrent neural network extends effectiveness of pixel neighborhood.

Disadvantages of RNN

It is very difficult to train RNN. If we use \tanh or relu as an activation function, then it won't process lengthy sequences.

41.1.4 Long Short-Term Memory Network (LSTM)

It is a recurrent neural network (RNN), which is modified so as to make it easier to remember past data in memory as shown in Fig. 41.3. It is used to overcome the vanishing gradient problem of RNN. In place of neurons, LSTM networks have memory blocks that are connected through layers. Back propagation is used for training. Three gates are present in a LSTM network [21–24]. The diagrammatic representation of the gates are given below.

Input Gate: determines the input value to be used to alter memory. Sigmoid function determines which values to let through 0–1 and \tanh function gives weightage to the values which are passed deciding their level of importance ranging from –1 to 1.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (41.6)$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad (41.7)$$

Forget gate: determines details that are to be deleted from the memory. Sigmoid function does this job. It looks at the previous state (h_{t-1}) and the content input (x_t) and outputs a number between 0 (omit this) and 1 (keep this) for each number in the cell state C_{t-1} .

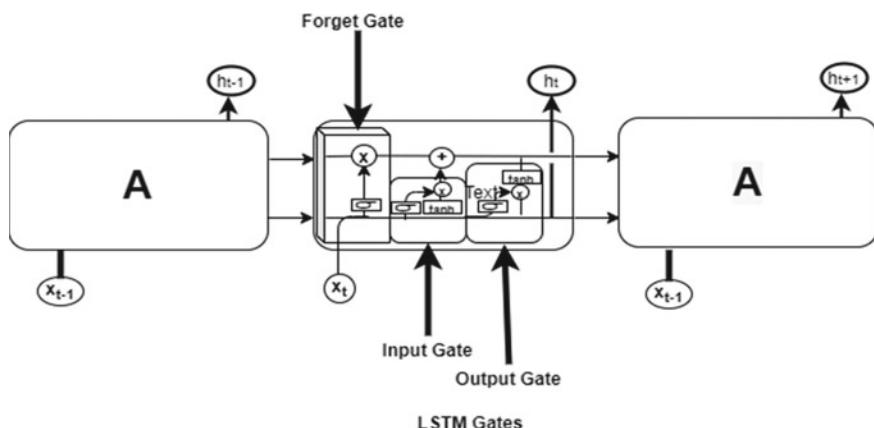


Fig. 41.3 Long short-term memory network

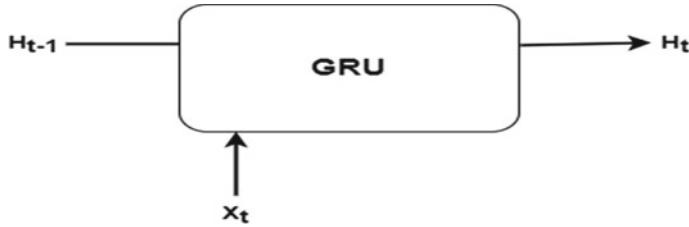


Fig. 41.4 Architecture of GRU

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (41.8)$$

Output gate: The output is determined by using input and memory of the block.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (41.9)$$

$$h_t = o_t * \tanh(C_t) \quad (41.10)$$

41.1.5 Gated Recurrent Unit (GRU)

The latest addition to the sequence modeling techniques after RNN and LSTM is the gated recurrent unit (GRU) as shown in Fig. 41.4. So, it offers improvement over LSTM and RNN. GRU is the advancement of the standard RNN, and it is a simplified version of LSTM. Gated recurrent unit uses different gates to control flow of information like LSTM. That is why they gave some improvement over LSTM and have a simple architecture. The GRU is different from LSTM as it has only a hidden state (H_t), whereas the LSTM has a separate cell state (C_t). Due to this simple structure, they are faster to train [24–26].

In GRU, there are two gates as opposed to LSTM. They are (1) reset gate and (2) update gate. Reset gate (Short-term memory)—it is accountable for the short-term memory (hidden state) of the network. Due to the sigmoid function, the value of r_t ranges from 0 to 1. The equation of the reset gate is given below.

$$r_t = \sigma(X_t * U_r + H_{t-1} * W_r) \quad (41.11)$$

Here, U_r and W_r are weight matrices.

41.2 Dataset Collection

We have collected the exchange rate dataset from International Monetary Fund. In this article, instead of using traditional technical indicators, we have used exchange rates of 23 economically sound countries as our predictors [4]. We have considered two datasets one from May 5th, 2003 to December 30th, 2019, i.e., before the COVID-19 pandemic, and the second dataset from January 3rd, 2005 to June 30th, 2021 which includes the exchange rates of different countries during COVID-19 pandemic. The data are then normalized using the required functions under the range from 0 to 1. After normalization, we have split both the dataset into two parts. The first 80% data is used for training the models, and the rest 20% data that is used for testing the model. We have analyzed 16 years of daily exchange rates of both the datasets using machine learning techniques to prepare the prediction model.

41.3 Experimental and Result Analysis

In this article, three models are used and in each model (ANN, LSTM, and GRU); prediction function is used to forecast exchange rate. Figs. 41.5 and 41.6 visualize the pattern of Indian rupees over the years for both the datasets. The corresponding graphs of the three models are being created to compare the actual and predicted values. Figures 41.7, 41.9, 41.11, 41.13, 41.15, and 41.17 depict the prediction results of Indian exchange rate before and after the COVID-19 pandemic. Figures 41.8, 41.10,

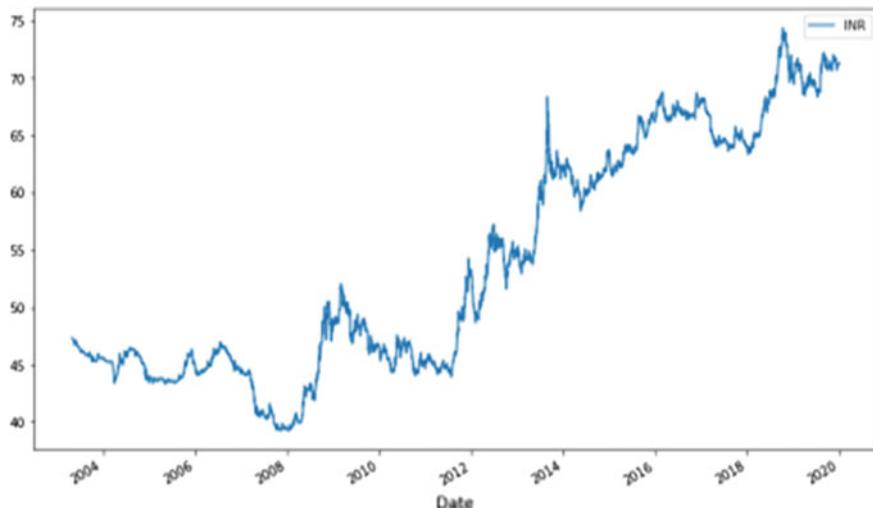


Fig. 41.5 Plotting dataset to visualize the pattern of prices over the years (excluding COVID-19 period)

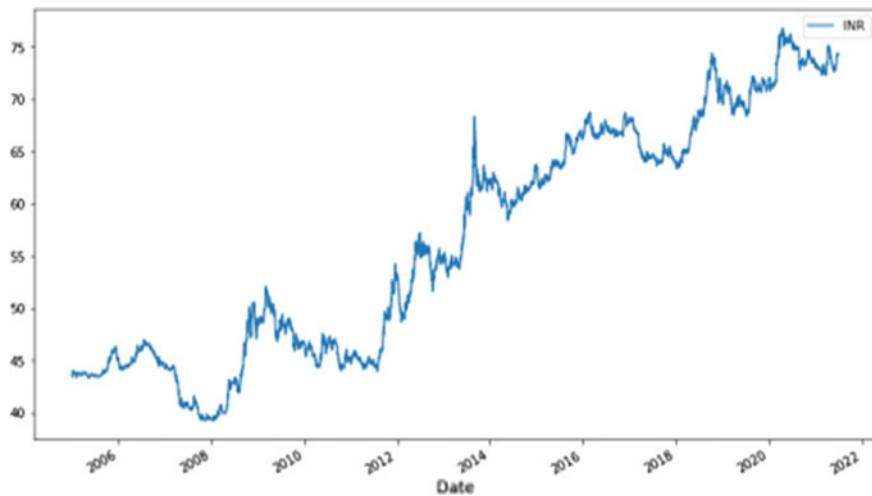


Fig. 41.6 Plotting dataset to visualize the pattern of prices over the years (including COVID-19 period)

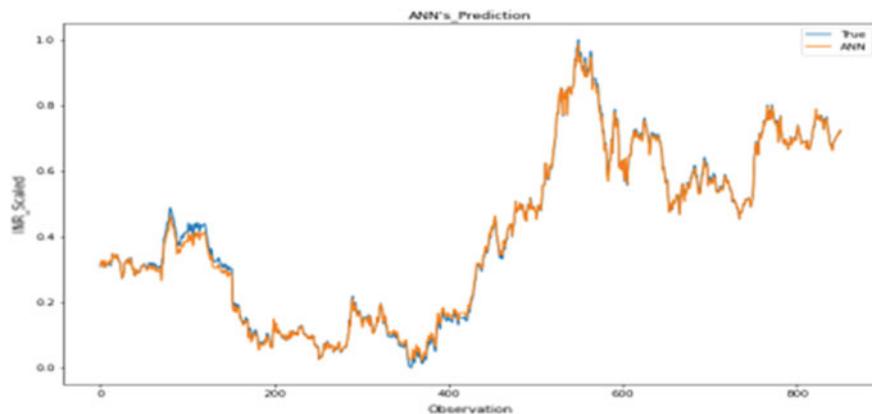


Fig. 41.7 Prediction of exchange rate using ANN (before COVID-19)

[41.12](#), [41.14](#), [41.16](#), and [41.18](#) represent the RMSE plot for both the testing dataset. The root mean square error (RMSE), mean absolute error (MAE), R^2 value, and adjusted R^2 value have been calculated to know each model's efficiency [9]. We have analyzed both the datasets, and their results have been compared based on the performance matrix. From the depicted result shown in Table 41.1, we have observed that the R^2 , adj R^2 , RMSE, and MAE values of GRU model for the first dataset, i.e., excluding the COVID-19 period is better than that of both ANN and LSTM model. In Table 41.2, we have shown the analysis of foreign exchange for the second dataset, i.e., including the exchange rate values during COVID-19. Experimental analysis of

both the dataset using machine learning techniques shows that MAE value and the RMSE value of GRU are less than both ANN and LSTM for both training and testing data. This shows that the error in GRU is less than that of the other two models. Also, the R^2 value of GRU is very closer to 1 that means the prediction of GRU is very accurate. So, we can conclude that GRU outperformed both ANN and LSTM. In Figs. 41.19 and 41.20, we have compared the performance of three models before and after the COVID-19 pandemic. We have analyzed from Figs. 41.19 and 41.20 that the predicting capabilities of machine learning models are not affected during this pandemic situation. Overall, we have analyzed that the machine learning techniques are performing better during this pandemic situation.

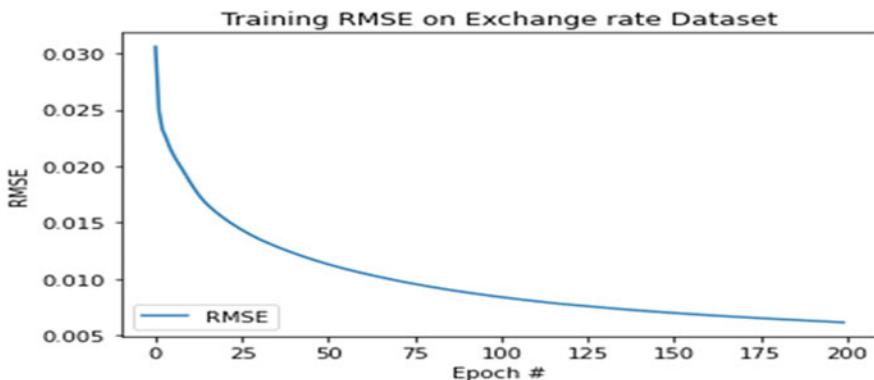


Fig. 41.8 RMSE plot for ANN (before COVID-19)

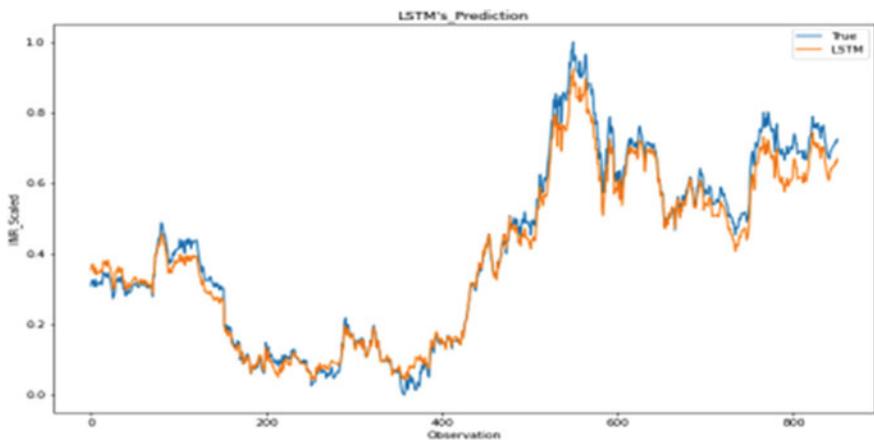


Fig. 41.9 Prediction of exchange rate using LSTM (before COVID-19)

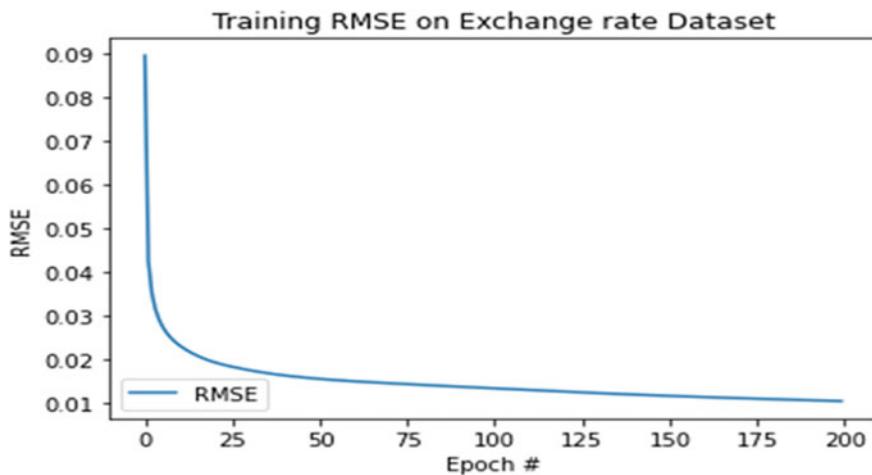


Fig. 41.10 RMSE Plot for LSTM (before COVID-19)



Fig. 41.11 Prediction of exchange rate using GRU (before COVID-19)

41.4 Conclusion and Future Scope

The conclusion is based on the considered study period and the input dataset. However, the results may differ according to the market, the study period and the currencies considered for the exchange rate. Therefore, conclusions of this analysis are subjective. We have analyzed that before and after COVID-19, GRU has been found to be the most correct and a hit algorithm within the time series prediction closely following by way of LSTM although the working procedure is quite comparable, and GRU requires less reminiscence because it uses less training parameters; accordingly, it is quicker than LSTM. We have also analyzed that the prediction

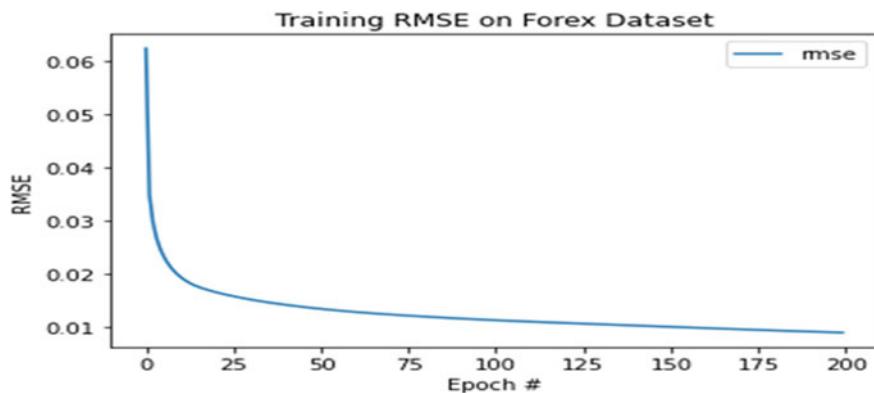


Fig. 41.12 RMSE plot for GRU (before COVID-19)



Fig. 41.13 Prediction of exchange rate using ANN (after COVID-19)

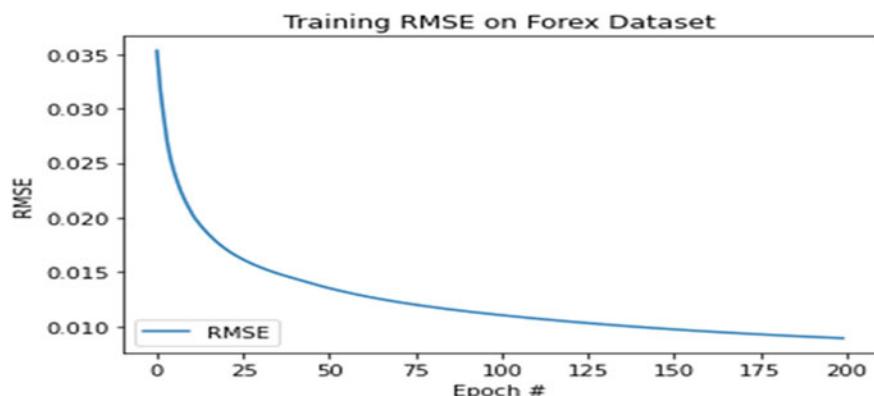


Fig. 41.14 RMSE plot for ANN (after COVID-19)

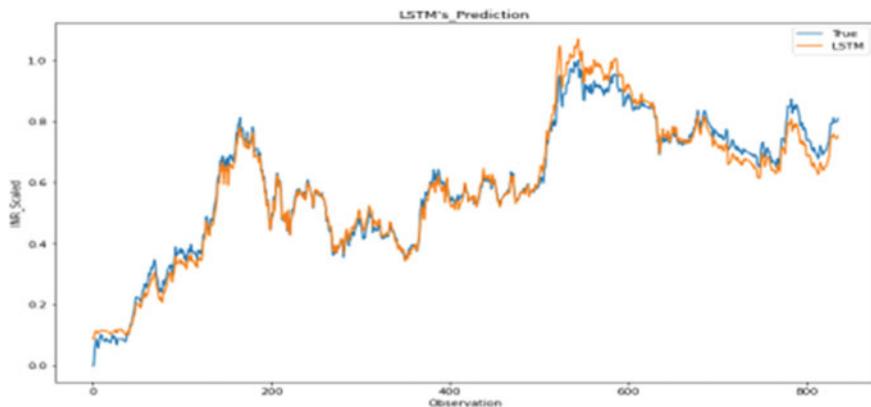


Fig. 41.15 Prediction of exchange rate using LSTM (after COVID-19)

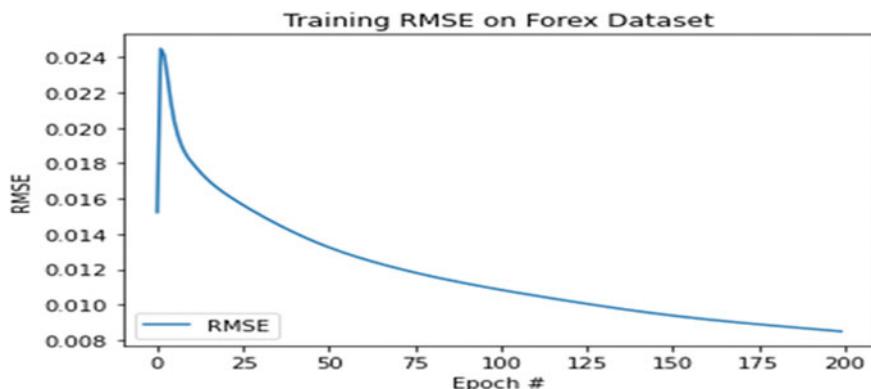


Fig. 41.16 RMSE plot for LSTM (after COVID-19)



Fig. 41.17 Prediction of exchange rate using GRU (after COVID-19)

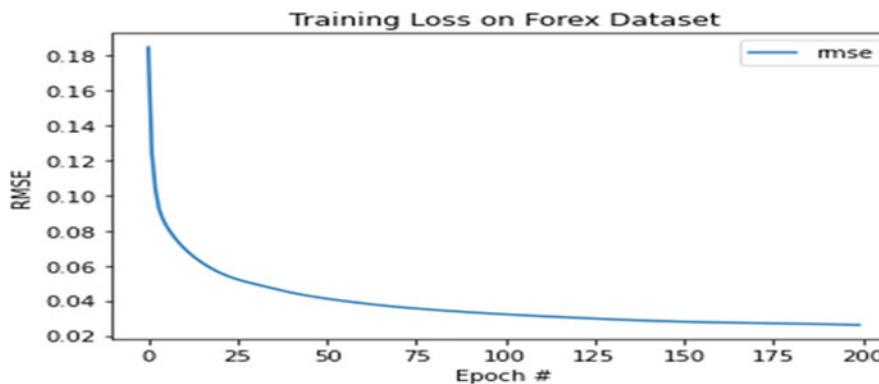


Fig. 41.18 RMSE plot for GRU (after COVID-19)

Table 41.1 Comparison of machine learning models before COVID-19

Model name	Training set				Testing set			
	R^2 value	MAE	RMSE	Adj. R^2 value	R^2 value	MAE	RMSE	Adj. R^2 value
ANN	0.994	0.018	0.023	0.994	0.937	0.048	0.064	0.935
LSTM	0.995	0.016	0.020	0.995	0.958	0.045	0.053	0.957
GRU	0.997	0.014	0.016	0.996	0.976	0.040	0.049	0.958

Table 41.2 Comparison of machine learning models after COVID-19

Model name	Training set				Testing set			
	MAE	RMSE	Adj. R^2 value	R^2 value	MAE	RMSE	Adj. R^2 value	R^2 value
ANN	0.998	0.011	0.015	0.998	0.993	0.016	0.019	0.992
LSTM	0.997	0.013	0.017	0.997	0.979	0.025	0.031	0.979
GRU	0.989	0.011	0.015	0.989	0.980	0.026	0.030	0.980

where MAE mean absolute error, RMSE root mean square error

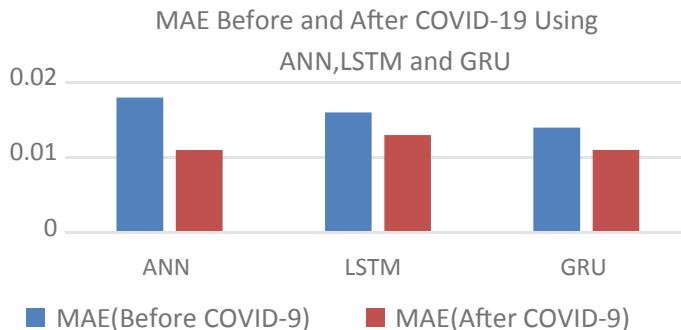


Fig. 41.19 MAE before and after COVID-19 using ANN

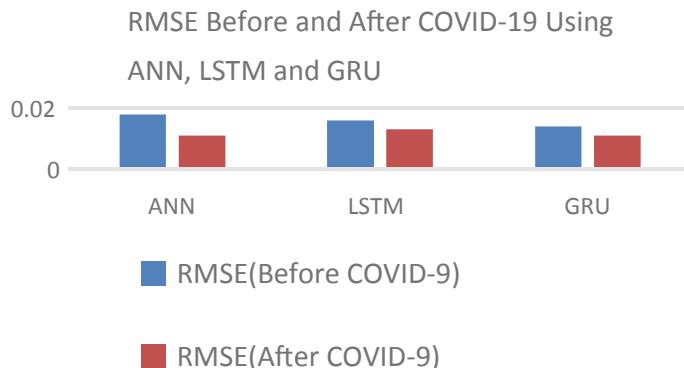


Fig. 41.20 RMSE before and after COVID-19 using ANN

capabilities of machine learning techniques are less affected during the pandemic situation. The possible future extension that can be made is to train the models for a short-term prediction and short dataset. This study can be further extended to predict the exchange rates of other countries by fine tuning the parameters of machine learning techniques as well.

References

1. Pandey, T.N., Jagadev, A.K., Mohapatra, S.K., Dehuri, S.: Credit risk analysis using machine learning classifiers. In: International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 1850–1854 (2017)
2. Pandey, T.N., Jagadev, A.K., Dehuri, S., Cho, S-B.: A novel committee machine and reviews of neural network and statistical models for currency exchange rate prediction: An experiment analysis. *J. King Saud Univ. Comput. Inf. Sci.* (2018)
3. Mahanta, A, Pandey, T.N., Jagadev, A.K., Dehuri, S.: Optimized radial basis functional neural network for stock index prediction. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 1252–1257 (2016)
4. Pandey, T. N., Jagadev, A.K., Dehuri, S., Cho, S.-B.: A review and empirical analysis of neural networks based exchange rate prediction. *Intell. Decis. Technol.* **1**(12), pp. 423–439 (2019)
5. Pandey, T.N., Priya, T., Jena, K.: Prediction of Exchange rate in a cloud computing environment using machine learning tools. *Intell. Cloud Comput.* 137–146 (2021)
6. Jhee, W.C., Lee, J.K.: Performance of neural networks in managerial forecasting. *Intell. Syst. Account. Fin. Manag.* **2**, 55–71 (1993)
7. Cao, L., Tay, F.: Financial forecasting using support vector machines. *Neural Comput. Appl.* **10**, 184–192 (2001)
8. Kaastra, I., Boyd, M.: Designing a neural network for forecasting financial and economic time-series. *Neuro-computing* **10**, 215–236 (1996)
9. Yao, J., Tan, C.L.: A case study on using neural networks to perform technical forecasting of forex. *Neuro-computing* **34**, 79–98 (2000)

10. Pandey, T.N., Giri, P.K., Jagadev, A.: Classification of credit dataset using improved particle swarm optimization tuned radial basis function neural networks. In: International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making (BITMDM), pp. 1–11. Springer, Heidelberg (2019)
11. Fathian, M., Kia, A.: Exchange rate prediction with multilayer perception neural network using gold price as external factor. *Manage. Sci. Lett.* **5**(1), 561–573 (2012)
12. Nayak, R.K., Mishra, D., Rath, A.K.: Optimized SVM-k-NN currency exchange forecasting model for Indian currency market. *Neural Comput. Appl.* **31**(7), 2995–3021 (2019)
13. Galeshchuk, S.: Neural networks performance in exchange rate prediction. *Neuro-computing* **172**, 446–452 (2016)
14. Dash, R., Dash, P.K., Bisoi, R.: A self-adaptive differential harmony search based optimized extreme learning machine for financial time series prediction. *Swarm Evolut. Comput.* **19**, 25–42 (2014)
15. Patra, B., Bhutia, S., Panda, N.: Machine learning techniques for cancer risk prediction. *Test Eng. Manag.* **83**, 7414–7420 (2020)
16. Patra, B.N., Bisoyi, S.K.: CFSES optimization feature selection with neural network classification for microarray data analysis. In: IEEE Xplore 2nd International Conference on Data Science and Business Analytics (ICDSBA), pp. 21–23. 45–50 (2018)
17. Patra, B.N., Jena, L., Bhutia, S., Nayak, S.: Evolutionary hybrid feature selection for cancer diagnosis. In: International Conference on Intelligent and Cloud Computing (ICICC-2019), Siksha ‘O’ Anusandhan Deemed to be University, pp. 16–17, 279–287 (2019)
18. Pradeep, K.D., Ravi, V.: Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network. *Appl. Soft Comput.* **58**, 35–52 (2017)
19. Jena, P.R., Majhi, R., Majhi, B.: Development and performance evaluation of a novel knowledge guided artificial neural network (KGANN) model for exchange rate prediction. *J. King Saud Univ.-Comput. Inf. Sci.* **27**(4), 450–457 (2015)
20. Kocak, C.: ARMA (p: q) type high order fuzzy time series forecast method based on fuzzy logic relations. *Appl. Soft Comput.* **58**, 92–103 (2017)
21. Chen, A.S., Leung, M.T.: Regression neural network for error correction in foreign exchange forecasting and trading. *Comput. Oper. Res.* **31**(7), 1049–1068 (2004)
22. Premanode, B., Toumazou, C.: Improving prediction of exchange rates using differential EMD. *Exp. Syst. Appl.* **40**(1), 377–384 (2013)
23. Jiang, P., Dong, Q., Li, P., Lian, L.: A novel high-order weighted fuzzy time series model and its application in nonlinear time series prediction. *Appl. Soft Comput.* **55**, 44–62 (2017)
24. Liu, S., Zhang, C., and Ma, J.: CNN-LSTM Neural network model for quantitative strategy analysis in stock markets. In: International Conference on Neural Information Processing, Springer, pp. 198–206 (2017)
25. Chawla, A., Lee, B., Fallon, S., and Jacob, P.: Host based intrusion detection system with combined CNN/RNN model. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp. 149–158 (2018)
26. Abbas, G., Nawaz, M., Kamran, F.: Performance comparison of NARX & RNN-LSTM neural networks for lifepo4 battery state of charge estimation. In: 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), IEEE, pp. 463–468 (2019)

Chapter 42

Effects of Binning on Logistic Regression-Based Predicted CTR Models



Manvik B. Nanda, Bhabani Shankar Prasad Mishra, and Vishal Anand

Abstract **Binning** is a process to group number of more or less continuous values into a smaller number of ‘bins’ based on certain criteria which results in handling high volume of data in computationally less expensive time. By adopting the process of binning we basically reduce the cardinality of the data. This paper proposes machine learning-based robust model which can work upon a high cardinality data-set and reproduce bins.

42.1 Introduction

Click Through Rate (CTR) prediction is the task of predicting the likelihood that something on a website (such as an advertisement) will be clicked [1–3]. Rapid development of the Internet and mobile devices, our daily activities connect closely to online services, such as online shopping, online news and videos, online social networks, and many more. At the same time online advertising has grown into a hundred-billion-dollar business since 2018, and the revenue has been increasing by more than 20% per year for 4 consecutive years. CTR prediction is critical in the online advertising industry, and the main goal is to deliver the right ads to the right users at the right time. Therefore, how to predict CTR accurately and efficiently has drawn the attention of both the academic and industry communities.

Rapid growth of technology gives rise to a high volume of unstructured data which needs to be filtered out so that a promising machine learning models can be built

M. B. Nanda
Happiest Minds Technologies, Bengaluru, India
e-mail: manvik.nanda@happiestminds.com

B. S. P. Mishra (✉)
School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India

V. Anand
Times Internet, Colombia, Noida, India
e-mail: vishal.anand@timesinternet.in

upon having high accuracy. For this, it is very important to understand the data and draw useful insights from the same.

42.2 Literature

Machine learning data analysis uses algorithms to continuously improve itself over time, but quality data is necessary for these models to operate efficiently.

Broadly machine learning models rely on four primary data types like numerical data, categorical data, time series data, and text data. While we are doing the CTR analysis we basically consider the factors like geographical location, type of advertisement, hour of the day, etc. which are mostly categorical data. So, in this paper we focus only on categorical data.

Categorical data is sorted based on different characteristics [4]. This can include gender, social class, ethnicity, hometown, the industry we work in, or a variety of other labels. While analysing these data type it is pretty clear that the attributes are non-numerical, so it is difficult to add them together, average them out, or sort them in any chronological order.

There are different ways to handle categorical variable which are enumerated as below.

42.2.1 One Hot Encoding

In this technique, the integer encoded variable is removed and a new binary variable is added for each unique integer value. The major problem in this is that number of distinct values in these features is high (explained in Fig. 42.1), multiplied with no of features leading to the ‘curse of dimensionality’. Thus, using one hot encoding technique to handle categorical variables lead to over-fitting and poor model build which is not robust.

Thus, before application of one hot encoding it is needed to find a way to bin similar values together, thus to be able to create buckets which are similar, using this we will be able to reduce the cardinality and then apply one hot encoding to handle the categorical features [5]. Example, red, green, and blue are converted as shown in Table 42.1.

The graph below shows the cardinality of various features which has been used for analysis in this paper, high cardinality of each feature leading to the ‘curse of dimensionality’ is clearly understandable from the figure.

Fig. 42.1 No of classes in each feature (plotted using sea born)

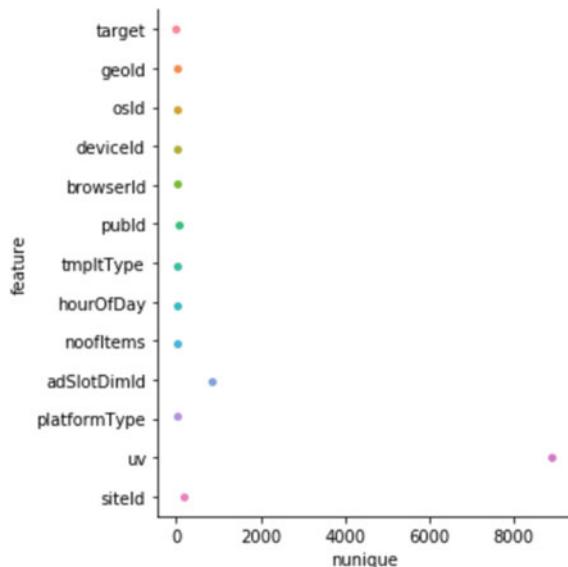


Table 42.1 Example of one hot encoding

	Red	Green	Blue
1	0	0	0
0	1	0	0
0	0	0	1

42.2.2 Target Encoding/Mean Encoding

Target encoding is a fast way to get the most out of categorical variables with little effort [6]. There can be a categorical variable x and a target y . y can be binary or continuous variable. For each distinct element in x it is required to compute the average of the corresponding values in y . Then each x_i is replaced with the mean. So basically mean encoding for each feature is number of true values of that particular class in the feature divided by the total no of values of that class in the feature list. The detail is explained in Fig. 42.2.

The major drawback of target encoding is: **over-fitting**. Indeed relying on an average value is not always a good idea when the number of values used in the average is low. It has to be kept in mind that the data-set used for training on is a sample of a larger set. This means that whatever artifacts found in the training set might not hold true when applied to another data-set (i.e. the test set). To solve this issue *additive smoothing* is used.

Mathematically this is equivalent to:

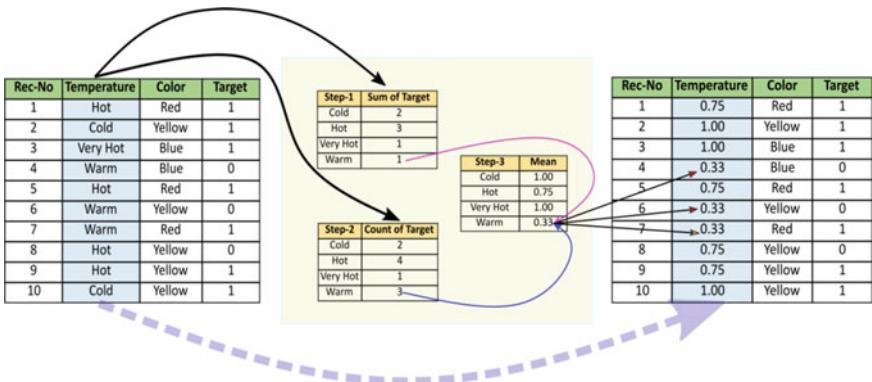


Fig. 42.2 Target encoding (towards data science—all about categorical variable encoding) [7]

$$\mu = \frac{n \times \bar{x} + m \times w}{n + m} \quad (42.1)$$

where

- μ is the mean we are trying to compute (the one that's going to replace our categorical values)
- n is the number of values you have
- \bar{x} is your estimated mean
- m is the ‘weight’ you want to assign to the overall mean
- w is the overall mean

In this notation m is the only parameter required to set. The idea is that the higher m is, the more rely on the overall mean w . If m is equal to 0 then it is required to compute the empirical mean only, and no smoothing is done, which is:

$$\mu = \frac{n \times \bar{x} + 0 \times w}{n + 0} = \frac{n \times \bar{x}}{n} = \bar{x} \quad (42.2)$$

When smoothing is performed it requires that there must be at least m values for the sample mean to overtake the global mean.

42.2.3 Feature Importance

Multiple features affect the calculation of CTR. The below graph represents the average clicks per feature based on various important parameters like Hour of Day, Geographical-Id, etc. is shown in Fig. 42.3.

Some of the problems of using the traditional approaches like manual calculation of CTR are:

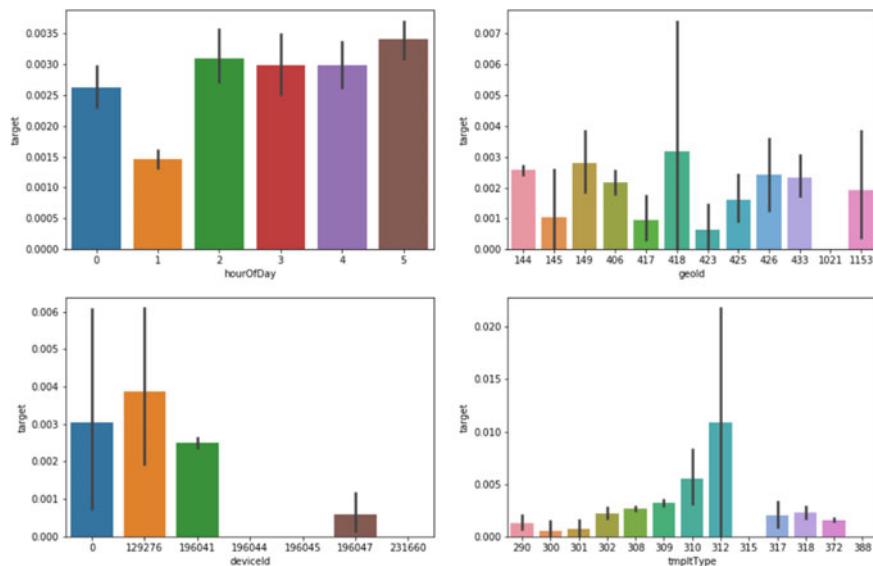


Fig. 42.3 Average clicks/feature (plotted using sea born)

- It does not standardize for the fact that CTR% changes when organic search position changes, which basically means that say your page gets higher on google, it will obviously have a higher CTR, so filtering it manually is not a good approach.
- Viewers reacts to ads differently, hence, a different distribution curve per ad, some ad will be left or right skewed, some will have a higher peak (higher kurtosis) and they will all have different mean and standard deviation.

The above enumerated disadvantages can be overcome by using machine learning approach.

1. Using ML approaches we can personalize ads, thus showing right kind of ad to the right user at the right time.
2. Both the revenue as well as the user experience increases, which in turn also increases the revenue or the product/website.

Thus, we go for ML approach for solving the CTR problem which is one of the attractive problems in the domain of business world and computer science.

42.3 Proposed Technique

The basic flow of the proposed technique is figured out in Fig. 42.4.

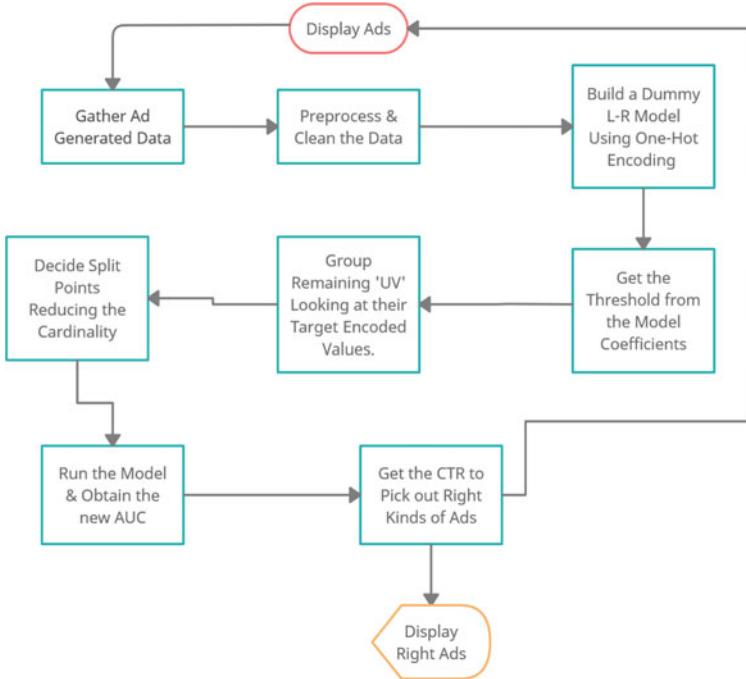


Fig. 42.4 Pipeline of the project

42.3.1 Algorithm

WHILE Website is running.

- (i) Gather the data that is generated indicating parameters like hour of the day, geographical location, when the user clicks on the particular ad
- (ii) Clean and preprocess the data, removing the NULL values and the ones which have a constant value throughout
- (iii) Build a dummy L-R model using one-hot encoding, this is the reference model whose AUC we need to try to achieve as it has no data loss.
- (iv) Get the threshold using the model coefficients, as it basically indicates the various probabilities of an ad getting clicked.
- (v) Group the reaming 'UV' looking at their Target Encoded Values which again is a parameter of an ad getting clicked.
- (vi) Decide split points, and bin the data using these split points thus reducing the cardinality.
- (vii) Run the model using this enhanced and the CTR to pick out the right kind of ads.

(viii) Display right ads.

42.3.2 Dummy Model

The idea is now to use the target encoded values and based on the model coefficients bin the data to merge the model coefficients and retain the accuracy. Binning is applied on a feature with high cardinality (here ‘UV’) which has 8500+ different values (Fig. 42.5; Table 42.2).

42.4 Experimental Detail

The below section explains about experimental setup, data-set used.

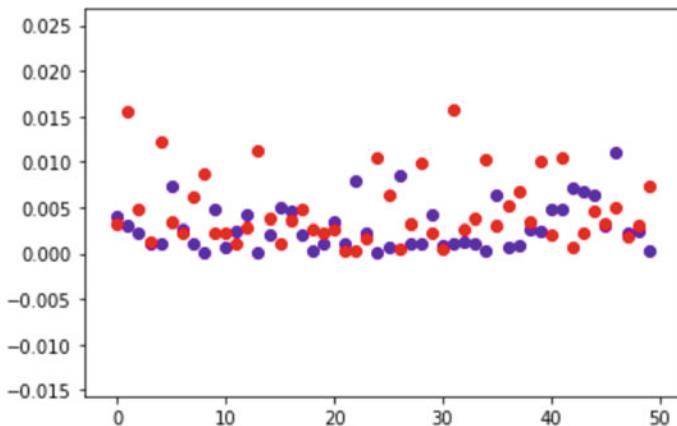


Fig. 42.5 Predictions for 100 random data-sets, Red-1; Purple-0 (plotted using Matplotlib)

Table 42.2 Dummy model result

Model used	Handled	Train_AUC	Test_AUC
Logistic regression	One hot encoding	0.80	0.75

42.4.1 Experimental Setup

Initially, we will be using data collected over a period of 6 h, and then later also extend it to 12 h to see how robust our model is.

Various data was collected from users (described below), each of which has its own importance when it comes to building the model, out of the data collected, we will be picking out some to be used in our model, and the ones which make the most intuitive sense to have in our project.

The project was run on the following environment:

OS: Microsoft Windows 10.

Processor: Intel(R) Core(TM) i5-7200U CPU @ 2.50 GHz, 2712 MHz, 2 Core(s), 4 Logical Processor(s).

RAM: 8 GB.

42.4.2 Data-Set Detail

After applying the prepossessing steps like removing the ones which have NULL values throughout or have a constant unique value the features in our data-set are shown in Table 42.3.

42.5 Comparison Detail

- (i) Getting the threshold from the model coefficients and try to group the remaining ‘UV’ looking at their Target Encoded values and place it in the suitable bin (Table 42.4).

Thus, we are able to retain the accuracy and at the same time decrease the Cardinality (Fig. 42.6).

- (ii) Cluster the model coefficients as done earlier but now on unsigned values so that similar ‘UV’ go into the same bins (Table 42.5).

Thus, we are not able to retain the accuracy using this method.

1. Task-1 and Task-2 were performed on a bigger data sample (12 h) as opposed to the initial (6 h) data and similar results were obtained.

42.6 Conclusion

Clearly using binning on the ‘UV’ feature gave us a great result as it retained the model accuracy. At the same time decreased our data-set cardinality thus building a

Table 42.3 Description of features in the data-set

S. No.	Feature	Description
1	Geoid	Used to measure precise surface elevations
2	Osid	Operating system id used to identify operating system and its and its key features or configurations
3	Deviceid	Used to identify the type of device
4	Browserid	Used to identify the type of browser used
5	Cityid	Used to identify the unique city
6	Stateid	Used to identify the unique state
7	Pubid	Publisher ID
8	tmpltType	The type of template used in the particular ad
9	hourOfDay	The hour of the day to identify the part of the day ad was clicked
10	Noofitems	No of items in the ad. space
11	Algoid	The type of algorithm used
12	Adslotdimid	Represents the dimension of the ad
13	Platformtype	Represent the type of platform
14	ModelDimid	The represents the model of the dimension of ad. space
15	itmClmbId	The Columbia id of the item
16	WebFamilyId	The family id of the type of web
17	UV	User view, it refers to page url. Every site has multiple pages and each page url is identified using unique id
18	Siteid	Uniquely identifies the site
19	Target	Ad. was clicked or not clicked (1 or 0)—The output feature to be predicted

Note All the features are categorical which needs to be handled before we build out model

Table 42.4 Analysis after binning (signed ‘UV’)

Cut-Off	Initial cardinality	Final cardinality	Train_AUC	Test_AUC
0.001	8400	5056	0.80	0.75
0.005	8400	3555	0.81	0.75
0.010	8400	2507	0.80	0.75
0.050	8400	720	0.80	0.75
0.090	8400	464	0.81	0.75
0.150	8400	300	0.81	0.75
0.300	8400	186	0.81	0.75
0.600	8400	130	0.81	0.74
0.900	8400	67	0.80	0.74
1.200	8400	12	0.80	0.73

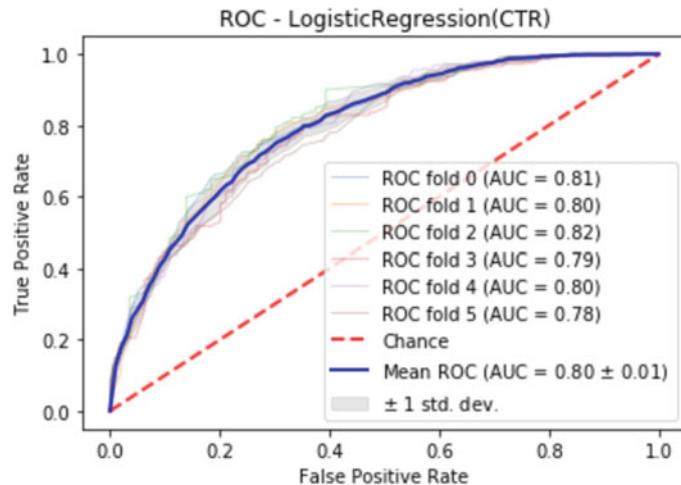


Fig. 42.6 Plot of train AUC (plotted using Matplotlib)

Table 42.5 Analysis after binning (unsigned ‘UV’)

No. of splits	Initial cardinality	Final cardinality	Train_AUC	Test_AUC
50	8400	50	0.82	0.73
100	8400	100	0.81	0.73
200	8400	200	0.81	0.73
250	8400	250	0.81	0.73
330	8400	330	0.82	0.73
500	8400	500	0.82	0.73
1000	8400	1000	0.82	0.73

robust model. This method can be further applied on the remaining features to test the same. Binning only makes sense with features of high cardinality; otherwise simply one hot encoding can be used. If binning is used with features of low cardinality, it may result in the loss of data to a huge extent.

References

1. Moneera, A., Maram, A., Azizah, A., AlOnizan, T., Alboqaytah, D., et al.: Click through rate effectiveness prediction on mobile ads using extreme gradient boosting. *Comput. Mater. Continua* **66**(2), 1681–1696 (2021)
2. Wang, F., Suphamitmongkol, W., Wang, B.: Advertisement click-through rate prediction using multiple criteria linear programming regression model. *Procedia Comput. Sci.* **17**, 803–811 (2013)
3. <https://paperswithcode.com/task/click-through-rate-prediction>

4. <https://towardsdatascience.com/feature-engineering-deep-dive-into-encoding-and-binning-techniques-5618d55a6b38>
5. https://en.wikipedia.org/wiki/Data_binning
6. <https://maxhalford.github.io/blog/target-encoding/>
7. <https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02>

Chapter 43

Proposed Multi-criterion Decision-Making Model—On Online Education System Perspective



Ishani Sengupta, Bhabani Shankar Prasad Mishra,
and Pradeep Kumar Mallick

Abstract Multi-objective linear programming (MOLP) multi-criterion decision-making problem has been solved first by using a LP modeller—PuLP written in Python and then solved it by LP method to set goal for each objective. Again, the MOLP problem has been solved in fuzzy environment to check the accuracy of most possible clarification of the goals. Here, in the proposed FLP model format, comparative weights close to the fuzzy operative term represent the relative worth of the objective functions and help to achieve the better goal. Furthermore, the projected activity based on E-learning technique is simply to be valid in that time when real world faces existent living problems (disaster/pandemic), which give improved result in the common sense that the values of objectives are adequately nearer to their goals. At last, one real model is used to exhibit the accuracy and utility of the projected process.

43.1 Introduction

In existent living circumstances, a decision creator always deals with different contradictory objectives. Then, it is essential to fix a goal for each objective of multi-objective linear programming (MOLP) problems.

Most MOLP problems of existent humankind were computationally troublesome. In cost-effective and corporal trouble of numerical encoding in general and in the linear programming problems, the coefficients in the models are supposed to be precisely identified. Still, in tradition, this statement is rarely satisfied by great popular

I. Sengupta (✉)
Infosys Limited, MCity, Chennai, India

B. S. P. Mishra · P. K. Mallick
School of Computer Engineering, KIIT University, Bhubaneswar, India
e-mail: bsmishrafcs@kiit.ac.in

P. K. Mallick
e-mail: pradeep.mallickfcs@kiit.ac.in

of existent living problems. Generally, the coefficients (some or all) are subjected to fault of capacity or they differ with marketplace environment [1, 2].

To defeat such a trouble, first the estimation of the target value for each objective plays a major role in several objective decision-making problems. Due to imprecise data of living world surroundings, the explanations are roughly not faithful, and also, it is extremely complicated to affect the existing process to locate the most favourable solution of linear programming (LP) problem in the common sense that there may be present a circumstances where a decision creator would like to construct a conclusion on the LP model, which absorbs the accomplishment of goals, where some of them may assemble the performance of the model and some may not [3, 4]. In such condition, the multi-objective linear programming (MOLP) model has been explained in fuzzy environment. The fuzzy set theory (FST) firstly was introduced by Zadeh [5]. The concept of fuzzy programming was originally introduced by Tanaka et al. [6].

The main principle of this paper is to set the closer to accurate target level for each objective of the LP model by using an LP modeller—PuLP written in Python and then solved the MOLP model by using Lindo software. Next, the MOLP model has been explained in fuzzy environment by using Lindo software to get the next better result.

Here, it has been noted that there are a number of fuzzy linear programming models where the weights are not controlled. The proposed procedure can guarantee the more significant fuzzy target, if the weights and also tolerances are different. A real example has been applied to exhibit the utility of the projected procedure, and the final outcome is explained by comparing with the outcome obtained from the existing model.

43.2 Decision-Making Procedure

See Fig. 43.1.

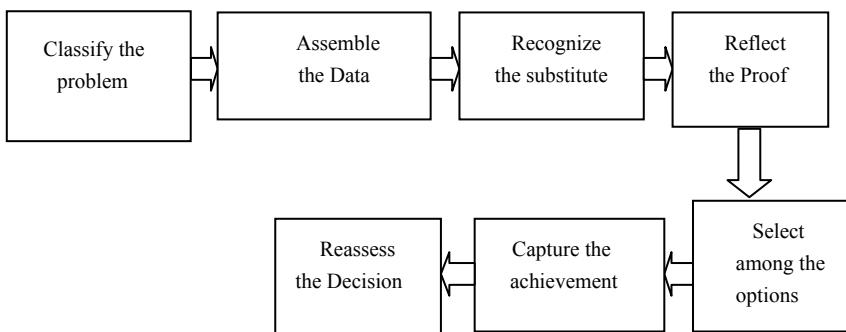


Fig. 43.1 Flow chart of decision-making procedure

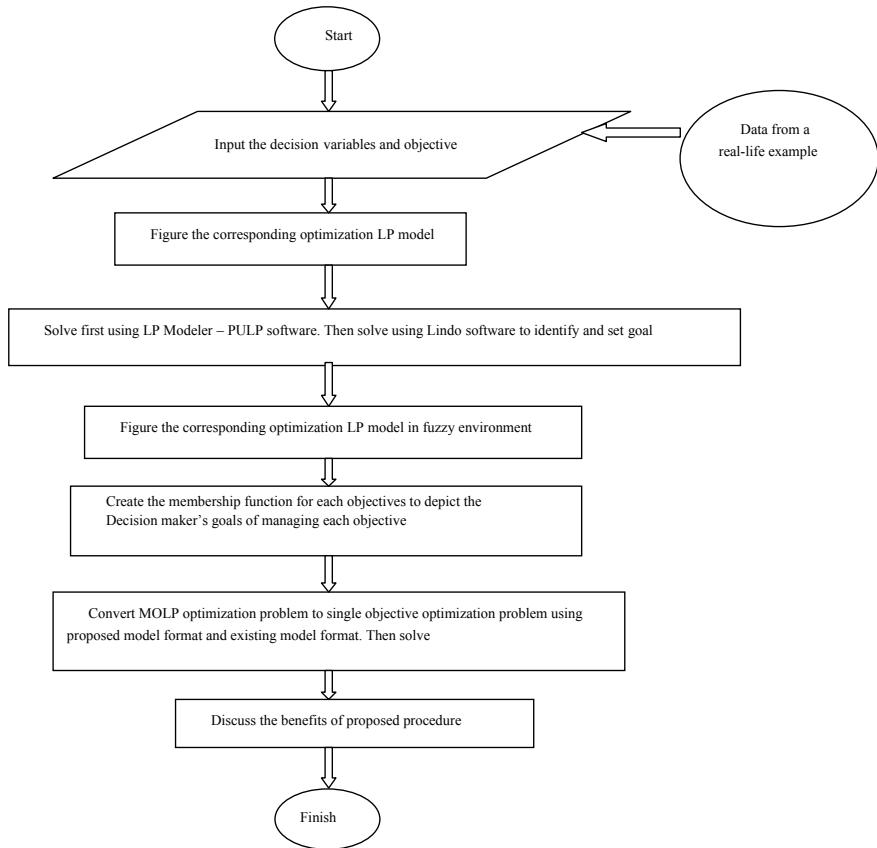


Fig. 43.2 Flow chart of proposed procedure

43.2.1 Proposed Procedure to Solve MOLP Problem

See Fig. 43.2.

43.3 Linear Programming Model Format

The general layout of the multi-objective linear programming (MOLP) models can be written as

$$\begin{aligned}
 & \text{Maximize/Minimize } P_i(y) = d_i(y) \\
 & \text{where } y \in Y = \{y \in R^n / Ay(\geq \text{ or } = \text{ or } \leq) b, y \geq 0, b^T \in R^m\}, \\
 & d_i^T \in R^n, i = 1, 2, \dots, I
 \end{aligned} \tag{43.1}$$

New linear programming model format can be written as

$$P_i(y)(\geq \text{ or } \leq)g_i$$

where $y \in Y = \{y \in R^n / Ay(\geq \text{ or } = \text{ or } \leq)b, y \geq 0, b^T \in R^m\}$,

$$d_i^T \in R^n, i = 1, 2, \dots, I \quad (43.2)$$

Here, g_i represents the goals attached to each objective function of the linear programming model.

Fuzzy programming model format is discussed as below having imprecise goals.

In MOLP model, if an inaccurate target level or goal g_i is set up to each of the objective $P_i(y)$, then these are named as imprecise or fuzzy goals. Then, the fuzzy goals are

- (i) $P_i(y) \gtrsim g_i$
- (ii) $P_i(y) \lesssim g_i$

The fuzzy multi-objective linear programming can be formulated as

$$\begin{aligned} P_i(y) &\gtrsim g_i \quad i = 1, 2, \dots, i_1; \\ P_i(y) &\lesssim g_i \quad i = i_1 + 1, \dots, I; \\ \text{Subject to } &Ay(\geq \text{ or } = \text{ or } \leq)b, y \geq 0 \end{aligned} \quad (43.3)$$

The membership function for the i th maximizing fuzzy goals can be expressed as

$$\mu_i(P_i(y)) = \begin{cases} 1, & P_i(y) \geq g_i \\ \frac{P_i(y) - l_i}{p_i}, & l_i \leq P_i(y) \leq g_i \\ 0, & P_i(y) \leq l_i \end{cases} \quad (43.4)$$

Again, the membership function for the i th minimizing fuzzy goals can be expressed as

$$\mu_i(P_i(y)) = \begin{cases} 1, & P_i(y) \leq g_i \\ \frac{u_i - P_i(y)}{q_i}, & g_i \leq P_i(y) \leq u_i \\ 0, & P_i(y) \geq u_i \end{cases} \quad (43.5)$$

where l_i is the lower tolerance limit, and u_i is the upper tolerance limit for the i th fuzzy goal. The tolerance $p_i(g_i - l_i)$ and the tolerance $q_i(u_i - g_i)$ both depend on decision creator.

Zimmermann's FLP Model

Maximize η

Subject to $\eta \leq \mu_i(P_i(y))$

$Ay(\geq \text{ or } = \text{ or } \leq)b, y \geq 0$

$$\eta \in [0, 1] \quad (43.6)$$

Explanation: We know that the main multi-objective function $d_i(y)$, $i = 1, 2, \dots, I$ is added to the constraints as a fuzzy goal in the Zimmermann's FLP model and the corresponding MOLP model with a new objective (η) is solved. When the MOLP model has optimal solutions, Zimmermann's FLP model may not always present the “best” solution. The cases that may arise: $d_i(y)$ may have different restricted values for the alternative optimal solutions or be limitless. Since all of the alternative optimal solutions have the same η , they have the same values for the new LP. Therefore, we test the value of $d_i(y)$ for all alternative optimal solutions. So, we propose a new fuzzy linear programming model format for eliminating these difficulties.

43.4 Proposed FLP Model

$$\begin{aligned} & \text{Minimize } (1 - \eta) \\ & \text{Subject to } \eta \leq (1 + \mu_i(P_i(y))) \\ & Ay(\geq \text{ or } = \text{ or } \leq) b, y \geq 0, \eta \in [0, 1] \end{aligned} \quad (43.7)$$

As we know that $\text{Maximize } (\eta) = \text{Minimize } (-\eta)$.

Now, suppose $(-\eta) = (1 - \eta)$ and $\eta \leq \mu_i(P_i(y))$ in Zimmermann's FLP Model can be written as $(-\eta) \geq -\mu_i(P_i(y))$,

$$\begin{aligned} & \text{then } 1 - \eta \geq -\mu_i(P_i(y)), \\ & \Rightarrow -\eta \geq -(1 + \mu_i(P_i(y))) \\ & \Rightarrow \eta \leq (1 + \mu_i(P_i(y))) \end{aligned}$$

Proposed FLP model with weighted approach

$$\begin{aligned} & \text{Minimize } (1 - \eta) \\ & \text{Subject to } w_i \eta \leq (1 + \mu_i(P_i(y))) \\ & Ay(\geq \text{ or } = \text{ or } \leq) b, y \geq 0 \\ & \eta \in [0, 1] \end{aligned} \quad (43.8)$$

The different weights are considered as $\sum w_i = 1$; $w_i \geq 1$; $i = 1, 2, \dots, I$.

The weights have been attached to fuzzy operator η to identify the better objective value which is very close to goal if the weights are varied.

43.5 Example

The structure of the higher education department in India is facing a major problem, due to the unavoidable challengeable real-life atmospheric situation. In that framework, Ministry of Human Resource Department (MHRD) in India makes a strategy to endorse their curriculum via on line education system to raise the enrolment rate of graduate students. To reach the goal, an innovative online education arrangement is supposed to be recognized in every universities/colleges. The online education arrangement holds Website, academic resources, graphic user interface (GUI), training apparatus, auditory/videotape equipment and distributed databases. Online education arrangement (OEA) be supposed to be allocated at various cities in India for nonstop smooth services, but each university cannot facilitate it in all places because of the constraint of investment accessibility. So, the university has to set up a speculation sketch to construct OEA in minimum four targeted cities; however, maximum six OEA can be organized. To attain consistency in education, the arrangement shall be recognized in three most important cities, namely Delhi, Chennai and Kolkata to bond each other with a basic three-hub-nodes ring trunking network. This set-up of OEA is shown in Fig. 43.1. Other nodes are to be precisely linked to one of the three-hub-nodes. Problem parameters are given as: six places recognized for OEA, fund requirements and the number of students in each town and are listed in Table 43.1.

There are two goals for this learning arrangement to happen:

Goal 1: At least 13,000 students are taken for this arrangement in six cities described in Table 43.1.

Goal 2: At least four OEA must be implemented by the university to fulfil the student's academic requirement in the better way possible.

As per resource restrictions of the universities, there are some constraints included to this arrangement:

- (a) At most ten staffs must be total personnel load for the scheduled function.
- (b) The total offered investment funds must not go over one crore rupees.
- (c) The implementation of basic ring trunking network should be maintained (Fig. 43.3).

Table 43.1 Problem parameters

OEA in cities	Decision variable	Funds requirements (crore)	Student's number	The number of personnel load
Delhi	y_1	0.1	4000	3
Gurgaon	y_2	0.2	2000	2
Kolkata	y_3	0.2	2500	3
Chennai	y_4	0.2	3000	3
Bangalore	y_5	0.2	1500	2
Bhubaneswar	y_6	0.3	1000	1

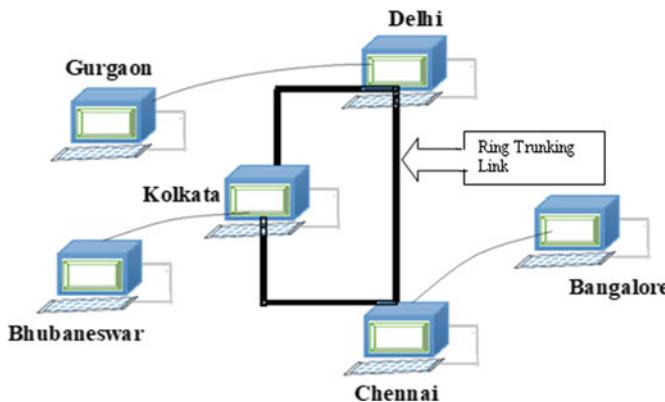


Fig. 43.3 Network of online education arrangement

The multi-objective linear programming models are represented as

Maximize Objective (1):

$$P_1(y) = 4y_1 + 2y_2 + 2.5y_3 + 3y_4 + 1.5y_5 + y_6$$

Maximize Objective (2):

$$P_2(y) = \sum y_i, \quad i = 1, 2, 3, 4, 5, 6$$

$$\text{subject to, } 3y_1 + 2y_2 + 3y_3 + 3y_4 + 2y_5 + y_6 \leq 10$$

(Personnel load constraint)

$$.1y_1 + .2y_2 + .2y_3 + .2y_4 + .2y_5 + .3y_6 \leq 1$$

(Funds requirement constraint)

$$y_1 + y_3 + y_4 = 3 \quad (\text{Basic ring trunking network constraint})$$

$$\text{where } y_i \geq 0, \quad i = 1, 2, \dots, 6. \quad (43.9)$$

The above MOLP model has been solved by using an LP modeller PuLP written in Python to fix target level/goal for each objective. The solutions are as in Table 43.2.

Besides, to achieve better goal/target level for each objective of the MOLP problem based on Eq. (43.9), the MOLP problem has been solved again by using an LP modeller PuLP written in Python Maximize Objective (1). Subject to personnel load constraint, funds constraint, basic ring trunking network constraint (43.10). Maximize Objective (2). Subject to personnel load constraint, funds constraint, basic ring trunking network constraint (43.11). The solution is as below

Table 43.2 Problem solution

$y_i = 0, i = 1, 2, 4, 5; y_3 = 3, y_6 = 1$
Value of objective (1) = 8.5
Value of objective (2) = 4

Table 43.3 Problem solution

Target level/goal (g_1, g_2)	MOLP problem in Eq. (43.12)	Infeasibilities
(8.5, 4)	$y_3 = 3, y_6 = 1, y_i = 0, i = 1, 2, 4, 5$	0
	Value of objective (1) = 8.5, Value of objective (2) = 4, 0	0
(13, 4)	$y_1 = 3, y_6 = 1, y_i = 0, i = 2, 3, 4, 5$	0
	Value of objective (1) = 13, Value of objective (2) = 4, 0	0

$$y_1 = 3, y_i = 0, i = 2, 3, 4, 5, y_6 = 1, \text{ Value of objective (1)} = 13 \quad (43.10)$$

$$y_i = 0, i = 1, 2, 4, 5, y_3 = 3, y_6 = 1, \text{ Value of objective (2)} = 4 \quad (43.11)$$

The target levels/goals for each objective are set as goal for Objective (1), (g_1) = 13 and Objective (2), (g_2) = 4.

Now, using LP model format in Eq. (43.2), the MOLP model in Eq. (43.9) can be written as

$$\begin{aligned} & \text{Objective (1)} \geq 13; \\ & \text{Objective (2)} \geq 4; \\ & \text{s.t. Personnel load constraint, Funds constraint, Basic ring} \\ & \text{trunking network constraint} \end{aligned} \quad (43.12)$$

The MOLP model in Eq. (43.12) has been solved now by Lindo software. The results have been arranged as in Table 43.3.

Next, the MOLP model in Eq. (43.12) has been resolved in fuzzy environment. Therefore, the several objectives FLP model has been represented as

$$\begin{aligned} & \text{Objective (1)} \gtrsim 13 \\ & \text{Objective (2)} \gtrsim 4 \\ & \text{s.t. Personnel load constraint, Funds constraint, Basic ring} \\ & \text{trunking network constraint} \end{aligned} \quad (43.13)$$

Now, the MOFLP model in Eq. (43.13) has been tried to be resolved by proposed FLP model in Eq. (43.7) by varying tolerance limits for the i th fuzzy goals ($i = 1, 2$). Table 43.4 summarizes the results.

In Table 43.4, it has been shown that by varying target level, the optimal value of both objectives could not meet their goal. Therefore, MOFLP model in Eq. (43.13) has been resolved by FLP model in Eq. (43.8). Table 43.5 analyses the corresponding results.

Table 43.4 Problem solution

Target level/goal (g_1, g_2)	Tolerances limits	Proposed FLP formulation in Eq. (43.7)
(13, 4)	(1, 1)	$y_1 = 3, y_i = 0, i = 2, 3, 4, 5, 6; \eta = 1, P_1(y) = 12, P_2(y) = 3$
	(5, 2)	$y_1 = 3, y_i = 0, i = 2, 3, 4, 5, 6; \eta = 1, P_1(y) = 12, P_2(y) = 3$
	(7, 1)	$y_1 = 3, y_i = 0, i = 2, 3, 4, 5, 6; \eta = 1, P_1(y) = 12, P_2(y) = 3$
	(10, 1)	$y_1 = 1, y_4 = 2, y_i = 0; i = 2, 3, 5, 6; \eta = 1, P_1(y) = 10, P_2(y) = 3$
(8.5, 4)	(1, 1)	$y_1 = 3, y_i = 0; i = 2, 3, 4, 5, 6; \eta = 1, P_1(y) = 12, P_2(y) = 3$
	(5, 2)	$y_1 = 3, y_i = 0; i = 2, 3, 4, 5, 6; \eta = 1, P_1(y) = 12, P_2(y) = 3$
	(7, 1)	$y_1 = 3, y_i = 0; i = 2, 3, 4, 5, 6; \eta = 1, P_1(y) = 12, P_2(y) = 3$

In the above, it has been shown that when target levels are (13, 4), weights attached to fuzzy operator η play an important role. Varying weights and tolerance limits, optimal value of two objective functions of MOLP model in Eq. (43.13) reached their target only when $w_1 = 5$ and $w_2 = 5$. On the other hand, when target levels are (8.5, 4), varying weights and tolerance limits, optimal value of two objective functions of MOLP model in Eq. (43.13) reached their target always. In this case, weights have no function. So, in front of decision-makers, there arises a confusion to set accurate target.

43.6 Discussion

Table 43.6 shows that the proposed FLP model in Eq. (43.8) gives better optimal value for both objectives without any difficulties.

The final solutions of the SOFLP model are discussed graphically in Fig. 43.4. Here, $P(y)$ represents objective (1): $P_1(y)$, and $Q(y)$ represents objective (2): $P_2(y)$.

Benefit of proposed procedure for solving real-life MOLP problem

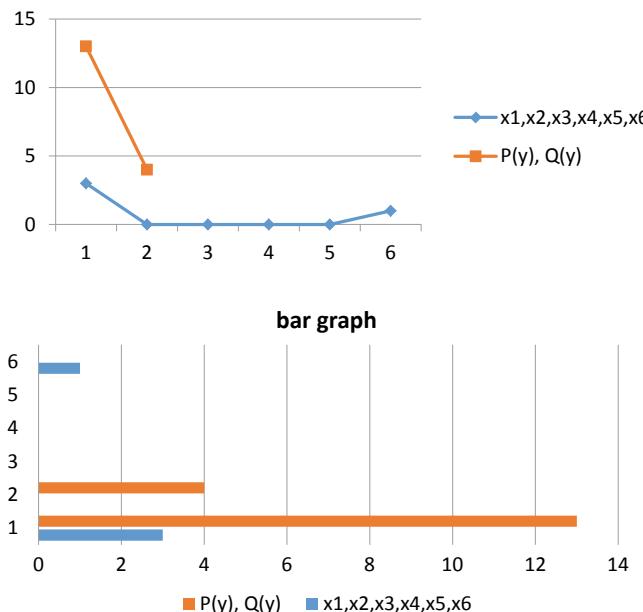
- We know that other than PuLP, better than this there are commercial solvers like Cplex or Gurobi which might be typically much faster but shall note that they come with their own Python-based modelling interfaces, whereas PuLP gives an advantage to build the model with an open source solver and also help switching to a commercial solver happening no change in the model code.
- So, a LP modeller—PuLP written in Python is used to solve a multi-objectives linear programming (MOLP) model to get the target level/goal. Here, both the objective functions have been taken along together with constraints. Next, solve

Table 43.5 Analysis of problem solution

Target level/goal (g_1, g_2)	Tolerance limits	Weights (w_1, w_2)	FLP model in Eq. (43.8)
(13, 4)	(1, 1)	(1, 5)	$y_3 = 3, y_6 = 1, y_i = 0, i = 1, 2, 4, 5; \eta = 0.40$
			$P_1(y) = 8.5, P_2(y) = 4$
		(5, 5)	$y_1 = 3, y_6 = 1, y_i = 0, i = 2, 3, 4, 5; \eta = 0.40$
			$P_1(y) = 13, P_2(y) = 4$
		(5, 4)	$y_1 = 3, y_2 = 0.5, y_i = 0, i = 3, 4, 5, 6; \eta = 0.40$
			$P_1(y) = 13, P_2(y) = 3.5$
	(5, 2)	(1, 5)	$y_3 = 3, y_6 = 1, y_i = 0, i = 1, 2, 4, 5; \eta = 0.40$
			$P_1(y) = 8.5, P_2(y) = 4$
		(5, 5)	$y_1 = 3, y_6 = 1, y_i = 0, i = 2, 3, 4, 5; \eta = 0.40$
			$P_1(y) = 13, P_2(y) = 4$
		(5, 4)	$y_1 = 3, y_2 = 0.5, y_i = 0, i = 3, 4, 5, 6; \eta = 0.40$
			$P_1(y) = 13, P_2(y) = 3.5$
(8.5, 4)	(1, 1)	(1, 5)	$y_3 = 3, y_6 = 1, y_i = 0, i = 1, 2, 4, 5; \eta = 0.40$
			$P_1(y) = 8.5, P_2(y) = 4$
		(5, 5)	$y_3 = 3, y_6 = 1, y_i = 0, i = 1, 2, 4, 5; \eta = 0.40$
			$P_1(y) = 8.5, P_2(y) = 4$
		(5, 4)	$y_1 = 2.25, y_4 = 0.75, y_6 = 1, y_i = 0, i = 2, 3, 5, \eta = 0.50$
			$P_1(y) = 12.25, P_2(y) = 4$
	(5, 2)	(1, 5)	$y_3 = 3, y_6 = 1, y_i = 0, i = 1, 2, 4, 5; \eta = 0.40$
			$P_1(y) = 8.5, P_2(y) = 4$
		(5, 5)	$y_3 = 3, y_6 = 1, y_i = 0, i = 1, 2, 4, 5; \eta = 0.40$
			$P_1(y) = 8.5, P_2(y) = 4$
		(5, 4)	$y_1 = 0.25, y_4 = 2.75, y_6 = 1, y_i = 0, i = 2, 3, 5, \eta = 0.50$
			$P_1(y) = 10.25, P_2(y) = 4$

Table 43.6 Proposed FLP model solutions

Target level/goal	Tolerances limits	Proposed FLP model in Eq. (43.8)	Zimmerman's FLP model in Eq. (43.6)
$g_1 = 13$ $g_2 = 4$	(1, 1)	$w_1 = 5, w_2 = 5$	
		$y_1 = 3, y_6 = 1, y_i = 0, i = 2, 3, 4, 5$	$y_1 = 3, y_6 = 1, y_i = 0, i = 2, 3, 4, 5$
		$\eta = 0.40, P_1(y) = 13, P_2(y) = 4$	$\eta = 1, P_1(y) = 13, P_2(y) = 4$
		Infeasibilities: 0	Infeasibilities: 0.1665335E-15



Here, $P(y)$ represents objective (1): $P_1(y)$ and $Q(y)$ represents objective (2): $P_2(y)$.

Fig. 43.4 Solutions of the SOFLP model

each objective function individually with constraints by LP modeller—PuLP written in Python. The values of goals in both cases are different.

- First set goals as $g_1 = 13, g_2 = 4$, then solve multi-objectives linear programming (MOLP) model in Eq. (43.12) by Lindo software. Next, setting goals as $g_1 = 8.5, g_2 = 4$ solve MOLP model in Eq. (43.12) by Lindo software. The optimal solutions in both cases meet their goal without any infeasibility. But there is no choice to set actual goal for MOLP model in Eq. (43.9).

- Hence, it is necessary to solve the MOLP model in Eq. (43.12) in fuzzy environment to set the accurate vague goals.
- The FMOLP model in Eq. (43.13) has been tried to be resolved by proposed FLP model in Eq. (43.7), but in this case, it was seen that the optimal solution for two objectives does not meet their goals.
- Hence, the FMOLP model in Eq. (43.13) has been resolved by FLP model in Eq. (43.8).
- In the FLP model (Eq. 43.8), weights act an important role to get good optimal solution for each objective function. Varying tolerances and weights when solved the MOFLP model in Eq. (43.13) it has been noticed that the optimal solution for both the objectives meet their goal (13, 4) accurately in only one case, whereas, not in the other case. What happens in the other case is, when goals are set as (8.5, 4) then the optimal solution for both the objectives meets their goals in two cases but are different. As a decision maker, could not get any constant results clearly, so decided to set the goal as (13, 4).
- Next, comparing the results obtained from Zymmerman's FLP model in Eq. (43.6) and proposed FLP model in Eq. (43.8), it has been shown that optimal solution for both the objectives meets their goals accurately in two cases but there are some infeasibilities present in the solution when solved by Zymmerman's FLP model.

43.7 Conclusion

In this paper, a multi-objectives linear programming (MOLP) model has been solved by proposed procedures, to get the appropriate success of goals, some of which are met and some not. Also, the necessity of weights in proposed FLP models acts an important role to get the better optimal solution, which assures the appropriate success of goals. The need for further research for explaining more real-life complicated decision-taking MOLP problems is essential.

References

1. Blank, J., Deb, K.: Pymoo: multi-objective optimization in python. *IEEE Access* **8**, 89497–89509 (2020). <https://doi.org/10.1109/ACCESS.2020.2990567>
2. Dong, J., Wan, S.: A new method for solving fuzzy multi-objective linear programming problems. *Iran. J. Fuzzy Syst.* **16**(3), 145–159 (2019). <https://doi.org/10.22111/ijfs.2019.4651>
3. Azadeh, A., et al.: A multi-objective fuzzy linear programming model for optimization of natural gas supply chain through a greenhouse gas reduction approach. *J. Nat. Gas Sci. Eng.* **26**, 702–710 (2015). <https://doi.org/10.1016/j.jngse.2015.05.039>
4. Andrade, R., Doostmohammadi, M., Santos, J.L., et al.: MOMO—multi-objective metabolic mixed integer optimization: application to yeast strain engineering. *BMC Bioinform.* **21**, 69 (2020). <https://doi.org/10.1186/s12859-020-3377-1>
5. Zadeh, L.A.: Fuzzy sets. *Inf. Comput.* **8**, 338–353 (1965)
6. Tanaka, H., Okuda, T., Asai, K.: On fuzzy-mathematical programming. *J. Cybern.* **3**(4), 37–46 (1973)

Chapter 44

Indoor Plant Health Monitoring and Tracking System



Nikhil Kumar, Sahil Anjum, Md Iqbal, Asma Mohiuddin,
and Subhashree Mishra

Abstract Indoor plants beautify homes and offices and bring us closer to nature. Decorating households with indoor plants is an amazing way to cope with the increasing problem of global warming, air pollution, and many other such environmental damages. Unfortunately, their need for regular watering makes it very difficult to take care of them. This makes the owners reluctant to travel to faraway places for long periods of time. This paper presents a solution to the above-discussed problem. The indoor plant health monitoring and tracking system will allow the owners to keep a track of their plant's health and allow them to observe different parameters like soil moisture, humidity, and temperature. They will be able to water the plants accordingly at will.

44.1 Introduction

Plants and greenery always add to the beauty of man-made environments. The addition of a hint of nature in our daily lives makes us calm and composed, while also adding to our productivity. Nature brings out the best in us, thus we want to stay close to nature even when we stay in the safety of our homes, or in offices while working [1]. Keeping a pot of a healthy green plant or shrub nearby, while working makes a person fresh and more productive. Having a habit of growing indoor plants has a number of benefits, like improving air quality and humidity for a healthy living environment, reduction in stress, fatigue, and anxiety. It may be a small step to tackle the rising issue of degradation of the environment. Indoor plants come in a broad spectrum of varieties that can suffice the demand of almost every person with different kinds of likings.

While having all the above-discussed benefits, indoor plants have only one drawback, that is, their demand for care does not allow their owners to travel for a long period of time. Having the need for moist soil, and a check on their environment (for some of the varieties), makes it difficult for the owner to leave their precious plant at

N. Kumar · S. Anjum · M. Iqbal · A. Mohiuddin · S. Mishra (✉)
School of Electronics Engineering, KIIT University, Bhubaneswar, India
e-mail: subhashree.mishrafet@kiit.ac.in

home while traveling. As a result, the owners end up asking their friends, family, or neighbors to take care of their plants in their stead. In the worst cases, they leave the plants to die at their homes or offices. For this issue, technology comes to the rescue. The automated plant monitoring system, consisting of a few sensors, like moisture sensor, temperature humidity sensor, etc., will keep a check on the plant's health and let the owner know about it [2, 3]. With the help of a water pump, the owner will be able to water the plant while also being able to adjust the light intensity for their plants at the scheduled time, or according to their wish. Internet of Things (IoT) and embedded system technology will be discussed in this paper that will make it possible for the owner to be able to connect to the indoor plant health monitoring and tracking system from anywhere in the world as long as there is an Internet connection available [4].

44.2 Literature Review

We are brought closer to nature by indoor plants in our homes and offices. There have been numerous research and survey projects designed to make indoor farming more efficient and easier. In 2013, Giri presented a paper on microcontroller-based drip irrigation systems [5], and the operating principle involved valve commutation based on soil moisture. This contributes to reducing soil erosion. Unfortunately, it was insufficient for dealing with practical challenges. Bansal in 2013 [6] demonstrated the use of microcontrollers with GSM modules integrated into wireless sensor networks. In this technology, factors such as soil and weather conditions are analyzed, but having poor connectivity makes this technique inapplicable in rural India. Devika in 2017 [7] introduced automation of plant watering through an Arduino board. A disadvantage of it was that it had limited resource sharing due to the fact that the information was restricted from being accessed globally. Ramu in 2013 [8] and Zareen in 2016 [9] came up with a set of techniques for incorporating a GSM module and 8051 microcontrollers into an irrigation system. In 2019 [10], Kumar introduced some more additional things, where the automation of watering and monitoring had been done by the NodeMCU, and data was uploaded on the Blynk App.

GSM is no longer the best option for our communications. We will greatly benefit from cloud computing [11] that allows us to maintain a vast database about the variety of soil types throughout India. It is also not necessary to use market-made applications like Blynk App to monitor the plants. This paper uses an application developed by us. By using the application, we can track the data of temperature, soil moisture, humidity, and light intensity from anywhere in the world. To access the data, an Internet connection is required. The intensity of the artificial light and the switching on or off of the water motor pump can be controlled with the application. This paper also focuses on the use of our own customized IoT board instead of market-made microcontrollers like NodeMCU. The customized IoT board is a real-time monitoring system based on IoT. The board is built with an ATmega328 and

Wi-Fi Nuttyfi. There are 14 digital input/output pins and eight analog input pins on the ATmega328 board, whereas Nuttyfi is built around ESP8266 12e.

44.3 System Architecture

The plant monitoring system will consist of a number of sensors that will be used to monitor the physical conditions of the plant in order to keep a check on the plant's health. The data captured by the sensors will be sent automatically to the cloud space, the access of which will be available to the user through a mobile application. The user will be able to know the live condition of the plant. After knowing the condition of the plant, the user will be able to switch to a water pump and adjust the light intensity according to the need of the plant. IoT will be used along with the embedded system technology. One more feature of the plant monitoring system was that an alarm will be sounded in the user's cell phone, to notify him in case if the physical conditions change a lot, below or above the normal acceptable range. This way, the user will be able to water the plant or adjust the light intensity if absolutely necessary.

44.3.1 Hardware Used

Temperature and Humidity Sensor (DHT11): DHT11 measures the relative humidity and temperature of the surroundings. By measuring the amount of water vapor present in the air, its saturation point yields the relative humidity in the air, and its thermistor measures temperature [12] (Fig. 44.1).

Soil Moisture Sensor: A soil moisture sensor FC28 is used to measure the moisture content of the soil. A conductor probe is located on both ends of the sensor. The change in resistance between the two conducting plates can be used to determine the moisture content in soil [13] (Fig. 44.2).

Mini Water Pump: This component will pump the water into the soil when commanded by the microcontroller. This could be done by setting up the relay to turn the water motor off automatically [14] (Fig. 44.3).

Fig. 44.1 Temperature and humidity sensor

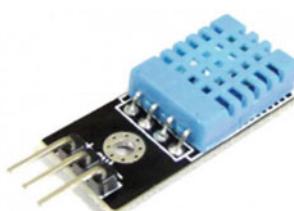


Fig. 44.2 Soil moisture sensor

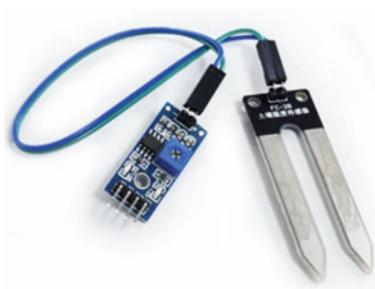


Fig. 44.3 Mini water pump



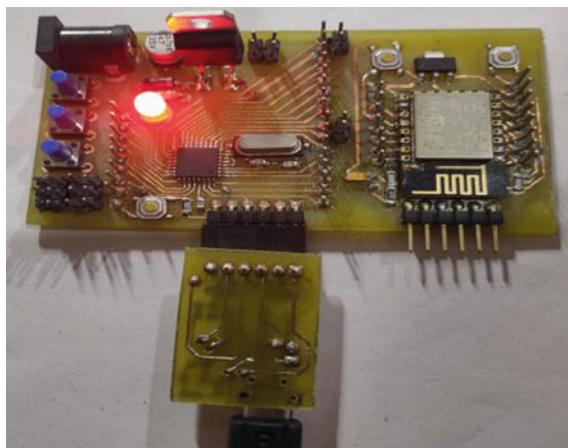
Light Intensity Monitoring Sensor (TEMT6000): This sensor is used to measure the intensity of light received by the plant. From this, we can know how much light is received by the plant. Plants need light for photosynthesis, and for that, a proper amount of light is necessary for indoor plants, which are mostly deprived of sunlight due to staying indoors. In case if the sunlight received by the plant is less than the threshold value for proper photosynthesis, we can give artificial light to the plant. The intensity of light can be controlled by the mobile application as well as by the microcontroller by analyzing the intensity of light (Fig. 44.4).

Relay Module: Generally, it performs the functions of a switch by electrical means. Here we have used the electromechanical relay. Relays can be broadly classified into two types: electromechanical (EMR) and solid-state relay (SSR) (Fig. 44.5).

Customized IoT Board: The customized IoT board consists of a detachable programmer board. The board consists of two microcontrollers, ATmega328p and ESP8266, which can be used individually and together as well as according to our needs. For using it together, we use Rx and Tx pins. They are used to receive and

Fig. 44.4 Light intensity monitoring sensor-TEMT6000



Fig. 44.5 Relay module**Fig. 44.6** Customized IoT board

transmit TTL serial data. They are connected with the corresponding ATmega328P USB to TTL serial chip (Fig. 44.6).

44.3.2 Customized Board Specification Table

See Table 44.1.

44.3.3 Software Used

Android App: We will be using an Android application for monitoring plant health. The temperature, humidity, soil moisture, and light intensity will be monitored. Depending on the level of soil moisture, we will be using a water pump button to provide water to the plant, and depending on the amount of light intensity received by the plant, we will adjust the artificial light intensity of the lamp with the help of the slider given in the application.

Table 44.1 Technical specification

S. No	Item/component	Specification/properties
1	Microcontroller	ATmega328P—8-bit AVR family microcontroller
2	Operating voltage	5 V
3	Recommended input voltage	7–12 V
4	Input voltage limits	6–20 V
5	Analog input pins	6 (A0–A5)
6	Digital I/O pins	14 (Out of which six provide PWM output)
7	DC current on I/O pins	40 mA
8	DC current on 3.3 V pin	50 mA

ThingSpeak: With ThingSpeak, we can visualize, estimate, and have a clear analysis of the live data streaming online in the form of a cloud. Additionally, we can store and acquire data using the HTTP protocol. With the help of the cloud-based database, it is possible to forecast future water requirements. A person who is unable to maintain their plants at home will find this very useful.

44.4 Applied Methodology

44.4.1 Flowchart

See Fig. 44.7.

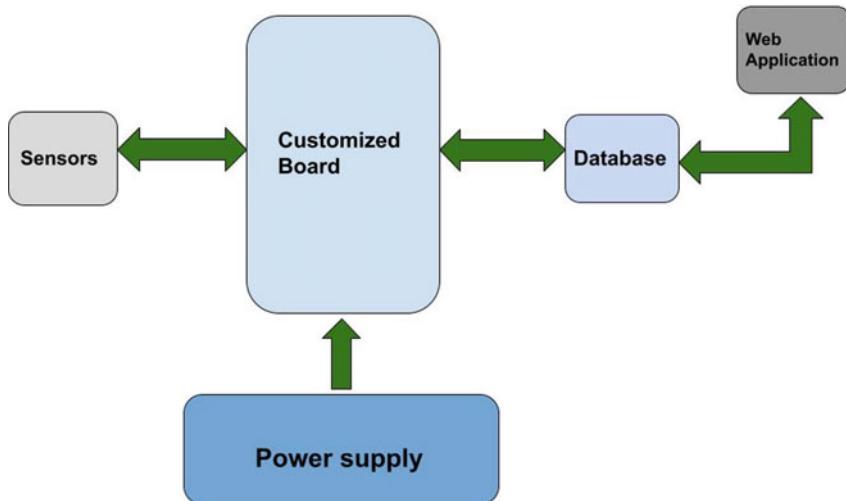
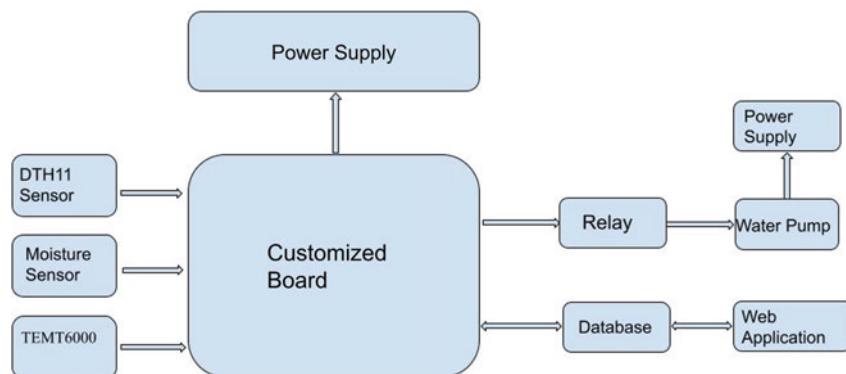
44.4.2 Block Diagram

See Fig. 44.8.

44.5 Discussion of Experimental Results

We will be using an android application for monitoring plants' health. In the attached screenshot of the application, we can see the temperature, humidity, soil moisture, and light intensity data of the plant's surroundings. Also, the water pump button to switch on/off the pump and a slider to adjust light intensity are visible in Fig. 44.9.

In Fig. 44.9, a screenshot of the application has been provided. On the left-hand side, we can see that the pump is off, light intensity adjusted to 29 foot-candle, soil moisture is 48% of 100% (means have to give water to the plant), humidity

**Fig. 44.7** Flowchart**Fig. 44.8** Block diagram

is 80 g m^{-3} , and temperature is 30°C . After watering the plant, and increasing the light intensity, on the right-hand side of Fig. 44.9, the pump is on, the light intensity has been increased to 39 foot-candle, soil moisture is 99% of 100%, humidity is 80 g m^{-3} , and temperature is 30°C .

The ThingSpeak platform allows viewers to visualize, estimate, and also analyze online data streams in real time. Figures 44.10, 44.11, and 44.12 show the data collected from the sensors for temperature, humidity, and moisture, respectively. On the right-hand side, we can see a live reading, and on the left-hand side, we can see a graph of different readings taken over a fixed interval of time. An individual can predict future requirements based on a plotted graph. It will be very useful for those

Fig. 44.9 Screenshot of the application

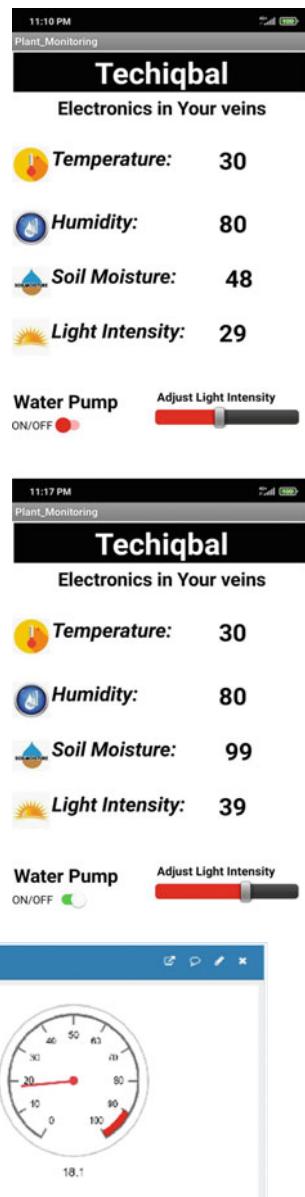


Fig. 44.10 Representation of temperature values

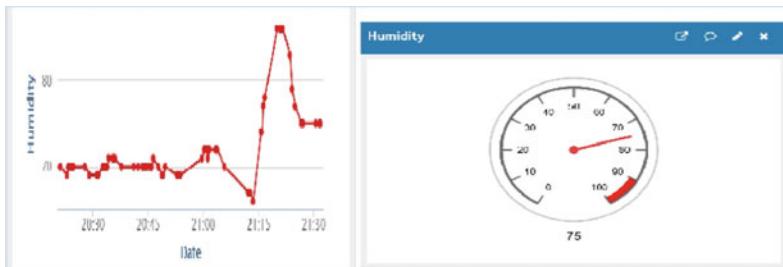


Fig. 44.11 Representation of humidity values



Fig. 44.12 Representation of moisture values

who are unable to care for their plants at home since the information will be readily available to them.

44.6 Conclusion

We conclude that by aiding the management of indoor plants, the “indoor plant health monitoring and tracking system” serves its purpose rightly. A person can track the data of temperature, soil moisture, humidity, and light intensity from anywhere in the world. The intensity of light and the switching on or off the water motor pump can be controlled with the application as per the collected data. It is possible to predict future water needs using the ThingSpeak database stored in the cloud. This system will be very useful, in case an individual is unable to look after their plants at home, as they will be well equipped with all the information. An Internet connection is mandatory for the indoor plant health monitoring and tracking system to communicate between IoT and embedded system devices.

References

1. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **29**(7), 1645–1660 (2013)
2. Kumar, V.S., Gogul, I., Raj, M.D., Pragadesw, S.K. and Sebastin, J.S.: Smart autonomous gardening rover with plant recognition using neural network. In: 6th International Conference On Advances In Computing and Communications, ICACC 2016, 6–8 September 2016, Cochin, India, Procedia Computer Science, vol. 93, pp. 975–981 (2016)
3. Kotkar, M., Sawant, A., Pathak, M., Kaur, S.: IOT based smart home garden watering system. *Int. J. Sci. Res. Eng. Manage.* **4**(4) (2020)
4. Thamaraimanalan, T., Vivekk, S.P., Satheeshkumar, G., Saravanan, P.:Smart garden monitoring system using IOT. *Asian J. Appl. Sci. Technol. (AJAST)* (2018)
5. Giri, M., Wavhal, D.N.: Automated intelligent wireless drip irrigation using linear programming. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **2**(1), 1–5 (2013)
6. Bansal, D., Reddy, S.R.N.: WSN based closed loop automatic irrigation system. *Int. J. Eng. Sci. Innov. Technol. (IIESIT)* **2**(3), 229–237 (2013)
7. Devika, C.M., Bose, K., Vijayalekshmy, S.: Automatic plant irrigation system using Arduino. In: IEEE International Conference on Circuits and Systems (ICCS), Thiruvananthapuram, India 20–21 Dec 2017
8. Ramu, M., Rajendra, C.H.: Cost-effective atomization of Indian agricultural system using 8051 microcontrollers. *Int. J. Adv. Res. Comput. Commun. Eng.* **2**(7), 2563–2566 (2013)
9. Zareen, S.G., Zarrin, K.S., Ali, A.R., Pingle, S.D.: Intelligent automatic plant irrigation system. *Int. J. Sci. Res. Edu.* **4**(11), 6071–6077 (2016)
10. Kumar, J., Gupta, N., Kumari, A., Kumari, S.: Automatic plant watering and monitoring system using NodeMCU. In: 2019 9th International Conference on Cloud Computing, Data Science and Engineering (Confluence), 2019, pp. 545–550. <https://doi.org/10.1109/CONFLUENCE.2019.8776956>
11. Vagulabranan, R., Karthikeyan, M., Sasikala, V.: Automatic irrigation system on sensing soil moisture content. *Int. Res. J. Eng. Technol.* **3**(1), 2012–2019 (2015)
12. Tianlong, N.: Application of single bus sensor dht11 in temperature humidity measure and control system. *Microcontrollers Embedded Syst.* **6**, 026 (2010)
13. Byoungwook Min, S.J.P.: A Smart Indoor Gardening System Using IoT Technology. Springer Nature Singapore Pte Ltd., (2018)
14. Durani, H., Sheth, M., Vagharia, M., Kotech, S.:Smart automated home application using IoT with Blynk App. IEEE (2018)

Chapter 45

Hybrid of Array-Based and Improved Playfair Cipher for Data Security



**K. R. Harini, D. Vijayaraghavan, S. Sushmidha, Vithya Ganesan,
and Pachipala Yellamma**

Abstract In any era, a secured communication is an essential part of human life. Data exchange is the act of transferring data from one node to another utilizing computer and communication technologies. This requires electronic or digital data to be transferred between two or more nodes, whatever the geographical location, medium or data contents are. The processing of data in a secured manner is the essential part in any transmission. Cryptography would help in this secure transmission for encoding the information or data in a format which cannot be understood easily and insensitive to brute force attack. The HABIP algorithm is proposed which does two levels of encryption. The encryption process simulates array-based encryption and improved Playfair cipher that severs the text to be transmitted into 26-bit string and encrypts it based on its position value in the root array. In next level, two successive characters are compared with two different key matrices. In this approach, high chances of repetition are reduced to 95% as each character gets encrypted as different character in every occurrence. HABIP, thus, improves the purity and security to a greater level than the contemporary encryption methods, and it is insensitive to brute force attack.

45.1 Introduction

The HABIP approach does two levels of encryption and decryption. Here, the first level of encryption process uses a root array that compares each character in the text with the root array and determines the new index value using the derived formula. In

K. R. Harini (✉) · D. Vijayaraghavan · S. Sushmidha

Department of Information Technology, St. Joseph's College of Engineering, Chennai, Tamil Nadu 600119, India

e-mail: rahaan6750@gmail.com

V. Ganesan

K L University, Vijayawada, Andhra Pradesh, India

P. Yellamma

Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Guntur, Andhra Pradesh, India



Fig. 45.1 Block diagram for encryption



Fig. 45.2 Block diagram for decryption

second level of encryption, two characters successively are encrypted using certain rules and text is encrypted again. This process is reversed for decrypting or decoding the encrypted text.

The above figure explains the process of converting the plain text to the encrypted text by the two-level encryption technique. The first level generates the first-level encrypted text. The encrypted text is sent to the second-level encryption technique. This level generates the final cipher text. Finally, this encrypted text is stored and transmitted (Fig. 45.1).

The above figure describes the process of decryption at the receiver, and it is the reverse of encryption. The proposed approach outputs the text transferred accurately. Thus, throughout the transmission, the purity and security of the data are maintained. The received text is processed to generate the first-level decrypted text, and the second-level decryption is performed to obtain the actual text transmitted (Fig. 45.2).

45.2 Literature Survey

(Aftab Alam [1]) The traditional Playfair cipher algorithm is modified, as the 5×5 matrix is changed to 7×4 matrix that includes two extra characters, '*' and '#'. A one-to-one correspondence is created due to these two characters. Unlike HABIP, special characters except '*' and '#' cannot be encrypted, or in other words, this method does not support special characters.

(Aftab Alam [1]) Here, the problem of odd number of characters in a text to be encrypted is solved by adding a letter 'X' at the last position of the text. Thus, making the size of text even makes encryption even easier in the case of traditional algorithm. Even then the text encrypted is not secured as the letter 'X' is repeated, and ambiguity arises if the last character is itself the letter 'X'.

(Basu [2]) In this approach, the contemporary Playfair cipher is altered and broadened. Here, the confusion i and j is excluded by extending the character set that

appeared in [3]. The generation of one—one cipher text and modification of matrix enhanced efficiency. This method is less immune to corruption of data as this also uses single key matrix.

(Shahil [4]) Here, a 10×9 matrix is deployed in the process of encrypting the text that is transmitted. Also, the process of encryption is done six times recursively. Even then, redundancy occurs as the same key matrix is used and overhead due to six iterations.

45.3 Root Array

Here, a root array is created that consists of 90 characters, consisting of all basic keyboard characters such as numbers, alphabet and special characters. It starts with space and special characters followed by numerals, upper case and the lower case alphabet. All values have an index value, respectively, from 0 to 89. We can append the parent array as per our needs and requirements, like mathematical and scientific symbols and alphabet of other languages. Here, the root array used to explain this technique is the most simple and basic. This array can be expanded according to the needs.

The starting index value can be determined and changed dynamically according to the text to be encrypted. The remainder obtained when diving the number of characters in the text by the number of words gives the starting index value.

For example, consider the text, Meet me at 9AM.

No. of characters = 15; No. of words = 4. So, $15\%5 = 0$. Thus, the starting value is 0 (Table 45.1).

Table 45.1 Root array

1		11	*	21	:	31	2	41	C	51	M	61	W	71	g	81	q
2	?	12	(22	'	32	3	42	D	52	N	62	X	72	h	82	r
3	~	13)	23	"	33	4	43	E	53	O	63	Y	73	i	83	s
4	!	14	-	24	,	34	5	44	F	54	P	64	Z	74	j	84	t
5	@	15	_	25	<	35	6	45	G	55	Q	65	a	75	k	85	u
6	#	16	=	26		36	7	46	H	56	R	66	b	76	l	86	v
7	\$	17	+	27	>	37	8	47	I	57	S	67	c	77	m	87	w
8	%	18	\	28	/	38	9	48	J	58	T	68	d	78	n	88	x
9	^	19		29	0	39	A	49	K	59	U	69	e	79	o	89	y
10	&	20	;	30	1	40	B	50	L	60	V	70	f	80	p	90	z

45.4 Encryption

45.4.1 First-Level Encryption

The root array is created as mentioned above. Now, the incoming message or data is spilt into a 26-bit character array. Index values for the 26-bit character array are assigned starting from 1 to 26. They represent their index position in the array. The spaces between the words are also included in the array such that the space occurs if the character stops at the 25th position. A character in the array is encrypted by adding its index value in the array with its index value in the root array, respectively, and the sum is modulated by 90, the size of root array. The corresponding character for the obtained result in the root table is the cipher text for this level. Similarly, any given paragraph can be encrypted by taking 26 characters at a time. This is the first level of encryption.

The formula used in the array-based approach is given by (Table 45.2).

$$C[i] = (P[i] + i) \bmod 90 + s$$

where

C cipher text1

Table 45.2 First-level encryption

Plain text	M	e	e	t		m	e		a	t		9	A	M
Index value i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Index in root array j	51	69	69	84	1	77	69	1	65	84	1	38	39	51
i + j	52	71	72	88	6	83	76	9	74	94	12	50	52	65
k = (i + j)%90 + s	52	71	72	88	6	83	76	9	74	4	12	50	52	65
P[k]	N	g	h	x	#	s	1	^	j	!	(L	N	a

Table 45.3 Generated key matrix 1(empty cell represents space)

c	i	p	h	e	r	8	7	+	\$
(?	~	!	@	#	%	^	&
*)	-	_	=	\	l	;	:	'
"	,	<		>	/	0	1	2	3
4	5	6	9	A	B	C	D	E	F
G	H	I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	X	Y	Z
a	b	d	f	g	j	k	l	m	n
o	q	s	t	u	v	w	x	y	z

Table 45.4 Generated key matrix 1 (empty cell represents space)

E	n	c	r	y	p	t	#	1	2
-		?	~	!	@	\$	%	^	&
*	()	-	=	+	\		;	:
'	"	,	<		>	/	0	3	4
5	6	7	8	9	A	B	C	D	F
G	H	I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	X	Y	Z
a	b	d	e	f	g	h	i	j	k
l	m	o	q	s	u	v	w	x	z

Table 45.5 Second-level encryption

Cipher Text1	Index in key matrix 1	Index in key matrix 2	Rule number	Cipher text 2
Ng	5,7	7,4	3	Ng
hx	7,6	8,7	3	jX
#s	0,7	8,2	3	1d
l^	8,0	1,8	3	m +
j!	7,8	1,4	3	Bu
(L	2,1	5,5	3	+ @
Na	5,7	7,0	3	Na

P plain text

i position value of the character in the string

P[i] position value of the character in the root array

s starting value of root array.

Plain text Meet me at 9 AM.

Cipher Text Nghx#sl^j!(LNa

45.4.2 Second-Level Encryption

In the second-level encryption, any two different key matrices are generated using two key values. First, the key values are filled and then the remaining characters, in order. The input, here, is the cipher text-1 obtained from the first level. Here, the cipher text is encrypted by taking two characters at a time. One character is marked in one matrix and the other in the second key matrix.

The following are rules that are designed for the encryption:

- **RULE-1:** If the two elements fall in the same row of the two matrices, then the immediate right element is taken and replaced.

- *RULE-2:* If the two elements fall in the same column of the two matrices, then the immediate below element is issued for replacement.
- *RULE-3:* If two elements fall in two different rows and columns, then the row values are paired, the column values are paired, and the corresponding values are replaced.

Step-1: Key matrix generation.

Key value 1: Encrypt#12- (Table 45.3).

Key value 2: cipher87 + \$ (Table 45.4).

Step 2: Two-matrix encryption.

Cipher Text1: Nghx#sl^j!(Lna. (Table 45.5).

Example CipherCipherText-2: NgjXldm + Bu + @Na.

45.5 Decryption

The cipher text encrypted by the two-level encryption technique is sent to the receiver. The receiver decrypts the cipher text to obtain the plain text using two levels of decryption.

45.5.1 First-Level Decryption

Considering the two key matrices used in encryption, the data is decrypted by taking two characters at a time and mapping them across the matrices.

The first level of decryption takes place in two steps, they are key matrix generation.

1. Two-matrix decryption.

The rules followed are as follows:

- *RULE-1:* If the two elements fall in the same row of the two matrices, then the immediate left element is replaced.
- *RULE-2:* If the two elements fall in the same column of two matrices, then the immediate above element is replaced.
- *RULE-3:* If two elements fall in two different rows and columns, then the row values are paired, the column values are paired, and the corresponding values are taken and replaced.

The empty cell in Tables 45.2 and 45.3 represents space character ("").

Step-1: Key matrix generation (Tables 45.6 and 45.7).

Step 2: Two-matrix decryption (Table 45.8).

Decrypted text: Nghx#sl^j!(Lna.

Table 45.6 Key value 1: Encrypt#12-

E	n	c	r	y	p	t	#	1	2
-		?	~	!	@	\$	%	^	&
*	()	-	=	+	\		;	:
'	"	,	<		>	/	0	3	4
5	6	7	8	9	A	B	C	D	F
G	H	I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	X	Y	Z
a	b	d	e	f	g	h	I	j	k
l	m	o	q	s	u	v	w	x	z

Table 45.7 Key value 2: cipher87 + \$(

c	i	p	h	e	r	8	7	+	\$
(?	~	!	@	#	%	^	&
*)	-	-	=	\		;	:	'
"	,	<		>	/	0	1	2	3
4	5	6	9	A	B	C	D	E	F
G	H	I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	X	Y	Z
a	b	d	f	g	j	k	l	m	n
o	q	s	T	u	v	w	x	y	z

Table 45.8 First-level decryption

Cipher text1	Index in key matrix 1	Index in key matrix 2	Rule number	Cipher text2
Ng	5,7	7,4	3	Ng
jX	7,8	6,7	3	hx
1d	0,8	7,2	3	#s
m +	8,1	0,8	3	l^
Bu	7,1	8,4	3	j!
+ @	2,5	1,5	3	(L
Na	5,7	7,0	3	Na

45.5.2 Second-Level Decryption

The decrypted text received is severed into strings of 26 bits. Referring the root array, the position value of each character in the array is determined. Thus, position value of the character in the string is subtracted from its index value in the root array, respectively. The modulus operation is performed on the difference obtained by size

Table 45.9 Second-level decryption

Decrypted text	N	g	h	x	#	s	l	^	j	!	(L	N	a
Index value i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Index in root array j	52	71	72	88	6	83	76	9	74	4	12	50	52	65
j-i	51	69	69	84	1	77	69	1	65	84	1	38	39	51
k = (j-i)%90 + s	51	69	69	84	1	77	69	1	65	84	1	38	39	51
R[k]	M	e	e	t		m	e		a	t		9	A	M

Text Meet me at 9AM.

of parent array, here 90. The character corresponding to the obtained result in the root array is identified. The result obtained is the actual text transferred.

Equation for second-level decryption:

$$R[i] = (D[i] - i) \bmod 90 + s$$

where

D deciphered text

R received text

I index value of the character in the root array

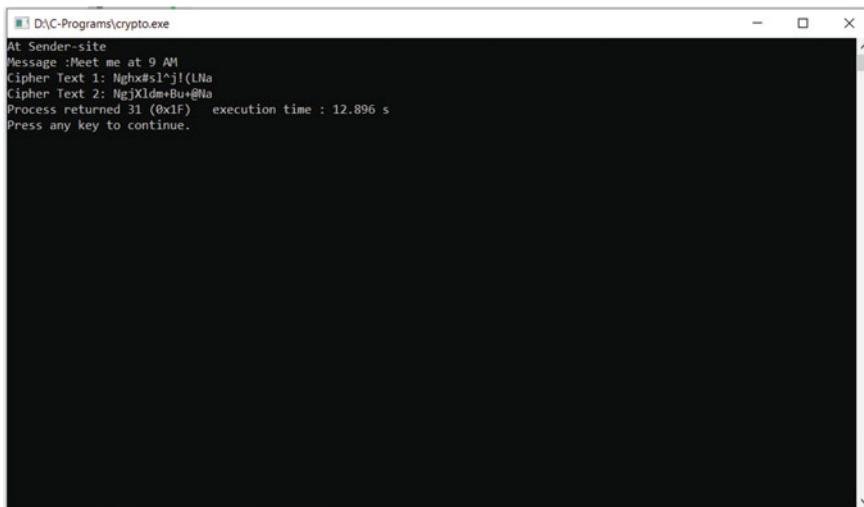
R[i] denotes the index value of the character in the parent table

s starting position value of root array.

Note: If $(D[i]-i)$ is negative, then add 90 (root array size) with the result (Table 45.9), i.e., $(D[i]-i) + 90$.

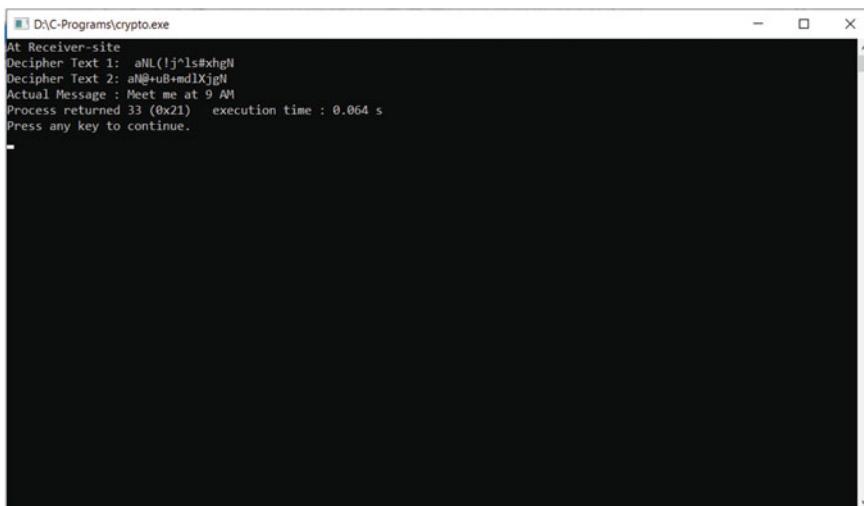
45.6 Analysis of the Proposed Approach

The suggested algorithm HABIP enhances the security as two levels of encryption take place. The first level itself encrypts the characters based on the position value. Thus, the same character gets re-acquired by various characters each and every time. Redundancy is efficiently reduced. And the process of encrypting using two key matrices improves the performance and the security at higher level. Also, the key value pair can be changed or replaced repeatedly. When it comes to brute force attack, it is impossible to crack the code, as it needs to check with $(90! \times 90!)$ possible key values. If the parent array's size increases, then it becomes even more difficult. This obviously proves the increased level of security provided by the HABIP method than the current techniques (Figs. 45.3 and 45.4).



```
D:\C-Programs\crypto.exe
At Sender-site
Message :Meet me at 9 AM
Cipher Text 1: Nghx#s1^j!@Lna
Cipher Text 2: NgjXldm+Bu+@Na
Process returned 31 (0x1F) execution time : 12.896 s
Press any key to continue.
```

Fig. 45.3 Sender side



```
D:\C-Programs\crypto.exe
At Receiver-site
Decipher Text 1: aLL(j^ls#xhgN
Decipher Text 2: aN@+uB+mD!XjgN
Actual Message : Meet me at 9 AM
Process returned 33 (0x21) execution time : 0.064 s
Press any key to continue.
```

Fig. 45.4 Receiver side

45.7 Output

45.8 Conclusion

The HABIP approach improves the security provided by the traditional techniques like Playfair cipher and Caesar cipher. It is made possible by using two variant key values and key matrices, and two levels of encryption. The first-level encryption of one character can have more than 90 (based on the need) different possibilities as the character's position in a text or a message varies repeatedly, but its position value in the root array is persistent. There is no repetition in this method. HABIP incorporates the process of encoding using two variant key matrices. If a character appears successively, then it can also be encrypted, whereas it is not possible in traditional methods which makes it stronger than the latter. Also, similar characters that occur in a text are encrypted differently each time as the characters have different position values. HABIP upholds the overall virtue of the text transferred and security in a superior way.

References

1. Alam, A.A., Khalid, B.S., Salam, C.M.: A modified version of playfair cipher using 7×4 matrix. Int. J. Comput. Theory Eng. **5**(4) (2013)
2. Basu, S., Ray, U.K.: Modified playfair cipher using rectangular matrix. Int. J. Comput. Appl. **46**(9), 0975–8887 (2012)
3. Bhattacharyya, S., Chand, N., Chakraborty, S.: A modified encryption technique using playfair cipher 10 by 9 matrix with six iteration steps. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **3**(2) (2014)
4. Shahil, T.: An efficient modification to playfair cipher. ULAB J. Sci. Eng. **5**(1) (2014)

Chapter 46

Modeling Internet of Things-Based Solution for Evading Congestion and Blockage in Waterways



**Madhuri Rao, Narendra Kumar Kamila, Sampa Sahoo,
and Kulamala Vinod Kumar**

Abstract The eight day long blockage that started on 23rd of March 2021 at the Suez Canal, which is one of the busiest waterways of the world, requires serious contemplation and efficient solutions. The incident could have been evaded by rigorous planning and essentially monitoring the health of ships while entering the canal with respect to weather conditions and minimizing the scope of human error. Additional waterways are infrastructures that cannot be built soon, and therefore, an Internet of Things-based (IoT) system that is proposed here could be a boon and also be implemented rather soon. This paper proposes an IoT-based system level design that could assist shipping industry in considering managing the dynamics of waterways more efficiently by forecasting possible risks related to weather systems and status of ships infrastructure including human error.

46.1 Introduction

The world is amidst a serious global recession owing to the current pandemic situation and accidents, and natural disasters make it even worse. The accident that happened on the Suez Canal affected businesses, supply chain systems, and also affected throughput of global world maritime trade negatively as more than 10% of global maritime trade passes through it [1]. A container ship fully loaded ran around the canal causing a blockage that lasted for seven days. One of the technical reason accounted was the decreased engine power accompanied with intensified high wind speeds that day in the canal, led the ship aground. The ship was carrying containers of 14,000 m² which converged it into an effective sail. Figure 46.1 depicts the satellite

M. Rao (✉) · K. V. Kumar

Department of Computer Science and Engineering, Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India

M. Rao · N. K. Kamila

Biju Patnaik University of Technology, Rourkela, Odisha, India

S. Sahoo

Departement of Computer Science, C. V. Raman Global University, Bhubaneswar, Odisha, India



Fig. 46.1 Satellite image of Suez Canal accident. *Source* [Cnes2021, Distribution ES Airbus DS)

image of the Suez Canal when the accident happened. It eventually led to a blockage of the waterway with line of ships waiting outside the canal for entry. Figure 46.2 depicts the satellite image of Suez Canal on February 1st, March 27th, and March 29th, 2021. The traffic jam was even visible from space however could not be avoided as it was not predicted.

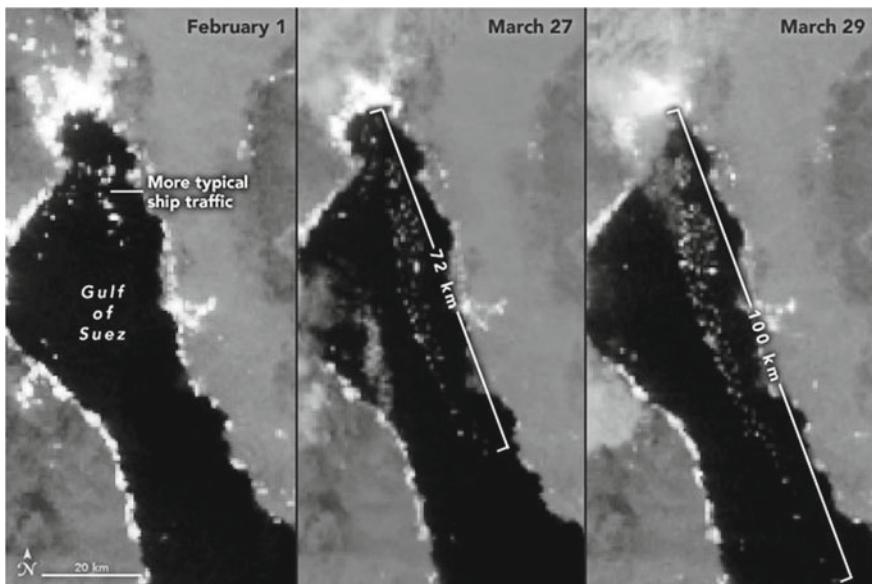


Fig. 46.2 Satellite image of traffic jam at Suez Canal. *Source* [go.nasa.gov/2PD6Mzg]

46.2 Background

Zhou et al. [2] have quantitatively analyzed the effect of wind and tide on how a ship behaves with respect to its size. However, they do not address issues such as human error and dynamics of the waterways. Hiekata [3] have carried out various studies of how well IoT can be deployed in maritime industry. According to [3], a ship is considered as a sociotechnical system which is a system of systems and is modeled using object process methodology. Though their research studies the impact of use of IoT in an individual ship, however, it does not explore the connections between other entities that play a crucial role in the maritime industry. Ships have automatic identification systems (AISs) that help in tracking ships movements and know its real-time location. Waterways that are crucial should monitor and authorize ship movements through it with respect to climate and weather conditions. Currently, waterways only monitor movements of ships ad traffic conditions but do not assess risk and vulnerability involved with its movement to a possible scale.

46.3 Internet of Things (IoT) Deployment Model for Forecasting

Choi et al. [4] have developed a framework for IoT-based container ship monitoring system. Deployment of Internet of Things (IoT) technology requires inter-connecting its various components such as devices or sensors, resource, controller service, database, Web service, and analysis component with an application. There are six types of IoT deployment models. While each model is useful in its own way, it is however sensible to deploy the system as per how it needs to function. The various stake holders of the tradition ship and waterway management system would need to access certain sections of the information collected by various sensors deployed on ships, on waterways and ports. Sea weather prediction information needs to be correlated with the ship condition and heath state of crew members in order to minimize the effect of any adverse situation when on sea. Zunnaid et al. [5] model a real-time Web application for monitoring ships and detect catastrophic event based on GPS tracker and water level indicators. Their model though a working prototype does not include Web sockets and does not address the vulnerability of weather imposed conditions. The heath condition of the cruise members can be monitored using sensors which can further help in reducing human errors to a considerable extent. In [6], a mental model is developed for an IoT-based harbor surveillance system, but the model does not consider traffic congestion issues.

Here, in this paper, an IoT deployment template for the proposed waterway management system is designed and is depicted as in Figure 46.3. The system is proposed to comprise of multiple sensor nodes installed on the ship to track the fuel status and engine status. Health monitoring sensors are proposed to be worn by crew members. Sea sickness and lack of sleep are some factors that affect heath

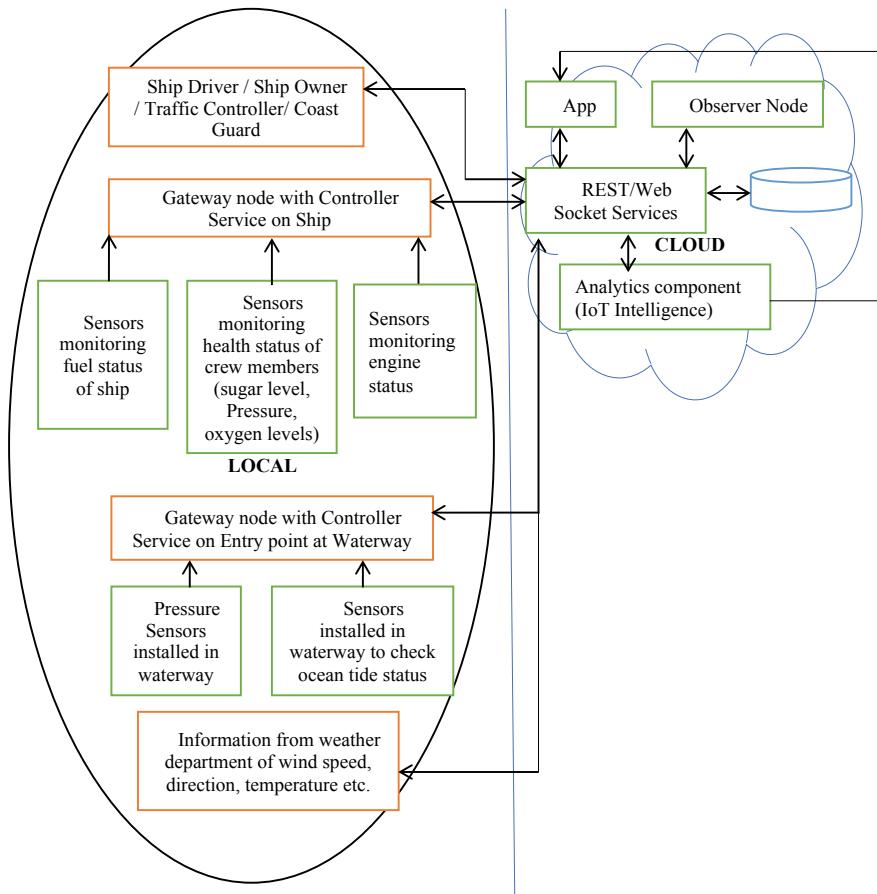


Fig. 46.3 Deployment template of IoT-enabled ship and waterway management system

of crew members that could lead to human error causing catastrophes. A member that is driving the ship and if not well can be thus identified. Prediction information from weather department is taken as the next input to the system. Here, direction and intensity of the wind and ocean tides are taken as input for predicting required engine power while crossing a waterway.

The AIS data of ships moving on the waterway are also available via cloud. Hence, the traffic controller of the waterway is now better equipped to deal with crisis like situation. A customized Web application can enable a ship owner, a coast guard, a traffic controller, and indeed the driver to take more rational decision with information collected from the various sensors. A centralized controller node monitors all the end sensors nodes. Figure 46.4 depicts how the proposed design of IoT-enabled ship and waterway management system can be deployed. The IoT deployment template explains how the various parts of the proposed system would be connected. Here,

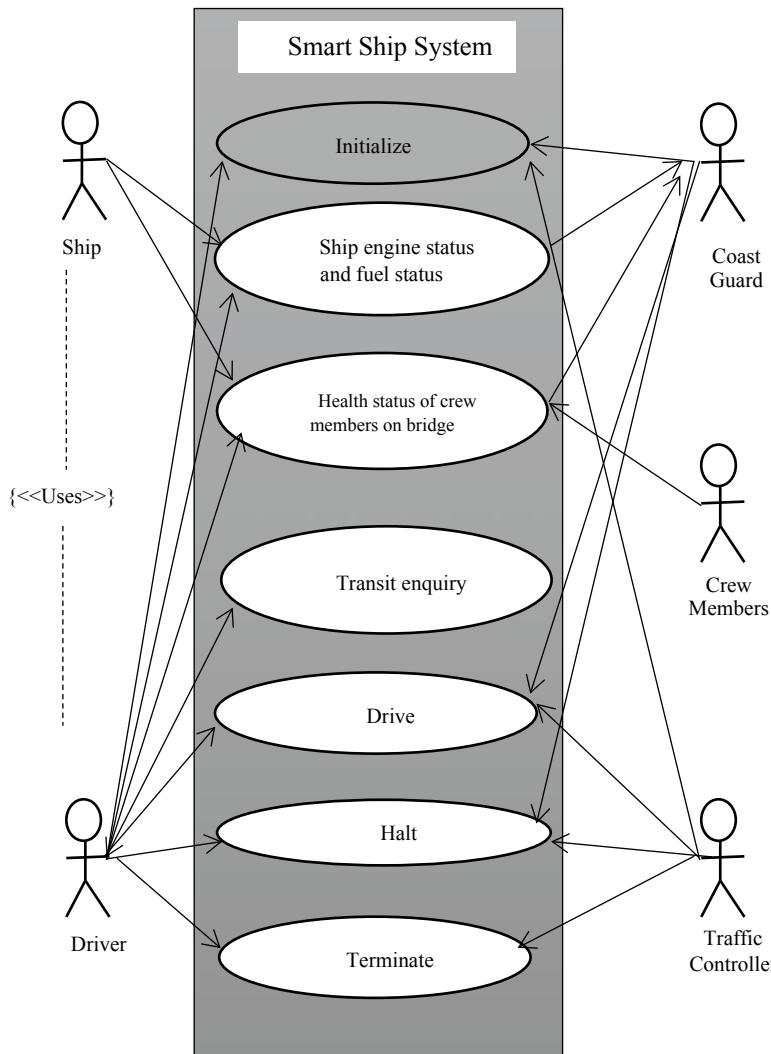


Fig. 46.4 Use case diagram of a smart ship system

there are certain tasks which need to perform locally, while certain tasks need to be performed on the cloud.

As per Figure 46.3, sensor nodes deployed on waterway and on ship collect data that is sent to the cloud via a gateway node that also has controller service. The gateway node enables sharing information to the cloud via a Web socket service or by using representational state transfer (REST) protocol. The weather details like wind speed, ocean condition, and wind direction are shared via the weather department of the given region via a REST/Web service technique. The observer

node located in cloud can access the information received from sensor nodes for review. The cloud can now develop intelligence of how the traffic conditions would be based on information accessed from weather department, status of fuel, status of ship engine, health of crew members. The dynamic nature of challenges imposed by environment and natural behavior of entities involved in this complex system can thus be addressed efficiently.

46.4 System Requirements

Though an IoT-based system is an efficient solution to manage waterways, but it requires various entities and users of the system to be well connected. A use case diagram is simple to depict the various users and how they interact with the system in an overview level. Kaiser et al. [7] have modeled the architecture of waterways management system as a finite state machine. The waterway management system they develop is however not equipped to deal with detection of errors and predict a possible fault. Kamolov et al. [8] develop an IoT-based ship berthing method using ultrasonic sensors. An empty berth is detected and notified to an arriving ship for anchoring. Yet again, weather conditions and human error are not considered in their model which is otherwise very amenable to implement. Hence, it is first essential to identify the needs and functions of each user and entities of the smart ship system.

We propose an IoT-based smart ship system embedded with various sensors could share the fuel status and engine condition with servers connected to waterways monitoring system. Figure 46.4 depicts the use case diagram of the proposed smart ship system. The users and entities of the system are the ship, driver, coast guard, crew members, and traffic controller. Each user has a set of functions or possible interactions with the system.

Figure 46.5 depicts the use case diagram of the proposed IoT-enabled waterways management system. Coast guard, ship driver, ship owner are the actors of the system. The actor ship receives recommendation from coast guard on what parameters need to be maintained during the course of its journey through the waterway. Required engine power to pass through the waterway would need a certain value which can be determined by understanding the combating forces of wind speed and direction.

The use of IoT in any application is specific in design and approach. Understanding its various components and how they are connected is utmost crucial. Use case diagram is the best when it comes to designing a specific system. For instance, a ship driver needs to start or initialize the ship, need to check fuel status and engine status, know the latest health status of crew members onboard, need to know the details of travel and therefore enquire of transit, drive, halt, and finally stop or terminate the ship. Transit enquiry involves information of what is loaded on the ship, details of passengers, expected arrival, and departure time.

The ship for instance carries numerous and various type of sensors. The data sensed by sensors deployed for monitoring fuel status, engine status, GPS details are required to be collected and analyzed when making an important decision with

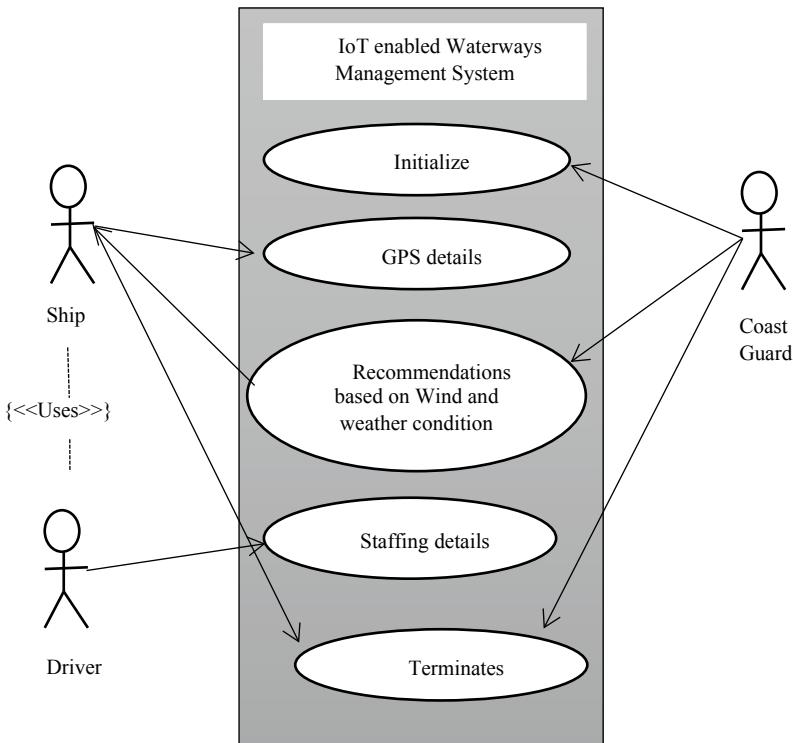


Fig. 46.5 Use case diagram of IoT-enabled waterway management system

respect to change of course or weather to halt at a given port for refueling. Similarly, sensors worn by crew members can help the ship doctor to know the health status of the crew members. A weak or ill member can be made to take rest and important tasks such as driving can be handed over to a fit person. Many accidents have happened due to human negligence which is directly dependent on the wellbeing of the person itself. The coast guard similarly has to make decisions about which ship needs to be allowed to pass through a canal or rather be halted. He can now also suggest what power the engine must hold in order to pass through the canal, based on the intelligence received from the cloud. The intelligence on these details is reached with the help of data received from ship, weather department, and from details of crew members.

46.5 Major Challenges and Opportunities

Designing and deploying an IoT-based system have many challenges to address. Security and updating the firmware of the sensor are also of concern. There are no

IoT regulations yet in place, and therefore, their security is often an open challenge. Lack of standards and compatibility between the devices cannot be ignored either. Bandwidth and connectivity issues can be overlooked only to an extent. While these are some of the challenges of an IoT system itself, for a waterway management system, there are other issues related to the region where deployed as well.

These are just a few challenges to name, but with a dynamic system as proposed the benefits are far appealing. A ship owner can track the movements of his ship, and a coast guard can ensure that his waterway has no traffic congestion. A situation as adverse as Suez Canal blockage can be avoided thereby ensuring supply chain systems are more efficient and reliable.

46.6 Conclusions and Future Works

The crisis that arose on Suez Canal is rather rare but also avoidable now with an efficient IoT-based waterways management system. Though sensors are already being used in ships, canals, and for weather monitoring, they are not being inferred together as of yet. A system that allows these sub-systems to be perceived together can help in reducing the impact of unforeseen crisis.

In this paper, a design of an IoT-based ship and waterways management system is presented. The various users of the system and their expectations from the system are explained using use case diagrams. A possible IoT deployment template is suggested which shows how the various users can interact from the system efficiently.

In our future work, we intend to simulate the proposed model with Web socket service as well.

References

1. Robinson, W.: News: Suez Canal blocked—traffic jam growing by the hour. News and Insights, Standard Club (25th March 2021)
2. Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S.P.: Impacts of wind and current on ship behavior in ports and waterways: a quantitative analysis based on AIS data. *Ocean Eng.* **213** (2020)
3. Hiekata, K., Wanaka, S., Mitsuyuki, T., Ueno, R., Wada, R., Moser, B.: Systems analysis for deployment of Internet of Things (IoT) in maritime industry. *J. Mar. Sci. Technol.* **26**, 459–469 (2021)
4. Choi, H.R., Moon, Y.S., Kim, J.J., Lee, J.K., Lee, K.B., Shin, J.J.: Development of IoT-based container tracking system for China's Belt and Road (B&R) initiative. *Marit. Policy Manage.* (2017)
5. Zunnaid, H., Hasan, W.U., Zaman, K.T., Haque, M.I., Aoyon, S.S.: Design and implementation of IoT based monitoring system for inland vessels using multiple sensor networks. In: 2nd International Conference on Smart Sensors and Application (ICSSA), pp. 38–43 (2018)
6. Rao, M., Kumar, N.K., Adhikari, N.: Evaluation of elicitation and specification of the requirements for an Internet of Things (IoT) system. In: Proceedings of Industry Interactive Innovations in Science, Engineering and Technology (I3SET2K19) (2020)

7. Kaiser, E., Austin, M., Papadimitriou, S.: Formal development and evaluation of narrow passageway system operations. *Eur. Transp.* **34**, 88–104 (2006)
8. Kamolov, A., Park, S.: An IoT based ship berthing method using a set of ultrasonic sensors. *Sensors* **14**(3), 5181 (2019)

Chapter 47

A Multi-objective Evolutionary Algorithm with Clustering-Based Two-Round Selection Strategy



M. Sri Srinivasa Raju, Kedar Nath Das, and Saykat Dutta

Abstract Existing multi-objective evolutionary algorithms can handle multi-objective optimization problems (MOPs) with regular Pareto fronts in which non-dominated solutions are distributed continuously over the objective space. This study proposes a clustering-based environmental selection for tackling MOPs with irregular Pareto fronts to address this issue. The main idea is to adaptively form clusters based on the solutions stored in the archive, and individual cluster will act as subpopulations. In each subpopulation, the PF will be regular in which will be easy to maintain diversity and accelerate convergence. The performance of the proposed algorithm is investigated in 19 different multi-objective test instances. Our results make evident the competitiveness of C2REA for multi-objective optimization, especially for problems with irregular Pareto fronts.

47.1 Introduction

Typically, the real-life optimization problems deal with multiple number of objectives often conflicting objectives. No single solution can optimize all the objective because of conflicting nature of objectives. So, a set of trade-off solutions are needed to represent among various conflicting objectives. Typically, a multi-objective optimization problem (MOP) can be described as follows:

$$\text{Maximize /Minimize } F(x) = (f_1(x), f_2(x), \dots, f_m(x))$$

$$\text{such that } x = (x_1, x_2, \dots, x_n) \in X, \quad X \subseteq \mathbb{R}^n$$

M. S. S. Raju (✉) · K. N. Das · S. Dutta

Department of Mathematics, NIT Silchar, Silchar, Assam 788010, India

K. N. Das

e-mail: kedarnath@math.nits.ac.in

where $F \subset \mathbb{R}^m$ is the objective vector which lies in the m -dimensional space, and the decision variable $x \in R^n$ lies in the n -dimensional space. The former one is known to be as objective vector space, and the last one is known to be as the decision variable space. The objective function F basically maps from n -dimensional space to m -dimensional space (i.e., $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$). A vector $a = (a_1, \dots, a_m)$ is said to dominate another vector $b = (b_1, \dots, b_m)$, denoted as $a \prec b$, if $\forall i \in \{1, \dots, m\}$, $a_i \leq b_i$ and for at least one $i \in \{1, \dots, m\}$, $a_i < b_i$. If there exist no $x \in X$ such that $F(x)$ dominates $F(x^*)$, then x^* is known as the Pareto-optimal points. In the multi-objective optimization studies, set of Pareto-optimal solutions is known as Pareto-optimal set (PS). The mapping of PS in objective space is known as Pareto front (PF).

During the last two decades, numerous multi-objective evolutionary algorithms (MOEAs) are developed to solve the MOPs. Because MOPs contain a number of Pareto-optimal solutions rather than a single optimal solution; population-based evolutionary algorithms are particularly well-suited to solve them. For many decision-makers, a reasonable approximation to the PF is extremely important so that they may assess several trade-off options before selecting the most suitable or preferable solution for their MOP. So, a large population is needed to represent the entire PF. However, this makes it difficult for a decision-maker to understand and make an informed choice on a chosen solution [1]. An EMO algorithm optimizes a MOP by pushing its population toward PF, which ensures convergence and diversity. Despite the fact that all the objectives are equally important, most modern EMO algorithms use a “convergence first, diversity second” approach. This means that when it comes for selecting the following generation’s population, solutions that contribute more to convergence than to variety are given preference. Multi-objective evolutionary algorithms can be classified into three primary categories as follows:

Dominance-based

In this group, developing a new dominance relationship is the first and most logical approach, as it will enhance selection pressure in MOEAs. Many modified dominance relationships have been proposed in recent years, such as extending the dominance area is α -dominance [2], CDAS [3], SDR [4], θ -Dominance [5]. Some other techniques are also fall under this category by modifying the diversity maintains criterion without changing the dominance relationship.

Decomposition-based

The decomposition-based methods decomposes a MOP into a sets of single-objective optimization problems that are optimized in a collaborator way using an evolutionary algorithm. To handle population diversity and convergence, these approaches employ a collection of predetermined, evenly distributed reference vectors. MOEA/D [6], MOEA/D-DE [7] are the first algorithm in this lineup. Apart from this, some of the algorithms used the dominance and decomposition collaborate manner. The Pareto dominance principle improves the convergence. Some popular algorithm in this category is NSGA-III [1], MOEA/DD [8], and RVEA [9]. Some other algorithms like MOEA/D-AWA [10], ANSGA-III [11] are also fall in this category where the uniform

reference vectors upgrades their position depending upon the problems. But, identification of inappropriate reference vectors and deciding their new position is still very challenging, which may cause of computational burden in generations.

Indicator-based

Indicator-based methods map the multi-dimensional values to a scalar values, instead of dominance criterion, which is used to assess the quality of the solutions. That individual scalar values carry the information-related convergence and diversity combined. IBEA [12] and HypE [13] are most famous MOEAs that fall under this category. Most of the indicator-based methods used hyper-volume. The main limitation of indicator-based approaches is the computing cost of the hyper-volume value. There are also some initiatives to reduce the computational cost and improve the quality of the indicator value. ISDE+ [14] indicator based is proposed by considering the combination of the sum of objectives and the SDE indicator, which are well known for their potential to faster convergence and diversity. A multi-indicator-based selection strategy 2REA [15] was proposed by Liang et al. which success to balance between convergence and diversity. A comprehensive survey can be found in [16–19].

However, when handling MOPs with irregular Pareto fronts, these approaches face issues in diversity maintaining. To enhance the diversity measure in the population, a cluster-based evolutionary algorithm is proposed in this article. When handling MOP with irregular PF, this approaches form clusters based on the discontinued parts of whole PF. After forming clusters, these clusters will be treated as subpopulation which will obviously has regular PF in their respective regions. So, maintaining diversity will became easy.

The main contribution of this article is highlighted as follows:

1. A new parameter-free clustering approach is proposed.
2. This clustering approach is incorporated in 2REA which results new algorithm, namely C2REA.

The reminder of this article is organized in the following manner. Section 47.2 defines some preliminaries and motivation for the current work. Section 47.3 presents the proposed algorithm in detail. Section 47.4 presents results and discussions. Finally, the conclusions and future work are discussed in Section 47.5.

47.2 Related Work and Motivation

47.2.1 Related Work

The 2REA was proposed in [15]. 2REA uses two-round environmental selection strategy to maintain diversity and convergence of the population. The merged population MerPop (=Parent+Offspring population) individuals will be selected one by

one as a result of the two indicators (i.e., distribution and convergence indicator) until number of selected solutions is N (required). In detail, the environmental selection starts by selecting best converged solution and stores in the empty population Pop (population for next generation). The adaptive position transformation (APT) distance of each non-dominated solution in MerPop ($d_{APT}(x, y)$) is computed in the first-round selection (where $d_{APT}(x, y)$ is the Euclidean distance of position transformed vectors $F(x)$ and $F(y)$). A larger d_{min} value indicates a smaller neighborhood density of the solution x . On the PF, the corner individuals are the most representative solutions. So, their d_{min} values are set to $+\infty$ such that they are certainly selected into Pop. All non-dominated solutions are sorted in descending order according to their d_{min} values. The topmost-ranked $N - |\text{Pop}|$ non-dominated solutions are moved to a temporary set. If the number of non-dominated solutions in the temporary set is less than $N - |\text{Pop}|$, all non-dominated individuals move to temporary set. The best converged solution s in the temporary set is shifted to Pop in the second-round selection. The two-round selection process is continued until Pop approaches the size of N .

47.2.2 Motivation

As discussed in Section 47.1, most of the approaches perform well on MOP with regular PFs. However, maintaining diversity will become issue for the most of the approaches while handling MOPs with irregular PFs. If we consider a disconnected front alone in a MOP with irregular PFs which have disconnected fronts, then it will be regular front. Based on this idea, the solutions presented in each disconnected part will form a cluster. These clusters will act as subpopulation with different sizes. Now, one may have doubt that how to set population size for each subpopulation. In this article, based on the Euclidean distance between any two corner solutions in each cluster, the length will be decided. In Figure 47.1, the clusters formed on different datasets with proposed clustering approach is presented. The different colors represent different clusters formed.

In the initial generation, the shape and characteristics of the problem are not known priori. So, introducing clustering-based environmental selection will become

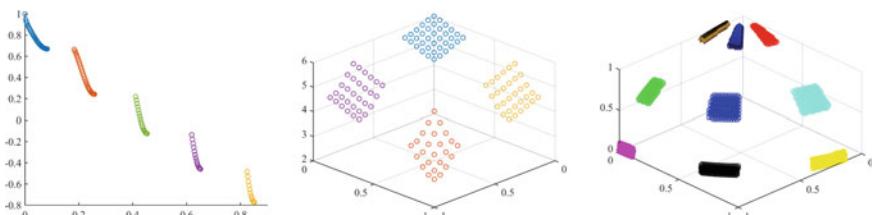


Fig. 47.1 Visualization of different clusters formed by the proposed clustering approach in different datasets

huge computational burden and may deteriorates the performance of the algorithm. To avoid this type of problems, the clustering-based environmental selection is introduced only when 80% of generations is completed. The proposed algorithm C2REA is explained in depth in the next section.

47.3 The Proposed Method

Algorithm 1 outlines the basic structure of the proposed algorithm, C2REA. C2REA starts with the randomly initialized population Pop. In every generation, an Offspring population is generated with the help of variation operators like binary crossover and polynomial mutation. Until the 80% of functional evaluations, best N solutions are selected from merged population R -based 2REA environmental selection. After that environmental selection based on cluster which are formed with the help of newly proposed clustering approach is carried out. The process of evolution is repeated until the set termination conditions are met. The key components of C2REA are discussed in the following sections.

Algorithm 1 General Framework of C2REA

Input: M - Number of Objective
 N - Population size
 $MaxEv$ - Maximum no. of evaluations
 $Archive$ - Empty Archive

Output: Pop - Final Population

1. $Pop \leftarrow InitializeRandomPopulation(N)$
2. **while** ($FE \leq MaxEv$)
3. $MP \leftarrow MatingSelection(Pop)$
4. $Offspring \leftarrow Variation(MP)$
5. $R \leftarrow Pop \cup Offspring$
6. **if** $FE \leq 0.8 * MaxEv$
7. $Pop \leftarrow EnvironmentalSelection(R, N)$
8. **else**
9. $[C, len] \leftarrow Clustering(Archive)$
10. $Pop \leftarrow ClusterBasedEnvironmentalSelection(C, len, Archive)$
11. **end**
12. $Archive \leftarrow ArchiveSelection(Archive, Pop)$
13. **end**

47.3.1 Clustering Strategy

The proposed clustering method (Algorithm 2) is a simple and efficient strategy. It starts by evaluating Euclidean distance between individual in the given dataset (Algorithm 2 line 1). After that select a random individual x in given set of data and cluster index as 1 along by inserting x index in the first cluster (Algorithm 2 line 2–3). Set the value of radius rad as maximum value of minimum distances in each row of the matrix $Dist$ and multiplied it with 2 (Algorithm 2 line 4–5). Insert the indices of the individuals in the $C(i)$ cluster which have the Euclidean distance less than rad and the set the distance of the indices in $Dist$ to inf so that selected individuals will not repeated again (Algorithm 2 line 7–10). If there exist no individuals which have Euclidean distance less than rad , then remove the selected individual from the archive and increment the cluster index i with and select a random individual (Algorithm 2 line 11–12). This process will be repeated until the length of the archive is 0.

Algorithm 2 Clustering

Input: $Archive$ - Archive
Output: C - Clusters
 len - No. of solutions need to be selected in each cluster

1. $Dist = EuclideanDistance(Archive, Archive)$
2. Select a random solution x from $Archive$ and set $i = 1$
3. $C(i) = x$
4. $rad = \max(\min(Dist, rows))$ \\ Finding maximum of minimum distances in each row
5. $rad = 2 * rad$
6. **while** $\text{length}(Archive) \sim = 0$
7. $C(i) = [C(i), r]$ \\ r is the set of indices which satisfies the condition $\underset{y \in Archive}{\text{Dist}}(C(i), y) \leq rad$
8. **if** $\text{isempty}(r) == 0$
9. $Dist(C(i), :) = inf$
10. **else**
11. $Archive(C(i)) = []$
12. Select a random solution x from $Archive$ and set $i = i + 1$
13. **end**
14. **end**

47.3.2 Clustering-Based Environmental Selection

Now, these clusters will act as L (no. of clusters) subpopulations and the required size of each subpopulation will calculated based on the Euclidean distance between two

corner solutions in each cluster (Algorithm 3 line 1–6). The value $\text{len}(i)$ indicates i^{th} subpopulation size. The corner solutions are found by selecting random individual in the cluster. After that find the individual which have largest Euclidean distance from randomly selected individual and mark it as 1st corner solution. The second-corner solution can be found by finding the individual which have largest Euclidean distance with the 1st corner solution. In each subpopulation, there will be more solutions than required. To select best individuals in each subpopulation, the environmental selection of 2REA has been employed which selects the individuals according to the required subpopulation size len (Algorithm 3 line 7–11).

Algorithm 3 Clustering Based Environmental Selection

Input: C - Clusters

Archive - Archive

Output: Pop - Next Generation Population

1. **for** $i = 1 : \text{length}(C)$
 2. $[a, b] = \text{findcorner}(C(i)) \setminus\setminus \text{Find any corner solution}$
 3. $\text{ed}(i) = \text{EuclideanDistance}(a, b)$
 4. **end**
 5. $t = \frac{N}{\text{sum}(\text{ed})}$
 6. $\text{len} = \text{round}(\text{len} * t)$
 7. **for** $i = 1 : \text{length}(C)$
 8. $\text{TempPop} = []$
 9. $\text{TempPop} = \text{EnvironmentalSelection}(\text{Archive}(C(i)), \text{len}(i))$
 10. $\text{Pop} = [\text{Pop}, \text{TempPop}]$
 11. **end**
-

47.4 Results and Discussions

The performance of the C2REA has been evaluated and compared with 8 state-of-the-art multi-objective evolutionary algorithms on two different multi-objective benchmarks DTLZ and ZDT [6]. All the compared algorithms and proposed algorithm were performed 30 independent runs on a PC with Intel (R) Core (TM) i5-7500 CPU @ 3.40 GHz and Windows 10 Pro 64-bit operating system with 8 GB RAM. For all two-objective test instances, maximum 250 generations are considered as stopping criteria. For three-objective test instances, 700 generations are considered as stopping criteria for DTLZ 1 and 1000 considered for DTLZ 3. For all other remaining problems (DTLZ 2, DTLZ 4-DTLZ7), it is set to 250.

In the article, the population size of all algorithms is set as 100 for two objective and 105 for three-objective benchmark problems. Variation operator like SBX and PM [20] with the distribution indices and probabilities set to $n_m = 20$, $p_c =$

1.0 , $n_c = 20$, and $p_m = 1/D$, respectively, are employed. To compare performance of the C2REA to the current state-of-the-art, a qualitative indicator such as the hyper-volume indicator [20] is employed. HV can assess the convergence ability as well as the diversity of solutions provided by different algorithms. The algorithm's supremacy is shown by a high HV value.

The proposed algorithm C2REA is compared against state-of-the-art algorithms like NSGA-II [20], IBEA [12], GrEA [21], MOEA/D-DE [7], MOEA/D-AWA [10], CAMOEA [22], ARMOEA [23], and 2REA [15] to demonstrate its efficacy. The experimental results of the benchmark suites are shown in Table 47.1 (mean and standard deviation values of HVs). Furthermore, we compared the suggested algorithm's performance to that of state-of-the-art algorithms using a statistical significance test with a confidence level of 0.05. The “+,” “=,” and “–” signs along with the HV values designate that the C2REA is statistically “worst,” “comparable,” or “better” with the state-of-the-art.

Table 47.1 Last row highlights the C2REA overall performance in terms of the number of cases in two-objective problems, where it is better, comparable, or worse than the equivalent state-of-the-art approach. In Table 47.2, the experimental results of all algorithm on three-objective problems are presented in terms of HV.

The proposed algorithm C2REA is better than or comparable to NSGA-II, IBEA, GrEA, MOEA/D-DE, MOEA/D-AWA, CAMOEA, ARMOEA, and 2REA in 94.73, 94.73, 94.73, 89.47, 68.42, 94.73, 73.68, and 94.73% in all 19 instances. Figures 47.2 and 47.3 present the plots of approximated PF by all compared algorithms.

The proposed algorithm significantly outperforms in all compared algorithms in two-objective ZDT2, ZDT4, DTLZ 2–6 and in three-objective DTLZ 4–7 test instances. There is no single instance that is statistically better than proposed algorithm in these test instance. MOEA/D-AWA performs consistently better than proposed algorithm in two-objective ZDT1, ZDT3, DTLZ1 and three-objective DTLZ 2–3 test instances.

47.5 Conclusions and Future Scope

In this article, a novel clustering-based environmental selection is proposed. The proposed approach is incorporated in 2REA framework. The proposed C2REA is compared with 8 state-of-the-art MOEAs including NSGA-II, IBEA, GrEA, MOEA/D-DE, MOEA/D-AWA, CAMOEA, ARMOEA, and 2REA on ZDT and DTLZ benchmark test suits. The empirical results show that C2REA is very comparative with the compared state-of-the-art algorithms. The proposed clustering-based environmental selection succeeds in obtaining best solutions with good trade-off between population diversity and convergence.

This article is mainly focused on solving MOPs. Providing a large number of solutions to the estimated PF makes it easier for decision-makers to understand and make an appropriate choice. However, this paper can be extended by incorporating the preference of the decision-maker. So, they can focus on the most convenient

Table 47.1 Comparison of HV and statistical results on ZDT and DTLZ test problems (—WIN, ==TIE, +LOSS)

#	M	D	ARMOEA	IBEA	NSGA-II	GrEA	MOEA/D-DE	MOEA/D-AWA	CAMOEA	2REA	C2REA
ZDT1	2	30	7.1991e-1 (1.08e-4)-	7.2007e-1 (1.07e-4)+	7.1884e-1 (2.20e-4)-	7.1502e-1 (9.41e-4)-	6.9262e-1 (8.34e-3)-	7.2017e-1 (1.32e-4)+	7.1922e-1 (2.27e-4)-	7.1990e-1 (1.30e-4)-	7.2001e-1 (6.85e-5)
ZDT2	2	30	4.4456e-1 (1.99e-4)=	4.4401e-1 (1.55e-4)-	4.4347e-1 (3.26e-4)-	4.4164e-1 (8.38e-5)-	4.2217e-1 (8.82e-3)-	4.3924e-1 (1.96e-2)=	4.4394e-1 (1.76e-4)-	4.4459e-1 (1.23e-4)=	4.4464e-1 (8.46e-5)
ZDT3	2	30	6.0737e-1 (2.72e-2)+	5.9817e-1 (1.20e-4)-	6.0219e-1 (1.62e-2)+	6.0318e-1 (2.27e-2)+	5.8550e-1 (1.56e-2)-	6.0170e-1 (1.61e-2)+	5.9851e-1 (4.60e-4)-	6.0220e-1 (1.62e-2)+	5.9939e-1 (1.56e-4)
ZDT4	2	10	7.1208e-1 (5.98e-3)-	6.5418e-1 (5.12e-2)-	7.1522e-1 (2.74e-3)-	5.6993e-1 (1.01e-1)-	3.0282e-1 (1.84e-1)-	7.1070e-1 (5.49e-3)-	7.1454e-1 (3.35e-3)-	7.1659e-1 (2.78e-3)-	7.1834e-1 (7.11e-4)
ZDT6	2	10	3.8751e-1 (9.77e-4)-	3.8761e-1 (1.16e-4)-	3.8748e-1 (4.99e-4)-	3.8606e-1 (1.80e-4)-	3.8867e-1 (3.53e-4)+	3.8827e-1 (2.88e-4)-	3.8724e-1 (6.92e-4)-	3.8837e-1 (2.88e-4)-	3.8855e-1 (1.02e-4)
DTLZ1	2	6	5.8234e-1 (2.50e-4)+	4.1452e-1 (2.22e-2)-	5.8127e-1 (3.29e-4)-	5.1632e-1 (9.17e-2)-	5.8260e-1 (3.40e-6)+	5.8240e-1 (2.08e-4)+	5.8160e-1 (4.58e-4)-	5.8206e-1 (1.59e-4)-	5.8219e-1 (8.97e-5)
DTLZ2	2	11	3.4721e-1 (1.47e-5)-	3.4622e-1 (2.41e-4)-	3.4649e-1 (1.95e-4)-	3.4389e-1 (6.94e-5)-	3.4699e-1 (3.38e-5)-	3.4725e-1 (3.44e-5)-	3.4667e-1 (1.75e-4)-	3.4726e-1 (6.89e-5)-	3.4731e-1 (4.65e-5)
DTLZ3	2	11	3.4576e-1 (1.04e-3)-	1.7280e-1 (6.60e-4)-	3.4539e-1 (9.92e-4)-	3.4304e-1 (8.40e-4)-	3.4541e-1 (9.25e-4)-	3.4644e-1 (9.13e-4)=	3.4561e-1 (1.32e-3)-	3.4583e-1 (9.37e-4)-	3.4662e-1 (3.90e-4)
DTLZ4	2	11	2.6176e-1 (1.23e-1)-	2.8672e-1 (1.10e-1)-	3.2095e-1 (7.80e-2)-	3.0185e-1 (9.59e-2)-	3.4673e-1 (6.66e-5)-	2.8744e-1 (1.10e-1)-	3.3816e-1 (4.67e-2)-	3.1308e-1 (8.86e-2)-	3.4731e-1 (4.95e-5)
DTLZ5	2	11	3.4721e-1 (1.75e-5)-	3.4606e-1 (2.57e-4)-	3.4653e-1 (1.46e-4)-	3.4389e-1 (6.67e-5)-	3.4698e-1 (2.91e-5)-	3.4724e-1 (2.41e-5)-	3.4664e-1 (2.14e-4)-	3.4729e-1 (9.62e-5)=	3.4733e-1 (5.03e-5)
DTLZ6	2	11	3.4721e-1 (2.19e-7)-	3.4268e-1 (6.96e-4)-	3.4634e-1 (2.13e-4)-	3.4382e-1 (5.47e-5)-	3.4721e-1 (1.19e-7)-	3.4724e-1 (2.27e-5)-	3.4717e-1 (1.27e-4)-	3.4740e-1 (7.80e-5)-	3.4744e-1 (4.18e-5)
DTLZ7	2	21	2.2943e-1 (2.71e-2)-	2.4043e-1 (1.22e-2)-	2.4271e-1 (5.96e-5)=	2.3944e-1 (8.21e-4)-	2.1263e-1 (3.30e-2)-	2.4262e-1 (2.99e-4)-	2.4270e-1 (4.53e-5)+	2.4273e-1 (5.82e-5)-	2.4273e-1 (2.51e-5)
\pm/\equiv		29/1	1110	110/1	1110	210/0	37/2	1110	1110	1/9/2	

Table 47.2 Comparison of HV and statistical results on DTLZ test problems (—WIN, ==TIE, +LOSS)

#	M	D	ARMOEA	IBEA	NSGA-II	GrEA	MOEA/D-DE	MOEA/D-AWA	CAMOEA	2REA	C2REA
DTLZ1	3	7	8.4414e-1 (2.56e-4)+	5.0628e-1 (5.88e-2)-	8.2517e-1 (4.21e-3)-	6.7754e-1 (1.48e-1)-	8.0724e-1 (1.77e-3)-	8.4043e-1 (3.35e-3)=	8.3927e-1 (1.25e-3)-	8.4044e-1 (1.56e-3)-	8.4138e-1 (6.99e-4)
DTLZ2	3	12	5.6261e-1 (1.14e-4)+	5.5856e-1 (1.30e-3)-	5.3450e-1 (5.09e-3)-	5.5983e-1 (6.14e-4)-	5.3189e-1 (1.49e-3)-	5.6433e-1 (3.34e-4)+	5.5204e-1 (1.85e-3)-	5.5941e-1 (8.03e-4)-	5.6016e-1 (4.32e-4)
DTLZ3	3	12	5.5917e-1 (2.47e-3)=	2.4812e-1 (1.95e-3)-	5.3249e-1 (5.37e-3)-	5.1579e-1 (7.03e-2)-	5.0267e-1 (9.67e-2)-	5.6227e-1 (2.48e-3)+	5.5241e-1 (4.20e-3)-	5.5849e-1 (1.38e-3)-	5.5990e-1 (9.46e-4)
DTLZ4	3	12	4.2211e-1 (1.53e-1)=	5.5891e-1 (1.14e-3)-	5.3633e-1 (3.23e-3)-	4.5985e-1 (1.31e-1)-	5.2380e-1 (2.39e-2)-	4.5996e-1 (1.28e-1)=	5.3757e-1 (8.44e-2)-	5.3162e-1 (7.35e-2)-	5.6026e-1 (5.12e-4)
DTLZ5	3	12	1.9952e-1 (9.22e-5)-	1.9903e-1 (1.73e-4)-	1.9923e-1 (1.68e-4)-	1.8831e-1 (4.51e-4)-	1.9550e-1 (6.85e-5)-	1.9608e-1 (3.64e-4)-	1.9923e-1 (1.59e-4)-	1.9980e-1 (9.21e-5)-	1.9989e-1 (5.02e-5)
DTLZ6	3	12	1.9991e-1 (1.41e-4)-	1.9631e-1 (1.00e-3)-	1.9958e-1 (1.59e-4)-	1.8766e-1 (5.80e-5)-	1.9594e-1 (2.17e-5)-	1.9595e-1 (3.02e-4)-	2.0000e-1 (9.44e-5)-	2.0001e-1 (6.89e-5)=	2.0004e-1 (3.57e-5)
DTLZ7	3	22	2.6668e-1 (1.96e-2)-	2.7544e-1 (1.01e-2)-	2.6688e-1 (5.57e-3)-	2.7043e-1 (7.32e-3)-	2.5978e-1 (2.34e-2)-	2.7320e-1 (7.73e-3)-	2.7567e-1 (1.42e-3)-	2.7977e-1 (1.10e-2)-	3.96e-4)
±=			23/2	07/0	07/0	07/0	23/2	07/0	07/0	06/1	

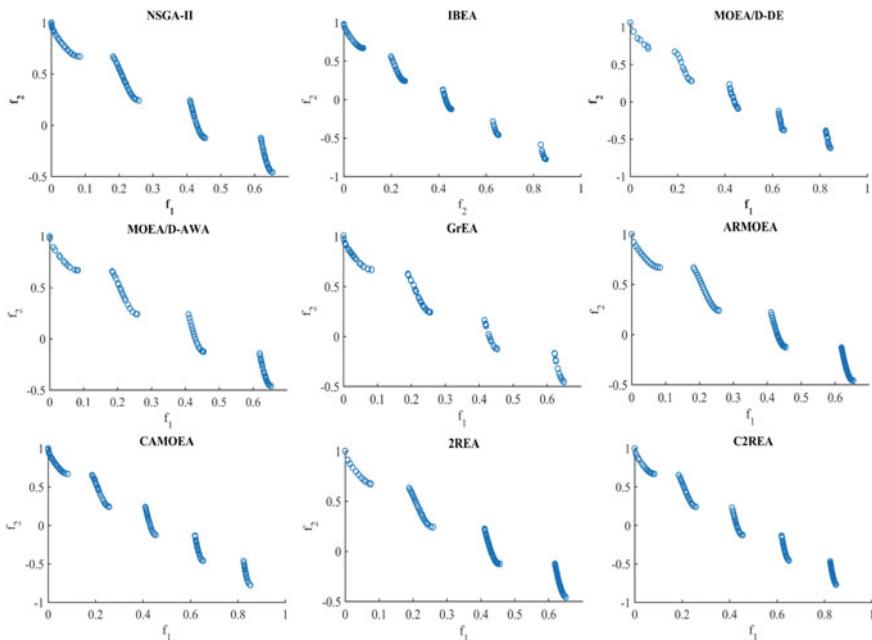


Fig. 47.2 Visualization of PF approximation by nine algorithms on the two-objective ZDT3

regions of the PF. The incorporation of the preference in the C2REA can be done by providing the user-defined cluster centers. Having user-defined preference in the algorithm makes the algorithm to avoid unnecessary computational efforts. This incorporation can help the decision-makers to improve the quality of decisions.

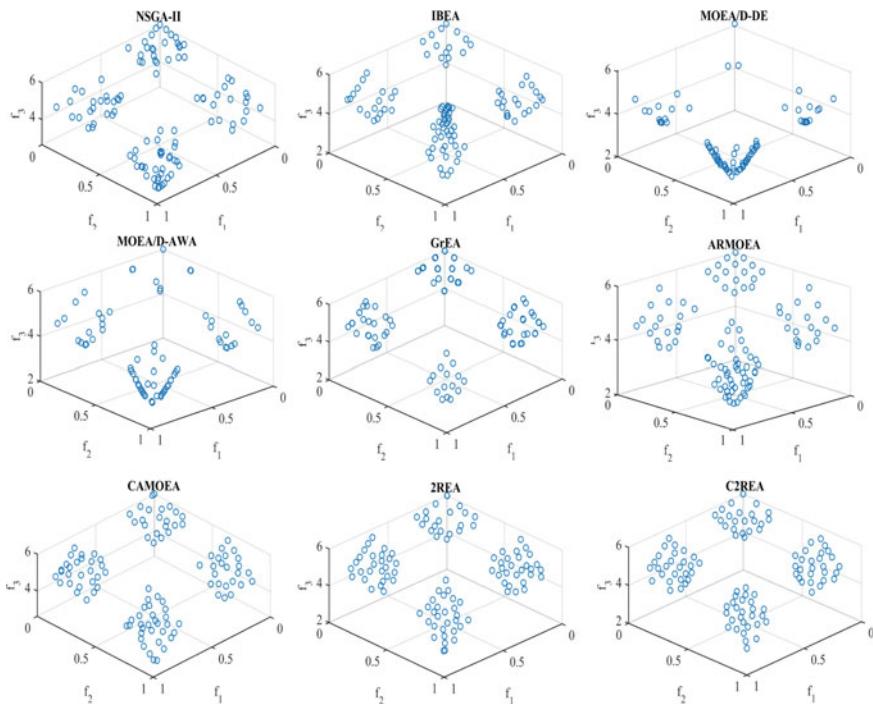


Fig. 47.3 Visualization of PF approximation by nine algorithms on the three objective

References

1. Deb, K., Jain, H.: An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, Part I: solving problems with box constraints. *IEEE Trans. Evol. Comput.* **18**(4), 577–601 (2014)
2. Ikeda, K., Kita, H., Kobayashi, S., Failure of pareto-based moeas: does non-dominated really mean near to optimal? In: Proceedings of the 2001 Congress on Evolutionary Computation, **2**, 957–962 (2001)
3. Sato, H., Aguirre, H.E., Tanaka, K.: Controlling dominance area of solutions and its impact on the performance of MOEAs. In: Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T. (eds.) Evolutionary Multi-Criterion Optimization. EMO 2007. Lecture Notes in Computer Science, vol. 4403. Springer, Berlin, Heidelberg (2007)
4. Tian, Y., Cheng, R., Zhang, X., Su, Y., Jin, Y.: A strengthened dominance relation considering convergence and diversity for evolutionary many-objective optimization. *IEEE Trans. Evol. Comput.* **23**(2), 331–345 (2019)
5. Yuan, Y., Xu, H., Wang, B., Yao, X.: A new dominance relation-based evolutionary algorithm for many-objective optimization. *IEEE Trans. Evol. Comput.* **20**(1), 16–37 (2016)
6. Zhang, Q., Li, H.: MOEA/D: a multi-objective evolutionary algorithm based on decomposition. *IEEE Trans. Evol. Comput.* **11**(6), 712–731 (2007)
7. Li, H., Zhang, Q.: Multi-objective optimization problems with complicated pareto sets, MOEA/D and NSGA-II. *IEEE Trans. Evol. Comput.* **13**(2), 284–302 (2009)
8. Li, K., Deb, K., Zhang, Q., Kwong, S.: An evolutionary many-objective optimization algorithm based on dominance and decomposition. *IEEE Trans. Evol. Comput.* **19**(5), 694–716 (2015)

9. Cheng, R., Jin, Y., Olhofer, M., Sendhoff, B.: A reference vector guided evolutionary algorithm for many-objective optimization. *IEEE Trans. Evol. Comput.* **20**(5), 773–791 (2016)
10. Qi, Y., Ma, X., Liu, F., Jiao, L., Sun, J., Wu, J.: MOEA/D with adaptive weight adjustment. *Evol. Comput.* **22** (2013) https://doi.org/10.1162/EVCO_a_00109
11. Jain, H., Deb, K.: An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach, Part II: Handling constraints and extending to an adaptive approach. *IEEE Trans. Evol. Comput.* **18**(4), 602–622 (2014)
12. Zitzler, E., Künzli, S.: Indicator-based selection in multiobjective search. In: Yao, X., et al. (eds.) *Parallel Problem Solving from Nature—PPSN VIII*. PPSN 2004. Lecture Notes in Computer Science, vol. 3242. Springer, Berlin, Heidelberg (2004)
13. Bader, J., Zitzler, E.: HypE: An algorithm for fast hypervolume-based many-objective optimization. *Evol. Comput.* **19**(1), 45–76 (2011)
14. Pamulapati, T.R., Mallipeddi, R., Suganthan, P.: ISDE+—An indicator for multi and many-objective optimization. *IEEE Trans. Evol. Comput.* (2018)
15. Liang, Z., Hu, K., Ma, X., Zhu, Z.: A many-objective evolutionary algorithm based on a two-round selection strategy. *IEEE Trans. Cybern.* (2019)
16. Yang, S., Li, M., Liu, X., Zheng, J.: A grid-based evolutionary algorithm for many-objective optimization. *IEEE Trans. Evol. Comput.* **17**(5), 721–736 (2013)
17. Dutta S., Das K.N.: A survey on pareto-based EAs to solve multi-objective optimization problems. In: *Advances in Intelligent Systems and Computing*, vol. 817. Springer, Singapore (2019)
18. Trivedi, A., Srinivasan, D., Sanyal, K., Ghosh, A.: A survey of multiobjective evolutionary algorithms based on decomposition. *IEEE Trans. Evol. Comput.* **21**(3), 440–462 (2017)
19. Ma, X., Yu, Y., Li, X., Qi, Y., Zhu, Z.: A survey of weight vector adjustment methods for decomposition-based multiobjective evolutionary algorithms. *IEEE Trans. Evol. Comput.* **24**(4), 634–649 (2020)
20. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**(2), 182–197 (2002)
21. Tian, Y., Cheng, R., Zhang, X., Cheng, F., Jin, Y.: An Indicator-based multiobjective evolutionary algorithm with reference point adaptation for better versatility. *IEEE Trans. Evol. Comput.* **22**(4), 609–622 (2018)
22. While, L., Hingston, P., Barone, L., Huband, S.: A faster algorithm for calculating hypervolume. *IEEE Trans. Evol. Comput.* **10**(1), 29–38 (2006)
23. Hua, Y., Jin, Y., Hao, K.: A clustering-based adaptive evolutionary algorithm for multiobjective optimization with irregular pareto fronts. *IEEE Trans. Cybern.* **49**(7), 2758–2770 (2019)

Chapter 48

An Efficient Evolutionary Technique for Solving Non-linear Fixed Charge Transportation Problem



Rajeev Das and Kedar Nath Das

Abstract One of the popular variants of fixed charge transportation problem is the non-linear fixed charge transportation problem (NFCTP). It has a ‘variable cost’ proportional to the quadratic of the amount of goods shipped from supplier to customer, and the other is ‘fixed cost’ independent of the amount shipped. This paper aims at providing an optimal technique for solving single-objective NFCTP. Based on the series hybridization of Jaya search and quadratic approximation operator, an efficient evolutionary algorithm is developed as a solution approach in this work. This new approach takes the advantage of well-balanced exploration and exploitation mechanism and exhibits good performance with respect to solution quality and computational run times. At the end, numerical instances are solved to illustrate the robustness of the proposed algorithm.

48.1 Introduction

This section comprises of two sub-sections. One of which is the problem formulation for non-linear fixed charge transportation problem (NFCTP), and the other is the literature review based on different mathematical variants and algorithmic approaches for solving fixed charge transportation problem (FCTP) including NFCTP.

48.1.1 Problem Formulation

As reported in [1], the mathematical model of NFCTP based on the notations from Table 48.1 is formulated as follows:

R. Das (✉) · K. N. Das

Department of Mathematics, National Institute of Technology, Silchar, Assam, India

K. N. Das

e-mail: kedarnath@math.nits.ac.in

Table 48.1 Notations for the model (1) of NFCTP

Indices	m	Number of suppliers
	n	Number of customers
	i	Supplier identifier, $i=1,2,\dots,m$
	j	Customer identifier, $j=1,2,\dots,n$
Parameters	a_i	Amount of a homogeneous product available at the i th supplier
	b_j	Demand of the j th customer
	c_{ij}	Unit transportation cost from the i th supplier to the J th customer
	f_{ij}	Fixed transportation cost from the i th supplier to the j th customer
Decision variables	x_{ij}	Amount of units shipped from i th supplier to the j th customer
	y_{ij}	1, if goods shipped from i th supplier to j th customer 0, otherwise

$$\left. \begin{array}{l} \min \sum_{i=1}^m \sum_{j=1}^n (c_{ij}x_{ij}^2 + f_{ij}y_{ij}) \\ \text{subject to} \\ \sum_{j=1}^n x_{ij} \leq a_i; i = 1, 2, 3, \dots, m \\ \sum_{i=1}^m x_{ij} \geq b_j; j = 1, 2, 3, \dots, n \\ x_{ij} \geq 0, y_{ij} = \begin{cases} 1, & \text{if } x_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \end{array} \right\} \quad (48.1)$$

In Model (1), the objective is to minimize the total transportation cost under the situation that the variable cost is directly in proportion to the quadratic of the amount shipped and a fixed cost applied only when a route is used. The first two constraints indicate the supply and demand constraints of the transportation network, respectively. Lastly, the real and binary variables used in the NFCTP model are defined by the third constraint.

48.1.2 Literature Review

The classical transportation problem (TP) is a simple linear programming problem (LPP) which can be easily solved using traditional solution approaches. The origin of TP dates back to the early 40s of the twentieth century where it got first proposed in [2]. Since then, a lot of studies from different perspective of numerous researchers have given rise to several variants of TP and their respective solution methods. For example, [3–6] are some of the recently published optimization methods for solving TPs.

One of the well-known variants of TP is the fixed cost transportation problem (FCTP). Basically, FCTP is formed of a ‘variable cost’ in proportion to the amount

of goods transported from supplier to customer, and a ‘fixed cost’ is independent of the amount transported. The NP-hard nature of FCTP [7] has drawn major attention of different researchers in the recent era. Quality works belonging to both single-objective FCTP and multi-objective fixed cost transportation problem (MOFCTP) have been performed since the last two decades. Many multi-criteria decision-making methods (MCDMs) have been introduced in this field due to the influence of social factors, environmental factors, time, demands etc. Some notable works of the twenty-first century based on complex mathematical formulations of FCTP (both single-objective and multi-objective) and solution techniques include [1, 8–20].

Non-linear fixed charge transportation problem (NFCTP) is an extension of the FCTP considered in the works of [9–12]. In each of their works, an improvement can be seen via different genetic algorithm (GA)-based approaches to reach the near optimal objective values of the set of problems considered. However, in this paper, three of their popular NFCTP’s with different problem sizes have been computationally analyzed to yield way better objective values using a newly proposed hybrid meta-heuristic. In order to achieve this, a series hybridization of Jaya algorithm [21] and quadratic approximation (QA) [22] operator is being employed in alternative generations. In addition, an S-type dynamic penalty function [23] has been chosen as the constraint handling technique.

The remaining sections of this paper are organized in a way that Sect. 48.2 illustrates the working mechanism of the proposed approach, Sect. 48.3 covers the application of the proposed approach through different numerical instances, and the results and discussion are included in Sect. 48.4. Finally, the conclusive remarks along with future scopes are presented in Sect. 48.5.

48.2 Proposed Approach

48.2.1 Motivation

Based on the recent works available in literature, Jaya algorithm [21] provides promising results in solving constrained and unconstrained optimization problems. Besides, its parameter-free nature makes it quite easy to implement. However, Jaya lacks in achieving the global optimum due to an imbalance of exploration and exploitation which results in getting stuck at some local optima. Therefore, the authors felt the necessity to choose an operator in order to improve the solution quality. Thus, the quadratic approximation (QA) [22] operator is chosen in this respect which had been previously addressed in other notable works like [24, 25]. The resultant hybrid algorithm reaps the combined advantages of inherent exploration and exploitation capacity of both Jaya and QA and guides towards faster convergence. Moreover, in order to handle the constraints of the NFCTP, an efficient dynamic penalty function [23] is added to this algorithm.

48.2.2 Quadratic Approximation Based Jaya Algorithm

The above facts encouraged the authors to design the new hybrid approach and its methodological steps are covered in this section. Jaya search starts with the initialization of population followed by functional evaluations. In the intermediate steps, an update equation is used to change the solutions and finally terminates with greedy selection procedure. However, the newly designed technique alternatively employs the steps of Jaya algorithm in the odd generations and QA in the even generations. So this new approach is termed as QA-based Jaya (JaQA) algorithm. The methodological steps of JaQA are outlined below:

- Step-1: Set the Population size (m), Number of decision variables (n) and the maximum number of generations (MaxGen), Current Generation (Gen=1).
- Step-2: Create the initial population randomly.
- Step-3: For odd value of Gen go to Step-4, else go to Step-5. Step-4: (Jaya)
 - i. Assess the values of fitness function in the population and pick the best and worst solutions.
 - ii. Next solution (X_{new}^k) to be updated by

$$X_{\text{new}}^k = X_{ij}^k + \text{rand}_1 * (X^k - |X_{ij}^k|) - \text{rand}_2 * (X^k - |X_{ij}^k|)$$

Here, X^k is the value of the best candidate solution for the j th variable, X^k is the value of the worst candidate solution of the j th variable, and rand1 and rand2 are two random numbers for the j th variable during the k th iteration in $[0, 1]$.

- iii. Implement Greedy Selection.
- iv. Go to Step 6.

- Step-5: (QA)
 - i. To generate the minima (child) of the quadratic surface passing through best point R1, and two random unique points R2 and R3 use

$$\text{child} = 0.5 \times \left(\frac{(R_2^2 - R_3^2)f(R_1) + (R_3^2 - R_1^2)f(R_2) + (R_1^2 - R_2^2)f(R_3)}{(R_2 - R_3)f(R_1) + (R_3 - R_1)f(R_2) + (R_1 - R_2)f(R_3)} \right)$$

Here, $f(R_1)$, $f(R_2)$ and $f(R_3)$ denote the respective functional values at R_1 , R_2 and R_3 .

- ii. Determine the value of $f(\text{Child})$ and replace this value with the worst fitness value $f(X^k)$ in the population.

- Step-6: Gen = Gen + 1. If (Gen > MaxGen) Terminate, else go to Step-3.
- Step-7: Report the best solution of the present population and exit.

The pseudo code of JaQA is presented in Fig. 48.1 below

Pseudo Code for JaQA
begin
Gen=1, MaxGen
Generate initial random population
While (Gen ≤ MaxGen) do
Fitness evaluation
If (Gen = Odd) implement Jaya Algorithm
If (Gen = Even) implement QA
Gen=Gen+1
end do
Record the best solution
end begin.

Fig. 48.1 Pseudo Code for the proposed hybrid approach JaQA

48.3 Numerical Instances

This section illustrates the application of the proposed approach JaQA with three numerical instances of NFCTP with different sizes. Table 48.2 displays the total supply/demand for these three instances. These problems are extracted from [11], and their near optimal values are compared with different approaches from the literature. These three NFCTPs with sizes 3×4 , 4×5 and 5×10 are mentioned in Tables 48.3, 48.4, and 48.5, respectively. The program code for JaQA was executed in Intel (*R*) Core (TM) i5-7500 CPU (3.40 GHz) processor with Python 3. 9.

Table 48.2 Total supply/demand ranges of fixed cost

Problem dimension ($m \times n$)	Total supply/demand	Range
3×4	100	[0, 50]
4×5	275	[50, 100]
5×10	1485	[150, 500]

Table 48.3 Unit costs in 3×4 NFCTP

(c_{ijk}, f_{ijk})	1	2	3	4	Supplies
1	(8, 60)	(8, 88)	(3, 95)	(5, 76)	50
2	(3, 86)	(5, 84)	(4, 70)	(8, 92)	20
3	(8, 67)	(4, 89)	(5, 99)	(3, 89)	30
Demands	20	40	30	10	

Table 48.4 Unit costs in 4×5 NFCTP

(c_{ijk}, f_{ijk})	1	2	3	4	5	Supplies
1	(8, 60)	(4, 88)	(3, 95)	(5, 76)	(8, 97)	57
2	(3, 86)	(6, 84)	(4, 70)	(8, 92)	(5, 76)	93
3	(8, 67)	(4, 89)	(5, 99)	(3, 89)	(4, 100)	50
4	(4, 86)	(6, 84)	(8, 70)	(3, 92)	(3, 88)	75
Demands	88	57	24	73	33	

48.4 Results and Discussion

For the 3×4 problem, JaQA yielded a near optimal cost value of 6878 with $x_{11}=11$, $x_{12}=8$, $x_{13}=21$, $x_{14}=10$, $x_{21}=5$, $x_{22}=11$, $x_{23}=4$, $x_{31}=21$, $x_{33}=5$ and other decision variables are all zeros.

The 4×5 problem reached a near optimal cost value of 25355 using JaQA with $x_{11}=21$, $x_{12}=22$, $x_{13}=2$, $x_{14}=12$, $x_{21}=27$, $x_{22}=13$, $x_{23}=16$, $x_{24}=17$, $x_{25}=20$, $x_{31}=16$, $x_{32}=10$, $x_{34}=22$, $x_{35}=2$, $x_{41}=24$, $x_{42}=12$, $x_{43}=6$, $x_{44}=22$, $x_{45}=11$, and other decision variables are all zeros.

And lastly, the NFCTP with size 5×10 is executed via JaQA to give a near optimal cost value of 245038 with $x_{15}=40$, $x_{17}=27$, $x_{19}=90$, $x_{21}=90$, $x_{22}=37$, $x_{23}=53$, $x_{28}=57$, $x_{29}=56$, $x_{32}=23$, $x_{33}=37$, $x_{34}=90$, $x_{41}=90$, $x_{42}=90$, $x_{44}=90$, $x_{45}=90$, $x_{46}=88$, $x_{49}=37$, $x_{410}=90$, $x_{51}=45$, $x_{54}=35$, $x_{57}=30$, $x_{58}=67$, $x_{59}=90$, $x_{510}=43$, and other decision variables are all zeros. All these three obtained values of near optimal costs are now compared with existing results from literature via different evolutionary approaches in Table 48.6.

The boldface letters in Table 48.6 signify better solution. Clearly, it can be observed from Table 48.6 that JaQA surpassed all the other approaches in terms of achieving a minimum optimal value cost of transportation. However, for the third problem (5×10), JaQA obtained the same optimal value as NFCTP-HGA [11].

From Table 48.6, it appears that NFCTP-HGA is the next best approach following JaQA for solving the aforementioned NFCTP's; therefore, a comparison of CPU run time (s) between NFCTP-HGA and JaQA is displayed in Table 48.7. Unlike NFCTP-HGA, JaQA showed impressive results in the case of computational run times and better results are marked bold.

48.4.1 Future Scope of JaQA

The effectiveness of JaQA is mainly shown here for single-objective NFCTP through numerical results. It converges with a faster rate, reducing computational effort, and hence, making it a hassle-free approach. However, JaQA is expected to generate equally better compromise solutions for MOFCTP as well. Several other parameters like transporting time, customer budget, deterioration rate of goods, safety factors

Table 48.5 Unit costs in 5×10 NFCTP

(c_{ijk}, f_{ijk})	1	2	3	4	5	6	7	8	9	10	Supply
1	(8160)	(4488)	(3295)	(5376)	(2297)	(1360)	(3199)	(5292)	(2481)	(6162)	157
2	(3451)	(3172)	(4265)	(8487)	(5176)	(3260)	(5280)	(1300)	(4354)	(5201)	293
3	(7167)	(4250)	(5499)	(3189)	(4340)	(2216)	(4177)	(3495)	(7170)	(3414)	150
4	(1386)	(2184)	(8370)	(1292)	(3188)	(1206)	(4340)	(6205)	(8465)	(2273)	575
5	(4156)	(5244)	(6460)	(3382)	(3270)	(4180)	(2235)	(1355)	(2276)	(1190)	310
Demand	225	150	90	215	130	88	57	124	273	133	

Table 48.6 Comparison of near optimal transportation cost

NFCTP	$m \times n$	st-GA [9]	st-GA [10]	pb-GA [12]	NFCTP-HGA [11]	JaQA
1	3×4	–	–	–	7229	6878
2	4×5	37,090	43,940	38,282	27,127	25,355
3	5×10	304,200	418,887	304,200	245,038	245,038

Table 48.7 Comparison of CPU run time (s)

NFCTP	$m \times n$	CPU run time (s)	
		NFCTP-HGA [11]	JaQA
1	3×4	120	1.28
2	4×5	120	5.05
3	5×10	480	10.11

etc., can also be added to the problem and then solved under multi-objective environment. Popular MCDM techniques like analytical hierarchy process (AHP) [26], PROMOTHEE [26], ELECTRE [26], TOPSIS [26] etc., can be used in conjunction with JaQA. Easy implementation of JaQA makes it very simple to program the code for the case of MOFCTP with MCDM techniques. Hopefully, this algorithm is expected to gain victory in other fields of decision-making problems with the same optimal quality of solutions.

48.5 Conclusion

Due to the difficulty of solving NFCTP's using traditional optimization techniques, a robust hybrid meta-heuristic is proposed in this paper. Jaya search and QA operator have been applied in alternative generations to reach the global optimum. Additionally, a dynamic penalty factor is applied for the constrained handling purpose. For a thorough comparison of results with respect to transportation cost and computational time, experiments with three popular NFCTP's are carried out. Observed results indicate that JaQA explicitly outperforms other renowned algorithms with faster convergence. This paper mainly focuses on single-objective NFCTP. However, as a future scope of research, the proposed algorithm can also be applied for MOFCTP or other decision-making problems with the popular MCDM methods.

References

1. Adlakha, V., Kowalski, K.: A simple algorithm for the source induced fixed charge transportation problem. *J. Oper. Res. Soc.* **55**(12), 1275–1280 (2004)
2. Hitchcock, F.: The Distribution of a product from several sources to numerous localities. *J. Math. Phys.* **20**, 224–230 (1941)
3. Ahmed, M.M., Tanvir, A.S.M., Sultana, S., Mahmud, S., Uddin, M.S.: An effective modification to solve transportation problems: a cost minimization approach. *Annu. Pure Appl. Math.* **6**(2), 199–206 (2014)
4. Babu, M.A., Das, U.K., Khan, A.R., Uddin, M.S.: A simple experimental analysis on transportation problem: a new approach to solve zero supply or demand for all transportation algorithm. *Int. J. Eng. Res. Appl.* **4**(1), 1344–1348 (2014)
5. Ahmed, M.M., Khan, A.R., Ahmed, F., Uddin, M.: A new approach to solve transportation problems. *Open J. Optim.* **5**, 22–30 (2016)
6. Das, K.N., Das, R., Acharya, D.B.: Least-looping-stepping-stone based ASM approach for transportation and triangular intuitionistic fuzzy transportation problems. *Complex Int. Syst.* (2021). <https://doi.org/10.1007/s40747-021-00472-0>
7. Hirsch, W., Dantzig, G.: The fixed charge problem. *Nav. Res. Logist.* **15**, 413–424 (1968)
8. Glover, F., Amini, M., Kochenberger, G.: Parametric ghost image processes for fixed charge problems: a study of transportation networks. *J Heuristics* **11**(4), 307–336 (2005)
9. Jo, J.B., Li, Y., Gen, M.: Nonlinear fixed charge transportation problem by spanning tree-based genetic algorithm. *Comp. Ind. Eng.* **53**(2), 290–298 (2007)
10. Hajighe-Kesheli, M., Molla-Alizadeh-Zavardehi, S., Tavakkoli-Moghaddam, R.: Addressing a nonlinear fixed-charge transportation problem using a spanning tree-based genetic algorithm. *Comput. Ind. Eng.* **59**, 259–271 (2010)
11. Xie, F., Jia, R.: Nonlinear fixed charge transportation problem by minimum cost flow-based genetic algorithm, *Comp. Ind. Eng.* (2012). <https://doi.org/10.1016/j.cie.2012.04.016>
12. Lotfi, M.M., Tavakkoli-Moghaddam, R.: A genetic algorithm using priority-based encoding with new operators for fixed charge transportation problems. *Appl. Soft Comput.* **13**, 2711–2716 (2013)
13. Yang, L., Feng, Y.: A bicriteria solid transportation problem with fixed charge under stochastic environment. *Appl. Math. Model.* **31**(12), 2668–2683 (2007)
14. Adlakha, V., Kowalski, K., Wang, S., Lev, B., Shen, W.: On approximation of the fixed charge transportation problem. *Omega* **43**, 64–70 (2014)
15. Kowalski, K., Lev, B., Shen, W., Tu, Y.: A fast and simple branching algorithm for solving small scale fixed-charge transportation problem. *Oper. Res. Perspect.* **1**(1), 1–5 (2014)
16. Dalman, H., Sivri, M.: Multi-objective solid transportation problem in uncertain environment. *Iran J. Sci. Technol. Trans. A. Sci.* **41**(2), 505–514 (2017)
17. Angulo, G., Vyve, M.: Fixed-charge transportation problems on trees. *Oper. Res. Lett.* **45**, 275–281 (2017)
18. Akbari, M., Molla-Alizadeh-Zavardehi, S., Niroomand, S.: Metaheuristic approaches for fixed-charge solid transportation problem in two-stage supply chain network. *Oper. Res.* (2017). <https://doi.org/10.1007/s12351-017-0332-7>
19. Majumder, S., Kundu, P., Kar, S., Pal, T.: Uncertain multi-objective multi-item fixed charge solid transportation problem with budget constraint. *Soft Comput.* **23**(10), 3279–3301 (2019)
20. Balaji, A., Nilakantan, J., Nielsen, I., Jawahar, N., Ponnambalam, S.: Solving fixed charge transportation problem with truck load constraint using metaheuristics. *Annu. Oper. Res.* **273**(1–2), 207–236 (2019)
21. Rao, R.: Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. *Int. J. Ind. Eng. Comput.* **7**, 19–34 (2016)
22. Mohan, C., Shanker, K.: A random search technique for global optimization based on quadratic approximation. *Asia Pac. J. Oper. Res.* **11**, 93–101 (1994)
23. Liu, J., Teo, K.L., Wang, X., Wu, C.: An exact-penalty function based differential search algorithm for constrained global optimization. *Soft Comput.* **20**, 1305–1313 (2016)

24. Deep, K., Das, K.: Performance improvement of real-coded genetic algorithm with quadratic approximation based hybridization. *Int. J. Intell. Defence Support Syst.* **2**(4), 319–334 (2009)
25. Bansal, J., Deep, K.: Quadratic approximation PSO for economic dispatch problems with valve-point effects. In: International Conference on Swarm, Evolutionary, and Memetic Computing. LNCS 6466, pp. 460–467 (2010)
26. Sabei, D., Erkoyuncu, J., Roy, R.: A review of multi-criteria decision making methods for enhanced maintenance delivery. *Procedia CIRP* **37**, 30–35 (2015)

Chapter 49

Solar PV Application in Aerospace Technologies



Saumya Ranjan Lenka, Sonali Goel, and Renu Sharma

Abstract In recent years, there has been great deal of interest in exploration of alternative fuels such as solar PV, other than jet fuel for aircraft propulsion in order to reduce the greenhouse gas (GHG) emissions. In this paper, a solar PV application in aerospace technologies has been described. The method is based on integration of photovoltaic (PV) system into the aircraft, thereby utilizing it to charge the battery. The high-altitude solar powered aircrafts are attempt to mitigate climate change. The paper focuses on the technological status of various solar power aircrafts which include photovoltaic systems, battery, and power management systems.

49.1 Introduction

The PV frameworks are intended to flexibly capacity to electrical load. The load might be of DC or AC type and relying on the application. While a PV panel creates power just in daylight time, a quantity of energy conservation arrangement is needed to manage the load in the non-daylight time.

This energy conservation is normally practiced through batteries. While the non-day-light time, the load may likewise be fueled by supplementary power sources. The fuel cell efficiency and mass are important characteristics of an energy storage system in aerospace application.

The solar-powered aircraft power gadgets are the photograph voltaic cells, maximum power point tracker (MPPT), and battery-powered [1]. The MPPT is the basic part of power conversation efficiency and different calculations for MPPT are accessible, each with various highlights and standard of activity.

Generally, the photovoltaic cells are fixed on both side wings of the aircraft. The motivation behind photovoltaic cell is to change over sunlight-based irradiance into electrical vitality at that point, move this capacity to drive framework and flight of airplane. Photovoltaic solar-based cells are arranged for material utilized in the

S. R. Lenka · S. Goel · R. Sharma (✉)

Department of Electrical Engineering, ITER, SOA (Deemed to be University), Bhubaneswar, India

e-mail: renusharma@soa.ac.in

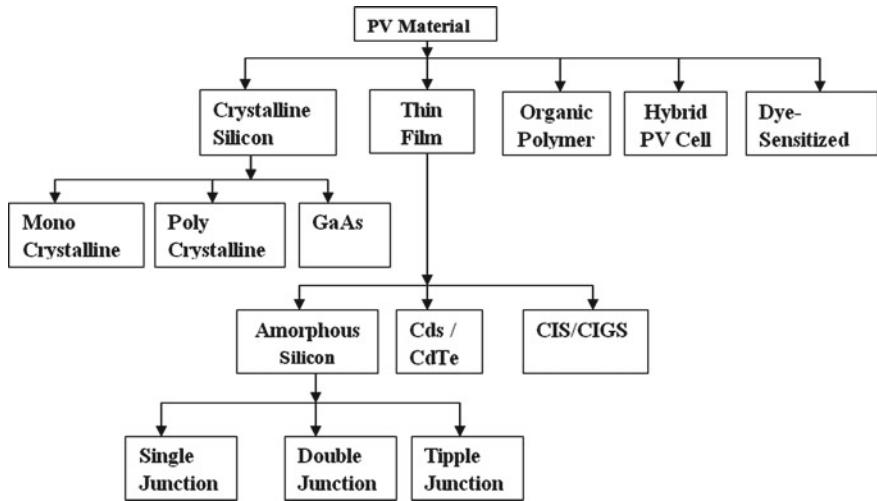


Fig. 49.1 Classification of photovoltaic cell based on PV material [4]

creation of solar powered cell. The arrangement is as per the following; (a) crystalline silicon thin film, (b) natural, (c) hybrid PV, and (d) desensitized photovoltaic cell.

The battery is attached to fuselage and inside the wing of airplane. The battery is charged only sunny time. It releases electrical power to the airplane and electronic equipment fitted in the aircraft at night [2]. The rechargeable battery in airplane fueled by solar-based vitality is essential to support ceaseless flight and high elevation and long continuance [3]. In normal environment, Li-ion and Li-Sulfur batteries are the most reasonable and promising innovation, albeit principal issues with respect to the work of batteries on board despite everything should be tended to. These comprise of wellbeing and extreme working conditions, particularly regarding fire and blast hazards, low working temperature, and high force rates.

The solar cell photovoltaic (PV) materials are grouped dependent on the resources. Classification of photovoltaic cell based on PV materials is shown in Fig. 49.1.

49.2 Photovoltaic: An Overview

Over the years, thin-film-solar-based cell advances have improved with high energy conversion efficiency. The improved application of thin film solar-based cells is a direct result of their capacity at higher temperature to create high efficiency than crystalline silicon cells. Quantum dots innovation can improve the productivity of solar-oriented cells. This is accomplished by expanding the band gap of solar-based cells [5]. Quantum dot innovation can get better the efficiency of solar-based cell by over 66%. This is accomplished by expanding the band gap of solar-based cells. Nanomaterials, for example, nanowires and nanoparticles have a bit of leeway in

Table 49.1 The different types of aircraft powered by crystalline silicon, battery and its performance [6]

Name of aircraft	Year	Photovoltaic			Battery	
		Name of solar cell	Conversion efficiency (%)	Output power (W)	Name of battery	Specific energy (Wh/kg)
Solar impulse II	2014/2016	Mono crystalline	18	260	Li-ion	260
Zepher7	2010	Amorphous silicon	19	–	Lithium-sulfur	400–600
Xihe	2009	–	16	–	–	–
Solar impulse I	2009	Mono crystalline	18	240	Li-ion polymer	240
So-Long	2005	Mono crystalline	18	220	Li-ion	220
Heliplatt	Start from 2004	Mono crystalline silicon	22	1500	–	–
Sky sailor	2004	Mono crystalline	18	84	Li-ion polymer	172.8
Helios	2001	Sun power mono crystalline	16	–	–	–
Pathfinder	1997	–	14.5	225	–	–
Gossamer penguin	1983	Mono crystalline silicon 3920	13.2	600	Nickel Cd	50
Sunrise II	1975	Mono crystalline silicon	16.1	580–600	Li-ion polymer	145
Sunrise I	1974	Mono crystalline silicon	11	400	Li-ion polymer	145

application than photovoltaic gadgets. The innovation can be utilized to deliver the airframe of the airplane not just restricted to the solar-powered cell (Tables 49.1 and 49.2).

49.3 Maximum Power Point Tracker (MPPT)

Maximum power point tracker tracks the point on a power I–V curve, that has the highest value of the product of its corresponding V and I, or the highest power output [7]. MPPT is a gadget that gets the ideal force from the PV, as the situation of MPPT

Table 49.2 Characteristics of common batteries

Characteristics	Lead acid	Ni–Cd	Ni–MH	Li-ion	Li–Po	Li–S	Zn air
Energy density (Wh/kg)	33–40	40–60	30–80	160	130–200	250–350	230
Energy/volume (Wh/L)	50–100	50–150	140–300	270	300	600	270
Power density (W/kg)	80–300	200–500	250–1000	1800	2800	2800	105
Recharge time (h)	8–16	1	2–4	2–3	2–4	x	10
Cycle eff (%)	82	80	70	99.9	99.8	99.8	X
Lifetime	x	x	x	24–36 mth	24–36 mth	24–36 mth	X
Life cycles	300	500	500–1000	1200	> 1000	> 1000	> 2000
Nominal voltage (V)	2	1.2	1.2	3.6	3.7	3.7	1.2
Operating temperature (°C)	15–25	–20 to 60	–20 to 60	–40 to 60	x	x	–25
Commercial use	1900/1970	1900/1950	1900	1991	1999	x	X
Cost per Wh (\$)	0.17	1.50	0.99	0.47	x	x	X

changes through variety in environmental conditions [8]. The MPPT finds the PV voltage and current reliable with MPP also lock its ideal force from the PV cluster to the framework [9]. The achievement of solar-based fueled airplane vitality is a basic issue [10].

MPPT methods are extremely proficient, with quick response, and are more confounded contrasted with the conventional techniques that are more straightforward, less expensive, and less productive, for example, P&O and incremental conductance methods [11].

49.4 The Adaptability of PV for Encapsulation on the Wing

Embodiment of PV cells on the wing of solar power-controlled airplane to twist on the airfoil is significant for PV cell charge efficiency. To guarantee PV cell is typified on solar power-controlled airplane and bowed on the airfoil adequately, the PV cell must be adaptable. The power device of the solar powered aircraft is shown in Fig. 49.2.

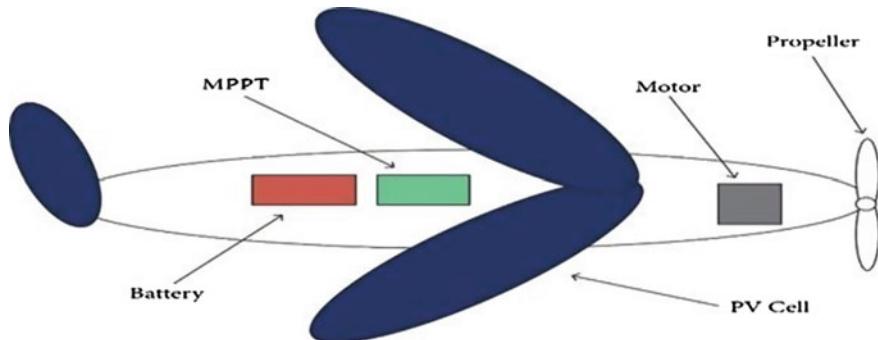


Fig. 49.2 The power device of solar-powered aircraft

49.5 Recent Technology Trends of Solar-Powered Aircraft

Solar-powered airplane is a framework whereby solar energy is utilized to control the drive segment of airplane. Solar-powered aircrafts can be streamlined by bringing down the force stacking to accomplish higher drive and streamlined efficiencies.

49.6 Conclusion

Solar-based energy has an enormous energy to be a significant part of an upcoming period without carbon energy range in aviation. In any case, techno-legitimate advances and achievements are important to defeat low-change proficiency and significant expense as of now accessible frameworks. A key end is that the innovation improvement pattern in solar-powered airplane frameworks can be evaluated dependent on individual technology frameworks with huge effect on the general airplane.

References

1. Cleave, V.: Solar power: a flight to remember. *Nature* **451**(7181), 884–886 (2008)
2. Gao X, Hou Z, Guo Z, Chen X, Reviews of methods to extract and store energy for solar-powered aircraft. *Renew. Sustain. Energy Rev.* **44**(109), 96–108 (2015)
3. Barbosa, R., Escobar, B., Sanchez, V.M., Hernandez, J., Acosta, R., Verde, Y.: Sizing of a solar/hydrogen system for high altitude long endurance aircraft. *Int. J. Hydrogen Energy* **39**(29), 16637–16645 (2014)
4. Tyagi, V.V., Rahim, N.A.A., Rahim, N.A., Selvaraj, J.A.L.: Progress in solar PV technology: research and achievement. *Renew. Sustain. Energy Rev.* **20**, 443–461 (2013)
5. Nozik, A.J.: Quantum dot solar cells. *Phys. E Low-dimension. Syst. Nanostruct.* **14**(1–2), 115–120 (2002)

6. Danjuma, S.B., Omar, Z., Abdullah, M.N.: Review of photovoltaic cells for solar-powered aircraft applications. *Int. J. Eng. Technol.*, 134 (2018)
7. Bouselham, L., Hajji, M., Hajji, B., Bouali, H.: A new MPPT-based ANN for photovoltaic system under partial shading conditions. *Energy Procedia* **111**, 924–933 (2017)
8. Wang, N., Wu, M., Shi, G.: Study on characteristics of photovoltaic cells based on MATLAB simulation. In: Power and Energy Engineering Conference, vol. 3, pp. 2–5 (2011)
9. Kobayashi, K., Takano, I., Sawada, Y.: A study on a two stage maximum power point tracking control of a photovoltaic system under partially shaded insolation conditions. *IEEE Power Eng. Soc. Gen. Meet.* **4**, 2612–2617 (2003)
10. Noth, Design of Solar Powered Airplanes for Continuous Flight, Ingénieur en Micro-technique Ecole Polytechnique Fédérale de Lausanne, Suisse Born (2008)
11. Ram, J.P., Babu, T.S., Rajasekar, N.: A comprehensive review on solar PV maximum power point tracking techniques. *Renew. Sustain. Energy Rev.* **67**, 826–847 (2017)

Chapter 50

Cloud Classification-Based Fine KNN Using Texture Feature and Opponent Color Features



Prabira Kumar Sethy and Sidhant Kumar Dash

Abstract Weather forecasting and alerts are issued by the meteorological department to alert citizens. Cloud image classification is a challenging task for the research community of meteorology and machine learning. Most of the researches for cloud classification are based on histogram and machine learning using texture features. In this manuscript, the opponent color feature and fine KNN are used to classify the 11 types of cloud images and achieved an accuracy of 99.5% and AUC of 1.0.

50.1 Introduction

The earth's energy balance is maintained through the water cycle, where the clouds play an important role [1–3]. “In weather forecasting and other climatic changes, the clouds are always considered the main factor [4].” The traditional and continual observation of the cloud's features are time-consuming and depend on the observer's experience. The digital imaging techniques and substantial development of hardware systems have made it possible for automatic and continuous observation of cloud conditions. The ground-based images have much more impact on feature analysis and high spatial resolution than satellite images [5].

In our work, we have cirrus cumulus stratus nimbus (CCSN) dataset is considered [6]. Many steps have been taken to identify the cloud type and features accurately and effectively to address this demanding issue [5–17]. “Buch et al. adopted pixel brightness, position, and texture features from the entire sky images and categorized them with the help of decision trees [7].” To extract the co-occurrence matrices, autocorrelation, law's features, edge frequency, primitive length, and texture feature were taken to recognize the cloud [8]. “Calbo and Sabburg considered the features of Fourier transform the statistical texture and threshold image to identify the whole-sky imager (WSI) and total sky imager (TSI) images [5].” “Heinle et al. extracted 12 statistical features to represent the texture and color of the image [9].” The sky conditions were divided into seven groups using the k-nearest-neighbor (KNN) classifier. The texture

P. K. Sethy (✉) · S. K. Dash

Department of Electronics, Sambalpur University, Jyoti Vihar, Burla, India

orientation of ground-based images and satellite images is clearly distinguished; the Gabor-based multiple features classification approach is used in conjunction with the support vector machine (SVM), and total accuracy of 88.3% is attained [10]. “Typical local binary patterns (LBPs) and weighted local binary patterns (WLBP) were presented, depicting the fusion of the variance of a local patch to increase the contrast to recognize the type of cloud [11].” “Cheng and Yu performed the block-based classification (BC) by combining the LBPs and other statistical features with the Bayesian classifier [12].” “Liu et al. employed seven structure features from the image edge to illustrate the structural characteristic of the infrared clouds [13].” Using the structure and texture features separately may not enhance the classification performance, as Zhuo et al. [14] explained. Structure and texture characteristics have thus been considered to cloud type classification with SVM. Xia et al. in [6] and Xiao et al. in [15] also suggested that the integration of all these features should be much more effective in order to identify the cloud type, including texture, color, and structure, and that the experiments should show that multiple features should be combined at a time. The physical features also have a significant role in representing the cloud types. “Kazantidis et al. proposed the complete cloud coverage, the solar zenithal angle, the existence of rain, and the visible percentage of the sun in sky images, to illustrate the physical property of the cloud [12].” Besides this, the sky camera image has been used to extract 12 image features. Again, another seven features of the cloud layer were extracted from the adopted random forests and ceilometers and combined for the classification described by Tato et al. [17].

“Li et al. proposed a discriminative approach based on bag of microstructures (BOMs) to better display the cloud pictures [18].” It has shown a competitive performance in recognizing the type of cloud. However, the disadvantage of BOMs is that they could not classify the visually complex cloud. Therefore, a duplex norm bounded sparse representation model was proposed in [19]. The model has been verified to extract the most notable patterns from the various categories of cloud and has naturally achieved a high degree of accuracy.

50.2 Material and Methods

The details of the dataset and adapted methodology are discussed in the appropriate subsection.

50.2.1 About Dataset

The CCSN dataset contains 2543 cloud images. According to the World Meteorological Organization’s genera-based classification recommendation, we divide into 11 different categories: Ac, Sc, Ns, Cu, Ci, Cc, Cb, As, Ct, Cs, and St. It is worth noting that contrails have consideration in our dataset. Representative sample images

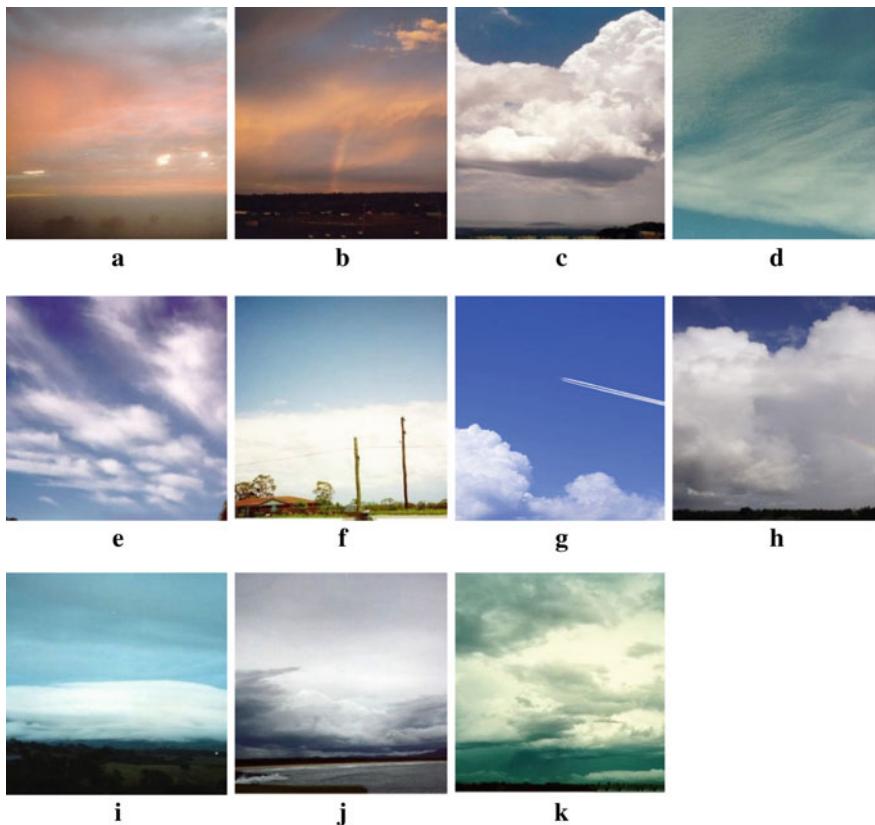


Fig. 50.1 Different cloud images, **a** cirrus, **b** cirrostratus, **c** cirrocumulus, **d** altocumulus, **e** altostratus, **f** cumulus, **g** cumulonimbus, **h** nimbostratus, **i** stratocumulus, **j** stratus, **k** contrail

from each category are shown below. Ci = cirrus; Cs = cirrostratus; Cc = cirrocumulus; Ac = altocumulus; As = altostratus; Cu = cumulus; Cb = cumulonimbus; Ns = nimbostratus; Sc = stratocumulus; St = stratus; Ct = contrail. All images are fixed resolution 256 × 256 pixels with the JPEG format. The samples of clouds are illustrated in Fig. 50.1.

50.2.2 Methodology

The 11 types of cloud images are taken for classification based on opponent color features and fine KNN. The opponent-process theory of color vision is one of the theories that helped develop our current understanding of sight. The theory suggests that different color complexes with opposing actions control sight color. Here, six opponent color features with three basic texture features like mean red, mean green,

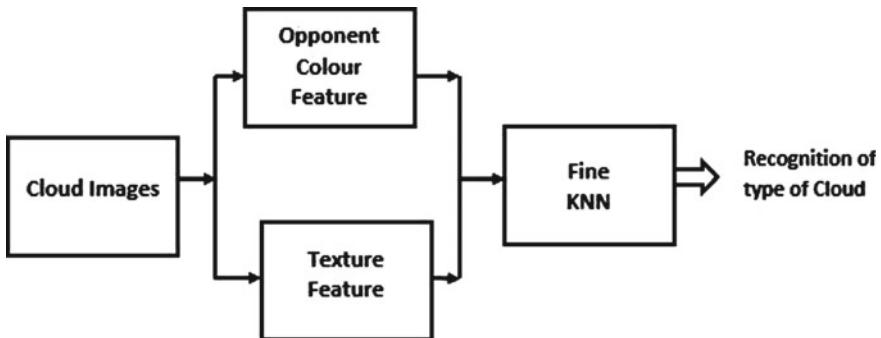


Fig. 50.2 Cloud classification-based fine KNN using the opponent color feature and texture feature

and mean blue are considered. Here, the fine KNN has taken these nine features and classify the 11 types of cloud images (Fig. 50.2).

The mathematical expressions of six opponent color features are given below.

$$\text{Opponent Feature 1} = \text{mean of Red} - \text{mean Green} \quad (50.1)$$

$$\text{Opponent Feature 2} = \text{mean Red} - \text{mean Blue} \quad (50.2)$$

$$\text{Opponent Feature 3} = \text{mean Green} - \text{mean Blue} \quad (50.3)$$

$$\text{Opponent Feature 4} = ((\text{mean Red} - \text{mean Green})/1.414) \quad (50.4)$$

$$\text{Opponent Feature 5} = ((\text{mean Red} + \text{mean Green} - 2 \times \text{mean Blue})/2.449) \quad (50.5)$$

$$\text{Opponent Feature 6} = ((\text{mean Red} + \text{mean Green} + \text{mean Blue})/1.732) \quad (50.6)$$

50.3 Result and Discussion

The proposed model is executed in HP Pavilion, window 10, 8 GB RAM in MATLAB 2019a. The performance of the proposed model is evaluated in terms of accuracy, AUC, TPR, FNR, PPV, and FDR. The proposed model's confusion matrix and ROC curve are illustrated in Figs. 50.3 and 50.4, respectively. Here, the index 1 to 11 are assigned for altocumulus, altostratus, cumulonimbus, cirrus, cirrostratus, contrail, cumulus, nimbostratus, stratocumulus, and stratus, respectively. Further, the model

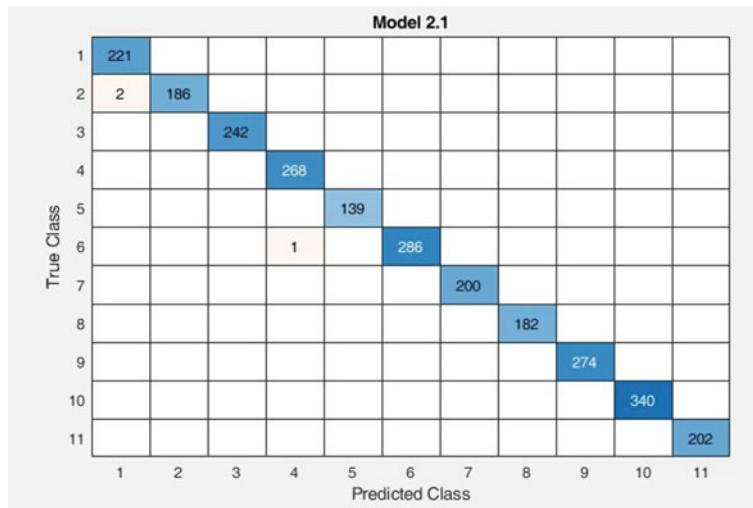


Fig. 50.3 Confusion matrix of proposed model

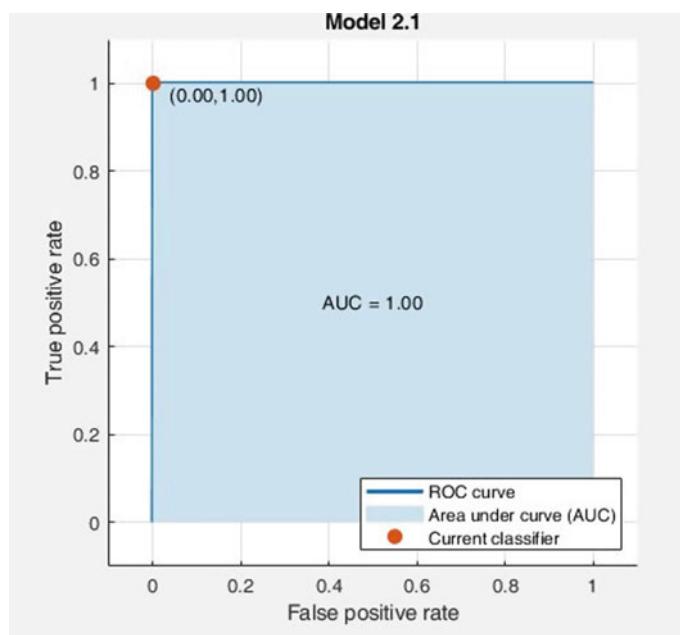


Fig. 50.4 ROC of proposed model

Table 50.1 Performance of proposed model in terms of TPR, FNR, PPV, and FDR

Cloud types	TPR (%)	FNR (%)	PPV (%)	FDR (%)
Ac	100	0	99.1	0.9
As	98.9	1.1	100	0
Cb	100	0	100	0
Cc	100	0	99.6	0.4
Ci	100	0	100	0
Cs	99.7	0.3	100	0
Ct	100	0	100	0
Cu	100	0	100	0
Ns	100	0	100	0
Sc	100	0	100	0
St	100	0	100	0

concerning each type of cloud is investigated in terms of TPR, FNR, PPV, and FDR (Table 50.1).

50.4 Conclusion

The cloud classification model based on fine KNN using texture feature and opponent color feature is proposed. This model achieved an accuracy of 99.5% and AUC 1.0. Again, the TPR values of each cloud type are more than 99%, and the FPR is almost 0. This proposed model is helpful for weather forecasting.

References

- Stephens, G.L.: Cloud feedbacks in the climate system: a critical review. *J. Clim.* **18**(2), 237–273 (2005)
- A.J. Teuling, C.M. Taylor, J.F. Meirink, et al.: Observational evidence for cloud cover enhancement over western European forests. *Nat. Commun.* **8**, 14065 (2017)
- B.A. Baum, P.F. Soulen, K.I. Strabala, et al.: Remote sensing of cloud properties using MODIS airborne simulator imagery during SUCCESS: 2. cloud thermodynamic phase. *J. Geophys. Res.-Atmosph.* **105**(9), 11767–11780 (2000)
- Houghton, J.T., Ding, Y., Griggs, D.J., et al.: Climate Change 2001: Scientific Basis. Cambridge University Press, Cambridge (2001)
- Calb'o, J., Sabburg, J.: Feature extraction from whole-sky ground-based images for cloud-type recognition. *J. Atmos. Oceanic Tech.* **25**(1), 3–14 (2008)
- Zhang, J.L., Liu, P., Zhang, F., Song, Q.Q.: CloudNet: Ground-based cloud classification with deep convolutional neural network. *Geophys. Res. Lett.* **45**, 8665–8672 (2018). <https://doi.org/10.1029/2018GL077787>

7. Buch, K.A.J., Sun, C.H., Orne, L.R.: Cloud classification using whole-sky imager data. In: Proceedings of the fifth Atmospheric Radiation Measurement (ARM), Science Team Meeting, San Diego, CA, USA (1995)
8. Singh, M., Glennen, M.: Automated ground-based cloud recognition. *Pattern Anal. Appl.* **8**(3), 258–271 (2005)
9. Heinle, A., Macke, A., Srivastav, A.: Automatic cloud classification of whole sky images. *Atmosph. Meas. Tech.* **3**(3), 557–567 (2010)
10. Liu, R., Yang, W.: A novel method using the Gabor-based multiple feature and ensemble SVMs for ground-based cloud classification. In: Proceedings of Seventh International Symposium on Multispectral Image Processing and Pattern Recognition (MIPPR2011), Guilin, China (2011)
11. Liu, S., Zhang, Z., Mei, X.: Ground-based cloud classification using weighted local binary patterns. *J. Appl. Remote Sens.* **9**(1), 095062 (2015)
12. Cheng, H.Y., Yu, C.C.: Block-based cloud classification with statistical features and distribution of local texture features. *Atmosph. Meas. Tech.* **8**(3), 1173–1182 (2015)
13. Liu, L., Sun, X., Chen, F., Zhao, S., Gao, T.: Cloud classification based on structural features of infrared images. *J. Atmosph. Oceanic Technol.* **28**(3), 410–417 (2011)
14. Zhuo, W., Cao, Z., Xiao, Y.: Cloud classification of ground-based images using texture-structure features. *J. Atmosph. Oceanic Technol.* **31**(1), 79–92 (2014)
15. Xiao, Y., Cao, Z., Zhuo, W., Ye, L., Zhu, L.: mCLOUD: a Multiview visual feature extraction mechanism for ground-based cloud image categorization. *J. Atmos. Oceanic Tech.* **33**(4), 789–801 (2016)
16. Kazantzidis, A., Tzoumanikas, P., Bais, A.F., Fotopoulos, S., Economou, G.: Cloud detection and classification with the use of whole-sky ground-based images. *Atmosph. Res.* **113**(1), 80–88 (2012)
17. Tato, J.H., Benítez, F.J.R., Barrena, C.A., Mur, R.A., Leon, I.G., Vazquez, D.P.: Automatic cloud-type classification based on the combined use of a sky camera and a ceilometer. *J. Geophys. Res.: Atmosph.* **122**(20), 11045–11061 (2017)
18. Li, Q., Zhang, Z., Lu, W., Yang, J., Ma, Y., Yao, W.: From pixels to patches: a cloud classification method based on a bag of micro-structures. *Atmosph. Meas. Tech.* **9**(2), 753–764 (2016)
19. Gan, J., Lu, W., Li, Q., et al.: Cloud type classification of total sky images using duplex norm-bounded sparse coding. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **10**(7), 3360–3372 (2017)

Chapter 51

Machine Learning-Based Diabetes Prediction Using Missing Value Impotency



Santi Kumari Behera, Julie Palei, Dayal Kumar Behera, Subhra Swetanisha, and Prabira Kumar Sethy

Abstract Diabetes is a chronic disease that has been impacting an increasing number of people throughout the years. Each year, it results in a huge number of deaths. Since late diagnosis results in severe health complications and a significant number of deaths each year, developing methods for early detection of this pathology is critical. As a result, early detection is critical. Machine learning (ML) techniques aid in the early detection and prediction of diabetes. However, ML models do not perform well with missing values in the dataset. Imputation of missing values improves the outcome. The article proposes an ensemble method with a strong emphasis on missing value imputation. Numerous ML models have been used to validate the proposed framework. The experimentation uses the Pima Indian Diabetes Dataset, which contains information about people with and without diabetes. TPR, FNR, PPV, FDR, overall accuracy, training time, and AUC are used to evaluate the performance of the 24 ML methods. The collected results demonstrate that subspace KNN outperforms with an accuracy of 85%. The collected data are confirmed systematically and orderly utilizing receiver operating characteristic (ROC) curves. Using missing value imputation for data pre-processing and classification has been shown to beat state-of-the-art algorithms in the diabetes detection sector.

S. K. Behera

Department of CSE, VSSUT Burla, Burla, Odisha, India

J. Palei · P. K. Sethy (✉)

Department of Electronics, Sambalpur University, Burla, Odisha 768019, India

J. Palei

e-mail: 19mscel05@suiit.ac.in

D. K. Behera

Department of CSE, Silicon Institute of Technology, Bhubaneswar, Odisha, India

S. Swetanisha

Department of CSE, Trident Academy of Technology, Bhubaneswar, Odisha, India

51.1 Introduction

“According to the World Health Organization (WHO), around 1.6 million people die each year from diabetes [1].” “Diabetes is a type of disease that arises when the human body’s blood glucose/blood sugar level is abnormally high. Type 1 diabetes, commonly known as insulin-dependent diabetes, is most frequently diagnosed in childhood [2].” In Type 1, the pancreas is attacked by the body’s antibodies, which then kill internal body parts and cause the pancreas to stop producing insulin. “Type 2 diabetes is often referred to as adult-onset diabetes or non-insulin-dependent diabetes [3].” Although it is more merciful than Type 1, it is nevertheless extremely damaging and can result in serious complications, particularly in the small blood vessels of the eyes, kidneys, and nerves [4]. Type 3 gestational diabetes [5] develops from increased blood sugar levels in pregnant women whose diabetes is not recognized earlier. “Some authors have created and validated a risk score for primary cesarean delivery (CD) in women with gestational diabetes. In women with gestational diabetes mellitus (GDM), a risk score based on nulliparity, excessive gestational weight gain, and usage of insulin can be used to determine the likelihood of primary CD [6].” “Ghaderi et al. worked on the effect of smartphone education on the risk perception of Type 2 diabetes in a woman with GDM [2].”

Diabetes mellitus is related to long-term consequences. Additionally, people with diabetes face an increased chance of developing a variety of health concerns. Glucose levels in the human body typically range between 70 and 99 mg per deciliter [1]. “If the glucose level is more significant than 126 mg/dl, diabetes is present. Prediabetes is defined as a blood glucose level of 100–125 mg/dl [7].”

Diabetes is influenced by height, weight, hereditary factors, and insulin [8], but the primary factor evaluated is blood sugar content. Early detection is the only approach to avoid difficulties. Predictive analytics strives to improve disease diagnosis accuracy, patient care, resource optimization, and clinical outcomes. Numerous researchers are conducting experiments to diagnose disease using various classification algorithms from ML approaches such as J48, SVM, Naïve Bayes, and decision tree. Researchers have demonstrated that machine learning algorithms [9, 10] perform better at diagnosing various diseases. Naïve Bayes, SVM, and decision tree ML classification algorithms are applied and assessed in work [8] to predict diabetes in a patient using the PIDD dataset.

ML techniques can also be utilized to identify individuals at elevated risk of Type 2 diabetes [11] or prediabetes in the absence of established impaired glucose regulation. Body mass index, waist-hip ratio, age, systolic and diastolic blood pressure, and diabetes inheritance were the most impactful factors. Increased risk of Type 2 diabetes was associated with high levels of these characteristics and diabetes heredity.

ML techniques aid in the early detection and prediction of diabetes. However, ML models do not perform well with missing values in the dataset. Therefore, this work emphasizes on the missing value imputation. The objectives of this work are as follows:

- Study the impact of missing value in the PIMA diabetes dataset.
- Performing missing value imputation by replacing the missing value with the mean value of the group.
- Designing an ensemble subspace KNN for classifying diabetes.
- Comparative analysis of various traditional classifiers against the ensemble classifier.

51.2 Related Works

The purpose of the article [12] is to illustrate the construction and validation of 10-year risk prediction models for Type 2 diabetes mellitus (T2DM). Data collected in 12 European nations (SHARE) are used for validation of the model. The dataset included 53 variables encompassing behavioral, physical, and mental health aspects of participants aged 50 or older. To account for highly imbalanced outcome variables, the logistic regression model was developed, each instance was weighted according to the inverse percentage of the result label. The authors used a pooled sample of 16,363 people to develop and evaluate a global regularized logistic regression model with an area under the receiver operating characteristic curve of 0.702. Continuous glucose monitoring (CGM) devices continue to have a temporal delay, which can result in clinically significant differences between the CGM and the actual blood glucose level, particularly during rapid changes. In [13], authors have used the artificial neural network regression (NN) technique to forecast CGM results. Diabetes can also be a risk factor for developing other diseases such as heart attack, renal impairment, and partial blindness. Kayal Vizhi and “Aman Dash worked on smart sensors and ML techniques such as random forest and extreme gradient boosting for predicting whether a person would get diabetes or not [14].” Mujumdar et al. [5] developed a diabetes prediction model that took into account external risk variables for diabetes in addition to standard risk factors such as glucose, BMI, age, and insulin. The new dataset improves classification accuracy when compared to the available PIMA dataset. The study [15] covers algorithms such as linear regression, decision trees, random forests, and their advantages for early identification and treatment of disease. The research study discussed the predictive accuracy of the algorithms mentioned above. Mitushi Soni and Sunita Varma [16] forecasted diabetes using ML classification and ensemble approaches. When compared to other models, each model’s accuracy varies. Their findings indicate that random forest outperformed different ML algorithms in terms of accuracy. “Jobeda Jamal Khanam and Simon Foo conducted research using the PIMA dataset. The collection comprises data on 768 patients and their nine unique characteristics. On the dataset, seven ML algorithms were applied to predict diabetes. They concluded that a model combining logistic regression (LR) and support vector machine (SVM) effectively predicts diabetes [1].” “Varga used NCBI PubMed to conduct a systematic search. First, articles that had the words “diabetes” and “prediction” were chosen. Next, the authors searched for metrics relating to predictive statistics in all abstracts of original research articles published in the field

of diabetes epidemiology. To illustrate the distinction between association and prediction, simulated data were constructed. It is demonstrated that biomarkers with large effect sizes and small P values might have low discriminative utility [17].” The article [18] attempts to synthesize the majority of the work on ML and data mining techniques used to forecast diabetes and its complications. Hyperglycemia is a symptom of diabetes caused by insufficient insulin secretion and/or use. For experimental purposes, Kalagotla et al. [19] designed a novel stacking method based on multi-layer perceptron, SVM, and LR. The stacking strategy combined the intelligent models and improved model performance. In comparison with AdaBoost, the proposed unique stacking strategy outperformed other models. Authors in [20] worked on a pipeline for predicting diabetes individuals using deep learning techniques. It incorporates data enhancement using a variational autoencoder (VAE), feature enhancement via a sparse autoencoder (SAE), and classification via a convolutional neural network.

51.3 Material and Methods

The details about dataset and adapted methodology are elaborated in appropriate subsection.

51.3.1 Dataset

The Pima Indians Diabetes Database is used in this paper. “The National Institute of Diabetes and Digestive and Kidney Diseases first collected this dataset. The dataset’s purpose is to diagnostically predict if a patient has diabetes or not, using particular diagnostic metrics contained in the dataset. Therefore, numerous limits on the selection of these examples from a broader database were imposed. All patients at this facility are females who are at least 21 years old and of Pima Indian ancestry. The datasets contain a variety of medical predictor variables and one outcome variable. The number of pregnancies, their BMI, insulin level, and age are all predictor variables of the patient [21].”

51.3.2 Proposed Model

This research focuses heavily on enhancing the outcomes and accuracy of diabetes detection. The proposed approach is depicted in Fig. 51.1. Numerous classical ML classifiers and related ensemble variation models are used to categorize disease as positive or negative. Numerous characteristics in the original dataset have entries of 0. According to the experts’ advice, these values must be considered a missing value, such as glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree

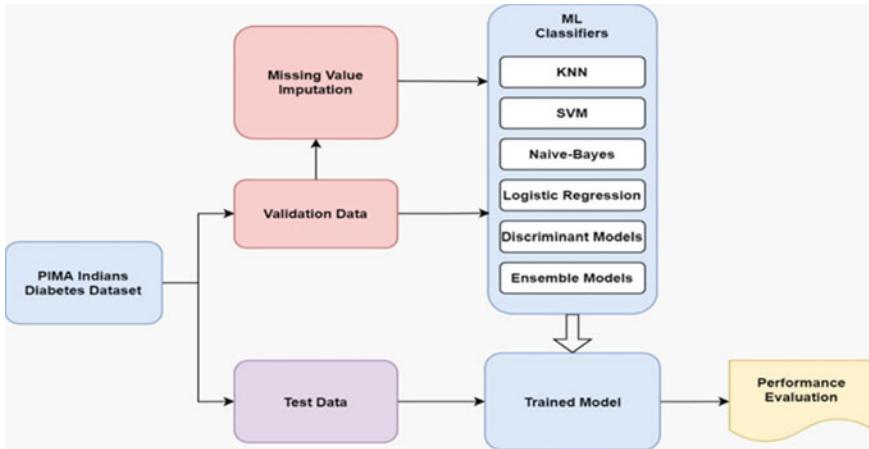


Fig. 51.1 Proposed framework for diabetes prediction

function, and age cannot be zero. Hence, the missing value is imputed by considering the mean value of the group. After that, the class label of the dataset is evaluated by taking two different values into account: Diabetic is set to 1 and non-diabetic by 0. Then, the dataset is divided into validation and test set. The validation data are used to train the classifier in two scenarios. In the first case, the classifier is trained with missing value and in another case by missing value imputation. The missing value imputation does not pre-process the test data, and it is passed to the model for prediction.

51.4 Results and Discussion

The classification performance in TPR, FNR, PPV, FDR, overall accuracy, training time, and AUC is evaluated using the most robust ML classifiers such as KNN, SVM, Naïve-Bayes, logistic regression, discriminant models, and ensemble models.

Table 51.1 depicts the performance of various models on the validation data without considering missing value imputation (MVI), whereas Table 51.2 represents performance by considering missing value imputation. From the data, it is clear that the AUC of all the models improves a lot in missing value imputation.

Table 51.3 represents the performance of the prediction on the test data. Again, subspace KNN performs better as compared to other classifiers.

From Table 51.3, it is clear that subspace KNN performed better than the other classifier. The confusion matrix of the ensemble subspace KNN model is depicted in Fig. 51.2. The AUC of ROC is shown in Fig. 51.3.

Table 51.1 Performance analysis on validation data without MVI

Validation model	Accuracy (validation)	Training time (s)	TPR	FNR	PPV	FDR	AUC
Fine KNN	100	0.28351	100	0	100	0	1
weighted KNN	100	0.21176	100	0	100	0	1
Subspace KNN (ensemble)	100	0.37183	100	0	100	0	1
Bagged trees (ensemble)	99.7	0.51479	99.65	0.35	99.8	0.2	1
Fine Gaussian SVM	98.8	0.20768	98.3	1.7	99.1	0.9	1
Fine tree	93.2	1.7373	92.3	7.7	92.75	7.25	0.98
Boosted trees (ensemble)	85.5	0.62198	83.1	16.9	84.6	15.4	0.95
Cubic SVM	87	0.41443	84.7	15.3	86.2	13.8	0.93
RUSBoosted trees (ensemble)	84.9	0.45368	86.05	13.95	83.4	16.6	0.93
Medium Gaussian SVM	82.7	0.2608	78.3	21.7	82.65	17.35	0.9
Medium tree	83.3	0.35821	80.2	19.8	82.35	17.65	0.89
Quadratic SVM	79.8	0.26857	74.9	25.1	79.25	20.75	0.87
Cosine KNN	79.8	0.20666	75.25	24.75	78.95	21.05	0.87
Medium KNN	78.5	0.26587	72.85	27.15	78.15	21.85	0.87
Cubic KNN	78.4	0.25979	72.95	27.05	77.8	22.2	0.87
Linear discriminant	78.4	0.3483	73.7	26.3	77.1	22.9	0.84
Coarse Gaussian SVM	78.4	0.1325	72.75	27.25	77.95	22.05	0.84
Logistic regression	78.3	0.73955	73.6	26.4	76.9	23.1	0.84
Linear SVM	77.3	0.39083	72.55	27.45	75.8	24.2	0.84
Quadratic discriminant	76.4	0.34544	72.1	27.9	74.4	25.6	0.83
Subspace discriminant (ensemble)	76.3	1.4926	69.4	30.6	76.25	23.75	0.83
Kernel Naive Bayes	75.4	0.61343	68.05	31.95	75.45	24.55	0.83
Gaussian Naive Bayes	76.2	0.30505	72.7	27.3	73.85	26.15	0.82
Coarse KNN	75.4	0.22108	66.9	33.1	77.45	22.55	0.82
Coarse tree	77.2	0.22121	72.3	27.7	75.75	24.25	0.74

Table 51.2 Performance analysis on validation data with MVI

With MVI (validation)	Accuracy (validation)	Training time (s)	TPR	FNR	PPV	FDR	AUC
Fine KNN	100	0.40763	100	0	100	0	1
Weighted KNN	100	0.22415	100	0	100	0	1
Subspace KNN (ensemble)	100	0.39983	100	0	100	0	1
Bagged trees (ensemble)	99.7	0.57935	99.7	0.3	99.7	0.3	1
Fine Gaussian SVM	97.8	0.2167	96.9	3.1	98.25	1.75	1
Boosted trees (ensemble)	96.5	0.75288	95.9	4.1	96.3	3.7	1
Fine tree	97.4	2.0716	96.6	3.4	97.65	2.35	0.99
RUSBoosted trees (ensemble)	94.3	0.47435	94.4	5.6	93.25	6.75	0.99
Medium tree	93.6	0.36612	91.9	8.1	94.05	5.95	0.98
Cubic SVM	87.4	0.49901	85.35	14.65	86.55	13.45	0.94
Coarse tree	88	0.25733	84.05	15.95	89.65	10.35	0.93
Kernel Naive Bayes	73.2	0.72157	62.1	37.9	81.55	18.45	0.93
Quadratic SVM	81.5	0.46121	77.5	22.5	80.7	19.3	0.89
Medium Gaussian SVM	81.4	0.2668	77.15	22.85	80.75	19.25	0.89
Medium KNN	80.5	0.26288	76	24	79.75	20.25	0.88
Cubic KNN	80.1	0.26259	75.35	24.65	79.4	20.6	0.88
Cosine KNN	78.4	0.25825	73.8	26.2	77.05	22.95	0.87
Coarse Gaussian SVM	78.6	0.15006	73.45	26.55	77.9	22.1	0.85
Linear SVM	78.4	0.51793	73.55	26.45	77.25	22.75	0.85
Logistic regression	78	1.0221	73.5	26.5	76.45	23.55	0.85
Linear discriminant	77.6	0.47924	73	27	76	24	0.85
Coarse KNN	76.6	0.22954	69.8	30.2	76.55	23.45	0.84

(continued)

Table 51.2 (continued)

With MVI (validation)	Accuracy (validation)	Training time (s)	TPR	FNR	PPV	FDR	AUC
Subspace discriminant (ensemble)	76.4	0.44237	69.8	30.2	76.15	23.85	0.84
Quadratic discriminant	70.3	0.36838	59.3	40.7	72.1	27.9	0.82
Gaussian Naive Bayes	71.1	0.37027	61.2	38.8	71.35	28.65	0.81

Table 51.3 Performance analysis on test data without MVI

Test	Accuracy (test)	TPR	FNR	PPV	FDR	AUC
Subspace KNN (ensemble)	85	88.45	11.55	85	15	1
Fine KNN	80	84.6	15.4	81.8	18.2	0.85
Weighted KNN	80	84.6	15.4	81.8	18.2	0.9
Fine tree	70	73.6	26.4	71.7	28.3	0.65
Medium tree	70	73.6	26.4	71.7	28.3	0.65
Medium Gaussian SVM	70	76.9	23.1	76.9	23.1	0.81
Boosted trees (ensemble)	70	63.75	36.25	66.65	33.35	0.75
Bagged trees (ensemble)	70	76.9	23.1	76.9	23.1	0.91
RUSBoosted trees (ensemble)	70	76.9	23.1	76.9	23.1	0.87
Coarse Gaussian SVM	65	69.75	30.25	68.75	31.25	0.73
Subspace discriminant (ensemble)	65	69.75	30.25	68.75	31.25	0.69
Linear discriminant	60	62.6	37.4	61.65	38.35	0.74
Logistic regression	60	62.6	37.4	61.65	38.35	0.71
Linear SVM	60	62.6	37.4	61.65	38.35	0.76
Quadratic SVM	60	69.25	30.75	73.35	26.65	0.76
Cubic SVM	60	69.25	30.75	73.35	26.65	0.63
Fine Gaussian SVM	60	69.25	30.75	73.35	26.65	0.96
Coarse KNN	60	65.95	34.05	65.95	34.05	0.75
Cubic KNN	60	65.95	34.05	65.95	34.05	0.72
Coarse tree	55	65.4	34.6	71.9	28.1	0.73
Quadratic discriminant	55	62.1	37.9	66.25	36.9	0.67
Gaussian Naive Bayes	55	62.1	37.9	63.1	36.9	0.68
Kernel Naive Bayes	55	62.1	37.9	63.1	36.9	0.71
Medium KNN	55	62.1	37.9	63.1	36.9	0.76
Cosine KNN	55	62.1	37.9	63.1	36.9	0.7

Fig. 51.2 Confusion matrix of subspace KNN on test data

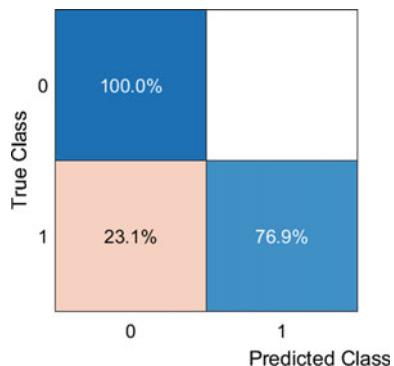
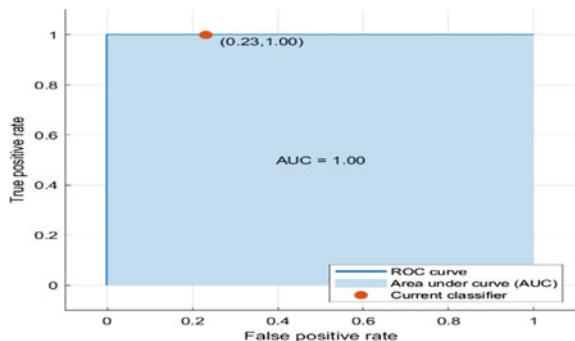


Fig. 51.3 ROC of subspace KNN on test data



51.5 Conclusion

Diabetes early identification is a big challenge in the health care business. In our research, we developed a system capable of accurately predicting diabetes. The purpose of this study is to propose an ensemble method based on in-depth ML techniques for diabetes prediction utilizing a well-known dataset called Pima Indian Diabetes. We used twenty-four different ML algorithms, including KNN, SVM, Naïve-Bayes, logistic regression, discriminant models, and ensemble models to predict diabetes and evaluate performance on various measures like TPR, FNR, PPV, FDR, overall accuracy, training time, and AUC. This work also emphasizes missing value imputation. In the validation data, overall accuracy improves to a great extent with missing value impotency, depicted in Table 51.2. Among all the proposed models, the subspace KNN is considered the most efficient and promising for predicting diabetes, with an accuracy of 85% in test data.

References

1. Khanam, J.J., Foo, S.Y.: A comparison of machine learning algorithms for diabetes prediction. *ICT Express* (2021). <https://doi.org/10.1016/j.icte.2021.02.004>
2. Ghaderi, M., Farahani, M.A., Hajiha, N., Ghaffari, F., Haghani, H.: The role of smartphone-based education on the risk perception of type 2 diabetes in women with gestational diabetes. *Health Technol. (Berl)* **9**(5), 829–837 (2019). <https://doi.org/10.1007/s12553-019-00342-3>
3. Mandal, S.: New molecular biomarkers in precise diagnosis and therapy of Type 2 diabetes. *Health Technol. (Berl)* **10**(3), 601–608 (2020). <https://doi.org/10.1007/s12553-019-00385-6>
4. Himsworth, H.P.: The syndrome of diabetes mellitus and its causes. *Lancet* **253**(6551), 465–473 (1949). [https://doi.org/10.1016/S0140-6736\(49\)90797-7](https://doi.org/10.1016/S0140-6736(49)90797-7)
5. Mujumdar, A., Vaidehi, V.: Diabetes prediction using machine learning algorithms. *Procedia Comput. Sci.* **165**, 292–299 (2019). <https://doi.org/10.1016/j.procs.2020.01.047>
6. Phaloprakarn, C., Tangjittgamol, S.: Risk score for predicting primary cesarean delivery in women with gestational diabetes mellitus. *BMC Pregnancy Childbirth* **20**(1), 1–8 (2020). <https://doi.org/10.1186/s12884-020-03306-y>
7. <https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284>
8. Sisodia, D., Sisodia, D.S.: Prediction of diabetes using classification algorithms. *Procedia Comput. Sci.* **132**(Iccids), 1578–1585 (2018). <https://doi.org/10.1016/j.procs.2018.05.122>
9. Kavakiotis, I., Tsavos, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* **15**, 104–116 (2017). <https://doi.org/10.1016/j.csbj.2016.12.005>
10. Okagbue, H.I., Adamu, P.I., Oguntunde, P.E., Obasi, E.C.M., Odetunmibi, O.A.: Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer. *Health Technol. (Berl)* **11**, 887–893 (2021). <https://doi.org/10.1007/s12553-021-00572-4>
11. Lama, L. et al.: Machine learning for prediction of diabetes risk in middle-aged Swedish people. *Heliyon* **7**, e07419 (2021). <https://doi.org/10.1016/j.heliyon.2021.e07419>
12. Gregor Stiglic, L.C., Wang, F., Sheikh, A.: Development and validation of the type 2 diabetes mellitus 10-year risk score prediction models from survey data. *Prim. Care Diabetes* **15**(4), 699–705 (2021)
13. Lebech Cichosz, O.S., Hasselstrøm Jensen, M.: Short-term prediction of future continuous glucose monitoring readings in type 1 diabetes: development and validation of a neural network regression model. *Int. J. Med. Inform.* **151**, 104472 (2021)
14. Vizhi, K., Dash, A.: Diabetes prediction using machine learning. *Int. J. Adv. Sci. Technol.* **29**(6), 2842–2852 (2020). <https://doi.org/10.32628/cseit2173107>
15. Muhammad Daniyal Baig, M.F.N.: Diabetes prediction using machine learning algorithms. *Lect. Notes Netw. Syst.* (2020). <https://doi.org/10.13140/RG.2.2.18158.64328>
16. Soni, M., Varma, S.: Diabetes prediction using machine learning techniques. *Int. J. Eng. Res. Technol.* **9**(09), 921–924 (2020). https://doi.org/10.1007/978-981-33-6081-5_34
17. Varga, T.V., Niss, K., Estampador, A.C., Collin, C.B., Moseley, P.L.: Association is not prediction: a landscape of confused reporting in diabetes—a systematic review. *Diabetes Res. Clin. Pract.* **170**, 108497 (2020). <https://doi.org/10.1016/j.diabres.2020.108497>
18. Jaiswal, T.P.V., Negi, A., Pal, T.: A review on current advances in machine learning based diabetes prediction. *Prim. Care Diabetes* **15**(3), 435–443 (2021)
19. Kalagotla, K., Satish Kumar, Gangashetty, S.V.: A novel stacking technique for prediction of diabetes. *Comput. Biol. Med.* **135**, 104554 (2021)
20. García-Ordás, M.T., Benavides, C., Benítez-Andrades, J.A., Alaiz-Moretón, H., García-Rodríguez, I.: Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Comput. Methods Programs Biomed.* **202**, 105968 (2021). <https://doi.org/10.1016/j.cmpb.2021.105968>
21. Pima Indians Diabetes Database. Available at: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

Chapter 52

A Detailed Schematic Study on Feature Extraction Methodologies and Its Applications: A Position Paper



Niharika Mohanty, Manaswini Pradhan, and Pradeep Kumar Mallick

Abstract Feature extraction is stated to be one of the most important aspects in the field of machine learning, image processing and pattern recognition. In these fields, as the dimension of the data increases, it becomes highly important to provide a reliable analysis for growing. The process of feature extraction mostly starts from basic dataset, and slowly, it builds features derived from the data to make an informative learning step for human reading. In order to present this process of learning, we thought of providing a position paper which will discuss all the criteria, information, methodology and existing work in the area of feature extraction in different fields and domain. A clear, descriptive analysis is presented that discusses the feature extraction methods for selecting significant features that will improve the quality of the results.

52.1 Introduction

With the advent of high-end science and technology, and methodologies, the data mostly acquired are really high in proportion. Hence, as a consequence, the analysis of data is no more complex. But the most challenging task that has always prevailed is the extraction of useful and meaningful features from the highly redundant set of features. This process of feature extraction plays a significant role for pattern recognition and machine learning. Dimensionality reduction is an utmost important topic, and it carries a much highlighted position in the view of feature extraction. The objective of this is to resolve the issue of “curse of dimensionality” that prevails in any high-dimensional feature. Usually, the high-dimensional feature poses a great pressure during the computational process for any type of machine learning models. Use of the concept of dimensionality reduction is mostly achieved by two famous concepts: feature selection and feature extraction.

N. Mohanty · M. Pradhan

P. G. Department of Information and Communication Technology, F. M. University, Balasore, India

P. K. Mallick (✉)

School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha, India
e-mail: pradeep.mallickfcs@kiit.ac.in

As stated earlier, feature plays a very important role in almost all the areas, that is, from data mining to machine learning. Concentrating to a greater extent on the mind and the concept of feature extraction, the assumption behind the descent is to transform the features and build a new set of features from the original band. Normally, this process typically incurs some information loss in the due course of action (process of dimensionality reduction) which is usually not reversible. In the area of pattern recognition and image processing, feature extraction is mostly tagged as a special form of dimensionality reduction. The use of feature extraction is very prominent in the pattern recognition, typically in the area of image processing. Applications related to it include document verification, character recognition, health insurance, script recognition, etc. Feature projection or very famously called feature extraction has the power to transform any original features set of new features set by the mode of transformation. It is considered as one of the prime steps in the process of classification. In the field of computer vision, where a large aggregation of images is used, the extraction process plays a critical part in transforming the enriched images into different characteristics which can be further used for the purpose of sorting and survival of the fittest.

The process of feature extraction mostly is caused after the preprocessing, but it need not forever be thus. As discussed earlier, the process of classification requires features to be extracted for constructing any proper pattern classification where the relevant information needs to be derived for characterizing each class accurately.

The paper is divided into the following sections: A thorough and detailed analysis of many papers is also discussed. Section 52.2 gives an overview about the work done by researchers in different domains. Section 52.3 describes the methodology. And lastly, Sect. 52.4 provides brief summary about the techniques used for feature extraction.

52.2 Literature Review

We know that machine learning and image processing are broad spectrum areas, and hence, when the dimension of the data grows, it becomes really important for providing reliable analysis. Hira and Gillies [1] depicted some of the famous feature extraction methodology for the extraction of many significant features in the microarray data. Nguyen et al. [2] proposed two feature extraction algorithms to solve the highway congestion problem. The proposed method called point-based and area-based methods proved to provide high levels of confidence in classifying the patterns. The point-based method by Nguyen et al. [3] involves some major steps like key point identification and feature vector formulation for representing the capability of the classifier on different traffic patterns. On the other hand, the area-based model approach by Krishnakumar et al. [4] provides a different highlight, where different discern congested patterns are used to represent different contour lines in any speed images. Ng et al. [5] used deep networks along with single descriptors which is a locally aggregated descriptor vector for the purpose of image retrieval. Choras [6]

implemented a feature extraction method that mapped image content with low-level features for content-based image retrieval and biometric systems that uses image content and image feature extraction. Features like shape, color, and texture were extracted to characterize the images.

El-gayar et al. [7] made a clear comparative analysis for low-level feature extraction on different computer vision problems like object recognition and other visionary objects. They discussed various feature-based and texture-based groups and assessed each on the basis of efficiency and performance. SIFT was presented by Lowe [8] extraction of features from images that can be different from scaling and rotation of the image. Later, Lowe et al. [9] proposed a PCA-SIFT algorithm that uses patch (gradient patch) instead of histograms for normalization. The PCA-based descriptors were quite robust and clearly distinctive to image deformation, but the extraction of the features was a bit slow. Vinay et al. [10] proposed an ORB-PCA-based feature extraction algorithm for face recognition that helps in dimensionality reduction of the descriptors which further improves the authentication performance over different existing algorithms. Li et al. [11] introduced a FAST feature point merged with SURF description matching technique that is basically used for real-time matching of the target. This proposed technique is said to be more fast and accurate than the existing SURF and FAST algorithm.

Batik image classification is quite needed to preserve the cultural heritage of Indonesia; hence, Azhar et al. [12] proposed a brain image classification technique using SIFT, bag of features and support vector machine to extract unique features of the batik images. A quite famous feature extraction technique based on a neural network model was presented by Oja et al. [13] where a 1D PCA was used and extended to multiple dimensions. Similarly, a 3-layer auto-associator network was introduced by Baldi and Hornik [14] that outperformed the PCA algorithm, in terms of performance, as in auto-associator network the performance is very good in nonlinear dimensionality reduction including the principal surfaces. Marinho et al. [15] proposed a feature extraction algorithm based on structural co-occurrence matrix with ECG signals. They proposed a method to detect cardiac arrest through the ECG signals, and the experimental results conducted upon it were mainly based on accuracy, sensitivity, specificity and computational cost. Another feature extraction method was introduced by Guler and Ubevli [16], where they have used spectral analysis of the ECG signal using discrete wavelet transformation technique. Using this technique, the statistical features like mean, mean energy, standard deviation and absolute mean ratio were extracted as a part of the task. Yu and Chen [17] proposed a feature extraction technique again based on DWT and Haar wavelet. Here, they used the first- and second-level detailed coefficient, where the coefficient provided the attributes like mean energy, signal coherence and signal morphology. These attributes in accordance with other features like QRS variance, mean RR intervals, etc., are used for extraction of a vector of 13 features. Yu et al. [18] proposed a feature extraction methodology based on RR interval and independent component analysis (ICA) where these independent components are calculated using the fast-ICA algorithm. Similarly, Ye et al. [19] implemented an eighth ordered Daubechies wavelet technique to extract coefficients from different levels of approximation coefficient.

This coefficient concatenated with the extracted coefficients and the dimensionality was reduced by PCA.

An unsupervised EEG feature extraction method is proposed by Sun et al. [20] based on echo state network for encoding the EEG signals to suitable features. Here, they have used an autoencoder and based on that an unsupervised EEG feature extraction method was introduced that provided an excellent performance along with a reduced computational complexity and higher efficiency. Han et al. [21] came up with an autoencoder model that was basically used for the feature extraction of the EEG signals. The authors also stated that their model was validated and trained with the AR coefficient, which is capable of differentiating EEG signals under varied physiological circumstances, and all the signals considered were linear in nature. Contrary to this, Wang et al. [22] stated that the above model introduced in [21] cannot independently extract and identify features without taking the highly complex and nonlinear EEG signals into account. Li et al. [23] proposed a feature extraction technique for EEG signals where Fourier spectral analysis was used for extraction and transformation of the EEG signals. Similarly, Sadati et al. [24] used discrete wavelet transform method for the extraction of features from EEG signals in terms of both time and frequency domains. It was also pointed out that EEG signal frequency would change with respect to time and over the time, and hence, Li et al. [25] proposed a short-time Fourier transform (STFT) for the purpose of calculation of the spectrum density. A feature extraction method based on the manifold is quite an effective and efficient technique for handling unsupervised classification task, but the major issue with this method is that it is quite sensitive to noise. So, Zhan et al. [26] thought of addressing this issue and came with a feature extraction method based on low-rank and sparsity preserving embedding of unsupervised learning which was quite robust to noise and was much better in terms of performance.

Liu et al. [27] introduced a flexible unsupervised feature extraction (FUFE) for the image classification task. Through this technique, they theoretically provided that PCA is a special case of FUFE and it provides a non-iterative algorithm to solve it. Also, the proposed model was very much suitable for handling nonlinear manifolds and characterize the local and global structures. Era of deep learning has started, and a feature extraction technique based on deep learning was introduced by Zhong et al. [28]; they proposed a hybrid model based on the premise of deep learning that combines multilevel feature extraction and learning to the task of recognition where identification of each image is assessed. They used a feature aggregation network (FAN) for extraction of CNN features by merging information at each layer. The validation of the CNN features was taken up by the recurrent comparative network (RCN). Gray and Tao [29] introduced the AdaBoost algorithm for selecting localized ensembling features in the domain of computer vision. They defined a feature space based on some assumption and later showed how class-specific representation object and the discriminative recognition model are learnt using the above model. Liao et al. [30] designed a feature descriptor using local maximal occurrence (LOMO) and cross-view quadratic discriminant analysis (XQDA) for maximizing the horizontal occurrence of local features. This method proved to be robust to illumination and viewpoint changes, and the same was observed from the experimental evaluation done

on four person re-identification databases. Likewise, a symmetry-driven method was introduced for the accumulation of features for the same feature re-identification database by Farenzena et al. [31]. They extracted features for color and textures of both symmetric and asymmetric types. The method used is quite robust against low resolution and illumination change, and it is usually used where the number of candidates varies continuously (considering only single frame image). Zhang et al. [32] proposed a manifold supervised machine learning method based on HIS feature extraction that is used for mapping fast and efficient nonlinear features. Apart from this, they also have designed a framework for extracting different topological networks in HSI data for dimensionality reduction. The results produced provided high classification accuracy. A two-dimensional algorithm based on regression model known as matrix regression preserving projection (MRPP) was developed by Luofeng et al. [33], for feature extraction. This technique is said to minimize the N-norm based on low-rank coefficient reconstruction error.

Shi et al. [34] proposed a feature extraction technique with a dynamic graph learning to solve issues like noise which reduces the performance affinity. Multi-view data are mostly exploited by the multi-view feature extraction, and this has been proposed by Yan et al. [35]; these methods are typically based on the concept of graph theory, where many graphs are constructed for representing data similarity. These graphs are then clubbed together into one, and then based on this, the features are extracted. But the major drawback of these methods is that the graph construction and the feature extraction are two separate processes, and hence, it leads to sub-optimal results. Roweis et al. [36] introduced a locally linear embedding (LLE) feature extraction method where the original structure (manifold) is reduced to a small amount of data dimensions. Again, this method is quite sensitive to the number of nearest neighbors which has a severe effect on the performance of the feature extraction method. Similarly, Wang et al. [37] proposed a projective unsupervised flexible embedding model with optimal graph called as PUF-E-OG for dimensionality reduction for image and video processing. The drawback of this method is that the learning process of the graph relies on the fixed graph which is quite unreliable. Krishnan and Athavale [38] discussed the feature extraction technique in the domain of biomedical signal processing on the basis of machine learning and artificial intelligence. Here, they also discussed the feature extraction technique ranging from basic A-to-D conversion to domain transformation.

Cepstrum analysis is a feature extraction technique which is mostly used in human speech signal analysis. This was proposed by Tabatabaei et al. [39], which use the cepstrum as the feature vector instead of the Fourier transform. These features are transformed into Mel scale and later to Mel frequency. This method cannot be used for capturing any transient pattern in the signal. Rather, they are mostly used for stationary signals. Nallapareddy et al. [40] have discussed a feature extraction technique called linear predictive coding (LPC), where the features extracted are linear prediction coefficient that helps in validating the values of the signals accurately. The authors have used the lattice prediction filter-based method with the LPC concept for coefficient generation, the results for the same are typically good, and it is quite robust to noise too. Morphological feature extraction method mostly involves physiological

properties, and it is a combination of basic mathematical functions and algorithms. Thiran et al. [41] used this technique for the purpose of disease classification using morphological feature extraction from digital histopathological images based on its shapes and size of the cells.

Aquino et al. [42] proposed a new feature extraction method based on the optic disk segmentation. They used the retinal database for the evaluation, and the results fetched offered an excellent success rate, quality and efficiency. Detection of arrhythmia can be done from ECG which is considered as a vital method for the purpose. Shi et al. [43] proposed a region-based feature extraction method ensembled with classifier for inter-patient heartbeat classification. The results generated by the method provided improved accuracy and sensitivity when compared with any other existing features extraction algorithm. A privacy preserving SIFT was presented by Hsu et al. [44] which had a domain encryption that uses Paillier cryptosystem. There the server had the capability for comparing two random ciphertexts. Similarly, a SecSIFT algorithm was introduced by Qin et al. [45, 46] where a preserving encryption algorithm was proposed. This ciphertext order matched the corresponding plain text and leaked some of the important crucial statistical information. Wang et al. [47] proposed a SIFT algorithm which is highly secured and leveraged the garbled circuit where the key point location is leaked. Also, Li et al. [48] studied a face recognition privacy preserving technique by using SIFT technique which itself is a part of the fuzzy classification recognition process. They proposed the method where the client encrypts the image data and sends the results off to the company. The method was found to be highly scalable, efficient and reliable than any other similar existing method. Sulatana and Shubhangi [49] proposed a feature extraction algorithm for encrypted images without revealing any information to cloud service providers. The complexity and the overhead incurred are basically upon the cloud and not upon the client.

Liu et al. [50] proposed an algebraic feature extraction method for image recognition. They used optimal discriminant criteria for extracting algebraic features from the sample images. This approach is said to have a good performance than any of the existing approach. Another feature extraction technique based on gray-level co-occurrence matrix is proposed by Mohanaiah et al. [51] for image datasets. The method proposed is said to have extracted statistical texture features (second order) for motion estimation of images, the results produced have high level of accuracy, and less computational time is incurred. A feature extraction method based on adaptive flow orientation for fingerprint images was introduced by Ratha et al. [52]. This technique was responsible for extracting structural features from the image set, and the result provided produced a high level of accuracy and is quite reliable too. Tsai et al. [53] made a comparison study combining various feature extraction methods like global and local block-based and region-based methods. The same was implemented on an image database, and the result produced stated that global and local block method outperforms the latter on. Zhu et al. [54] considered a LBP, Gabor wavelet and edge orientation histogram along with SURF descriptor for extraction of local features from the image database which is also said to improve the overall efficiency

of the system. Zhou et al. [55] proposed a locality image representation called hierarchical Gaussianization which is a mix of Gaussian mixture model and Gaussian map for locality feature extraction information. In computer vision and image processing, and analysis, edge detection is one of the basic vital steps for edge feature extraction. Cui et al. [56] implemented some existing feature extraction techniques to remove noise from the image dataset, and they found that from all the techniques (that they have used), binary morphology provides better results as compared to the rest. Yuille et al. [57] proposed a technique for detecting and describing features for faces using deformable templates for detecting specified set of parameters that enables to get a clear idea about the shape of the features. These templates are subject to change in size and other parameters for matching with the data. A feature extraction method for offline recognition of segmented characters was developed by Trier et al. [58]. The method was based on the representation of the characters like binary characters, gray-level characters, etc. They focused on discussing the feature extraction properties based on reconstructability, distortion and variability of the characters. Hermansky et al. [59] proposed a feature extraction method for conventional hidden Markov model (HMM) system in word recognition system by using neural network discriminative feature with Gaussian mixture distribution modeling.

Li et al. [60] proposed a maximum margin criterion (MMC) which further gives rise to another method called a LDA. The new feature extracted from this feature extractor method uses varied types of constraint and is quite independent of the non-singularity in within a class scatter matrix. Gorodetsky and Samoylov [61] proposed feature extraction method for machine learning where the features are automatically extracted for classification or recognition. They assume that the large scale of learning data is available and they are represented in object form. In existing techniques, this is not so. Rather in existing techniques for feature extraction, they are mostly oriented toward flat table. Fukuma et al. [62] presented a clear study on the disease classification and feature extraction on glioma pathology images where automatic image analysis methods are used for nuclei segmentation and labeling for histopathology data. Priyanka and Suresh [63] have presented a feature extraction method for crop disease identification using histogram of oriented gradient (HOG) algorithm from the crop image dataset. After processing the image, they extracted the features using HOG algorithm which was later classified using SVM classifier for the measurement of accuracy. Ranjan et al. [64] proposed a technique for detecting and classification leaf disease by extracting feature by RGB conversion in the hue saturation format. They used the neural network method for the purpose of classification and accuracy measurement. Dhaygude and Kumbhar [65] proposed a feature extraction method where the color co-occurrence method is used and texture statistics is used for disease analysis.

Hrishkesh et al. [66] tried using the HIS hue image method for disease identification in leaf, where the feature extraction is done in the segmented disease area. This technique helped in giving a clear discrimination of diseased spots by extracting color and size. Dony and D'Souza [67] provided a study on the existing feature extraction techniques which will help in understanding the disease and detection of the disease in agricultural products. Dessouky and Elrashidy [68] used five different

types of optimization technique on two proposed Alzheimer disease feature extraction techniques for getting the optimum number of features that give higher accuracy. Andavarapu and Vatsavayi [69] used co-occurrence of histogram of oriented gradient (CoHOG) for detection of human. They used weights with the CoHOG for calculating the weighted co-occurrence matrix for feature vector. Alzughabi and Chaczko [70] used a feature extraction method for human detection model in video frames using HOG and local binary pattern features. These are later used to generate a feature vector, and SVM is used for the purpose of detection.

52.3 Feature Extraction

52.3.1 *Methodologies*

52.3.1.1 Principal Component Analysis (PCA)

PCA is a famous linear unsupervised dimensionality reduction technique where the premise is to use orthogonal transformation methodology for converting a set of correlated features/variables into linearly uncorrelated variable. This technique was introduced by Pearson and was often used for multivariate data analysis. These uncorrelated variables are typically called as principal components. PCA tries to search for a subspace where the reconstruction error (average) of the data (training) is lessened. As stated before, the main components are usually less or may be equal to the number of original variables. These components are mostly hard to interpret. This statistical technique aims at reducing the dimension of the data, and it runs essentially on the principle of factoring for the extraction of pattern in a linear arrangement. But there is a constraint to this very famous technique. PCA is quite prone and sensitive to outliers. Hence, many studies use some kind of formulation for the optimization to this problem of PCA.

52.3.1.2 Linear Discriminant Analysis (LDA)

LDA is a generalized version of Fisher's linear discriminant, which is mostly a statistical and pattern recognition method for finding linear features for separating two or more classes. It is another popular dimensionality reduction/feature extraction method before classification overtook it. LDA is closely associated with analysis of variance (ANOVA), regression, PCA and factor analysis as they all consider a linear combination of variables for explaining any data. This method possesses the ability to retrieve back and avoid the misclassification problem by employing linear decision boundaries for maximization between-class to within-class ratio. Usually, discriminant analysis plays an utmost important part in the creation of multivariate statistics. This technique has several applications in different fields ranging from

health science to industry to social scientific discipline and others. Again, the limitation of this technique is that the distribution (probability) is assumed to be Gaussian for two classes which are further assumed to be of equal covariance matrix.

52.3.1.3 Gabor Feature

Gabor features or Gabor jet, Gabor banks or multiresolution Gabor feature refers to a method which is typically used in the field of computer vision, face recognition, image processing, etc. This technique considers pieces of local information, which are further combined to recognize an object of concern. This technique is mostly based on the 2D Gabor filter function, which is a product of Gaussian function and Euler function. Gabor filter function typically gives an optimal solution to both spatial and frequency-related fields for extracting features locally. They can also be used as a directional extractor of feature as they capture energy-related value from multiple direction related to the visual system. But one of the major constraints of this method is that it has quite complex in comparison and the graphical intervention to be trained manually.

52.3.1.4 Scale-Invariant Feature Transform (SIFT)

SIFT is a feature extraction algorithm used frequently in the area of computer vision and image processing for detection of features in the images. These typically working processes of the SIFT method involve extraction of the images from a lot of existing reference images that would be further preserved in a database. Afterward, the target is compared with new image individually and the matched features are drawn out. This method is applied for finding stable and salient feature from an icon. As mentioned earlier, the algorithm used in this technique typically converts an image data into a local feature vector known as SIFT descriptor. The features have the capability to transform geometrically for scaling and rotation. It is one of the efficient locally invariant feature descriptors available. One of the major issues with this technique is that the SIFT features are very prone and are subject to change in illumination and noise which sometimes make them inconvenient to use.

52.4 Comparative Analysis

The comparative analysis is described in Table 52.1.

Table 52.1 Comparative analysis

S. no.	Authors	Findings
1	Hira and Gillies [1]	Provided survey on different existing feature extraction techniques used in microarray data
2	Nguyen et al. [2]	Point-based and area-based feature extraction method introduced for highway congestion problem
3	Nguyen et al. [3]	Key point identification and feature vector formulation were used for identification of traffic patterns
4	Krishnakumar et al. [4]	Area-based model approach was used for discern congested pattern for analyzing contour images in speed images
5	Ng et al. [5]	Used deep networks for image retrieval
6	Choras [6]	Proposed a feature extraction method for content-based image retrieval and biometric systems
7	El-gayar et al. [7]	Comparative analysis was performed taking various feature extraction methods for object recognition
8	Lowe [8]	Proposed a feature extraction method for image data; they did not consider scaling and rotation for extracting features
9	Lowe et al. [9]	PCA-SIFT algorithm introduced on image dataset Robust and clearly distinctive to image deformation Slow in nature
10	Vinay et al. [10]	ORB-PCA-based feature extraction algorithm for face recognition Performance is far better than existing techniques
11	Li et al. [11]	FAST feature point merged with the SURF description matching method proposed Fast and accurate than the existing technique
12	Azhar et al. [12]	Introduced a brain image classification technique using SIFT, bag of features and support vector machine to extract unique features of the batik images
13	Oja et al. [13]	1D PCA was used for extracting features
14	Baldi and Hornik [14]	3-layer auto-associator network was proposed Outperformed PCA Performance is too good with nonlinear dimensionality reduction

(continued)

Table 52.1 (continued)

S. no.	Authors	Findings
15	Marinho et al. [15]	Introduced structural co-occurrence matrix with ECG signal features extraction technique Computational cost is less, and accuracy is said to be high than the existing technique
16	Guler and Ubevli [16]	Used spectral analysis of the ECG signal using discrete wavelet transformation technique Features based on statistics like mean, mean energy, standard deviation, etc., were extracted
17	Yu and Chen [17]	Proposed feature extraction technique again based on DWT and Haar wavelet
18	Yu et al. [18]	Feature extraction methodology based on RR interval and independent component analysis (ICA)
19	Ye et al. [19]	PCA was used for dimensionality reduction in accordance with the eighth ordered Daubechies wavelet technique for extracting coefficients
20	Sun et al. [20]	Autoencoder was used to propose an EEG feature extraction technique for echo state network
21	Han et al. [21]	Autoencoder model was used and trained with AR coefficients for differentiating EEG signals Linear signals were only considered
22	Wang et al. [22]	To support [21], highly complex and nonlinear EEG signals were used for identifying features
23	Li et al. [23]	Fourier spectral analysis was used for feature extraction
24	Sadati et al. [24]	Discrete wavelet transform method for the extraction of features from EEG signals
25	Li et al. [25]	Short-time Fourier transform (STFT) was used for feature extraction and for calculating spectral density
26	Zhan et al. [26]	Proposed a low-rank and sparsity preserving embedding for unsupervised learning method for handling unsupervised classification task Robust to noise and better performance
27	Liu et al. [27]	Proposed flexible unsupervised feature extraction (FUFE) for the image classification task
28	Zhong et al. [28]	Feature aggregation network (FAN) and CNN were used for extracting feature from image dataset
29	Gray and Tao [29]	AdaBoost algorithm for selecting localized ensembling features was used in the domain of computer vision

(continued)

Table 52.1 (continued)

S. no.	Authors	Findings
30	Liao et al. [30]	Feature descriptor using local maximal occurrence (LOMO) and cross-view quadratic discriminant analysis (XQDA) was proposed for re-identification database
31	Farenzena et al. [31]	Similarly, for re-identification database feature extraction another symmetry-driven feature extraction method was introduced
32	Zhang et al. [32]	Proposed a HIS feature extraction that was used for mapping fast and efficient nonlinear features
33	Luofeng et al. [33]	Introduced a 2D algorithm based on regression model known as matrix preserving projection (MRPP)
34	Shi et al. [34]	Another feature extracting technique based on dynamic graph learning was proposed Reduced noise and increased the performance
35	Yan et al. [35]	Introduced a multi-view feature extraction for multi-view data based on the concept of graph theory
36	Roweis et al. [36]	Locally linear embedding (LLE) feature extraction method was proposed for dimensionality reduction
37	Wang et al. [37]	Proposed a projective unsupervised flexible embedding model with optimal graph called as PUFE-OG for dimensionality reduction for image and video processing
38	Krishnan and Athavale [38]	Proposed a feature extraction technique in the domain of biomedical signal processing on the basis of machine learning and artificial intelligence
39	Tabatabaei et al. [39]	Used cepstrum as the feature vector instead of the Fourier transform for transforming into Mel scale and later to Mel frequency
40	Nallapareddy et al. [40]	Proposed a feature extraction technique called linear predictive coding (LPC), where the features extracted are linear prediction coefficients for validating the signals
41	Thiran et al. [41]	Used morphological feature extraction from digital histopathological images for disease classification
42	Aquino et al. [42]	Proposed a feature extraction method based on the optic disk segmentation where retinal database was used

(continued)

Table 52.1 (continued)

S. no.	Authors	Findings
43	Shi et al. [43]	Region-based feature extraction method ensembled with classifier for inter-patient heartbeat classification was proposed
44	Hsu et al. [44]	Privacy preserving SIFT algorithm was used where Paillier cryptosystem was further used for extraction of features from the ciphertext
45	Qin et al. [45, 46]	SecSIFT algorithm was proposed that is a preserving encryption algorithm
46	Wang et al. [47]	Introduced SIFT algorithm that is highly secured and leveraged the garbled circuit where the key point location was leaked
47	Li et al. [48]	Studies face recognition privacy preserving technique by using SIFT technique
48	Sulatana and Shubhangi [49]	Introduced a feature extraction algorithm for encrypted images without revealing any information to cloud service providers
49	Liu et al. [50]	Feature extraction method for image recognition was proposed that used optimal discriminant criteria
50	Mohanaiah et al. [51]	Gray-level co-occurrence matrix-based feature extraction method was proposed for image dataset for statistical feature extraction
51	Ratha et al. [52]	Feature extraction method based on adaptive flow orientation for fingerprint images was proposed
52	Tsai et al. [53]	Introduced a comparative study for combining various feature extraction methods like global and local block-based and region-based methods
53	Zhu et al. [54]	LBP, Gabor wavelet and edge orientation histogram along with SURF descriptor for extraction of local features from the image database
54	Zhou et al. [55]	Proposed a locality image representation called hierarchical Gaussianization for locality feature extraction information
55	Cui et al. [56]	Used feature extraction techniques to remove noise from the image dataset, where binary morphology provided better results
56	Yuille et al. [57]	Proposed a technique for detecting and describing features for faces using deformable templates
57	Trier et al. [58]	Proposed feature extraction method for offline recognition of segmented characters

(continued)

Table 52.1 (continued)

S. no.	Authors	Findings
58	Hermansky et al. [59]	Introduced feature extraction method for conventional hidden Markov model (HMM) system in word recognition system
59	Li et al. [60]	Maximum margin criterion (MMC) along with LDA was used for feature extraction
60	Gorodetsky and Samoylov [61]	Proposed feature extraction method for machine learning where the features are automatically extracted for classification or recognition
61	Fukuma et al. [62]	Performed feature extraction on glioma pathology images where automatic image analysis methods are used for nuclei segmentation
62	Priyanka and Suresh [63]	Used a feature extraction method for crop disease identification using histogram of oriented gradient (HOG) algorithm
63	Ranjan et al. [64]	Used a technique for detecting and classification of leaf disease by extracting feature by RGB conversion in hue saturation format
64	Dhaygude and Kumbhar [65]	Color co-occurrence method was used for features extraction for disease analysis
65	Hrishkesh et al. [66]	Use HIS hue image method for disease identification in leaf
66	Dony and D'Souza [67]	Made a comparative study of the existing feature extraction technique for detection of the disease in agricultural products
67	Dessouky and Elrashidy [68]	Five different types of optimization technique on two proposed Alzheimer disease feature extraction techniques for getting optimum number of features
68	Andavarapu and Vatsavayi [69]	Co-occurrence of histogram of oriented gradient (CoHOG) for detection of human
69	Alzughabi and Chaczko [70]	Introduced a feature extraction method for human detection model in video frames using HOG and local binary pattern features

52.5 Conclusion

In this work, we proposed to summarize types of feature extraction techniques and same methodology of doing it. Last but not least, we discussed many of the research contribution in different fields and in varied aspects of feature extraction. We thought of giving a clear outlook about the types of feature extraction existing from time domain to region domain in applications like disease detection, pattern recognition and image processing, signals (ECG and EEG), etc. Each technique is discussed, and

each has its own pros and cons. Hence, stating which one is the best is quite unmanageable. In the future, we would be working on some of these features extraction methodologies based on our application and domain.

References

1. Hira, Z.M., Gillies, D.F.: A review of feature selection and feature extraction methods. *Appl. Microarray Data Adv. Bioinform.* **2015**, 1–13 (2015)
2. Nguyen, T.T., Krishnakumaria, P., Calverta, S.C., Vub, H.L., Lint, H.V.: Feature extraction and clustering analysis of highway congestion. *Transp. Res. Part C* **100**, 238–258 (2019)
3. Nguyen, H.N., Krishnakumari, P., Vu, H.L., Lint, H.V.: Traffic congestion pattern classification using multi-class svm. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pp. 1059–1064 (2016)
4. Krishnakumari, P., Nguyen, T., Heydenrijk-Ottens, L., Vu, H.L., van Lint, H.: Traffic congestion pattern classification using multiclass active shape models. *J. Transp. Res. Board* **2645**, 94–103 (2017)
5. Ng, J.Y.H., Yang, F., Davis, L.S.: Exploiting local features from deep networks for image retrieval. In: Proceedings of IEEE International Conference Computer Vision Pattern Recognition, DeepVision Workshop (CVPRW), pp. 53–61 (2015)
6. Choras, R.S.: Image feature extraction techniques and their applications for CBIR and biometrics systems. *Int. J. Biol. Biomed. Eng.* **1**(1), 6–16 (2007)
7. El-Gayar, M.M., Soliman, H., Meky, N.: A comparative study of image low level feature extraction algorithms. *Egypt. Inform. J.* **14**(2), 175–181 (2013)
8. Lowe, D.: Distinctive image features from scale-invariant key points. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
9. Ke, Y., Sukthankar, R., PCA-SIFT: A more distinctive representation for local image descriptors. In: Proceedings of Conference Computer Vision and Pattern Recognition, pp. 511–517 (2004)
10. Vinay, A., Kumar, C.A., Shenoy, G.R., Murthy, K.N.B., Natrajan, S.: ORB-PCA Based feature extraction technique for face recognition. *Procedia Comput. Sci.* **58**, 614–621 (2014)
11. Lia, A., Jiang, W., Yuana, W., Daia, D., Zhang, S., Wei, Z.: An improved FAST+SURF fast matching algorithm. *Procedia Comput. Sci.* **107**, 306–312 (2017)
12. Azhara, R., Tuwohingidea, D., Kamudia, D., Sarimuddina, Suciati, N.: Batik image classification using SIFT feature extraction, bag of features and support vector machine. *Procedia Comput. Sci.* **72**, 24–30 (2015)
13. Oja, E.: A simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15**(3), 267–273 (1982)
14. Baldi, P., Hornik, J.: Neural networks and principal component analysis: learning from examples without local minima. *Neural Netw.* **2**(1), 53–58 (1989)
15. Marinho, L.B., de Nascimento, N.M.M., Wellington, J., Souza, M., Gurgel, M.V., Rebouças Filho, P.P., de Albuquerque, V.H.C.: A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification. *Fut. Generat. Comput. Syst.* **97**, 564–577 (2019)
16. Güler, İ., Übeyli, E.D.: ECG beat classifier designed by combined neural network model. *Pattern Recogn.* **38**, 199–208 (2005)
17. Yu, S.N., Chen, Y.H.: Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. *Pattern Recogn. Lett.* **28**, 1142–1150 (2007)
18. Yu, S.N., Chou, K.T.: Integration of independent component analysis and neural networks for ECG beat classification. *Exp. Syst. Appl.* **34**, 2841–2846 (2008)
19. Ye, C., Coimbra, M.T., Kumar, B.V.: Arrhythmia detection and classification using morphological and dynamic features of ECG signals. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1918–1921 (2010)

20. Sun, L., Jin, B., Yang, H., Tong, J., Liu, C., Xiong, H.: Unsupervised EEG feature extraction based on echo state network. *Inf. Sci.* **475**, 1–17 (2019)
21. Han, M., Sun, L.: EEG signal classification for epilepsy diagnosis based on AR model and rvm. In: 2010 International Conference on Intelligent Control and Information Processing, pp. 134–139 (2010)
22. Wang, L., Xue, W., Li, Y., Luo, M., Huang, J., Cui, W., Huang, C.: Automatic epileptic seizure detection in EEG signals using multi-domain feature extraction and nonlinear analysis. *Entropy* **19**(6), 222 (2017)
23. Polat, K., Gne, S.: Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast fourier transform. *Appl. Math. Comput.* **187**(2), 1017–1026 (2007)
24. Sadati, N., Mohseni, H.R., Maghsoudi, A.: Epileptic seizure detection using neural fuzzy networks. In: 2006 IEEE International Conference on Fuzzy Systems, pp. 596–600 (2006)
25. Li, Y., Liu, Q., Tan, S.-R., Chan, R.H.M.: High-resolution time-frequency analysis of EEG signals using multiscale radial basis functions. *Neurocomputing* **195**, 96–103 (2016)
26. Zhan, S., Wu, J., Han, N., Wen, J., Fang, X.: Unsupervised feature extraction by low-rank and sparsity preserving embedding. *Neural Netw.* **109**, 56–66 (2019)
27. Liu, Y., Nie, F., Gao, Q., Gao, X., Han, J., Shao, L.: Flexible unsupervised feature extraction for image classification. *Neural Netw.* **115**, 65–71 (2019)
28. Zhong, W., Jiang, L., Zhang, T., Ji, J., Xiong, H.: Combining multilevel feature extraction and multi-loss learning for person re-identification. *Neurocomputing* **334**, 68–78 (2019)
29. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European Conference on Computer Vision, pp. 262–275 (2008)
30. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
31. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person reidentification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2360–2367 (2010)
32. Zhang, P., He, H., Gao, L.: A nonlinear and explicit framework of supervised manifold-feature extraction for hyperspectral image classification. *Neurocomputing* **337**, 315–324 (2019)
33. Luofeng, X., Ming, Y., Ling, W., Feng, T., Guofu, Y.: Matrix regression preserving projections for robust feature extraction. *Knowl.-Based Syst.* **161**, 35–46 (2018)
34. Shi, D., Zhu, L., Cheng, Z., Li, Z., Zhang, H.: Unsupervised multi-view feature extraction with dynamic graph learning. *J. Vis. Commun. Image Retrieval* **56**, 256–264 (2018)
35. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(1), 40–45 (2007)
36. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)
37. Wang, W., Yan, Y., Nie, F., Yan, S., Sebe, N.: Flexible manifold learning with optimal graph for image and video representation. *IEEE Trans. Image Proces.* **99**, 2664–2675 (2018)
38. Krishnan, S., Athavale, Y.: Trends in biomedical signal feature extraction. *Biomed. Signal Process. Control* **43**, 41–63 (2018)
39. Tabatabaei, T.S., Krishnan, S., Guergachi, A.: Emotion recognition using novel speech signal features. In: 2007 IEEE International Symposium on Circuits and Systems, pp. 345–348 (2007)
40. Nallapareddy, H., Krishnan, S., Kolios, M.: Parametric analysis of ultrasound backscatter signals for monitoring cancer cell structural changes during cancer treatment. *Cancer Acoust.* **35**(2), 47–54 (2007)
41. Thiran, J.P., Macq, B.: Morphological feature extraction for the classification of digital images of cancerous tissues. *IEEE Trans. Biomed. Eng.* **43**, 1011–1020 (1996)
42. Aquino, A., Gegúndez-Arias, M.E., Marín, D.: Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques. *IEEE Trans. Med. Imaging* **29**(11), 1860–1869 (2010)

43. Shi, H., Wang, H., Zhang, F., Huang, Y., Zhao, L., Liu, C.: Inter-patient heartbeat classification based on region feature extraction and ensemble classifier. *Biomed. Signal Process. Control* **51**, 97–105 (2019)
44. Hsu, C.Y., Lu, C.S., Pei, S.C.: Image feature extraction in encrypted domain with privacy-preserving sift. *IEEE Trans. Image Process.* **21**(11), 4593–4607 (2012)
45. Qin, Z., Yan, J., Ren, K., Chen, C.W., Wang, C.: Towards efficient privacy-preserving image feature extraction in cloud computing. In: The ACM International Conference, pp. 497–506 (2014)
46. Qin, Z., Yan, J., Ren, K., Chen, C.W., Wang, C.: Secsift: Secure image sift feature extraction in cloud computing. *ACM Trans. Multimed. Comput. Commun. Appl.* **12**(4s), 65–75 (2016)
47. Wang, Q., Hu, S., Ren, K., Wang, J., Wang, Z., Du, M.: Catch me in the dark: Effective privacy-preserving outsourcing of feature extractions over image data. In: IEEE INFOCOM 2016 - IEEE Conference on Computer Communications, pp. 1–9 (2016)
48. Li, P., Li, T., Yao, Z.A., Tang, C.M., Li, J.: Privacy-preserving outsourcing of image feature extraction in cloud computing. *Soft. Comput.* **21**(15), 4349–4359 (2017)
49. Sultana, S.F., Shubhangi, D.C.: Privacy preserving LBP based feature extraction on encrypted images. In: 2017 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–10 (2017)
50. Yong, K.L., Chen, Q., Tang, J.Y.: Image texture feature extraction using GLCM approach. *Pattern Recogn.* **26**(6), 903–911 (1993)
51. Mohanaiah, P., Sathyaranayanan, P., GuruKumar, L.: Image texture feature extraction using GLCM approach. *Int. J. Sci. Res. Publ.* **3**(5), 290–294 (2013)
52. Ratha, N.K., Chen, S., Jain, A.K.: Adaptive flow orientation-based feature extraction in fingerprint images. *Pattern Recogn.* **28**(11), 1657–1672 (1995)
53. Tsai, C.F., Lin, W.C.: A comparative study of global and local feature representations in image database categorization. In: Proceedings of 5th International Joint Conference on INC, IMS & IDC, pp. 1563–1566 (2009)
54. Zhu, J., Hoi, S., Lyu, M.: Near-duplicate keyframe retrieval by nonrigid image matching. In: Proceedings of ACM MM, pp. 41–50 (2008)
55. Zhou, X., Cui, N., Li, Z.: Hierarchical gaussianization for image classification. In: Proceedings of ICCV, pp. 1971–1977 (2009)
56. Cui, F., Zou, L., Song, B.: Edge feature extraction based on digital image processing techniques. In: 2008 IEEE International Conference on Automation and Logistics, pp. 1–10 (2008)
57. Yuille, A.L., Hallinan, P.R.W., Cohen, D.S.: Feature extraction from faces using deformable templates. *Int. J. Comput. Vis.* **8**(2), 99–111 (1992)
58. Due, Ø., Anil, T., Jain, K., Taxt, T.: Feature extraction methods for character recognition—a survey. *Pattern Recogn.* **29**(4), 641–662 (1996)
59. Hermansky, H., Ellis, D.P.W., Sharma, S.: Tandem connectionist feature extraction for conventional HMM systems. In: Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 6–12 (2000)
60. Li, H., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. *IEEE Trans. Neural Netw.* **17**(1), 157–165 (2006)
61. Gorodetsky, V., Samoylov, V.: Feature extraction for machine learning: logic–probabilistic approach. In: JMLR: Workshop and Conference Proceedings The Fourth Workshop on Feature Selection in Data Mining, pp. 55–65 (2010)
62. Fukuma, K., Surya Prasath, V.B., Kawanaka, H., Aronow, B.J., Takase, H.: A study on feature extraction and disease stage classification for Glioma pathology images. In: 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 10–20 (2016)
63. Priyankha, J.J., Kumar, K.S.: Crop disease identification using a feature extraction HOG Algorithm. *Asian J. Appl. Sci. Technol. (AJAST)* **1**(3), 35–39 (2017)
64. Ranjan, M., Rajiv Weginwar, M., Joshi, N., Ingole, A.B.: Detection and classification of leaf diseases using artificial neural network. *Int. J. Tech. Appl.* **13**–20 (2015)
65. Dhaygude, S.B., Kumbhar, N.P.: Agricultural plant leaf disease detection using image processing. *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.* **2**(1), 1–10 (2013)

66. Kanjalkar, H.P., Lokhande, S.S.: Feature extraction of leaf diseases. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **3**(1), 1–5 (2014)
67. Priya, P., D’souza, D.A.: Study of feature extraction techniques for the detection of diseases of agricultural products. *Int. J. Innov. Res. Electr. Electron. Instrum. Control. Eng.* **3**(1), 4–8 (2015)
68. Dessouky, M.M., Elrashidy, M.A.: Feature extraction of the Alzheimer’s disease images using different optimization algorithms. *J. Alzheimer’s Dis. Parkinsonism* **6**(2), 1–11 (2016)
69. Andavarapu, N., Vatsavayi, V.K.: Weighted CoHOG (W-CoHOG) feature extraction for human detection. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving, pp. 273–283 (2010)
70. Alzughabi, A., Chaczko, Z.: Human detection model using feature extraction method in video frames. In: 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–8 (2016)

Chapter 53

Image Colorization Using CNNs



R. I. Minu, S. Vishnuvardhan, Ankit Pasayat, and G. Nagarajan

Abstract This paper presents a survey on the current technological processes that are used to colorize images with little or no human intervention, using image processing and machine learning techniques. We also analyze the more recent deep- learning-based approaches, specifically those using convolutional neural networks (CNNs), which are specially optimized mathematical models for image processing and manipulation. By diving deep into the mathematical and statistical theory behind the implementation of these models, we have managed to decipher the essence of the process of colorization. Here, in this paper we present a comparative study of different approach and then propose changes that may theoretically increase the accuracy and color-reproducibility of these models.

53.1 Introduction

The term ‘colorization’ refers to the process of converting a grayscale (black-and-white) image into a colored one, without any prior information about the colors in the image. Essentially, image colorization is a problem of information recovery, where the information in the color channels of the input image is ‘lost’. The colorizer then proceeds to recreate this lost information from the data that it has encountered during training.

Historically, the process of colorization was almost entirely done manually, and still is, to some extent. Old photographs and memoirs are contracted to digital artists

R. I. Minu (✉)

SRM Institute of Science and Technology, Kattankulathur, India
e-mail: minur@srmist.edu.in

S. Vishnuvardhan

Site Reliability Engineer, CRED, Bangalore, India

A. Pasayat

Infosys, Bangalore, India

G. Nagarajan

Sathyabama Institute of Science and Technology, Chennai, India

who painstakingly add color to the black-and-white or sepia images, often at an exorbitant fee. In the present day, automatic colorization systems have reached the point, where humans cannot spot the difference between a base real object and a colorized test image. In reference papers [1–3], we can see examples of systems that are early proof-of-concepts and those that are complete algorithmic implementations complete with baseline testing.

Theoretically, a simple statistical model will be able to compute the most probable color tuple value (R, G, B) for a given pixel in a grayscale image, based on the images that it has previously encountered. But, in practice, such naïve implementations seldom merit any importance. Statistical colorizers often produce results that are not coherent and unpleasing to the human eye, due to the predictability of the system.

However, if a statistical colorizer is trained using several different images of an object in an angle that is, if the training images and the input image are visually similar to a large extent, then the colorization produced is satisfactory. In practice, it is not possible to obtain multiple different images of an object, enough to train a basic statistical colorizer.

Thus, we move toward the next innovation in the field of image processing to help us improve our theoretical model further—object classification. By using an object classifier as the base compute for our colorizer system, we can optimize the system to color an object in a way, no matter how it appears in the input image. For example, by classifying the objects in the background as trees or foliage, we can proceed to color them shades of green without losing legibility in the output image. The challenge in image colorization occurs with the fact that objects with different properties may possess multiple colors. The colorizer system must be impervious to radically different information in the training dataset, pertaining to the same object class.

For example, a plastic ball may appear in the training images as blue, red or any other color. The color decision made by the colorizer system about a similar ball in a testing image must be coherent to one of the images in the dataset. Furthermore, the object ‘ball’ may have an implicit relation to another object ‘bat’, such that an image containing both a bat and a ball, would be colorized in such a way that the ball is colored appropriately with respect to the bat. Thus, simple object classification cannot help us improve the accuracy of our hypothetical system much. Here, we employ another powerful image processing algorithmic paradigm—edge detection. Edge detection refers to the process of computing the vectors corresponding to relatively bold lines or curves (explicit edges) and areas with stark differences in color/contrast (implicit edges). By computing the edges in a training image instead of blindly trying to classify the objects present in the image, we tend to create more accurate associations between vectors and their corresponding colorizations. This is highly sought after in mathematical models such as neural networks due to their reception of faint associations with pronounced results. Neural networks are a mathematical learning model that is based on the structure of the biological nerve cells ‘neurons’, which are able to retain and react to information that is perceived by the sense organs. Neural networks are essentially layering of cells that hold a decimal value, and are affected by values of the cells in the previous layer and influence the next

layer. One kind of neural network that is commonly used for image processing is the convolutional neural network (CNN). They are also called shift-invariant networks or space-invariant networks. CNNs contain neurons arranged in a three-dimensional grid with only a part of them being exposed to the previous and the next layer, this fraction of the neuron layer is known as the receptive field.

53.2 State of Art

Convolutional neural networks are highly suitable for processing visual data due to their implicit structure and characteristics. One of the most popular implementations of CNNs for image processing is the VGG16 by Simonyan et al. The VGG16 is used for object classification and its architecture and weights used are freely available for general use in their website. The model achieves a 92.7% top-5 test accuracy in ImageNet, the standard dataset for image processing and training tasks. It is a collection of 14 million images that can be obtained without copyrights for research and academic use. As is apparent from the name, the VGG16 network contains 16 layers that are designed to manipulate and retain information from the training images in [4] order to detect macroscopic objects and subjects (Fig. 53.1).

At this point, we start to analyze the work done by Zhang et al., in their ECCV 2016 submission ‘Colorful Image Colorization’. They have used a custom architecture which is known as AlexNet and achieved positive results in their so called ‘colorization Turing Test’ in which human participants are asked to differentiate between a real image and an artificially colorized one. Zhang et al. have managed

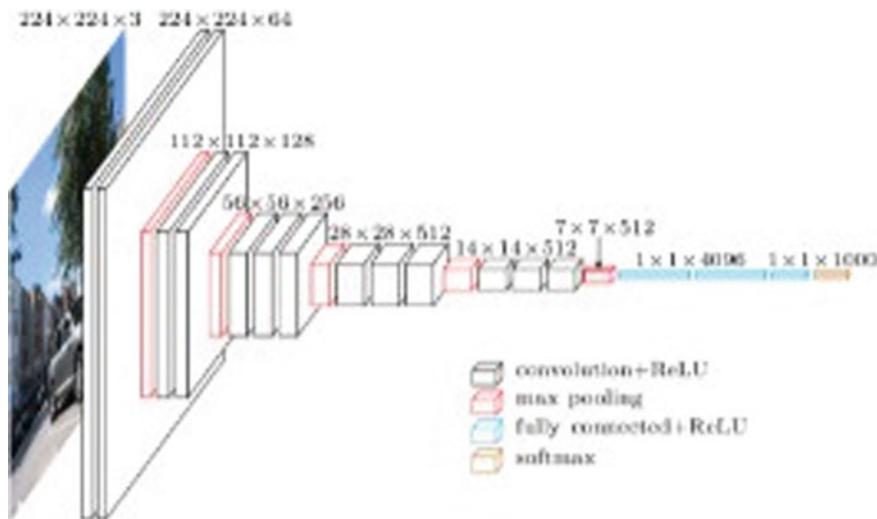


Fig. 53.1 Macro-architecture of the VGG16 network

to solve the object-color redundancy problem sufficiently, the fundamental fact that an object may have multiple plausible colorizations. For achieving this, they have incorporated a unique loss metric which is re-weighted during training time instead of being static.

This allows their model to use the entire information in the training dataset instead of converging on a few highly contributing images. Finally, they have taken the annealed mean of the distribution to produce the final colorization.

Here, the parameter T, closely relates to the color temperature in the output image that is, lowering T tends to produce images with more reddish and yellowish hues and increasing T produces images with pronounced blues and greens. By trial and error, they have chosen the T-value 0.38, which gives satisfactory results even though it is on the warmer end of the color spectrum. This may be due to the overall color temperature of the dataset being on the slightly warmer side. Next, we look at the work by Baldassare et al., in their research titled ‘Image Colorization using CNNs and Inception-ResNet-v2’. They have employed a pre-trained feature extractor, namely, the Inception ResNet v2, to cut down on training and implementation time.

The Inception ResNet models are an incremental model with each subsequent version having significant upgrades over the previous ones, in terms of accuracy and general reproducibility. The current version of the ResNet is v4, however, at the time of their paper, Baldassare et al. have used the v2 model. The paradigm used by them is also very different, as they have considered the CIE $L \times \alpha \times \beta$ color space instead of the traditional (R, G, B) space. By doing this, they have framed the colorization problem as one where there exists a function f, such that given only the Luminance (L) channel of an image, f can compute the α, β channels with enough accuracy such that the entirety of the tuple color space is coherent.

In the (R, G, B) color space, edge and feature information is divided between all 3 channels equally, whereas in the $L \times \alpha \times \beta$ channel, the edge and feature information is limited to the L channel alone. This takes it easier for information recovery systems such as colorizers for locating vital feature information locally.

Use of feature extractors: In Baldassare et al., they have used a pre-trained feature extractor to validate color information generated by their primary CNN. They achieve this by using a system of dual-channel decoders and encoders. The source grayscale image which initially only contains the Luminance (L) channel is down-sampled to a computation friendly 224×224 dimension. One copy of this down-sampled image is fed to the pre-trained feature extractor Inception-ResNet, which produces a vector containing the coherent information about the image. This information is then ‘fused’ with the output of the encoder in the primary system, which convolutes a down-sampled image through 8 layers.

The fusion causes the primary system to recognize features that are not immediately apparent from the L channel of the image alone. The fused result is then fed to the decoder, which is essentially a reverse convolution network, designed to amplify certain features in images. The results from the decoder are up-sampled to produce the colored version of the grayscale image. Each of the module in the system has a ReLU activation function, which is just a Linear Unit function, except for the last layer, which is hyperbolically tangential.



Fig. 53.2 A comparison between different colorization models

The work by Baldassare et al. generates images with high accuracy and photorealism. However, due to the relatively low volume of the dataset used (about 60,000 out of the 14 million images in ImageNet), it performs better when some features are present in the image. The model is better at colorization of the sea and foliage, whereas it fails to color human clothing and skin tones properly.

Figure 53.2 shows a side-by-side comparison of images that have been colored by [5–8] reference papers, and has been taken from Baldassare et al. Here, we can see that the statistical ground truth difference is the least in the images colored by the Inception ResNet v2, but if the ground truth image was not present, more human subjects would likely pick the images by the AlexNet and the ones by Larsson et al. to be the real images. The AlexNet performs spectacularly when the image contains human subjects and background vegetation. The Inception ResNet colors vegetation considerably better than the AlexNet but lacks satisfactory coloring of human subjects.

53.3 Proposed System

Every convolutional layer is a set of 2 or 3 repeated convolutional and ReLU layers [9] ending in a BatchNorm layer. The neural net does not contain pool layers. Resolution

changes happen due to spatial down sampling or up sampling between convolutional blocks (Fig. 53.3).

Apart from the visual part of colorization, we also look into how colorization can be seen as a first step toward representation learning. This model is similar to an automatic encoder [10], with the input and output being separate image channels, or more concisely a cross channel encoder (Fig. 53.4).

For the evaluation of the feature representation obtained through this cross channel encoding, we require two separate tests on the model. First, we check the task generalization capability of the features by storing the learned representation and training classifiers to linearly classify objects on already seen data. Second comes the finer adjustments to the network based on the PASCAL dataset for the classification as well as segmentation and detection. Here, we not only test held-out tasks, we also test the learned representation on the data generalization. For an accurate comparison with older feature learning models, we retrain an AlexNet net on the task, using this complete process, for 450 k loops.

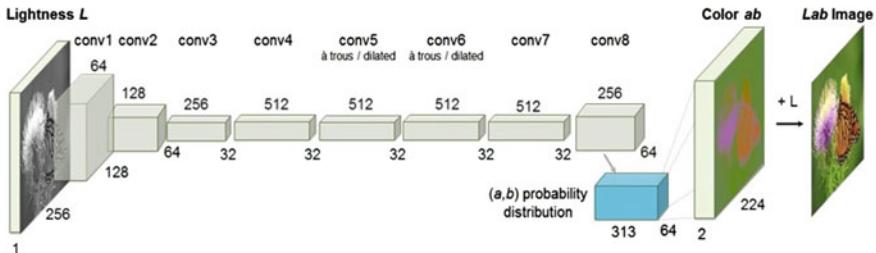


Fig. 53.3 System architecture

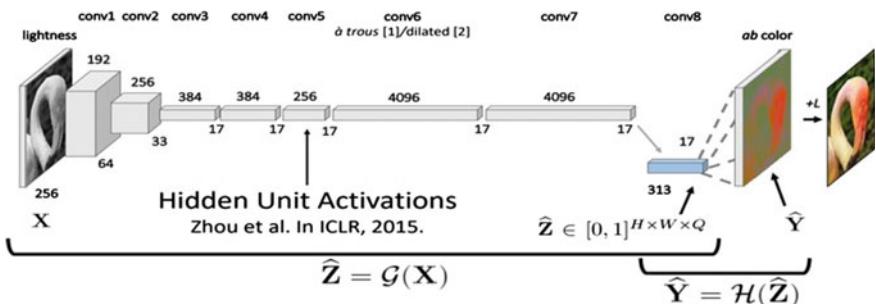


Fig. 53.4 Cross channel encoder

53.4 Result

By reviewing and studying reference papers related to image processing and colorization, we have gathered vital information regarding colorization processes and techniques. We have learned about the strengths and weaknesses of each technique and described 2 of the techniques in detail. Here are some of the observations from the reference papers:

1. The $L \times \alpha \times \beta$ color space is objectively better for training colorization systems due to the localized nature of information in LAB channel images.
2. Using a pre-trained feature extraction model or an object classification model as a secondary source of information from the source image can cut down on training time and resources required considerably.
3. Down sampling of the source images to a standardized small size is essential to reduce computational complexity, even if it reduces the amount of detail that can be produced in the final images.
4. Up sampling or decoding of the encoded channel images leads to unwanted noise and artifacts in the output images that reduce visual appeal.
5. The computational complexity of the overall system will be directly proportional to the number of neurons in each layer of the CNN.

So, we propose several improvements that can be made to increase the accuracy or colorization plausibility of the models described above. In the first model by Zhang et al. [3], the training set was vast, but the impact of each individual image was negligible, thus the overall accuracy of the system could theoretically be increased by weighting the input training images by their relevance or by using a separate feature-weight distributor.

This can lead to more accurate colorizations in abstract objects with no context. Their choice of T-value and learning rate is optimal and the only other improvements in their models can be made by replacing the architecture with a more complex one. However, the addition of dimensions in the model leads to reduction in information distribution within the model and requires careful optimizations and tweaking to perfect.

In the second paper, by Baldassare et al., the Inception ResNet v2 [5, 11] can be replaced by the v4, the latest version which boasts better accuracy and information recovery metrics. The training dataset, which is a small subset of the ImageNet dataset can be expanded to cover a larger number of images while reducing the impact of each image by lowering the learning rate and f_α values.

This will help the model encompass more of the object spectrum and reduce color patching when unknown objects and features are encountered in an image (Fig. 53.5).

Another area of improvement common to all models discussed in this paper is the up-sampling of the encoded images. Traditional up-sampling of small images causes noise and artifacts in images. Instead of using the up-sampling methods of basic image processing, we can use Deep Learning-based Super Sampling (DLSS). DLSS technology has been introduced by Nvidia at the end of 2018 for up-sampling

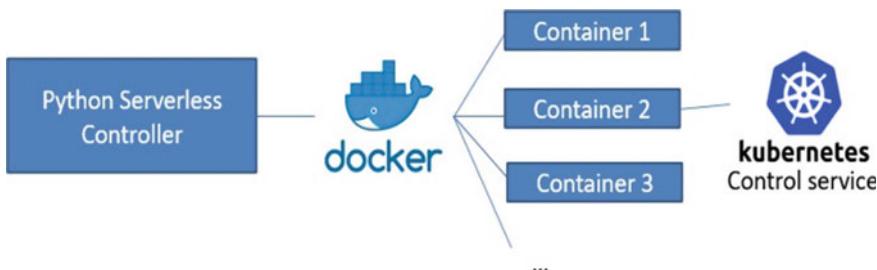


Fig. 53.5 Training architecture

images with deep learning. Currently, DLSS is being used by Nvidia in high-end video games and visual effects for creating the illusion of detail when there is none. DLSS technology can be applied to colorization to improve the quality of the output image considerably. However, this requires one to train an entirely separate system for super sampling images based on feature detection [12].

After super sampling of the encoded image is performed, a static color correction can be applied to account for the loss of color information during the process of DLSS. This color correction can range from a simple transform to a color analysis-based intelligent filter. This color correction will allow output colorized images to be of high color contrast as well as detail.

53.5 Performance Evaluation

We followed the human ‘Turing Test’ methodology for testing our results. We gathered 15 participants (friends, relatives and colleagues) and had them guess out 4 sets of two images each, which was the ground truth image and which was colorized by our model. The results are shown in Table 53.1.

Test Sample Instances

- **Against flora and fauna (Fig. 6a):** Our model showed at par results to other discussed colorization models including Richard Zhang’s model.
- **Against skies and monuments (Fig. 6b):** Our models performed significantly better because of its feature extraction technique and no down-scaling which generated a more vibrant image when compared to the dull ground image.

Table 53.1 Turing test accuracy

Number of correct guesses	23
Number of human fooled	37
‘Turning Test’ accuracy	61.66%

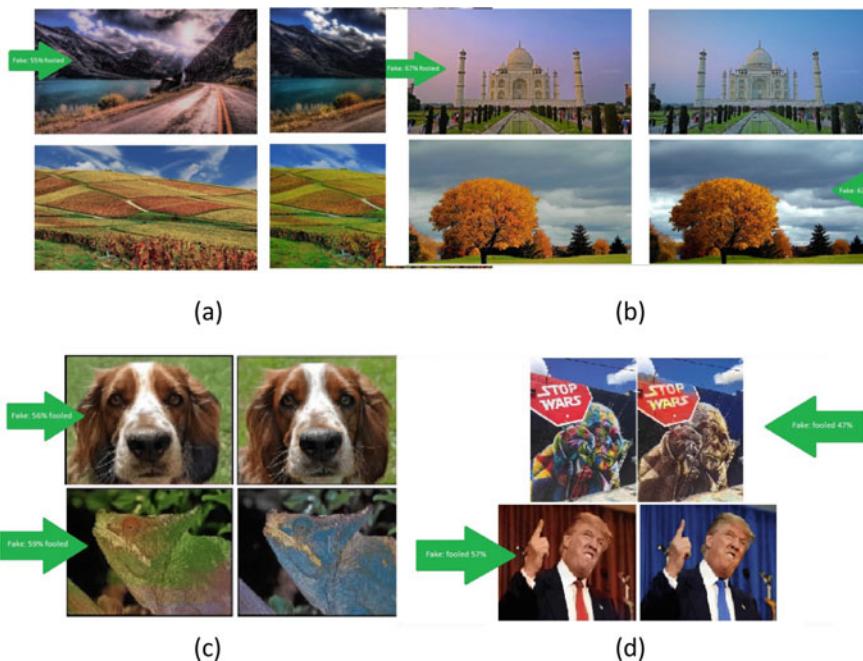


Fig. 53.6 **a** Flora and Fauna, **b** Skies and monuments, **c** Animals and reptiles, **d** Humans and humanoid subjects

- **Against humans and humanoid subjects (Fig. 6d):** Our model fared well in this category and was up to some extent able to separate facial features and skin tones.
- **Against animals and reptiles (Fig. 6c):** Sadly, our model contained bias when tested against animals and reptiles. The model misplaced facial colors (especially mouths). In case of snakes and other reptiles, the results were plausible because of the wide variety of colorful reptiles found in nature.

53.6 Conclusion

Image colorization is a complex problem because it deals with missing information in the input images. By using a myriad of techniques, we can recover this lost information to produce colorization that is plausible to the human eye. Convolutional neural networks are an emerging tool for processing and colorizing images. Advances in the architecture of CNNs used and optimization of the training methodologies will help to improve the accuracy of the colorization produced. Furthermore, promising techniques such as DLSS, hardware-based super sampling and intelligent color correction will help us produce images that are visually similar to the ground truth images.

References

1. Daly, R., Zhao, T.: CNN Assisted Colorization of Gray-scale Images
2. Hwang, J., Zhou, Y.: Image Colorization with Deep Convolutional Neural Networks. Stanford University (2016). Technical Report [Online]. Available: <http://cs231n.stanford.edu/reports2016/219Report.pdf>
3. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European Conference on Computer Vision, pp. 649–666. Springer, Cham (2016)
4. Nie, D., Ma, Q., Ma, L., Xiao, S.: Optimization based grayscale image colorization. Pattern Recogn. Lett. **28**(12), 1445–1451 (2007)
5. Baldassarre, F., Morín, D.G., Rodés-Guirao, L.: Deep Koalarization: Image Colorization Using cnns and inception-resnet-v2 (2017). arXiv preprint [arXiv:1712.03400](https://arxiv.org/abs/1712.03400)
6. Hussein, A.A., Yang, X.: Colorization using edge-preserving smoothing filter. SIViP **8**(8), 1681–1689 (2014)
7. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to greyscale images. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 277–280 (2002, July)
8. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: ACM SIGGRAPH 2004 Papers, pp. 689–694 (2004)
9. Chen, T., Wang, Y., Schillings, V., Meinel, C.: Grayscale image matting and colorization. In: Asian Conference on Computer Vision, pp. 1164–1169 (2004)
10. Kekre, H.B., Thepade, S.D.: Color traits transfer to grayscale images. In: 2008 First International Conference on Emerging Trends in Engineering and Technology, pp. 82–85. IEEE (2008, July)
11. Visshwak, J.J., Saravanakumar, P., Minu, R.I.: On-the-fly traffic sign image labeling. In: 2020 International Conference on Communication and Signal Processing (ICCP) (pp. 0530–0532). IEEE (2020, July)
12. Nagarajan, G., Minu, R.I.: Multimodal fuzzy ontology creation and knowledge information retrieval. In: Proceedings of the International Conference on Soft Computing Systems, pp. 697–706. Springer, New Delhi (2016)

Chapter 54

Music Generation Using Deep Learning



R. I. Minu, G. Nagarajan, Rishabh Bhatia, and Aditya Kunar

Abstract In recent years, recurrent neural network models have defined the normative process in producing efficient and reliable results in various challenging problems such as translation, voice recognition and image captioning, thereby making huge strides in learning useful feature representations. With these latest advancements in deep learning, RNNs have garnered fame in computational creativity tasks such as even that of music generation. In this paper, we investigate the generation of music as a sequence-to-sequence modelling task. We compare the performance of multiple architectures to highlight the advantages of using an added attention mechanism for processing longer sequences and producing more coherent music with a better quality of harmony and melody. The dataset used for training the models consists of a suitable number of MIDI files downloaded from the Lakh MIDI dataset. To feed the MIDI files to the neural networks, we use piano roll as the input representation. The recurrent neural networks learn long-term temporal specific structures and patterns present in musical data. Once we are done with training, we can generate the musical data using a primer of musical notes.

54.1 Introduction

A human brain can easily interpret sequential data due to our ability to utilize the contextual information provided by the understanding of previous inputs. Traditional neural networks are plagued by the loss of comprehension of precious preceding

R. I. Minu (✉)

SRM Institute of Science and Technology, Chengalpattu, India
e-mail: minur@srmist.edu.in

G. Nagarajan

Sathyabama Institute of Science and Technology, Chennai, India

R. Bhatia

Senseforth.ai, Bengaluru, India

A. Kunar

Delft University of Technology, Delft, Netherlands

information [1]. Recurrent neural networks evade this issue by using loopings to allow for the persistence of information and are widely used for natural language processing problems.

Music is after-all the ultimate language. Many well-known artists have created compositions with creativity and unwavering intent. Musicians like Johann Sebastian Bach are renowned for their proliferent talent in generating music that contains a very strong degree of underlying musical structure [2]. Our work is an attempt to develop and compare recurrent neural network models which concisely convey the idea of tunefulness and consonance that can also similarly learn such musical structure. We do so by providing a purposeful policy to embody notes in music by using piano roll matrices as our input and target representations and build interesting networks involving the use of multi-layered LSTMs and encoder-decoders with and without an added attention layer. Therefore, our goal is to evaluate generative models that harnesses the power of a recurrent neural network models to produce musical pieces with tunefulness and consonance that is pleasing to the ear just like as if it is composed by humans.

Our goal is to use the best and latest innovations in deep learning for the generation of music. We are very excited to find new tools for music creation. This will enable anyone with little musical knowledge to be able to create their own complex music easily and quickly.

54.2 Related Work

54.2.1 *Melody-RNN*

Melody-RNN is developed at Google's open source project Magenta. Project Magenta is constantly working on finding new techniques to generate art and music, possibly generating pleasing music without any human intervention needed. The Melody-RNN is built as a two-layered LSTM network [3]. At present, Melody-RNN comes in three variations. First is the normal two-layered LSTM network, which utilizes one hot encoded note sequences extracted from MIDI melodies as input and output targets to the LSTM; second is Lookback-RNN, it initiates customized inputs with labels to let the model comfortably extract patterns which take place across first and second bars; the final one is attention RNN, it contains an added attention layer to inform the model to focus on important aspects of past data without the need to save all of that data in the RNN cell's current state.

54.2.2 *Biaxial-RNN*

Biaxial-RNN is Daniel Johnson's impressive RNN music composition undertaking [4]. Its blueprint outlines properties such as:

- Creation of polyphonic music based on the choice of meaningful collection of notes or chords.
- Orderly repetition of identical notes, i.e. playing A# two times is fundamentally different than keeping A# pressed for two beats.
- Understanding tempo: The ability to create music-based on a predefined time signature
- Note-independent: Being able to transpose up and down musical notes with same framework of network.

An important point to be noted is that plenty of the RNN-based music generation techniques are independent of time but dependent in notes. However, if we go up one octave, we must also similarly wish that the model produces a near similar musical piece rather than something extremely divergent. The Biaxial-RNN model contains dual-axes (time axis and note axis) to help permit historical information to be properly utilised for both the time axis and note axis.

54.2.3 *WaveNet*

The WaveNet model has been heavily influenced by earlier works at Google, precisely a model named Pixel-RNN that was also created by Google's DeepMind division. WaveNet is special because it uses raw audio input as the training data and even in a single second of the sound, there are more than 16,000 samples and therefore making a single prediction conditioned on all the preceding samples is an extremely daunting task. WaveNet [5] circumvents this issue by using a fully convolutional neural network along the time axis, wherein the convolutional layers are dilated such that its receptive field can expand exponentially with increasing depth and cover hundreds of time steps within each stack of computation in the network. The output is therefore more meaningful and is far better than the latest text-to-speech models, diminishing the distance between human-level performances by more than 50%. During the training phase, the input data is actual waveforms. After the training is completed, we can perform sampling from the network's generated probability distribution to create authentic music. Developing up samples like this at each time step is very computationally straining, but it is equally important for producing highly convincing sounding music. The WaveNet model also has to have a large amount of training data so that it can learn to produce good quality musical pieces.

54.3 System Architecture

We have implemented four system architectures as follows (Fig. 54.1).

Recurrent neural networks face a problem while training on long sequences; they are not able to remember long-term dependencies. They face the problem of vanishing gradients because in recurrent neural networks we use back propagation through time (BPTT) for calculating gradients. In BPTT after each time step the gradients become smaller or larger as we propagate backwards through the network because recurrence brings repeated multiplication which causes gradients to become very small or very large, leading to the layers at the starting of the neural networks being updated very less compared to the layers at the end of the network. The solution to this is to use the long short- term memory (LSTM) model [6].

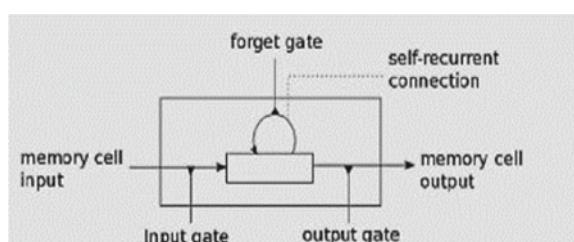
The core concept in LSTMs is to make sure information resides within protected memory cells which are not damaged or affected by the normal flow of data in the recurrent network. The dilemma of writing to, reading from and forgetting, via the operation of the gates, are controlled at the meta-level and is learned along the training process. Thus, the gates have their own weights which are differentiable and enables for back propagation through time and allows for the normative process of learning—hence, every LSTM cell adapts in keeping its weights on the basis of its input data to reduce the overall loss (Fig. 54.2).

We utilized a long short-term memory (LSTM) recurrent neural network (RNN) with two layers and dropout to create a regularized design to predict the next musical connotation for each time step. This architecture lets the user choose the different hyper parameters such as sequence length, sizes of each batch, and rates for learning. This whole work was performed using the TensorFlow library.

54.3.1 LSTM with Attention

A very challenging problem in utilizing machine learning to produce long sequences, such as music, is creating long-term structure. Long-term structure come easily to people, but it is difficult for neural networks. It seems that in absence of long-term structure, the music generated by neural nets are far off and random. Humans translating a rather long sentence usually pay special attention to the word that they

Fig. 54.1 LSTM architecture



```
Using TensorFlow backend.
corpus length: 193
total # of values: 78
nb sequences: 58
```

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 20, 128)	105984
dropout_1 (Dropout)	(None, 20, 128)	0
lstm_2 (LSTM)	(None, 128)	131584
dropout_2 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 78)	10062
activation_1 (Activation)	(None, 78)	0
Total params: 247,630		
Trainable params: 247,630		
Non-trainable params: 0		

Fig. 54.2 LSTM model architecture

are presently translating [7]. LSTM neural networks perform this functionality via an added attention mechanism, stressing on important portions of the sequential data they are provided with (Figs. 54.3 and 54.4).

The attention mask is created using a mechanism based on the content. The attention layer throws an inquiry concerning important aspects to focus on. Every input sequence is multiplied to the resulting inquiry to create a tally that entails the quality of matching the resulting inquiry. These tallies are converted into an attention mask

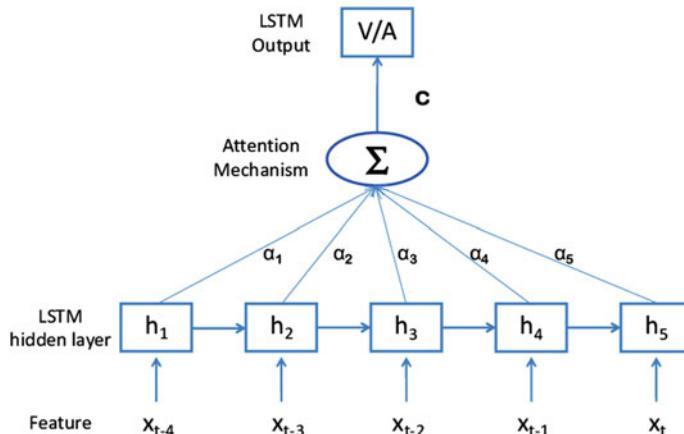


Fig. 54.3 LSTM with attention

Layer (type)	Output Shape	Param #
lstm_29 (LSTM)	(None, None, 64)	29184
batch_normalization_15 (BatchNorm)	(None, None, 64)	256
dropout_15 (Dropout)	(None, None, 64)	0
lstm_30 (LSTM)	(None, 64)	33024
repeat_vector_15 (RepeatVector)	(None, 50, 64)	0
AttentionDecoder (AttentionDecoder)	(None, 50, 49)	61824

Fig. 54.4 LSTM with attention model architecture

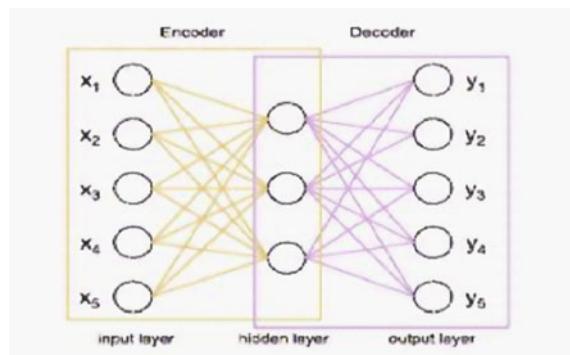
using the SoftMax function to produce the attention distribution. The mask is then fed to the output layers to produce the next note for each time step. This whole work was performed using the TensorFlow library.

54.3.2 Encoder-Decoder

An encoder-decoder architecture [2] is a special type of neural net with a set of hidden layers and an important constraint that we have used is that the output nodes are of same dimension as that of input nodes. The decoder is meant to use the latent representation or embedding of the input layer produced by the encoder. Training an encoder-decoder network is performed using conventional supervised learning. In reality, the power of the encoder-decoder lies in its ability to learn the encapsulated information for all the input elements in a more compressed latent representation. The hidden layer is made such that it has lesser units as compared to the dimensions of the input data, the encoder's job is to compress data and the decoder's job is to produce, as well as possible, the best next step prediction using the compressed data [8]. This enables the encoder-decoder network to uncover important and useful features for encoding the data into latent variables. The latent variables extracted are also called as embedding vectors. Once the net is trained, extracting features from an input is simple. A feed forward pass of the network allows us to obtain all of the activations of the output layer (Figs. 54.5 and 54.6).

In our model, we have utilized an long short-term memory (LSTM) of two layers with dropout and batch-normalization to produce an encoder that converts our data into its latent representation and similarly a decoder which enables us to make predictions of the subsequent note for each time step based on the embedding produced by the encoder with the help of repeat vector layer to pass the embedding vector with the appropriate shape to the decoder [9]. The time distributed layer makes it possible for our model to distribute the predictions calculated over discrete time

Fig. 54.5 Encoder-decoder architecture



Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, None, 64)	29184
batch_normalization_1 (Batch Normalization)	(None, None, 64)	256
dropout_1 (Dropout)	(None, None, 64)	0
lstm_2 (LSTM)	(None, 64)	33024
repeat_vector_3 (RepeatVector (None, 50, 64))		0
lstm_3 (LSTM)	(None, 50, 64)	33024
batch_normalization_2 (Batch Normalization)	(None, 50, 64)	256
dropout_2 (Dropout)	(None, 50, 64)	0
lstm_4 (LSTM)	(None, 50, 64)	33024
batch_normalization_3 (Batch Normalization)	(None, 50, 64)	256
dropout_3 (Dropout)	(None, 50, 64)	0
time_distributed_1 (TimeDistributed (None, 50, 49))		3185
Total params:	132,209	
Trainable params:	131,825	
Non-trainable params:	384	

Fig. 54.6 Encoder-decoder model architecture

steps and produce the output with the specified sequence length. This whole work was performed using the TensorFlow library along with Keras.

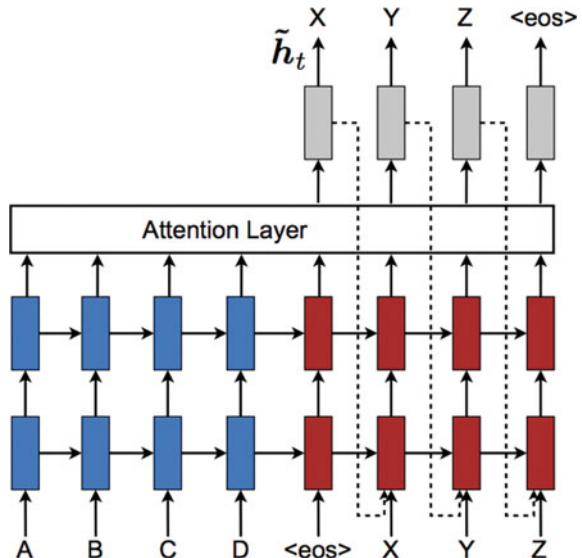
54.4 Encoder-Decoder with Attention

The encoder-decoder design for recurrent neural networks such as LSTMs is turning out to be a very useful approach for a wide variety of sequence-to-sequence prediction tasks in the field of natural language processing [10]. It is however burdened by input data of high sequential length (Fig. 54.7).

Attention technique is intended as a way to overcome this drawback of the encoder-decoder architecture in which an embedding of the input data to a vector of fixed length is decoded [9] at each output time step therefore increasing the ability of the model to learn a lengthy sequence (Fig. 54.8).

As an alternative of creating a single fixed context embedding of the input data, the attention layer creates a vector of context that is fitted especially towards every output prediction. The decoder then outputs one result at a time, which is combined

Fig. 54.7 Encoder-decoder with attention



```

Seq2Seq(
    (encoder): EncoderRNN(
        (gru): GRU(49, 256, batch_first=True)
    )
    (decoder): Decoder(
        (attention): Attention(
            (attn): Linear(in_features=512, out_features=256, bias=True)
        )
        (rnn): GRU(305, 256)
        (out): Linear(in_features=561, out_features=49, bias=True)
        (dropout): Dropout(p=0.5)
    )
)
  
```

Fig. 54.8 Encoder-decoder with attention model architecture

with more layers corresponding to the attention weights until it outputs a (y_{hat}), i.e. the prediction for the output at the present time step. This added attention layer basically calculates the measure of how each encoded sequence suits the decoder's prediction at each time step and helps the model generalise better to longer sequences [10]. Our deep learning implementation was done in pytorch.

54.5 Implementation

In this segment, we shall discuss training the network and how we produce the musical content (Figs. 54.9 and 54.10).

- A. *Dataset*—The Lakh MIDI dataset houses 176,581 MIDI files. The objective of this dataset is to promote large-scale music information retrieval.
- B. *Data preprocessing*—First we downloaded a suitable number of MIDI files from the Lakh MIDI dataset. Using the Mido library defined in python we extracted the tempo (beats per minute) and resolution (ticks per second) of each



Fig. 54.9 Flow diagram of music generation

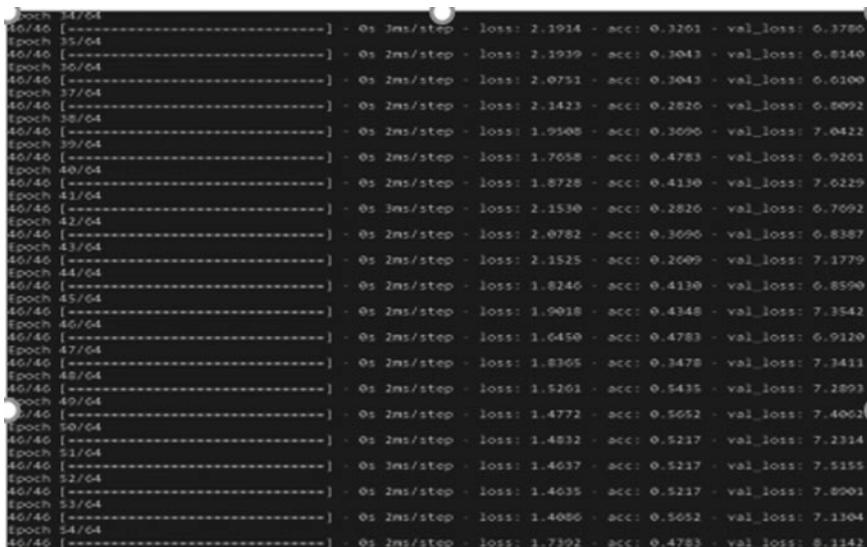


Fig. 54.10 Training of the model in our system

track. With this information we crafter a piano roll matrix of shape—(number of possible notes{49}, sequence of time steps{50}) for each track. This piano roll representation was used throughout all our models. We then created input and target sequences from the piano roll matrices such that assuming a sliding window of 50, the first 50 columns are fed to the model as the input and the next 50 are the targets which the model tries to capture. In such a manner we form the dataset on which the model is trained. A point to be noted about piano roll representation is that it allows for capturing of multiple notes to be played simultaneously and is therefore an ideal form of input representation to use for generating music.

- C. Training—Our model comprises of different architectures involving LSTMs and auto encoders with and without an added attention mechanism as specified in the above section. The model needs to be trained for extracting the true underlying conditional probability distribution of the musical connotations that can be played for a particular step in time, it is essentially governed based on the notes played in preceding time steps. The predictions of the model at each time step t is the probability of playing a musical note at that time step t based on prior note choices. Hence, the network is basically estimating the maximum likelihood of each note in a training sequence within the bounds of the conditional distribution. As we are generating polyphonic music, i.e. multiple notes being on at the same time, this is a multilabel classification problem and therefore we utilize the binary cross entropy loss function. We have executed our training in tensorflow, keras and pytorch.

Once training has completed, and the probability distribution is absorbed from the model and we perform sampling from this distribution which in turn allows us to generate new sequences. The network generates a unique sequence which is projected one step in time into the future. The respective prior inputs for every time step is used to help the LSTM layers compose the note in the next time period. A primer melody is provided as a sample based on which the next notes are picked from the distribution learnt.

- D. Data Post Processing—The output of our model is in piano roll representation. This output is converted once more to MIDI using the python Mido library. The output notes are pruned at each time step to be configured with respect to the tempo, resolution and user preferred sequence length. The MIDI file is then saved in the project folder and ready to be played by a media playback application such as windows media player.

54.6 Performance Evaluation

For measuring the performance evaluation for our models, we are using the binary cross entropy loss on the training and test set to see whether the model is learning sequence prediction correctly [11].

Our problem is of multilabel classification as for each time step there can be multiple notes that could be on at each time step so our target vector is a vector of 0's and 1's with a fixed number of labels C. This task is treated as C different binary and independent classification problems where each output neuron decides whether a class will be present in our output.

Binary cross entropy loss is also called sigmoid cross entropy loss as we apply a sigmoid activation on our output before calculating the cross entropy loss which is calculated as

$$\text{CE} = \sum_{i=1}^{c'=2} t_i \log(f(s_i)) = -t_1 \log(f(s_1)) - (1 - t_1) \log(1 - f(s_1))$$

For each class we are calculating the binary cross entropy as the output of each class is independent of the other classes. With the sigmoid function the output of the network represents a Bernoulli probability distribution for our class C_j .

After training our model we can set a threshold value for converting our output into the final representation containing a vector of 1's and 0's by setting all the values greater than or equal to the threshold as 1 and all the values less than the threshold as 0 (Fig. 54.11).

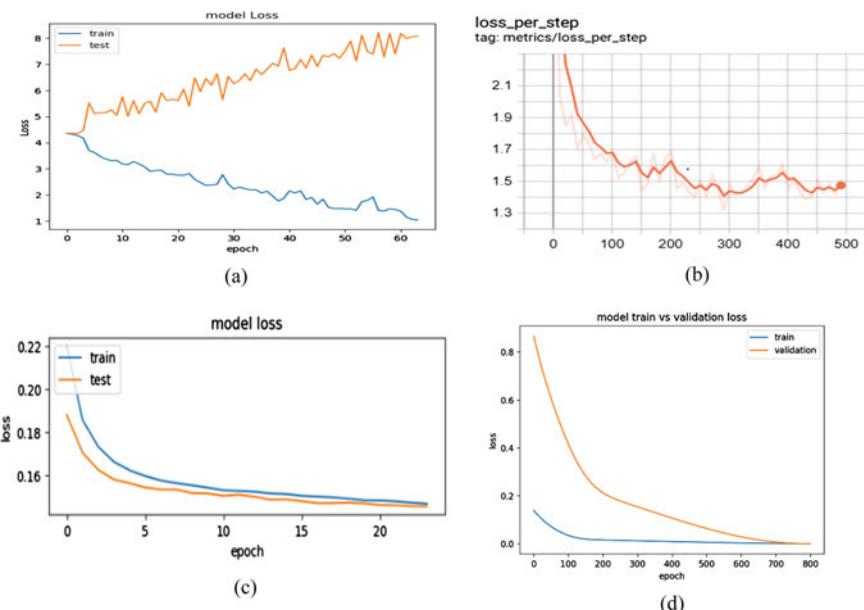


Fig. 54.11 **a** Loss plot for LSTM model. **b** Loss plot for LSTM with attention. **c** Loss plot for encoder-decoder. **d** Loss plot for encoder-decoder with attention

$$P(c_j|x_i) = \frac{1}{1 + \exp(-z_j)}.$$

We can clearly see that the loss is decreasing with more complex models when trained with the same dataset, the LSTM model has the highest validation loss and when we add attention to the model the loss decreases considerably so we can conclude that more complex models like LSTM with attention and encoder-decoder with attention give better performance than simpler models.

We also measure the perplexity of the models which is the weighted geometric average of the inverse of the probabilities, the formula for perplexity is

$$\exp\left(\sum_x p(x) \log_e \frac{1}{p(x)}\right)$$

It gives a measure of how well your model is able to predict the next sequence given the initial sequence and the complexity of the probability distribution learnt. We found that when we used simple LSTM the value of perplexity was high 3.5 as it should be and when we used the more complex model with attention the value of perplexity decreased to 1.2 so our model is good at predicting the next sequence.

We also performed listening studies on the output produced and concluded that the model which was trained for longer durations produced better output in most cases and complex models such as encoder-decoder with attention generally performed better quality than simpler models such as vanilla LSTM.

54.7 Conclusion

In this paper, we have described multiple architectures for generating polyphonic music. Through this project we are able to use deep learning methods to create new music which is trained on existing music. We observed that adding an attention layer in both vanilla LSTMs and encoder-decoder-based architectures lower the loss and improve the accuracy of the model for longer sequences and improve the model's ability at extracting useful features focusing on the important aspects in musical data. The models which we have used is able to generate music with both tunefulness and consonance and can be considered pleasant to the ear as if created by a human being. The models provide a good balance between local and global structures present in the data. The main application that we hope to achieve for this project is to foster the creative process using machine learning and discover new and exciting rhythms and patterns in music which can improve the quality of music generated by humans.

Compliance with Ethical Standards

We have not received any fund for this work of research. Also there is no human or animal study involved in this work. It is a work done by ourselves. None of the

three author received any fund for this project. Ethical approval: This article does not contain any studies with human participants performed by any of the authors.

References

1. Roberts, A., Engel, J., Raffel, C., Hawthorne, C., Eck, D.: (Submitted on 13 Mar 2018 (v1), last revised 30 Jul 2018 (this version, v4)). A Hierarchical Latent Vector Model for Learning Long Term Structure in Music. Cornell University Library [arXiv:1803.05428J](https://arxiv.org/abs/1803.05428)
2. Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., Norouzi, M.: Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. Cornell University Library. Submitted on 5 Apr 2017. <https://arxiv.org/abs/1704.01279>
3. Huang, S., Wu, R.: Deep Learning for Music. Cornell University Library. Submitted on 15 Jun 2016. [arXiv:1606.04930K](https://arxiv.org/abs/1606.04930)
4. Yang, L.-C., Chou, S.-Y., Yang, Y.-H.: MIDINET: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation. Cornell University Library. Submitted on 31 Mar 2017. [arXiv:1703.10847Y](https://arxiv.org/abs/1703.10847)
5. Hadjeres, G., Pachet, F., Nielsen, F.: DeepBach: A Steerable Model for Bach Chorales Generation. Cornell University Library. Submitted on 3 Dec 2016 (v1). [arXiv:1612.01010M](https://arxiv.org/abs/1612.01010)
6. Jaques, N., Gu, S., Turner, R.E.: Douglas Eck Generating Music by Fine-Tuning Recurrent Neural Networks with Reinforcement Learning. research.google.com/en/pubs/archive/45871.pdf
7. Boulanger-Lewandowski, N., Bengio, Y., Vincent, P.: Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription (2012). arXiv preprint [arXiv:1206.6392\(2012\)](https://arxiv.org/abs/1206.6392)
8. Tikhonov, A., Yamshchikov, I.P.: Music Generation with Variational Recurrent Autoencoder Supported by History (2017). arXiv preprint [arXiv:1705.05458](https://arxiv.org/abs/1705.05458)
9. Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: sequence generative adversarial nets with policy gradient. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1 (2017)
10. Nayebi, A., Vitelli, M.: Gruv: Algorithmic music generation using recurrent neural networks. Course CS224D: Deep Learning for Natural Language Processing (Stanford) (2015); Kumar, N.H., Ashwin, P.S., Ananthakrishnan, H.: MellisAI-An AI generated music composer using RNN-LSTMs. Int. J. Mach. Learn. Comput. **10**(2) (2020)
11. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint [arXiv:1409.0473](https://arxiv.org/abs/1409.0473) (2014).

Chapter 55

Detection of Heart Disease Using Data Mining



G. Kalaiarasi, M. Maheswari, M. Selvi, R. Yogitha, and Prathima Devadas

Abstract Data mining is performed on vast databases for separating concealed examples by utilizing the various methodology from factual examination, AI, and database innovation. In addition, clinical data mining is a critical research field because of its significance in the advancement of different applications in prospering human services area. While condensing the death happening around the world, coronary illness has all the earmarks of being the main source. The distinguishing proof of the chance of coronary illness in an individual is a confused errand for clinical professionals since it requires long periods of experience and extraordinary clinical tests to be led. Right now, data mining grouping calculations like decision tree, naïve Bayes, and random forest are tended to and used to build up an expectation framework so as to investigate and foresee the chance of coronary illness. The fundamental goal of this research work is to recognize the best characterization calculation appropriate for giving most extreme exactness when order of ordinary and strange individual is completed. Accordingly, anticipation of the loss of lives at a previous stage is conceivable. Evaluation was performed with the assistance of coronary illness benchmark dataset from UCI AI archive. It is discovered that random forest calculation performs best with 81% accuracy when compared with different methods of calculations for coronary illness expectation.

55.1 Introduction

Heart diseases are otherwise called cardiovascular sicknesses which happen because of unfortunate way of life, smoking, liquor, and high admission of fats which may cause hypertension, diabetics, and strokes. The World Health Organization dissected that many worldwide demise is because of the heart disease. It has been demonstrated that a decent way of life and early identification are some of the choices for the counteraction of coronary illness. One option in contrast to early location should

G. Kalaiarasi (✉) · M. Maheswari · M. Selvi · R. Yogitha · P. Devadas

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

e-mail: kalaiarasi.cse@sathyabama.ac.in

be possible with a PC. The utilization of the PC can give wellbeing administration enhancements. AI is one of the most troublesome innovations of this age. It is a piece of data science wherein the PC frameworks are made to gain from the various detail indexes based on designs created from the datasets. It fundamentally empowers a machine to learn on its own based on some valuable calculations explicitly produced for it. AI has its rule intensely utilized in all the divisions including producing social insurance, research and development, and so forth. It is likewise an ongoing pattern on the planet and is shaping a significant part of software engineering because of its capacity to mechanize things. Some eminent instances of AI usage includes the online chess table game or the online music player which takes a shot at a specific system, i.e., it learns the examples wherein the contribution for instance was furnished when tuned in to a specific tune each day at a particular time; then, after certain days, the music player learns the example, and naturally, it shows us the melody which we might want to be played. This is really a progressively definite idea dependent on AI which is called profound learning.

Cardiovascular breakdown is similarly an aftereffect of coronary disease, and difficulty in breathing can happen when the heart ends up being too weak to even think about blood circulation. Some heart problems occur without any side effects by any means, particularly in grown-ups and people with diabetes. The term ‘inborn coronary illness’ covers a scope of conditions; however, the general side effects incorporate perspiring, elevated levels of weariness, quick heartbeat, breathing, shortness of breath, and chest torment. These manifestations probably won’t create any problem until an individual is of 13 years. In these sorts of cases, the analysis turns into a complex undertaking, requiring incredible experience and high aptitude. A danger of a coronary failure or the chance of the coronary illness if recognized and treated early can enable the patients to be safe and less risky with proper managing of administrative measures. As of now, the medicinal services industry has been creating immense measures of data about individuals/patients, and the corresponding determination reports are particularly observed for the forecast of coronary failures around the world. At the point when the data about coronary illness is tremendous, the AI methods can be executed for the investigation.

55.2 Related Works

There are many steps for loading of data to get a better performance for managing data analysis of slowing down, loading likelihood data issues, and knowledge about data loading.

WHO conveys that death rate is high throughout the world mainly because of the cardio vascular disease, as it always pairs up with few other critical health issues like hypertension, diabetics, and so on [1]. Thus, the risk increases and almost high percent of death is due to the various categories of CVD’s. Habitual activities like unhealthy food and no physical activity that leads to obesity adds on to the risk on CVD’s [2]. Updated statistics report on the causes and healthy habitual changes that

can be taken to prevent CVDs has been explained deeply in different chapters in this report by the American Heart Association (AHA) [3]. A decision support system is created using the genetic algorithm hybrid with back propagation technique to predict the heart diseases. With the use of the hospital information system and the prediction technique, the goal is achieved in evaluating and predicting. With this system, large data can be analyzed quickly and makes the job easier for healthcare professionals in diagnosis of CVD's.

Various data mining techniques are used on the medical information to predict the heart diseases [4]. Here, comparison between the k-means clustering algorithm and map reduce algorithm is made to find the efficient algorithm that would extract the needed data from the medical information system. The greater part of the examination works has been actualized with a few data disclosure strategies for foreseeing coronary ailment analysis alongside the fluffy rationale, neuro fluffy, profound learning calculations, and counterfeit neural system. In [5], backpropagation calculation is utilized for learning and testing neural system. The neural system loads were weighted with the assistance of streamlining method called hereditary calculation. So, this multi-layered system had twelve data hubs, two yield hubs, and ten concealed hubs. The coronary illness hazard factor depends on the quantity of data layer. Weight and bias is refreshed and recorded with the assistance of system preparing function. MATLAB R2012a, Global Optimization Toolbox, and the Neural Network Toolbox were utilized for implementation. 50 patient's hazard segments were made, and the exactness results for preparing set was gotten with 96.2 and precision results for testing set with 89%.

In [6], utilized data extraction systems, for example, ID3, naïve Bayes, REPTree Simple Cart, and Bayes Net were used for deciding the events of 'coronary failures.' The data index was gathered from emergency clinic and specialists who were experts in South Africa. Eleven characteristics were considered from the dataal collection. That showed restraint Id, cholesterol, tobacco utilization, gender, cardiogram, blood pressure, rating of heartbeat, fasting sugar, age, chest torment and Alcohol consumption. The presentation execution had been finished with the apparatus called WEKA in the investigation of coronary illness events. WEKA device was utilized in deciding, examining, and recognizing the examples. The precision results acquired were J48 with 99.0741, REPTREE with 99.222, naïve Bayes with 98.148, Bayes Net with 98.57, and straightforward CART calculation with 99.0741. Among these, Bayes Net calculation created best outcomes when contrast and the naïve Bayes calculation [7, 8].

In [9] applied different order calculations for sickness forecast model in diagnosing the HD. Two kinds of models were utilized and analyzed, i.e., essential model is single model, and auxiliary model is the joined model which is called as half and half model; the two models are utilized to prepare the data. Data examination was finished utilizing these two models. This survey clearly provides better understanding on the recent techniques that can be used in predicting the risk related to heart disease. In [10], enhanced deep neural network learning was developed to help the healthcare professionals in easy diagnosis and prognosis of the heart diseases. The developed model predicted around 303 new patient cases that were predicted to

have the coronary heart disease at the Cleveland Clinic Foundation. In [11] built up a compelling keen clinical choice emotionally supportive network dependent on data mining systems. The creators considered to underscore on finding the proper classifier that can possibly provide proper precision by using data mining calculations like Guileless Bayes, support vector machine, and logistic regression and the exactness with 75, 80, and 79% individually [12, 13].

Uyar and Ilhan [14] proposed genetic algorithm (GA) based trained recurrent fuzzy neural networks (RFNN) to recognize heart diseases. The creators utilized absolutely 297 cases of patient data. Among these, 252 were taken for preparing and 45 for testing. The creators contrasted this RFNN approach and ANN-fuzzy-AHP approach. By breaking down the testing set, 97.78% precision was achieved as the result in the above-said calculation. Uyar suggested that meta-heuristic methodology with prepared fluffy neural systems approach is utilized for preparing the data index. The dataset for heart disease was taken from the UCI machine learning repository [15], which comprises of 4 databases from sources Cleveland, Hungary, Switzerland, and the VA Long beach. Certain attributes are taken for the learning and prediction process that is closely related to the diagnosis of heart diseases [16]. Guidelines to understand the map reduce function with the tutorial from the link [17]. ANN integrated with the fuzzy analytic hierarchy process provided a higher performance in prediction of the diagnosis of heart diseases. The usual 13 attributes were taken into consideration. Fuzzy_AHP technique was utilized to process the global weights for the attributes dependent on their individual contribution. Then, the global weights that represent the contributions were used to train an ANN classifier for the prediction of HF risks in patients. The results from the proposed method resulted in a 4.4% higher prediction performance.

The forecast technique for coronary illness with the help of neural network technique has been proposed in [18]. It has for the most part three layers, for example, data layer, concealed layer, and yield layer. The data is given to the info layer, and the outcome is acquired in the yield layer. At that point, the genuine yield and the normal yield are analyzed. The back engendering has been applied to discover the blunder and to change the weight between the yield and the past concealed layers. Once the back spread is finished, at that point, the forward procedure is begun and proceeds until the mistake is limited. KNN is a non-parametric technique which is utilized for order and relapse. Contrasted with other AI calculation, KNN is the most straightforward calculation. This calculation comprises K-wardrobe preparing models in the element space. This is based on client characterized steady. The test data are grouped by allocating a steady worth which is generally constant among the K-preparing tests closest to the point. Writing shows the KNN has the solid consistency result. Choice tree fabricates arrangement models as a tree structure. It splits the dataset into smaller subsets, gradually evolving a related choice tree. The choice tree utilizes a top-down methodology technique. The foundation of the choice tree is the data index, and the leaf is the subset of the data index [19, 20].

55.3 Proposed System Methodology

The motivation of this proposed system is to develop a coronary disease desire model for the estimate of occasion of coronary ailment. Also, this system recognizes the chance of coronary illness in a patient. This work is based on utilizing naïve Bayes, decision tree, and random forest. Subsequently, these three calculations are assessed at various levels and kinds of assessment systems. This will give specialists and clinical experts to find the prevalent one. This will be the best technique for predicting the heart diseases.

The coronary illness forecast can be performed based on the strategy presented in Fig. 55.1 which indicates the exploration technique for developing a grouping model for the expectation of the heart sicknesses in patients. The model structures a principal method for doing the coronary illness expectation utilizing any AI systems. So as to make expectations, a classifier should be developed based on the records and later produce a characterization model which is taken care of with another obscure record, and the forecast is made. The examination technique of this exploration incorporates the characterization calculations, for example, assessment utilizing cross approval and assessment utilizing rate split. In the cross approval, the preparation and testing data is separated from the coronary illness utilizing a few overlays, for example, ten folds and where every crease is recursively utilized for preparing and testing by trade in the dataset for testing and preparing. In the rate split, the preparation and testing data is separated in level of data, for example, 80 and 20%, where the 80% is taken for preparing and 20% for testing. Right now, preparing stage incorporates preparing the three arrangement calculations to be specific naïve Bayes, decision tree, and random forest utilizing the coronary illness dataset, and a characterization model is manufactured. All the three calculations are depicted in the areas given beneath.

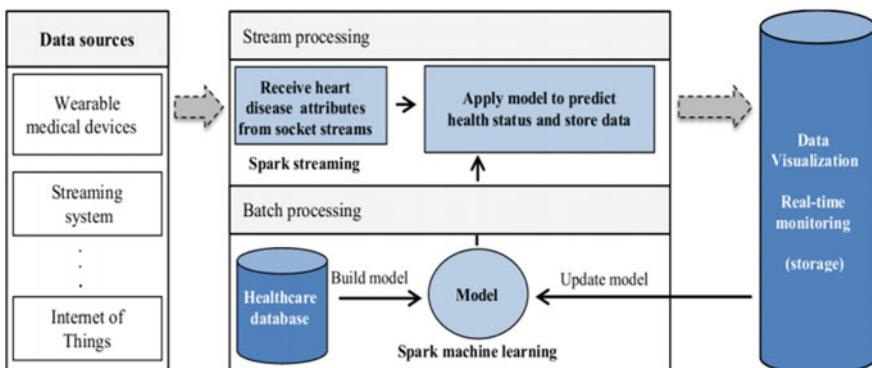


Fig. 55.1 Overview of the proposed system

55.4 Algorithm Used

55.4.1 Decision Tree

It is one of the most impressive and well-known calculations. Choice tree algorithm falls under the class of regulated learning calculations. It works for each constant what's more as absolute yield factors. Choice tree manufactures arrangement or changes the models with in the kind of tree structure. It divides the data collection into many small subsets, while at the steady time, a related choice tree is steadily formed. A call hub comprises of two branches and a leaf hub. The highest choice hub in a tree that refers the best indicator is referred to as root hub. Choice trees handle clear cut just as numerical data. The standards are found out individually utilizing the preparation data. Whenever a standard is found, the secured rules are evacuated by tuples.

55.4.2 K-Nearest Neighbor (KNN)

K-nearest neighbor is AI calculation that stores all cases related to trained data focusing in multi-dimensional space. At the point when an obscure scatter data or any kind of data is received, it dissects the nearest k number of occasions spared (closest neighbors) and returns the most widely recognized class as the expectation and for real worth data. Thus, restore the mean of k closest neighbors. To calculate the weighted closest neighbor, it loads every k neighbors based on their separation and provide necessary load to the nearest neighbors.

55.4.3 Support Vector Machines

A classifier sorts the data collection by setting an ideal hyper plane between data. This classifier is picked as it is unbelievably adaptable in the quantity of various kernelling capacities that can be applied, and this model can yield a high consistency rate. Bolster vector machines [10] are well known and is based on AI calculations. They were exceptionally regular from the time they were formed inside the nineties and still be the go-to method for a high performing algorithmic principle.

55.4.4 Naïve Bayes

Naïve Bayes is a measurable classifier that will not expect any oppression among qualities. Bayes classifier has basically two steps:

- Training Step: Given a class of data, the technique assesses the parameters of a likelihood dissemination known as the earlier likelihood from the preparation data.
- Prediction Step: For obscure test data, the strategy figures the back likelihood of the dataset which is having a place with each class. The strategy at long last groups the test data dependent on the biggest back likelihood from the set.

55.5 Results and Discussion

This section presents the results.

Figure 55.2 represents the database and the attributes considered in this paper.

In Fig. 55.3, it is seen that the table has the count, mean, std, min, 25%, 50%, 75%, and max. That means in the table, count of the each attribute, standard deviation, and min values are mentioned.

In Fig. 55.4, it can be checked whether the attributes are correlated or not in the matrix form. If the correlation occurs, we will see the correlated attributes in the matrix. The attributes that are not correlated the value of the correlation is 0. So because of that, we are using correlation matrix.

Figure 55.5 represents the histogram structure for the attributes taken. In this, the min and max values can be seen in the graph for the attributes.

By Fig. 55.6, it is clearly seen that if the variable has the outlier, it will reduce the performance. If it does not have the outlier, the performance is going to be good. So to find out that the performance, it is the best to find the presence of outliers.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 55.2 Dataset taken

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	54.366337	0.683168	0.966997	131.623762	246.264265	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Fig. 55.3 Describing dataset

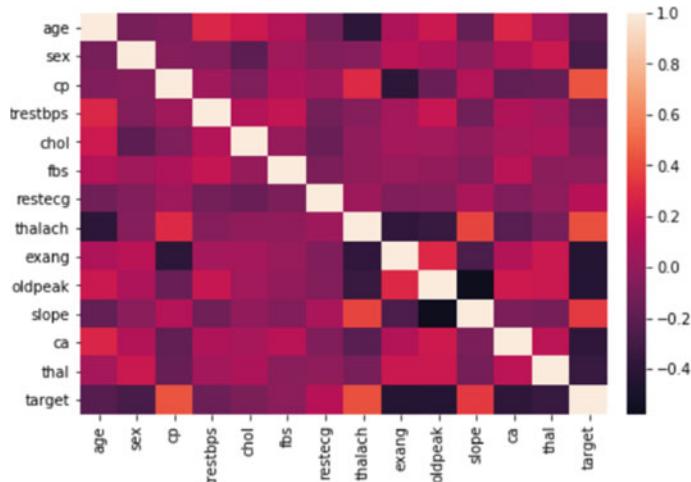


Fig. 55.4 Correlation matrix

55.6 Conclusion

The rate of heart beat surpasses the limit and crosses beyond its limit. Coronary illness is convoluted, and every second an individual gets affected and suffers to survive. By considering different data cleaning and mining methods, this is used to fabricate a dataset suitable for data mining. The general target of the work is to anticipate all the more precise event of coronary illness utilizing data mining systems. Right now, the UCI data archive is utilized for playing out the similar investigation of three calculations, for example, random forest, decision trees, and naïve Bayes. It has been understood that random forest gives ideal outcomes as contrast with decision tree and naïve Bayes. This work can be extended to create an effect in the precision of the decision tree and Bayesian classification in future for a better outcome in the process of applying hereditary calculation so as to diminish the genuine data for procuring the ideal subset of property that is sufficient for coronary illness forecast. The computerization of coronary illness forecast utilizing genuine constant data from human services associations and offices can be manufactured by utilizing large data. Based on the collected and maintained data, examination of the patients can be monitored continuously.

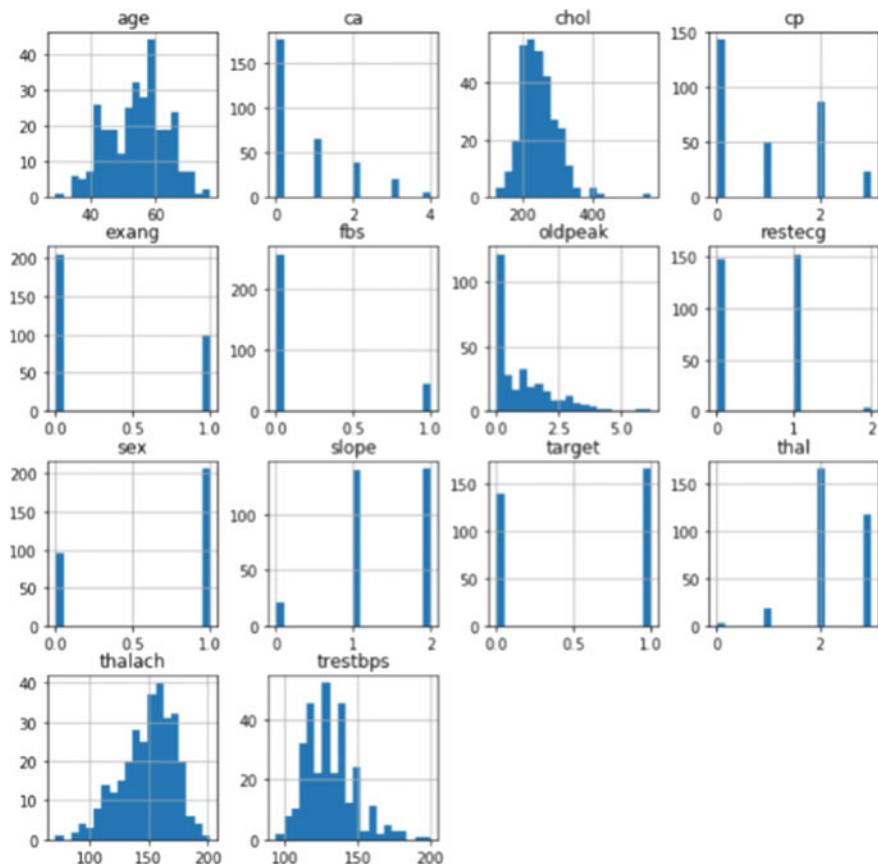


Fig. 55.5 Histogram representation

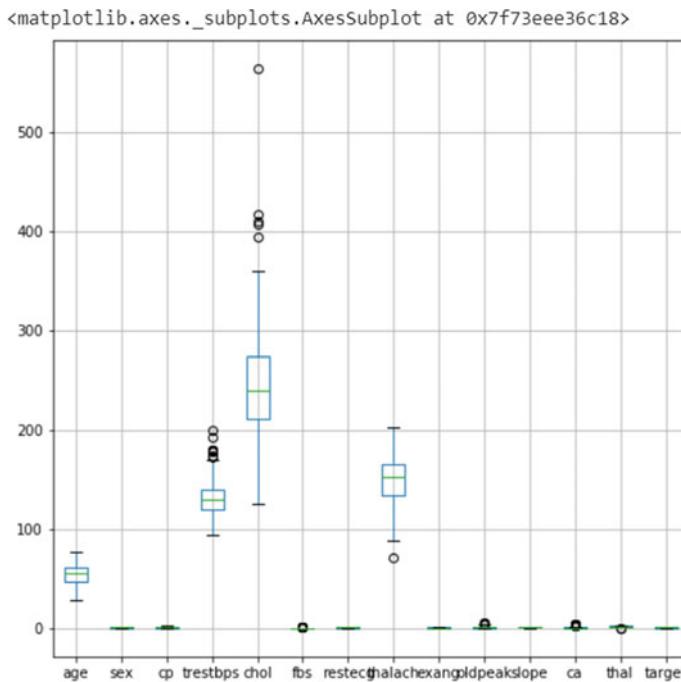


Fig. 55.6 Box plot of the data

References

- WHO Cardiovascular Diseases: http://www.who.int/cardiovascular_diseases
- Benjamin, E.J., Blaha, M.J., Chiuve, S.E., Cushman, M., Das, S.R., Deo, R., de Ferranti Muntner, P.: American Heart Association statistics committee and stroke statistics subcommittee. Heart disease and stroke statistics. Update: a report from the American Heart Association. *Circulation* **135**(10), 146–603 (2017)
- Dewan, A., Sharma, M.: Prediction of heart disease using a hybrid technique in data mining classification. In: 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 11 Mar 2015, pp. 704–706. IEEE
- Bagavathy, S., Gomathy, V., Rani, S.S., Sujatha, K., Bhuvana, M.K.: Early heart disease detection using data mining techniques with hadoop map reduce. *Int. J. Pure Appl. Math.* **119**(12), 1915–1920 (2018)
- Amin, S.U., Agarwal, K., Beg, R.: Genetic neural network based data mining in prediction of heart disease using risk factors. In: IEEE Conference on Information & Communication Technologies, 11 Apr 2013, pp. 1227–1231
- Shilaskar, S., Ghatol, A.: Feature selection for medical diagnosis: evaluation for cardiovascular diseases. *Expert Syst. Appl.* **40**(10), 4146–4153 (2013)
- Nagarajan, G., Thyagarajan, K.K.: A machine learning technique for semantic search engine. *Procedia Eng.* **38**, 2164–2171 (2012)
- Nagarajan, G., Thyagarajan, K.K.: Rule-based semantic content extraction in image using fuzzy ontology. *Int. Rev. Comput. Softw.* **9**(2), 266–277 (2014)
- Purusothaman, G., Krishnakumari, P.: A survey of data mining techniques on risk prediction: heart disease. *Indian J. Sci. Technol.* **8**(12), 1 (2015)

10. Miao, K.H., Miao, J.H., Miao, G.J.: Diagnosing coronary heart disease using ensemble machine learning. *Int. J. Adv. Comput. Sci. Appl.* **7**(10), 30–39 (2016)
11. Dbritto, R., Srinivasaraghavan, A., Joseph, V.: Comparative analysis of accuracy on heart disease prediction using classification methods. *Int. J. Appl. Inf. Syst.* **11**(2), 22–25 (2016)
12. Nagarajan, G., Minu, R.I.: Fuzzy ontology based multi-modal semantic information retrieval. *Procedia Comput. Sci.* **48**, 101–106 (2015)
13. Nirmalraj, S., Nagarajan, G.: Biomedical image compression using fuzzy transform and deterministic binary compressive sensing matrix. *J. Ambient Intell. Humaniz. Comput.* 1–9 (2020)
14. Uyar, K., İlhan, A.: Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia Comput. Sci.* **1**(120), 588–593 (2017)
15. UCI Machine Learning Repository: <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>
16. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
17. Samuel, O.W., Asogbon, G.M., Sangaiah, A.K., Fang, P., Li, G.: An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert Syst. Appl.* **1**(68), 163–172 (2017)
18. Dangare, C., Apte, S.: A data mining approach for prediction of heart disease using neural networks. *Int. J. Comput. Eng. Technol. (IJCET)* **3**(3) (2012)
19. Nirmalraj, S., Nagarajan, G.: An adaptive fusion of infrared and visible image based on learning of sparse fuzzy cognitive maps on compressive sensing. *J. Ambient Intell. Humaniz. Comput.* 1–11 (2019)
20. Indra, M.R., Govindan, N., Satya, R.K.D.N., Thanasingh, S.J.S.D.: Fuzzy rule based ontology reasoning. *J. Ambient Intell. Humaniz. Comput.* **12**(6), 6029–6035 (2021)

Chapter 56

Smart Agriculture Framework Implemented Using the Internet of Things and Deep Learning



R. Aishwarya, R. Yogitha, L. Lakshmanan, M. Maheshwari, L. Suji Helen, and G. Nagarajan

Abstract In the recent world, the Internet of things (IoT) is a rising trend among a variety of real-world applications which tends to collect real-time data from a variety of sensors which are connected together with an Internet of things (IoT) chip. Our proposed model aims for the implementation of a smart agriculture framework which is implemented using an Internet of things architecture built on a system that comprises of various sensors and layers which can collect data based on various parameters. Our proposed system is expected to revolutionize the agriculture heralding in the era of smart agriculture where the farmers can rely on smart sensors and intelligent systems to perform accurate predictions based on the data collected. Smart agriculture is efficient on resource usage, scalability, and flexibility it offers along with the automation that it offers by implementing various layers in the architecture.

56.1 Introduction

Internet of things is envisioned to be an interconnected network of various devices connected to an Internet connection which allows the devices to share real-time data with each other and perform data operations which allows the user to draw inferences from the data gathered by the sensors. According to a report by IDC, IoT devices are expected to generate almost 79.4 ZB which speaks volumes about the potential. Unlike other fields, agriculture has not seen any evolutionary change in regards to acceptance of technology ever since the last century which can potentially increase

R. Aishwarya (✉) · R. Yogitha · L. Lakshmanan · M. Maheshwari · L. Suji Helen · G. Nagarajan
Sathyabama Institute of Science and Technology, Chennai, India

L. Lakshmanan
e-mail: lakshmanan.cse@sathyabama.ac.in

L. Suji Helen
e-mail: sujihelen.cse@sathyabama.ac.in

G. Nagarajan
e-mail: gnagarajan.cse@sathyabama.ac.in

the yield and make the whole agriculture machinery and processes much more efficient [1]. In our proposed model, a software platform and a data platform have been proposed which can be used to collect data from the ground while also allowing the farmers to manage their crop production. The data platform will be connected to a sensor node which is capable of transferring data to the cloud framework which can be visualized using a software platform. This will allow the farmers to judiciously utilize the resources available with the artificial neural network to recommend the best crops for the given season utilizing the data collected from various nodes and help the farmer manage the crops.

56.2 Behind the Scenes

Behind the scenes, our proposed framework will be implemented using the concepts of the Internet of things and artificial neural networks. The total weighted input is transformed into the output using the activation function. The neural network is built to function exactly like a human brain by acquiring knowledge through learning with the knowledge of the network stored in the interconnection strengths known as synaptic weights [2]. The origins of Internet of things was laid down when the idea of networking things using the Internet was conceptualized [3] which allowed multiple devices and gadgets connected to the Internet to share data in real-time and allow algorithms to generate vital statistics. Along with the Internet of things, we are implementing an artificial neural network which is inspired by how the human brain works [4–6].

56.3 Architecture Setup for the Proposed Framework

The primary objective of developing the proposed framework is to implement a cost-effective automated model which can present better statistics and vital data to the farmer by utilizing data from the soil pH value, the predicted climate of the area, the moisture level, and the soil salinity. In Fig. 56.1, we have explained the block diagram of our system architecture which we will implement to collect data and process the data in our network [7, 8].

56.3.1 Physical Layer

The physical layer is the most important layer in our proposed framework and allows for the collection of better datasets which would be then transported to the network layer, after which it will be processed by an ANN to generate the final outputs to be shown to the user.

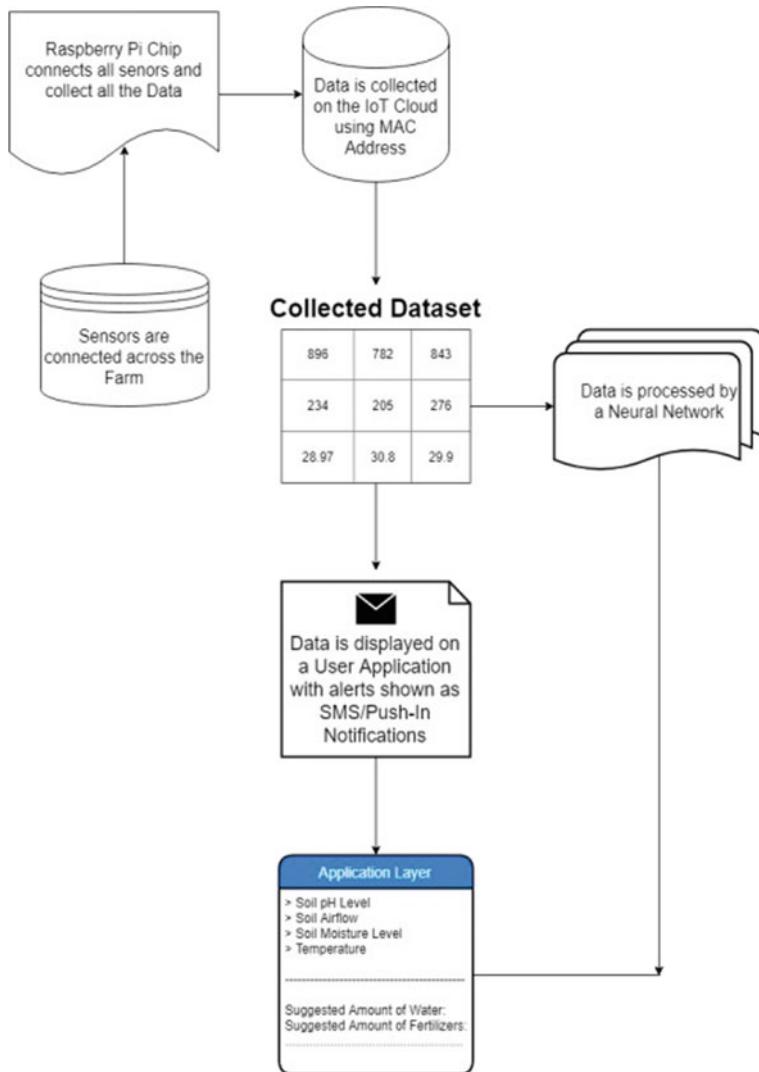


Fig. 56.1 A block diagram of the system architecture

The various sensors utilized in the physical layer are detailed below in brief:

1. Soil Moisture Sensor: Soil moisture sensor is utilized to estimate the volumetric water content in the soil.
2. Air Flow Sensor: The air flow sensor is utilized to measure the air permeability in either a fixed position or mobile mode.
3. Electrochemical Sensor: The electrochemical sensors are tasked to gather the chemical composition and data from the soil.

4. Optical Sensors: The optical sensors can be utilized using remote sensing satellites who are able to generate data about the moisture content of the soil, the humus content, and the contour form in a much better way to infer [9].

Along with this, we will implement a Raspberry Pi 3 microcontroller chip to connect all the sensors, and a Wi-Fi module (further discussed in the next section) and a power source will be utilized to generate the dataset in real-time [10].

56.3.2 Networking Layer

The networking layer consists of an ESP8266 Wi-Fi module which allows for a constant Internet connection to the Raspberry Pi Kit and ensures that the dataset is being gathered in real-time. Figure 56.1 shows the block diagram of smart agriculture framework. In this, using existing dataset, we predict the data is displayed on a user application with alerts in SMS or through PUSH message. Basic network layer table is given below.

56.3.3 Application Layer

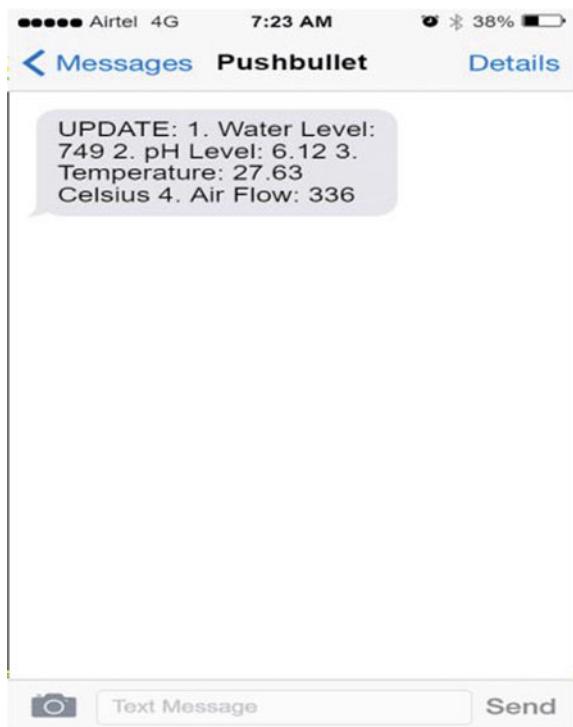
The application layer consists of two parts: the data layer and the processing layer which we will discuss separately. In the data layer, the data will be stored onto an IoT cloud. We will implement a standard and quite popular back propagation model where a weight adjustment property would be utilized to ensure minimum error function with our algorithms. With a two-and-fro forward and backward propagation, the weight and bias associated with the neural network would be adjusted [11].

The processed data would be displayed using a mobile/desktop application where the user can see all the processed data and valuable insights into how much resources he should be utilizing. In Fig. 56.2, we have showcased an example of how SMS can be sent to the user regarding the inputs from his farm directly onto his smartphone.

56.4 Proposed Methodology

Our proposed methodology includes the implementation of an artificial neural network which we have found beneficial in applications like agriculture where the input dataset is large, and through techniques like backpropagation, intelligent insights can be gained on the dataset. In our proposed framework, we will utilize two hidden layers with four inputs in the form of soil moisture (S), soil air flow (A), chemical composition of the soil (C), and finally, the humus content and temperature in the region (O). The input dataset is normalized by a normalization factor which

Fig. 56.2 An SMS from the user application displaying the real-time data collected from the sensors



can be set depending on the situation. To help achieve the backpropagation global minima and it will reduce the error, momentum (μ). Its helps to achieve the required validation for the neural network during the training phase.

The minimum error threshold has been kept at 0.01. If the threshold is achieved by the given value will be used to validate the neural network and if it is not validated or the global minima are not achieved, then the process is iterated once again and vice versa unless the desired output is not achieved. In Fig. 56.3, we explain the proposed flowchart of the system which depicts the control flow in the network and how the system will respond to the given dataset and process it to give the best possible outcome associated with the data. In Fig. 56.4, we explain our concept of a neural network with two hidden layers which will implement our data processing and final output generation [8]. The application would also consist of an emergency report system which would allow the users to quickly act upon issues like excessive adding of fertilizers or water logging the crops which prove to be detrimental, not only for the crop but also for the farming area as well as it loses its economic value in the long run due to being deprived of nutrients and excessive salivation [12].

Our proposed framework “Smart Agriculture implemented using the Internet of Things and Deep Learning” is designed to tackle such issues by allowing the farmers to have an intelligent insight on the farming activity by the aid of a neural network and IoT architecture which can gather real-time data from various parameters [11].

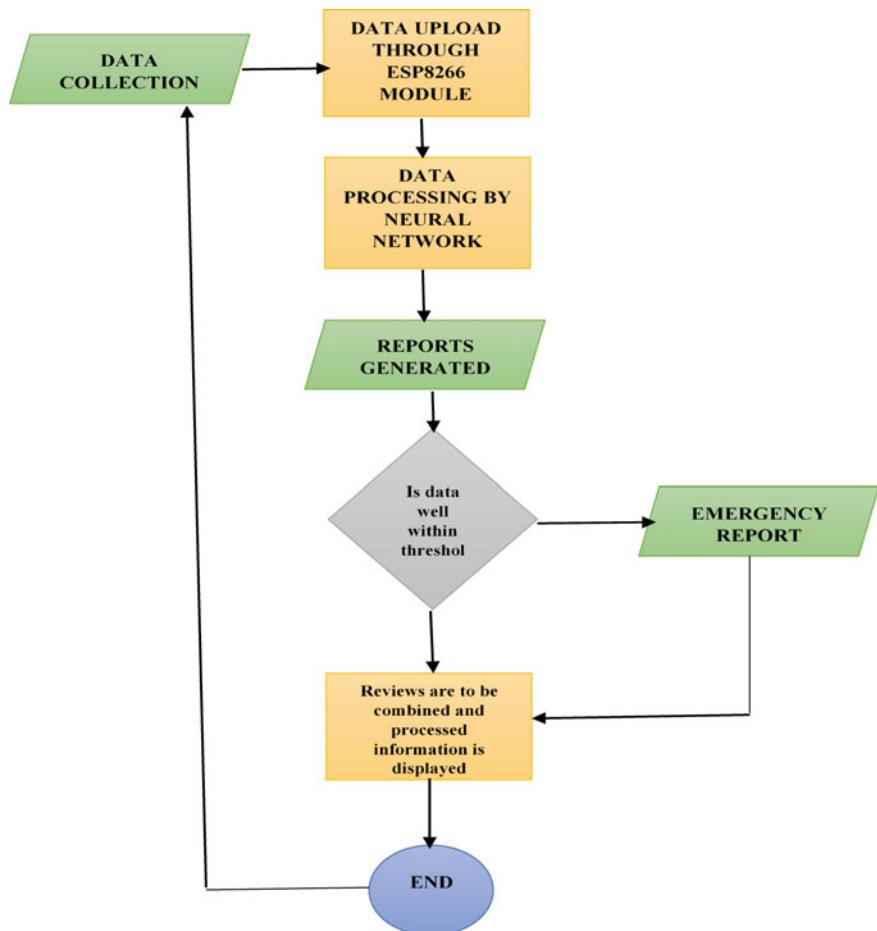
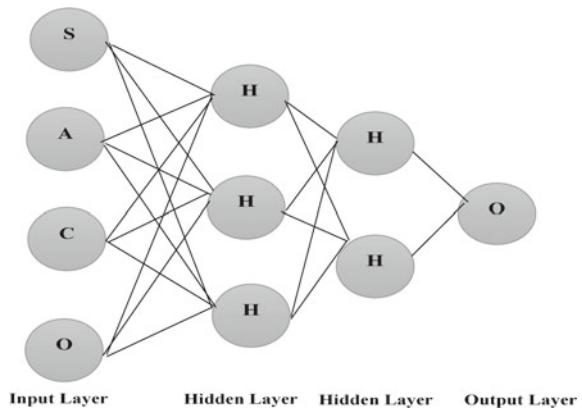


Fig. 56.3 Proposed flowchart for the system

Fig. 56.4 Layers of neural network



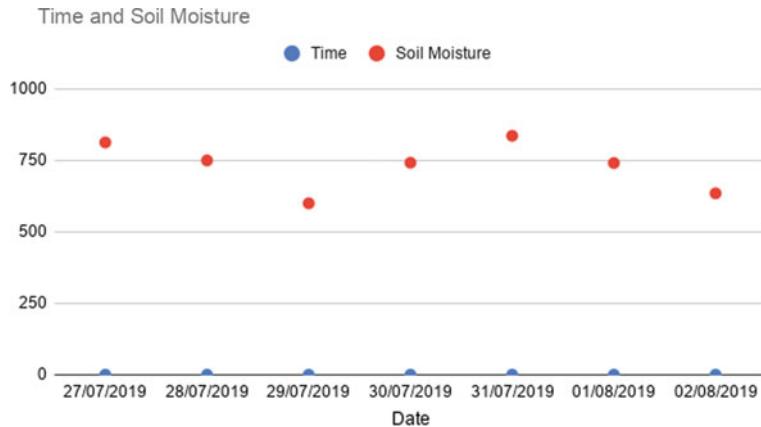


Fig. 56.5 Soil moisture that has been collected across a week across the different time frame

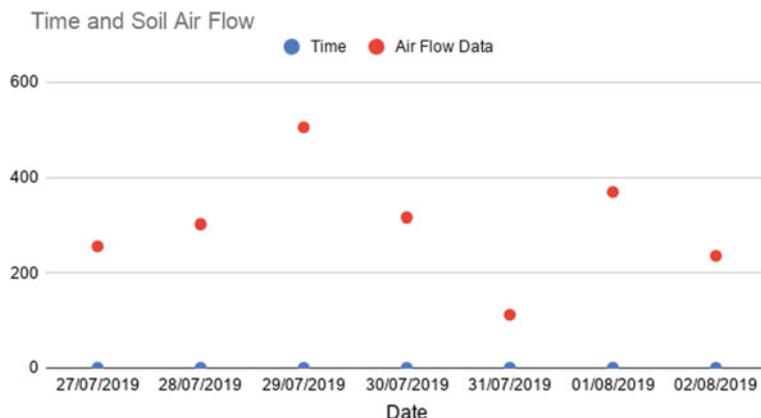


Fig. 56.6 Soil air flow that has been collected across a week across the different time frame

The results have been finally depicted in Figs. 56.5, 56.6, 56.7, and 56.8 where our IoT sensors were implemented to collect the data from the farm-field in real-time and are plotted accordingly to present the data in the best way possible.

56.5 Merits in Proposed Framework

In the proposed model, the main and useful merits such as automation, secure, real-time monitoring, versatile, and cheap. In terms of automation, the IoT architecture requires negligible human interference during functioning. In terms of secure, the login ID and password secures the application from foreign intrusion [5, 6]. In

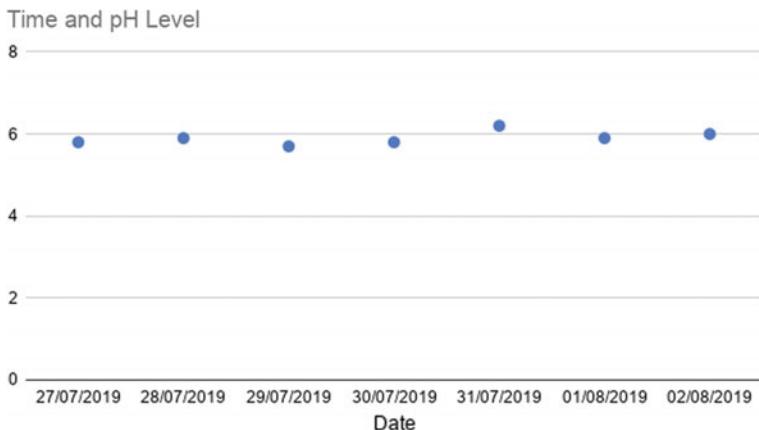


Fig. 56.7 Soil pH level that has been collected across a week across the different time frame

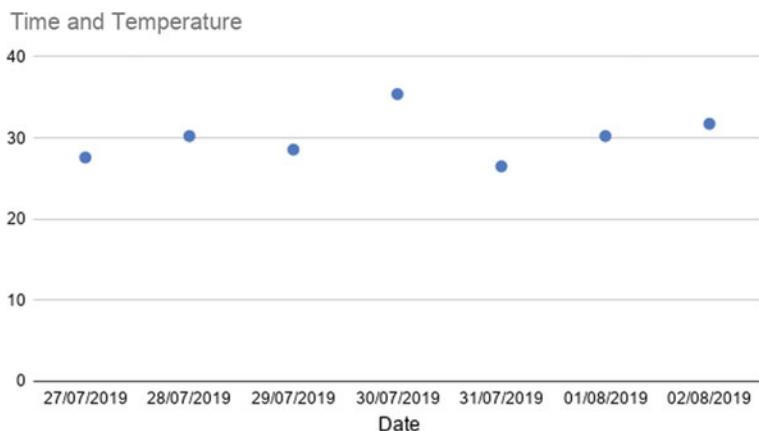


Fig. 56.8 Temperature level that has been collected across a week across the different time frame

terms of real-time monitoring, the architecture utilizes various sensors and a versatile microcontroller chip for the purpose of real-time monitoring of data which is an advantage over the older processes where the devices were used for a one-time analysis of the soil and crop [13]. The advantage of a versatile system as such is that multiple nodes can be connected to the neural network in future which will input further data which allows for better optimization of the predictable output. Compared to other precision farming tools, our architecture has been designed to incur the minimum cost for the user during the installation period [14].

56.6 Conclusion

In our proposed framework “Smart Agriculture implemented using the Internet of Things and Deep Learning,” we have introduced a novel technique to implement the procedure for smart farming with the use of sensors and IoT architecture to build a system which can collect real-time data from the farms and then create a dataset which can finally be fed onto a neural network where the data can be made sense off and processed to create the best fit for the output data which can be displayed on an application layer which can be utilized by the farmers to get a better insight of the farming activities. The adoption of our proposed system would allow bringing the farming community nearer to the modern-day technology to reap the best possible effects from the technological advances in society.

References

1. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of Things: a survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tutor.* **17**(4), 2347–2376 (2015)
2. Fatani, A., Kanawi, A., Alshami, H., Bensenouci, A., Brahim, T., Bensenouci, A.: Dual pH level monitoring and control using IoT application. In: Proceedings on 15th Learning and Technology Conference (L & T), pp. 167–170 (2018)
3. Ayaz, M., Ammad-Uddin, M., Sharif, Z., Mansour, A., Aggoune, E.-H.M.: Internet-of-Things (IoT)-based smart agriculture: toward making the field talk. *IEEE Access* **7**, 129551–129583 (2019)
4. Nagarajan, G., Minu, R.I.: Wireless soil monitoring sensor for sprinkler irrigation automation system. *Wireless Pers. Commun.* **98**(2), 1835–1851 (2018)
5. Nagarajan, G., Thyagarajan, K.K.: A machine learning technique for semantic search engine. *Procedia Eng.* **38**, 2164–2171 (2012)
6. Nirmalraj, S., Nagarajan, G.: Fusion of visible and infrared image via compressive sensing using convolutional sparse representation. *ICT Express* (2020)
7. Ahmed, N., De, D., Hussain, I.: Internet of Things (IoT) for smart precision agriculture and farming in rural areas. *IEEE Internet Things J.* **5**(6), 4890–4899 (2018)
8. Zhan, W., Chen, Y., Zhou, J., Li, J.: An algorithm for separating soil and vegetation temperatures with sensors featuring a single thermal channel. *IEEE Trans. Geosci. Remote Sens.* **49**(5), 1796–1809 (2011)
9. Kamath, R., Balachandra, M., Prabhu, S.: Raspberry Pi as visual sensor nodes in precision agriculture: a study. *IEEE Access* **7**, 45110–45122 (2019)
10. Holland, K.H., Lamb, D.W., Schepers, J.S.: Radiometry of proximal active optical sensors (AOS) for agricultural sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **5**(6), 1793–1802 (2012)
11. Ding, T., Liu, Z., Li, B.: Study the relationship between the flow-ratio of air and solution (FRAS) and the dehumidify capability and optimize the FRAS in the solar dehumidification system. In: Proceedings of the Second International Conference on Mechanic Automation and Control Engineering, pp. 6199–6202 (2011)
12. Elijah, O., Rahman, T.A., Orikumbi, I., Leow, C.Y., Nour Hindia, M.H.D.: An overview of Internet of Things (IoT) and data analytics in agriculture: benefits and challenges. *IEEE Internet Things J.* **5**(5), 3758–3773 (2018)

13. Aishwarya, R., Yogitha, R., Kiruthiga, V.: Smart road surface monitoring with privacy preserved scheme for vehicle crowd sensing. *J. Comput. Theor. Nanosci.* **16**(8), 3204–3210 (7) (2019)
14. Aishwarya, R., Deepika, S., Sowjenya, K.: Assessment & forecast of air quality data based on neural network. *Int. J. Pharm. Res.* **12**(1), 1397–1402 (2020)

Chapter 57

Effect of COVID-19 on Stock Market Prediction Using Machine Learning



J. Kalaivani, Ronak Singhania, and Shlok Garg

Abstract Share market is a chaotic and ever-changing place for making predictions, as there is no defined procedure to evaluate or forecast the value of a share of a company. Methods like time series, technical, statistical and fundamental analysis are used to predict the price of a share. However, these methods have not proven to be very consistent and precise for making predictions. COVID-19 has further deteriorated the chances to find such a tool as the markets have taken a huge hit in the first quarter of 2020. In this paper, support vector machine and multiple regression algorithms will be implemented for predicting stock market prices. Our aim is to find the machine learning algorithm which can predict the stock prices most accurately before and during the pandemic. The accuracy for every algorithm will be compared and the algorithm which is the most accurate would be considered ideal.

57.1 Introduction

The stock market is one of the oldest and most complex ways to do business. Stock market is considered to be an arrangement between two consulting parties who buy and sell equity shares of companies. The parties can be investors or traders.

Stock market is volatile, complicated and ever changing, thus making it very tedious to make reliable predictions. Therefore, making predictions for stock prices have been a field of curiosity for investors and traders alike.

Predicting stock prices are the process of forecasting the price of a share in the stock market. Large profits can be made by successful prediction of stock prices. For

J. Kalaivani (✉) · R. Singhania · S. Garg

Department of Computing Technologies, SRM Institute of Science and Technology,

Kattankulathur, Chennai, India

e-mail: kalaivaj@srmist.edu.in

R. Singhania

e-mail: rs7591@srmist.edu.in

S. Garg

e-mail: sp1821@srmist.edu.in

this to happen, the participants need to have a proper understanding of the trends and patterns of stock prices. They must be able to predict where to invest in order to maximize their profit. Market risk needs to be curtailed to a minimum to ensure least risk in investment. This can be done by strongly correlating with forecasting errors. This is the basis behind every investment.

There are multiple factors that affect stock prices, such as political, cultural and socio-economic. All these factors contribute towards making the stock prices unreliable and extremely hard to predict with a great deal of consistency. Features like daily stock prices and new developments within the firm will be used to accurately predict the future stock prices by implementing machine learning algorithms.

In this paper, prior data relating to the stock prices of a company will be considered and machine learning techniques will be implemented on the data to forecast share prices. The stock market data of a long range of time will be combined and be used to predict future prices on the basis of the same using machine learning techniques like support vector machines, linear, lasso, ridge and Bayesian ridge regression algorithms.

The global economic chaos caused by the COVID-19 pandemic has resulted in large scale unemployment and a rapid decline in business activity worldwide. As a result, stock markets also faced the brunt of this pandemic in early 2020, however, some companies managed to make a rapid recovery in the second quarter.

COVID-19 has made this already tedious job of predicting the future stock prices more complicated as the world economy has taken a huge hit in these tough times. Making accurate predictions for stock prices during COVID-19 have been a huge concern for investors and traders alike.

There have been constant changes in the stock market almost daily, with the world unable to cope with this never seen before situation. Making predictions in stock prices with such skewed and inconsistent data are what has made this process even tougher.

Our aim is to find the machine learning algorithm amongst the ones mentioned above which can predict the stock prices most accurately before and during the pandemic. The accuracy for every algorithm will be compared and the algorithm which is the most accurate for all the predefined time periods will be considered ideal.

57.2 Related Work

Subhadra and Chilukuri [1] focus on proving that random forest is the most accurate and efficient machine learning algorithm for predicting share prices. The aim was to select the most suitable attributes that provide the best result.

Kunal and Agarwal [2] propose to forecast the stock prices with a great deal of accuracy by using pre-existing algorithms and open-source libraries. The prediction model relied mainly on numbers and sometimes deviated from the actual world.

Shah [3] used most of the well-known machine learning algorithms and compared the results in accurately predicting the stock market prices. The algorithms used were linear regression, support vector machines, decision stump and boosting.

Reddy [4] proposes a prediction model using a machine learning algorithm. The model was used to predict the future stock prices using time series analysis. The algorithm used in this paper is support vector machine (SVM). The dataset was trained to make accurate predictions for the prices of any stock.

Maini and Govinda [5] focus on making reliable predictions of future stock prices. Machine learning algorithms like random forest and support vector machine were used. The aim was to predict if the price of a particular stock would increase or not.

Jibing and Sun [6] propose a new model for prediction of stock prices. In this model, use of logistic regression to forecast stock prices for the upcoming month based on data from the present month was done. This prediction model brings novelty because this does not consider data from the past.

Rustam et al. [7] use machine learning algorithms to predict the number of possible people who can get infected with COVID-19 during a particular time frame. This predictive model uses machine learning algorithms like logistic regression and SVM to make these predictions.

Usha and Sarkar [8] focus on making predictions of stock prices for companies listed on BSE SENSEX using machine learning algorithms. The main algorithm used is logistic regression. According to the paper, logistic regression helps provide better accuracy for predictions.

57.3 Methodology

57.3.1 Dataset

The dataset being used for our model is taken from www.quandl.com, a premier dataset platform. The dataset taken is for SENSEX, which is the index for stocks in BSE. 17 years of data from 2004 to 2020 were extracted. The attributes extracted from the dataset are opening and closing prices, volume and extremes of the prices of the stock.

The extracted dataset was divided into four parts:

- a. Training dataset containing data from the year 2004 to September 2019.
- b. Pre-COVID testing dataset containing data from October 2019 to December 2019.
- c. Early-COVID testing dataset containing data from March 2020 to May 2020 where there was a steep decline in the stock market.
- d. Mid-COVID testing dataset containing data from June 2020 to August 2020 where there was a recovery in the stock market.

Table 57.1 Technical indicators

S. No.	Technical indicator	Definition
1	Exponential moving average (EMA)	Average price with more weightage to recent data
2	Relative strength index (RSI)	Measures the magnitude of recent changes
3	Commodity channel index (CCI)	Difference between current and historical average
4	Moving average convergence divergence (MACD)	Shows relationship between two moving average prices
5	Stochastic indicator (STOCH)	Compares the current price with historical extremes
6	Know sure thing (KST)	Interprets rate of change of price
7	True strength index (TSI)	Used to determine trends and trading signals
8	Percentage price oscillator (PPO)	Percentage change between two moving averages
9	Donchian channels (DC)	Used to measure volatility of the market

One of the novelties of this dataset is that it provides both training and testing datasets. This gives a better evaluation of our trained model [9–13].

57.3.2 *Technical Indicators*

Technical indicators are analytical or pattern-based signals produced by the attributes in the dataset such as opening and closing prices, volume and extremes of a stock in the market. Future price movements and trends of a particular stock can be forecasted by analyzing historical data [14–16].

These indicators help in making more accurate predictions by increasing the number of features which are being used to calculate the predicted price of a stock.

Some of the technical indicators used in our model are as in Table 57.1.

57.3.3 *Implementation*

After data extraction and addition of technical indicators, the relevant features were selected. This was done to increase the accuracy by eliminating unnecessary features.

The data extracted were then cleaned to get rid of incomplete, inaccurate and irrelevant data. This decreases the discrepancies and increases the overall productivity.

To scale the data, the min–max scale was implemented. Scaling was done to avoid the algorithm from considering smaller values as lower values and greater values as higher, regardless of their unit. This is performed on both the training and testing datasets.

The training dataset, i.e. 2004 to September 2019 was used to train the model for each algorithm.

All the algorithms were applied to each testing datasets, i.e. pre-COVID, early-COVID and mid-COVID. Closing stock prices for each of them were predicted. The predicted closing prices were compared with the actual stock prices to determine the accuracy (Fig. 57.1).



Fig. 57.1 Model training workflow

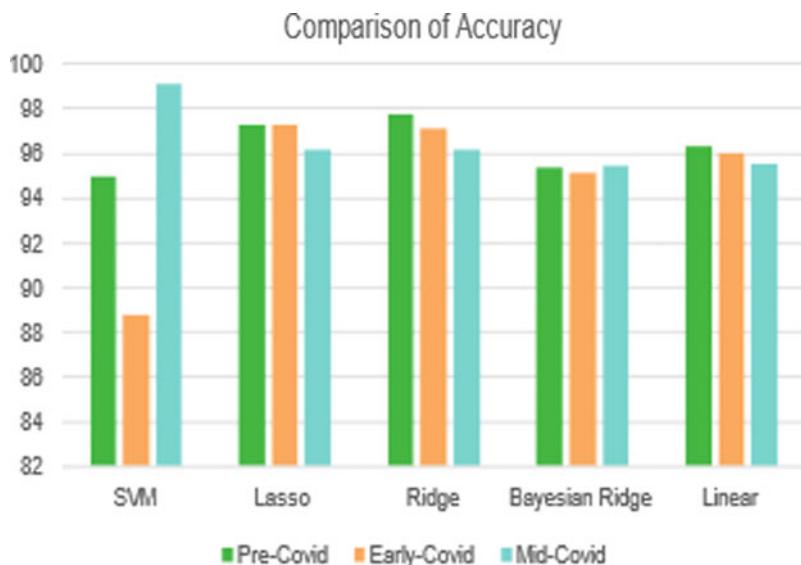


Fig. 57.2 Comparison of accuracies

57.3.4 Comparison

After the training and testing of the models with the BSE SENSEX dataset, accuracies were calculated for each algorithm for each time period. The accuracies found after the implementation of the model were plotted for the algorithms—lasso, ridge, Bayesian ridge, linear regression and support vector machine.

The graph given compares the results that were obtained to find the most suitable algorithm (Fig. 57.2).

Upon examination of the above graph, it was observed that even though SVM performed the best during the mid-COVID time period, it did not fare well during the onset of COVID. On the other hand, the regression algorithms were consistent at all times.

57.4 Conclusion

In the stock market world, the prediction of stock prices is the ultimate aim of traders. The ability to predict these prices from historical data is the most sought-after skill. This skill becomes even more valuable when the market is volatile, as was the case during the onset of COVID-19.

The effectiveness of the model is directly proportional to the potential for profits for the user. Machine learning algorithms can be used to accomplish the desired

results. In this analysis, an attempt was made to determine the most suitable machine learning algorithm for this purpose.

In this analysis, it was deduced that regression algorithms yielded consistent and accurate results for all time periods, however, support vector machines proved to be the most accurate during the mid-COVID time period, when the market saw a recovery. It was also observed that there was a significant decline in the accuracy of SVM in the early-COVID phase. This indicates that SVM does not perform adequately when the data are volatile and the market is unstable.

Thus, it can be concluded that regression techniques are well suited for the purpose of making stock market predictions and yield better results even in times of uncertainty.

References

1. Subhadra, K., Chilukuri, K.C.: Stock market prediction using machine learning methods. *Int. J. Comput. Eng. Technol.* **10**(3), 2019 (2020)
2. Kunal, P., Agarwal, N.: Stock market analysis using supervised machine learning. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 197–200. IEEE (2019)
3. Shah, V.H.: Machine learning techniques for stock prediction (2007)
4. Reddy, V.K.S.: Stock market prediction using machine learning. *Int. Res. J. Eng. Technol.* **5**(10) (2018)
5. Maini, S.S., Govinda, K.: Stock market prediction using data mining techniques. In: 2017 International Conference on Intelligent Sustainable Systems (ICISS), pp. 654–661. IEEE (2017)
6. Jibing, G., Sun, S.: A new approach to stock price prediction based on a logistic regression model. In: 2009 International Conference on New Trends in Information and Service Science, pp. 1366–1371. IEEE (2009)
7. Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B., Aslam, W., Choi, G.S.: COVID-19 future forecasting using supervised machine learning models. *IEEE Access* (2020)
8. Usha, A., Sarkar, R.: Application of logistic regression in assessing stock performances. In: 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), pp. 1242–1247. IEEE (2017)
9. Fieß, N.M., MacDonald, R.: Towards the fundamentals of technical analysis: analysing the information content of high, low and close prices. *Econ. Model.* **19**(3), 353–374 (2002)
10. Huang, W., Nakamuri, Y., Wang, S.-Y.: Forecasting stock market movement direction with support vector machine. *Comput. Oper. Res.* **32**(10), 2513–2522 (2005)
11. Nagarajan, G., Minu, R.I.: Multimodal fuzzy ontology creation and knowledge information retrieval. In: Proceedings of the International Conference on Soft Computing Systems, pp. 697–706. Springer, New Delhi (2016)
12. Nagarajan, G., Minu, R.I., Muthukumar, B., Vedanarayanan, V., Sundarsingh, S.D.: Hybrid genetic algorithm for medical image feature extraction and selection. *Procedia Comput. Sci.* **85**, 455–462 (2016)
13. Simpson, S.V., Nagarajan, G.: A table based attack detection (TBAD) scheme for Internet of Things: an approach for smart city environment. In: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 696–701. IEEE (2021)

14. Indra, M.R., Govindan, N., Satya, R.K.D.N., Thanasingh, S.J.S.D.: Fuzzy rule based ontology reasoning. *J. Ambient Intell. Humaniz. Comput.* 1–7 (2020)
15. Dhanalakshmi, A., Nagarajan, G.: Convolutional neural network-based deblocking filter for SHVC in H. 265. *SIViP* **14**, 1635–1645 (2020)
16. Sajith, P.J., Nagarajan, G.: Optimized intrusion detection system using computational intelligent algorithm. In: *Advances in Electronics, Communication and Computing*, pp. 633–639. Springer, Singapore (2021)

Chapter 58

Real-Time Elderly Multiple Biological Parameter Safety Monitoring System Based on mCloud Computing: Telemedicine Era



Sawsan D. Mahmood, Raghda Salam Al Mahdawi, and Shaimaa K. Ahmed

Abstract The patients' lives are impacted when patients do not get prompt and thorough medical examinations every day in clinics and hospitals alike, and real-time parameter values are not easily calculated. It can be daunting for hospitals to keep track of their patient's symptoms on a regular basis. In addition, elderly patients cannot be monitored continuously. To cope with cases like these, biosensors for healthcare purposes have been integrated with real-time machine learning technology. The technology-based on machine learning assist specialists in identifying patients for counseling as well as identifying and predicting diseases. The article's aim is to implement and develop a machine learning-based real-time intelligent health management system that gathers the data of the sensors from a wireless body sensor network and sends it to a historical clinical data-based predictive model. It enables specialists to remotely track patients, and if appropriate, take periodic decisions. Wearable sensors were used to evaluate a series of five parameters, including an ECG_rate, pulse_rate, pressure_rate, temperature-rate, and location detection. The machine employs two loops to do this. From the client-side, the transmission circuit is with the patient, and the reception circuit (from the server-side) is managed by the specialists or caregiver.

58.1 Introduction

Healthcare has gotten a lot of coverage in the last ten years. The key aim was to provide a dependable management structure that would allow healthcare workers to

S. D. Mahmood (✉)

College of Engineering, University of Tikrit, Tikrit, Iraq

e-mail: Sawsan.d.mahmood@tu.edu.iq

R. S. Al Mahdawi · S. K. Ahmed

Department of Computer Engineering, University of Diyala, Baqubah, Diyala 32001, Iraq

e-mail: raghdasalam@uodiyala.edu.iq

S. K. Ahmed

e-mail: Shaimaa_khamees@uodiyala.edu.iq

monitor patients who were ill or went about their daily lives. Due to new technology, hospital monitoring systems and caregiver services have recently been among the most significant innovations. There is now a need for a more contemporary approach. In the traditional approach, healthcare professionals play a vital part. They would pay a visit to the patient's ward in order to get the requisite prescription and advice. There are two major issues with this strategy. First, healthcare staff must be available at all hours at the patient's location, and second, the patient must stay in the hospital for a period of time, with biomedical instruments at his or her bedside. Patients are given awareness and facts about disease diagnosis and treatment in order to address these two issues. Second, a patient tracking device that is both accurate and accessible is needed. To boost the above situation, technologies can be used more intelligently by combining machine learning methods with various sensors. Wearable electrodes come into contact with the human body and monitor the physiological parameters. ECG sensors, temperature sensors, pulse controllers, and other sensors are available on the market today. Sensor prices vary depending on their scale, flexibility, and accuracy. In our method, we use a Raspberry Pi to monitor patient parameters (ECG, temperature, heart rate, pulse, etc.) [1, 2]. There are a variety of sensors available. This sensor gathered information, such as biometric data, which was then sent to the Raspberry Pi and then to the server. Depending on the application, sensitive information can be transmitted wirelessly using different solutions such as Wi-Fi, 3G, GSM, Bluetooth, and so on. Only registered staff has access to the data contained in a database and viewable on a Website. Physicians, risk control agencies, patients, and their family members may be given permission. The system even makes it simple for the specialist to view the patient's previous history from the memory data [3, 4].

58.2 Related Literature

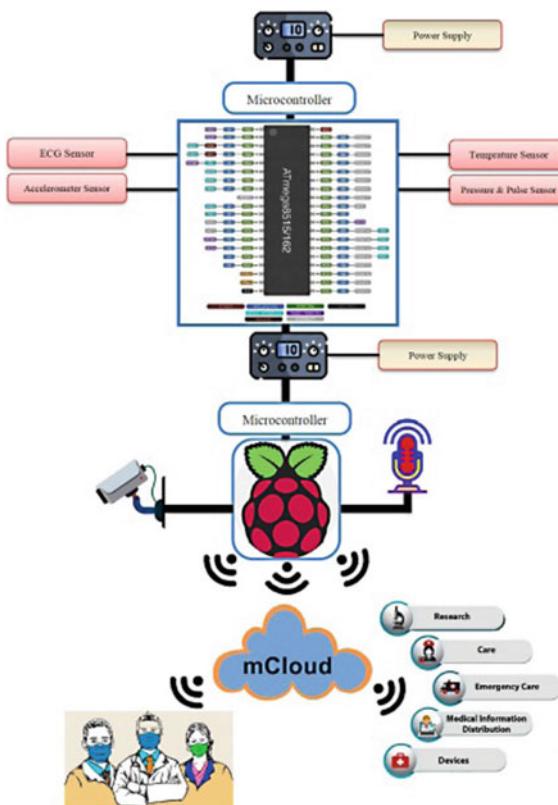
The IoT-based health management system has been developed with RaspPi and Arduino. Uplenchwar and Vedalankar, the mechanism consists primarily of three stages, the transmission, the process unit and the receiver. The transmission end is mostly comprised of biosensors which collect possible signals from the body of a patient [5]. Wearable sensors were used to evaluate a range of five factors, such as ECG, pulse, weight, temperature, and location. The Arduino and RaspPi are wired to the sensors. RaspPi serves as server and sends data to a given URL after it is linked to the Internet. In any electronic computer, like laptops or smartphones connected to the same network, critical parameters may be viewed and tracked. But there is no live tracking and data storage facility for this device. Baker et al. published an Internet research paper on smart healthcare things: innovations, threats, and opportunities. Data from the attached sensor node was supplied to the central node. It processes material that helps to make a decision and then forward it to an external website. Machine learning algorithms allow patterns in previously unknown medical data to be detected, treating plans and diagnosis can be delivered and individual healthcare providers and individual patients can be recommended. In order to promote

the application of computer education in broad datasets [6], mCloud computing storage systems must be built. With a RaspPi 2, Kirankumar and Prabhakaran have introduced a cheap Web-based human health surveillance system. Here are metrics of patient's wellbeing, for determining whether or not patients are drinking alcohol, ECG sensors, sound sensors, patient exhaustion level sensors, and live video cameras [7]. The RaspPi 2 microcontroller collects all these parameters and views them via the Putty SSL client on the PC. The RaspPi wired Wi-Fi module allows the device to be connected to the Internet through a Wi-Fi network. This allows the patient or child to be tracked entirely remotely on a Web page. The device drawback is that mCloud computing cannot classify a single specialist for a sensor data consultation. Introducing a healthcare infrastructure where a lot of data is available. Bhardwaj et al., this involves electronic medical reports containing either organized or unstructured data. It may include a range of facts and definitions from but not limited to the size of the patient, to temperature and also to general conditions such as headache and stomach pain. Standardized health records can be divided into a database. Use machine learning efficiently to help specialists diagnose almost perfectly, choose safer drugs for the patients [8]. Gurjar and Sarnaik have created an Internet of things device for heart attack monitoring by detecting heartbeat: IoT. The pulse and temperature can be periodically detected with the aid of a sensor. For both parameters, the specialist should set the threshold. When the limit exceeds these thresholds, the device sends server alerts via Wi-Fi. A compact system, which saves the chance of cardiac arrest because you can check it at home, afford it, and track temperature and heart rate with one watch, is the key benefit of this system. One individual seated in the server room is monitoring the patient. It also assists in the surveillance of hospitals. The device limitation requires medical reports that are not accessible on the Internet, no clear tracking, and noise problems with data received [9]. Kalamkar et al. introduced a method for human health surveillance through IoMT and RaspPi 3. Multiple sensors connecting to the RaspPi 3, which tracks all the information from the sensors to the IoT Website wirelessly. This is usually sent continuously to the hospital's Web server. An SMS will be sent to the specialist in case of any crisis in the operation of calm later. The key benefits of this device are to provide immediate information for those affected, to speed up and broaden communications coverage and promote freedom to increase the quality of life of the patient, to enable emergency monitoring and minimize deaths in cases of an injury. The system's drawback is the absence of data management requirements and the inability to transmit live videos [10–12].

58.3 Methodology

Most of the machines are made of Arduino UNO and RaspPi, two control panels. Arduino UNO is serially connecting by an ECG, acceleration and pulse sensors to a temperature sensor, a speed sensor. The Pi, board Type B is connecting to the microphone and the camera. The Arduino data was sent to the RaspPi and sent to the

Fig. 58.1 Health surveillance system scheme



SQL database of the DJANGO program. Live monitoring is provided on site. When an SVM classification occurs, classify sensor data and record an emergency situation. If the datasets differ, the response will be sent to the appropriate practitioner via the SMS WAY 2 program. It also warns the fire department if there is an emergency. The health monitoring system graph as seen in Fig. 58.1.

58.3.1 Implementation of SVM

A regular backup the primary aim of the classification vectors is to balance the medical data providing by a hyperplane “fit best.” Then, feed the classifier to see the “foreseen” class after the hyperplane is collected. This makes this algorithm particularly desirable for use, but you will use it in some situations.

In general, the function list is saved in a big X variable. Depending on the data form, we use 0, 1, 2, or 3 for our ratings, often referred to as “goals.” Looking at our dataset, we can see that there are “small” coordinate pairs and “higher” number

coordinates. Then set 0 pairs with lower coordinates and 1 for pairs with higher functions. There are marks. These are marked. The following is a description of label compression data.

def Pressure:

```
x = [[90, 60], [80, 50], [85, 65], [95, 65], [90, 50], [80, 40], [88, 58], [120, 80],  
[115, 75], [120, 75], [120, 70], [120, 78], [115, 75], [115, 80], [121, 81], [122,  
82],  
[123, 83], [125, 85], [128, 86], [130, 85], [135, 88], [140, 90], [140, 95], [145,  
90],  
[140, 95], [150, 100], [150, 95], [150, 105]]  
y = [0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3]
```

To define classifier:

```
clf = svm.SVC(kernel = linear, C = 1.000)
```

The kernel will be linear and C will be 1.0. That is a good parameter default. We are calling next:

```
clf.fit(X, y)
```

The study takes place from here. Since we have this dataset, it should be almost immediate. Figure 58.2 shows a clinician's prediction with the SVM.

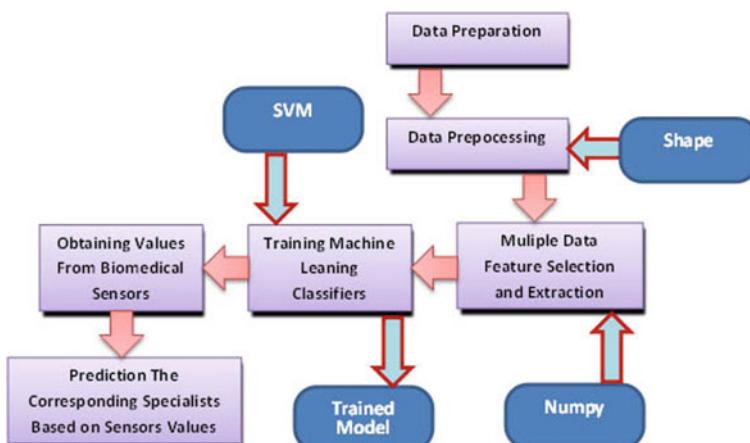


Fig. 58.2 Flowchart of SVM

58.4 HW Description

The ATMEGA8515-16PU microcontroller is the Arduino/Genuino UNO board. It has the specification, Package PDIP40; Program Memory 8 k; SRAM 512; EEPROM 512; I/O Pins 35; Timers 1 × 8, 1 × 16; A/D Comparator; SPI Yes; I²C Yes; PWM 3; USART Yes; Oscillator 16. The unit may be attached with a USB or an AC-to-DC converter or a startup battery to the computer. The ECG tracking sensor package is a fully assembled ECG leading end with a minimal 3.5 V, limited 170 A supply requirement. This architecture allows an integrated microwave to produce the achieved signals. The “AD8232” is a low-cost heart function measuring board. This electrical working of the core is an EKG and an output. The monitors “AD8232” can be very noisy and can be used as an op-amp for a basic PR and QT time indicator. For outdoor projects, a serial output is shown to process and view external circuits. Pressure of the blood is shown. The reading is systolic, pulsated, and diastolic. Compact style that suits as a wrist watch. The machine is easy to use with a handle that stops pumping. Gyrometer MPU-6050 3 Axis, GY-521 MPU6050 3 Axis Gyrometer Analog Sensors, MPU-6050 3 Axis Gyrometer Module, 3 axis accelerometer. MPU6050. Stromversorgung 3–5 V. The modes of communication are standard protocols for IIC communications. A 16-bit A/D converter and a 16-bit data output are included in the chip. Gyroscope spectrum: 250 500 1000 2000 or s. The size of graphic meters is from 30 to 4000 dps. RaspPi is a credit card device connecting to a computer screen or a TV with a standard mouse and keyboard. It helps seniors to join to experiment and learn to code in Scratch and Python languages. C920 HD is a portable Web-cam and a compact Web-cam. It is 70 × 18 × 30 mm and can comfortably fit into the palm of your hand. The Logitech C920 Web-cam head can be rotated to perfect vision in various directions (Fig. 58.3).

58.5 SW Description

An open-source prototyping platform for code writing and downloading is simple to use, software and hardware-based. Each Arduino board is open-source implementation that allows users to build and finally meet their unique requirements. Mike Thompson and Peter Green created Raspbian for the first time. In June 2012, the original construction was finished. There has been already ongoing growth of the operating system. ARM CPUs of the RaspPi line are significantly optimized for the lower performance. Django is a Python language open-source Web frame that follows the architectural structure of a shape presentation template (MVT). The main purpose of this software is to build sophisticated Websites easily powered by databases. Django emphasizes component reusability and connectivity, low coding, low coupling, and fast Web page growth. Python is used in both files and data models of setup.

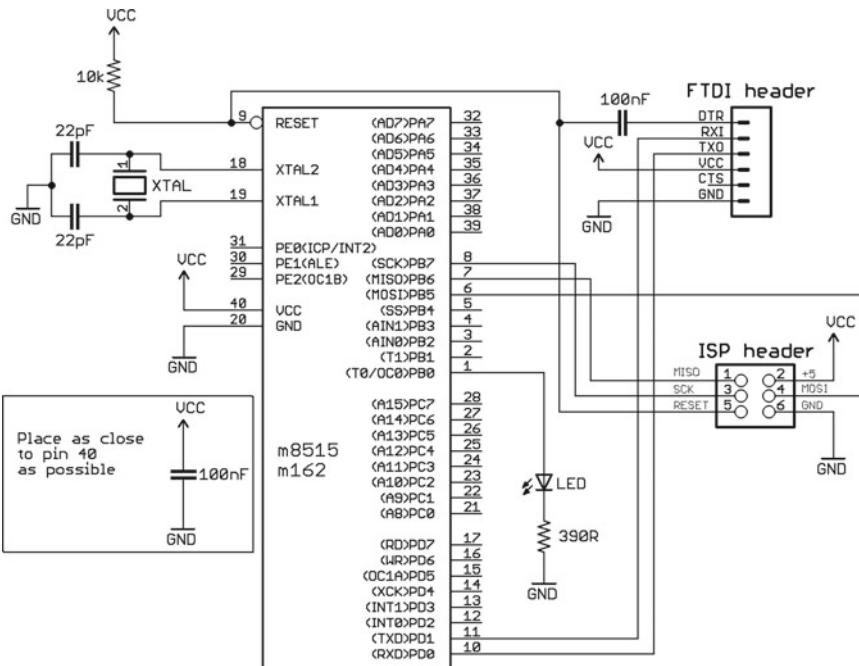


Fig. 58.3 Major core minimal setup

58.6 Discussion the Findings

Turn on and on the computer, then turn the blood pressure monitor on and the device will start running and transferring data to the Django server. The register and password pages make up the Web page. Using the email id and password, the licensed physician will access the registry. After signing in, the specialist will review the patient's physical parameters and details, as well as track the patient in real time by using the camera and microphone. He can even prepare drug orders for his patients. The objects that a patient presents to a specialist are registered as audio files, which the specialist can listen to by clicking the download audio button. It consists of the administrator, who has the ability to add new patients as well as remove them from the Web page. Figure 58.4 shows the various live Web pages for patient monitoring in real time.

58.7 Concluding

The planned device uses the RaspPi as a contact tool for real-time contact between the patient and the expert, which also incorporates a camera module, which is a

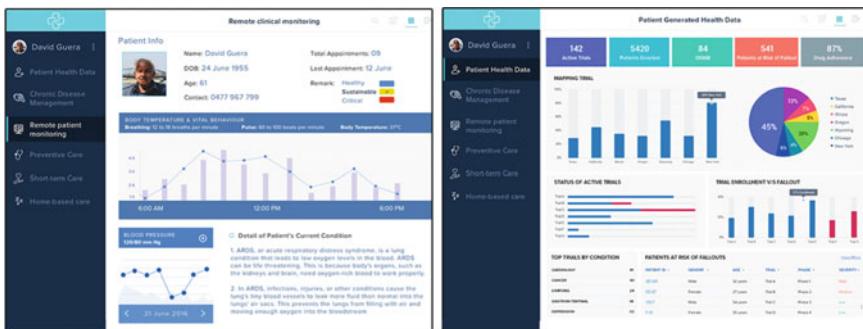


Fig. 58.4 Multiple live interfaces for daily elderly monitoring

revolutionary innovation that goes beyond the current framework. The RasPi board system is a fast computing system that can also be used in the modern age. The RasPi sends sensor data to specialists through the Internet, which can be very useful in the medicine field and enable specialists to retain a close eye on health of patients. Any health problems improvement may be immediately detected and communicated to a single individual, specialist, or emergency service via SMS. The project's aim is to link patients in India's rural areas with a specialist in clinics and major cities, enabling them to access high-quality care.

References

1. Gómez, J.: Patient monitoring system based on Internet of Things: a survey. *Comput. Netw.* **54**, 2787–2805 (2016)
2. Malokar, S.N., Mali, S.D.: Patient monitoring system based on Internet of Things using Raspberry Pi: key features, application and open issues. *Comput. Commun.* **54** (2016)
3. Pioggia: Personal health system architecture for stress monitoring and support to clinical decisions. *Comput. Commun.* **35**, 1296–1305 (2017)
4. Delmastro, F.: Pervasive communications in healthcare. *Comput. Commun.* **35**, 1284–1295 (2017)
5. Uplenchwar, K., Vedalankar, A.: IoT based health monitoring system using Raspberry Pi and Arduino. *Int. J. Innov. Res. Comput. Eng.* **5**(12) (2017)
6. Baker, S.B., Xiang, W., Atkinson, I.: Internet of Things for smart healthcare: technologies, challenges, and opportunities. *IEEE Trans.* **5** (2017)
7. Kirankumar, Prabhakaran: Design and implementation of low cost web based human health monitoring system using Raspberry Pi 2. In: International Conference on Electrical, Instrumentation and Communication Engineering (2017)
8. Bhardwaj, R., Nambiar, A.R., Dutta, D.: A study of machine learning in healthcare. In: IEEE Annual Computer Software and Applications Conference (2017)
9. Gurjar, A.A., Sarnaik, N.A.: Heart attack detection by heartbeat sensing using Internet of Things: IoT. *Int. Res. J. Eng. Technol.* **05**(03) (2018)

10. Kalamkar, P., Patil, P., Bhongale, T., Kamble, M.: Human health monitoring system using IoT and Raspberry pi3. *Int. Res. J. Eng. Technol.* **5**(3) (2018)
11. Bansal, R., et al.: *J. Phys. Conf. Ser.* **1963**, 012170 (2021)
12. Sharma, S., Singh, J., Obaid, A.J., Patyal, V.: Tool-condition monitoring in turning process of Fe-0.75Mn-0.51C steel with coated metal carbide inserts using multi-sensor fusion strategy: a statistical analysis based ingenious approach. *J. Green Eng.* 2998–3013 (2021)

Chapter 59

Intelligent Transportation Systems (ITSs) in VANET and MANET



Sami Abduljabbar Rashid, Lukman Audah, and Mustafa Maad Hamdi

Abstract The ad hoc vehicle network (VANET) has been described as one element of smart transport systems (ITSs). VANET is an on-the-flight network focused on vehicle availability on roads and ground service facilities such as stations. This infrastructures on the roadside have to provide connectivity services, in particular where there are not enough cars on the roads for efficient communication. The VANET nodes which now provide on-road vehicles and base stations along the roads that may be used in vehicles (V2V) and/or vehicle-to-infrastructures for connectivity (V2I). In contrast, ITS has provided road users with a broad variety of safety and unsafety applications that VANET can use to its best advantage in implementing an accessible and reliable ITS. However, the basic fundamental issue in VANET is that an efficient contact chance is only possible between two fast moving cars when a secure relation is formed between them, but only in a few seconds. The goal of this article is to present the various issues that are considered as challenging aspects in VANETs and to provide the future research gaps in this field.

59.1 Introduction

In the Malaysian highways and road systems, almost every corner of open areas is the greatest in the country. In Malaysia, the overall number of licensed motor vehicles rose 15 million at the end of 2005, with the motorcycle taking the largest share of the vehicle at 7 million (47%) and the passenger cars taking the closest part at 6.5 million at the end of 2005 (43%) [1]. The highest effect of fast motorization is the rising road deaths and pollution of vehicles. With increasing automotive ownership, the need for transport to wherever has been emphasized and traffic congestion has been stepped up, and drivers have been more exposed to deadly road

S. A. Rashid (✉) · L. Audah · M. M. Hamdi

Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia

L. Audah

e-mail: hanif@uthm.edu.my

accidents. In Malaysia, the vehicle density ratio was 71 vehicles per km road in 1999, in 1994, from a simple 46 [2]. Ad hoc vehicle network (VANET) was recognized as part of intelligent transport systems components (ITS) [3]. VANETs are an on-the-flight network focused on vehicle availability on roads and ground service facilities such as stations. This infrastructures on the roadside must provide connectivity services, where there are not enough cars on the roads for efficient communication. The VANET nodes which now provide on-road vehicles and base stations through the road that may be used in vehicles (V2V) and/or vehicle-to-infrastructures for connectivity (V2I). There are three types of communication networks, vehicle-to-vehicle communication (inter-vehicle Communication), vehicle-to-roadside communication (vehicle-infrastructure communication), and roadside-to-roadside communication (inter-infrastructure communication). ITS, in the contrary, has offering users of roads a wide variety of protection and security applications, which VANET can use to the maximum to deploy a successful ITS. However, VANET's specific fundamental issue is that there is only a possibility of successful two moving vehicles contact when stable connections between them are established [3]. The propagation (or routing) of message is thus a problem for many ITS applications' safety and security. The difficulty of developing a stable yet confident communication strategy is mainly based on the rigorous ITS security applications QoS specifications [4]. Considering the trends where people work and communicate when on the move in new global lifestyles, these lifestyles must constantly be strengthened to increase connectivity, comfort, and efficiency. When on the roads, VANET can give road users a kind of mobility help, they have equal, similar, or equivalent access, comfort, and productivity [5]. It is possible to communicate from a road user to a central office and vice versa, via multi-hop connection and through the road's base stations [6]. In addition, a paradigm for secure communication (or routing) technology must be developed critically for ITS to be deployed effectively with some levels of QoS. If the drivers and the relevant authority, including police departments, RTD, fire bridget, and road bodies can efficiently relay those smooth signals, then road accident can be cut down to a very low level. The most reliable connections between two communication vehicles can be accomplished in an efficient period, determining the radius, speed, distance, and orientation of their coverage. The QoS can be calculated with many parameters such as latency, pass, and packet lost rate that have been observed in intelligent transport systems by conveying a message from a sender node to the last receiver node [7].

The remaining of the article is presented as follows. In Sect. 59.2, we present the related work. Next, we present the contributions in Sect. 59.3. Afterward, standard, security/privacy, and QoS in VANET networks are presented in Sect. 59.4. Lastly, VANET and MANET challenges are presented in Sect. 59.5. The summary and conclusion are given in Sect. 59.6.

59.2 Related Work

Intelligent transport systems (ITS) provide targeted road safety and efficiency services and applications, especially in vehicle communications (IVC). The acceleration of population and economic development of urban centers calls for efficient mechanisms to schedule, reduce, and control traffic incidents and take the right decisions to enhance driver and bystander mobility. Vehicle networks enable transport management through the exchanging of messages in real time among cars; drivers can receive updates on the transit status, collisions, and other road-related eventualities, enabling them to choose the best routes, to avoid congestion and road incidents. Implementing services in vehicles requires robust performance assessment, design, and device activity on the street as nodes. Communication between these nodes is carried out using the roadside equipment's called as intelligent transport systems (ITSs) [8]. This exchange will take place between vehicles and vehicles. Roadside units are connected to the backbone network scattered across roads to facilitate connectivity. These units are in the short distance from cars, i.e., between 100 and 300 m [9]. In any case, if the car cannot connect to a disadvantaged roadside unit, other vehicles linked to an interconnection mobile roadside can be joined [10–12]. The correspondence in VANETs can, therefore, be categorized in three distinct forms vehicles to cars communicated by roadways known as roadside vehicle communication, and lastly, inter-road communication that allows the connection to the network, facilitating the communication between vehicles on the highway. Inter-vehicles communication between two vehicles which form inter-vehicle communications [8]. Dedicated short-range communications, i.e., DSRC in the range of 5.85–5.925 GHz is the dedicated 75 MHz spectrum allocated by the Federal Communications Commission for the VANET [13].

This allows wireless contact between VANET individuals without any single point of entry. Any vehicle can, therefore, be a sender, recipient, and router to transmit data in the ITS. This connectivity is feasible by installing VANET GPS and on-board systems on the vehicles and on highways, which enables communication across an ad hoc network that is short-range wireless [14]. The next section addresses the basic mobile ad hoc network philosophy and then details the available features and benefits of car interactions. Mobile networking is one of the most important innovations for computational advancement and is the key technology for the creation of an ad hoc vehicle network. It discusses the fundamental principle of MANETs. In different methods, such as infrastructure-based and ad hoc network communication can be conducted between two cellular mobile units. Wireless mobile networks depend on wireless definition and decent service support and link with fixed network infrastructure access points in these mobile devices. Examples such as GSM, UMTS, WLL, and WLAN are distinguishing example [15]. In these days, wireless networking and mobile devices have become widely available for self-organizing networks that do not require pre-established infrastructure. Ad hoc network building blocks are self-contained nodes for transferring knowledge. The simultaneous applicability of these nodes includes all end systems and routers. Ad hoc networks can be categorized as

statically ad hoc and mobile ad hoc in two groups. When a mobile ad hoc network is established, the location of a static ad hoc network has become a part of the network. There is, therefore, generally known as a mobile ad hoc network as MANET [16]. This provides the base to which vehicles link as an ad hoc vehicle network (VANET). VANET is a MANET with vehicles being the communicating nodes. A network that is built without a fixed network architecture or centralized administration is called a MANET and has mobile wireless nodes that are dynamically network compatible. These mobile nodes are reinforced with wireless transmitters and receptors by antennas that broadcast antennas or combine antennas. Between these communication nodes, a multi-hop or ad hoc network is built based on their coverage patterns for transmission and reception, co-channel interference levels, and transmission power levels. The ad hoc network topology will differ from time to time depending on the location of the traveling nodes and change the receipt parameters [17]. The volume editors, usually the program chairs, will be your main points of contact for the preparation of the volume.

59.3 Contributions

The contributions of this review are given as follows.

1. It provides an overview of the standard, security/privacy aspect in VAENTs.
2. It presents the various issues that are considered as challenges in this area.
3. It presents the research gaps and future works of researchers in the topic of VANETs and intelligent transportation systems.

We present a block diagram of the layout of the review article and its various discussed parts as shown in Fig. 59.1.

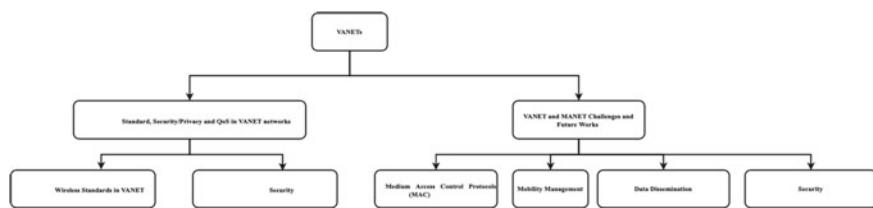


Fig. 59.1 Block diagram of the layout of the article and its main discussed points

59.4 Standard, Security/Privacy, and QoS in VANET Networks

This section provides an overview of the VANETs standard, security, privacy in VANETs. We present the wireless standards in VANETs in Sect. 59.4.1. The security aspect is given in Sect. 59.4.2.

59.4.1 Wireless Standards in VANET

IEEE 802.11 is based on standard IEEE 802.11 (most often referred to as Wi-Fi). It works in bands free of charge. Versions b and g have been expanded to include regular laptops and PDAs. Depending on the basic edition, it has a range of a few hundred meters and a speed up to 54 Mbps. The updated 802.11n edition aims to boost transmission speeds to 500 Mbps. Security is used in the initial protocols and increased in the revision in 802.11i [18]. At the same time as the IEEE standard of 802.11b, the IEEE standard was accepted. The regular 802.11a works on the 5 GHz band with a transmission speed of 54 Mbps. It cannot communicate with standard 802.11b equipment except when equipment is available and complies with both standards. Since electrical devices like microwaves, cableless devices and other appliances use the 2.4 GHz band, the 5 GHz band represents a typical 802.11a advantage, since less interference is present. The use of this band, however, is also disadvantageous because it limits the use of 802.11a to just a sight line, requiring the installation of a larger number of points of entry. It also ensures that teams that use this standard cannot join the 802.11b standard because their waves are consumed more quickly [19] is under standardization and will be responsible for promoting vehicle connectivity, also referred to as wireless access to the vehicle environment WAVE. WAVE is an IEEE 802.11a development with physical and MAC improvements to improve its behavior and to support smart transport systems (ITSs).

Similarly, the WAVE will be used to improve the dedicated short-range communications (DSRCs), another US Department of Transport standardization initiative and a large number of car manufacturers whose aim is to develop a national vehicle communications network [20]. WAVE is aimed at increasing transfers between 100 and 500 m, mostly in the short term. IEEE802.11a is a modulation technique that uses OFDM but has transmission speeds of 3, 4.5, 6, 9, 12, 18, 24, and 27 Mbps in 10 MHz. It uses 52 BPSK, QPSK, 16-QAM, or 64-QAM modulated undercover devices. In the channeling sector: 6 service channels (SCH) and one control channel of the 10 MHz non-overlapping channels of 5.9 GHz band: (CCH). The CCH is used as a channel to identify nearby vehicles first before connections are established [21].

59.4.2 Security

In addition to the standard wireless networking susceptibility, ad hoc networks have their own security issues. The lack of unified control capability raises security concerns and seems to be vulnerable to number of attacks when the mobile host enters and establishes the network with different network topology that varies dynamically. Naturally, the protection problem requires priority for the effective transmission of data. Due to the increasing sensitivity and vulnerability to external threats, wireless connectivity across nodes and changed network topology of mobile networks cannot be applied to the ad hoc networks because of their infrastructure-based routing functions [22]. Wireless safety in ad hoc networks is mostly difficult to achieve, due to the vulnerability of connections, insufficient physical security of each node, rare networking design, dynamically shifting topology, lack of a certification authority, and the absence of a central monitoring. This underlines the need to detect interferences, deter interferences, and related countermeasures. Wireless connections between nodes are extremely vulnerable to outbreaks that involve the following threats, passive eavesdropping, Active interfering, leakage of secret information, data tampering, data tampering, impersonation, and message replay or distortion [22]. Eavesdropping may provide an opponent with access to classified distortion [22].

59.5 VANET and MANET Challenges and Future Works

Decentralized nature, self-organization, and self-management pose the biggest obstacles in developing VANETs and MANETs, as the ability to drive vehicles is great. In addition, all conversations are carried out in short distance. For VANETs and MANETs, these special features pose significant challenges as

59.5.1 Medium Access Control Protocols (MACs)

Due to rapid changes in the topology and form of operation, MAC protocols in VANETs should be designed more importantly; as circulation messages in the control channel of vehicles are split into two key categories, they are categorized according to how they were created. Periodic notifications (safety signals) created to help vehicles respond to their environment by transmitting the current state of the vehicle to surrounding vehicles such as velocity, trajectory, or location. In the case of safety apps, this form of transmitted message may be used to enhance the content of the signals by both drivers so that immediate or hazardous conditions such as the crossing, the crash, and blind merge alerts are avoided.

59.5.2 *Mobility Management*

Besides fewer inter-vehicle connectivity networks, Internet resources will expand VANET applications. Internet gates (IGWs), which are mounted along the roadside, provide connectivity (RUs). However, Internet integration includes a corresponding support for versatility. Because VANET vehicles are extremely mobile, their IGWs also adjust as they receive Internet connectivity facilities. Thus, it is useful to provide mobility systems that consider the mobility of a car.

59.5.3 *Data Dissemination*

VANETs commonly use a mix of diffusion, multicast, unicast communications, and depending upon the kind of packets, we need to transmit between vehicles as opposed to other networks. The vehicle can transmit messages in all directions to any vehicle (everyone) or can be sent to a group or vehicle behind them (one-many). Each vehicle transmits information about itself and its other vehicles. Other vehicles, in the meantime, are provided with this information and are updated; accordingly, at that phase, the receiving vehicles will delay the communication until the following period.

59.5.4 *Security*

There are two reasons why the security issue is relevant. The first is to improve people's trust in the information processed and the information exchanged through those processes, through their explosive growth in computing systems and networks. However, this has enhanced the need to secure data and resources from access to other organizations and to protect these networks from network attacks. Second, the frameworks for cryptography have been developing and they can be implemented on these networks so that the intended individuals are able to safely encrypt and decrypt data. As security is a crucial component of VANET, the vehicles ad hoc network's striking features provide both threats and safety opportunities, as opposed to conventional (wired) networks where nodes need physical connections to network lines or to connect over multiple defense lines such as firewalls and gateways. VANET uses the wireless media so that attacks on the wireless network can be carried out in all directions. If it does not have such security measures, it offers high opportunities to be targeted. Connection attacks ranging from passive to active attack, replaying message, leakage of message, contaminated communication, and manipulation of message can therefore occur. This all suggest that VANET has no simple defensive line and that any node must be arranged for the various types of attacks. Thus, VANETs should have a distributed architecture, without centralized administration, in order to achieve a high level of survival and scalability, which should of course be

seen as a high mobility nature for VANET's, because pre-confidence is not possible to rely on in such networks. The distinctive features of VANETs present a range of new critical safety issues such as the decentralized infrastructure of peer-to-peer networks, wireless media networking, large-scale density, the importance of locations in vehicles, and the complex topology of networks.

59.6 Conclusion

Vehicle ad hoc network (VANET), one of the intelligent transport systems components has been established (ITS). VANET is an on-the-fly network focused on vehicle available on roads and road service facilities such as base stations. The road system is required if there is insufficient vehicles on the road to efficiently interact. The VAT nodes now provide vehicles for road and base stations that can be operated in vehicle to car (V2V) communications and/or vehicle communications with infrastructure (V2I). In contrast, ITS has provided road users with a wide range of safety and protection apps which VANET can powerfully use to ensure secure and effective ITS.

Future work is to investigate in the various issues and sub-problems in VANETs such as clustering and routing in 3D environment such as tunnels and bridges in smart cities. An additional future work is to survey the research work in the area of Internet of vehicles, its applications, and its relation with smart and intelligent transportation systems.

References

1. Department of Statistics Malaysia: Compendium of Environment Statistics: Malaysia. Percetakan Nasional Malaysia Bhd., Putrajaya (2006)
2. Shariff, N.M.: Private vehicle ownership and transportation planning in Malaysia. In: International Conference on Traffic and Transportation Engineering (ICTTE 2012), IPCSIT, vol. 26. IACSIT Press, Singapore (2012)
3. Yang, Y., Bagrodia, R.: Evaluation of VANET-based advanced intelligent transportation systems. In: Proceedings of the Sixth ACM International Workshop on Vehicular InterNET-working, pp. 3–12. ACM, New York, NY (2009)
4. Paolucci, M., Sacile, R.: Agent-Based Manufacturing and Control Systems. CRC Press, Florida (2004)
5. Mendel, J.M.: Type-2 fuzzy sets and systems: an overview. IEEE Comput. Intell. Mag. 2(1), 20–29 (2007)
6. Eichler, S., Schroth, C., Eberspächer, J.: Car-to-car communication. In: VDE Kongress—Innovations for Europe, Aachen (2006)
7. Abuelela, M.: A framework for incident detection and notification in vehicular ad-hoc networks. Citeseer (2011)
8. Eze, E.C., Zhang, S., Liu, E.: Vehicular ad hoc networks (VANETs): current state, challenges, potentials and way forward. In: 2014 20th International Conference on Automation and Computing (ICAC). IEEE (2014)

9. Saini, M., Singh, H.: VANET, its characteristics, attacks and routing techniques: a survey. *Int. J. Sci. Res. (IJSR)* (2015)
10. Dötzer, F.: Privacy issues in vehicular ad hoc networks. In: International Workshop on Privacy Enhancing Technologies. Springer (2005)
11. Ogah, C.P.A., et al.: Privacy-enhanced group communication for vehicular delay tolerant networks. In: 2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies. IEEE (2015)
12. Zeadally, S., et al.: Vehicular ad hoc networks (VANETS): status, results, and challenges. *Telecommun. Syst.* **50**(4), 217–241 (2012)
13. Sam, D.: A time synchronized hybrid VANET to improve road safety (2016)
14. Hadiwardoyo, S.A.: An overview of QoS enhancements for wireless vehicular networks. *Netw. Complex Syst.* **5**(1), 22–27 (2015)
15. Domínguez, F.J.M.: Improving vehicular ad hoc network protocols to support safety applications in realistic scenarios (2011)
16. Ahmed, E.F., et al.: Work in progress: LEACH-based energy efficient routing algorithm for large-scale wireless sensor networks. *J. Telecommun. Electron. Comput. Eng. (JTEC)* **10**(1–5), 83–87 (2018)
17. Aldabbas, H.: Securing data dissemination in vehicular ad hoc networks (2012)
18. Regan, K., Poupart, P., Cohen, R.: Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2006)
19. Sen, S.: Reciprocity: a foundational principle for promoting cooperative behavior among self-interested agents. In: Proceedings of the Second International Conference on Multi-Agent Systems, pp. 322–329 (1996)
20. Zhang, J., Cohen, R.: Trusting advice from other buyers in e-marketplaces the problem of unfair ratings. In: Proceedings of the Eighth International Conference on Electronic Commerce (2006)
21. Tran, T.: A reliability modelling based strategy to avoid infinite harm from dishonest sellers in electronic marketplaces. *J. Bus. Technol. (GBT)* **1**(1), 69–76 (2005)
22. Mukherjee, R., Banerjee, B., Sen, S.: Learning mutual trust. In: Trust in Cyber-Societies, pp. 145–158. Springer-Verlag (2001)
23. Jie, Y., et al.: Dynamic defense strategy against DoS attacks over vehicular ad hoc networks based on port hopping. *IEEE Access* **6**, 51374–51383 (2018)

Chapter 60

Lexicon-Based Argument Extraction from Citizen's Petition in Arabic Language



Sura Sabah Rasheed and Ahmed T. Sadiq

Abstract Argument extraction is the task of identifying arguments, along with their components in text. Arguments can be usually decomposed into a petition and one or more premises justifying it. In this paper, an approach to extract the arguments has been proposed. The proposed approach based on an Arabic lexicon included the main words which play an important role in arguments extraction. Text mining classical stages have been applied with the lexicon tool. The dataset has been collected from the Citizen Affairs Department in the service departments of the capital, Baghdad, Iraq, including more than 5000 petitions. The experimental results show that the proposed approach has a (91.5%) successful ratio in arguments extraction from the collected dataset.

60.1 Introduction

E-government is the use of information technologies that have the ability to be an effective mediator between persons, governmental, and non-governmental institutions. The aforementioned technologies can play an important role in equipping people with various government services, developing interaction with different sectors, facilitating people's access to the information they are allowed to have, and finally increasing the efficiency of government administration. It is possible to summarize the benefits resulting from implementing e-government by combating corruption, achieving transparency, reducing waste of time and money, achieving comfort for citizens, and increasing state revenues [1]. E-government has emerged to be the main factor in providing public services to citizens in a distinct, dynamic, effective, clear, and accountable manner. The main goal of e-government is to provide services to the community through the deep use of information technology and

S. S. Rasheed (✉) · A. T. Sadiq

Computer Science Department, University of Technology, Iraq, Baghdad, Iraq

A. T. Sadiq

e-mail: Ahmed.T.Sadiq@uotechnology.edu.iq

networks, taking into account the acquisition of community confidence, which ultimately leads to improving the interactive relationship between the government and citizens. E-government provides the citizens with many important benefits that have an urgent need, which is to reduce the time and cost to complete various government transactions in addition to their availability throughout the week and around the clock [2, 3]. The main node in the process of progressing in the type of services provided by e-government is through understanding the needs of the citizenry by governments, which has a great impact on reducing expenditures. Unfortunately, collecting this important information is very scarce [4].

In a normal life, argumentation between persons and groups are used in order to prove or disprove various arguments, through a torrent of facts and evidence. In recent decades, the number of people using social media has steadily increased, arguing and presenting various arguments to influence each other's opinions which make determining the arguments in these texts very important issue. The argument can be expressed simply and summarized as a petition (or premise) supported by the reasons associated with it [5, 6]. The process of arguments exploration with their main components within the text represents the arguments mining. Different algorithms specialized in mining for arguments attempt to identify the main elements of the argument which are petitions and premises in texts which are written in different disciplines and languages. In addition to the above, there is a lot of valuable information that can be extracted from different texts by using sentiment analysis techniques, which are a quick way to find out the direction of human feelings about a specific issue. It is not hidden to any one the importance to know the public opinion of people towards a specific issue in general and specifically in the field of policy-making and e-government [7].

Extracting information from a collection of texts is done by specific techniques which are working on textual data. The collection of these techniques is called text mining (TM), which is integrates with some procedures of natural language processing (NLP) that is specialized with text structures. NLP tools require large amounts of textual data, so text mining techniques impose severe restrictions on them. Actions taken by text mining on textual bases can be expressed as: a knowledge or information clustering and summarization, explore structures between different elements and groups, explore hidden links between textual components, general review for a large document [8, 9]. In order to find structures and links within the text, TM uses different procedures for this purpose. Some TM techniques require performing the classification process of the target text and then using the traditional methods of data mining (DM), while other techniques explore the entire target text [10, 11]. There is a collection of information extracted using TM techniques such as: set of words that operates as tags associated with each document, extract word sequences within a target document [12, 13].

In this paper, an approach to extract the arguments has been proposed, where it is based on an Arabic lexicon included the main words which play an important role in arguments extraction. The classical stages of text mining have been applied with the lexicon tool.

The rest of this paper is organized as follows. In Sect. 60.2 related works are discussed. In Sect. 60.3, a brief explanation of argument extraction and its principles is presented. In Sect. 60.4, the proposed system included the design and implementation is introduced. Experimental setup and results are stated in Sect. 60.5. Finally, we conclude with a summary in Sect. 60.6.

60.2 Related Works

Jasim et al. proposed the concept of argument extraction and differentiation depending on the supervised learning, where their prototype is used with Arabic texts that contain legal information which is one of the firsts in this field. The arithmetical model used in this work is collect and collaborates among the basic bulk of Arabic legal text [14].

Kadhim et al. proposed the concept of information extraction and analysis depending on the unsupervised learning, where their prototype is used with Arabic texts that contain legal information. The type of extracted information from the legal Arabic text is divided into two types; the first one is called the valuable qualities while the second one is called the worthy information [15].

Palau et al. proposed the concept of arguments finding process in the legal text as a categorization process. The model used in this work relies on analyzing the relationships between sentences as a basis for more accurate exploration of arguments, which leads to increasing the general accuracy for about 8% compared with another models [16].

Moens et al. proposed the concept of arguments extraction from legal texts as a classification process, which is practiced on a set of explained arguments. A collection of properties related to legal texts are evaluated which are: syntax, semantics, linguistic vocabulary, and conversation [17].

Poudyal et al. proposed the concept of arguments extraction from legal texts depending on new clustering algorithm. The main problem with this model is that the argument for a sentence may be related to and part of another argument. To solve this problem, fuzzy logic are used which is represented by fuzzy c-means (FCM) clustering algorithm [18].

60.3 Argument Extraction

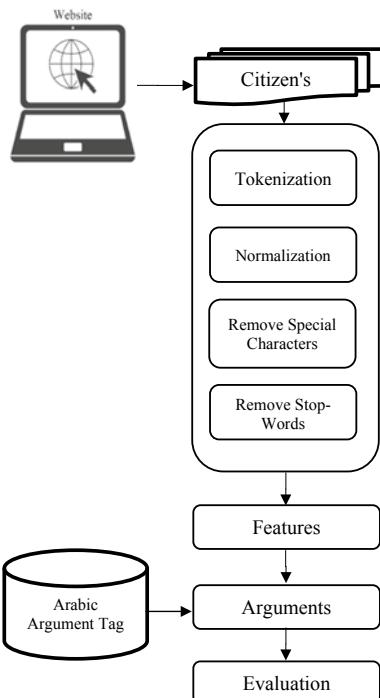
The personal conviction is an important and spiky phenomenon, where it is built from a collection of human's conclusion factors such as evaluation, perspectives or their opinions against some matters. Worse than that human's conclusion not always possible to conclude that it is true or false. Anyway, people tend to take other people's conclusions if they are supported by strong arguments. From the foregoing, we can touch the reason for the interest in extracting arguments from the different text types

[19]. The basic components of argument are construction and conclusion, where the argument extractions are dealing with exploring, specifying, and mining the so called components from the text of natural languages. Argument extraction is an advanced stage of a machine's understanding for natural human languages [20].

The issue of arguments can be touched in people's daily lives, and they apply a logical approach to reach conclusions and their causes [6]. We can observe the widespread use of arguments in many fields such as on economics, law, politics, sports, and culture. In the politics domain for example, politicians, their advisors, and their assistants are busy spending a lot of time searching for and analyzing different types of information in the hope of arriving at a suitable set of arguments that support their petitions or refute the petitions of others in specific cases. Therefore, the argument extraction proposed to facilitate these thorny issues through reaching to the basic ingredients of arguments (i.e., premises and petitions) and their relations. Furthermore, politicians can be supported in accepting a specific policy by extracting the arguments expressed in public statements to support or refute some issues that are compatible or inconsistent with the policy [6, 21, 22]. In another context, and specifically in the field of law, specialists (i.e., judges and lawyers) spend a lot of time and effort in searching and analyzing information related to the defendants' cases, in order to help them extract convincing arguments that support their petitions or refute the petitions of others. Whereas, the argument extraction are used to simplify these complex issues. Furthermore, professionals in the field of law can be supported in accepting or rejecting the collection of arguments [14].

60.4 Proposed System

The proposed system included the design and implementation of a website to receive citizens' petitions and process those petitions automatically to extract the arguments from each application. The process of arguments extraction from the application is an important factor and represents credibility and evidence of the citizen's need to implement his petition. An arguments extraction from the texts of citizens' petitions is an important feature for decision-makers and for adding strength to e-governance. An arguments extraction process from the texts of citizens' petitions is not an easy process, but it became possible depending on several factors such as; the support of text mining methods, some capabilities of the lexicon style and finally implementing a set of logical steps. Figure 60.1 represents the block diagram of the proposed system that extracts arguments from the texts of citizens' petitions. The first stage includes the preprocessing, which contains several steps such as tokenization, normalization, remove punctuations and finally remove stop words. The second stage is the features extraction, which involves extracting the important characteristics that play an important role in the argument's extraction stage. The third stage is the arguments extraction which is the most important one, because it represents the outcome of the previous stages, and based on what will be obtained it is decided the success or failure of the entire process. Finally, the system evaluation stage, by which the whole

Fig. 60.1 Proposed system

process is evaluated. In the following, we will deal with the aforementioned stages in some detail.

60.4.1 Preprocessing Phase

This phase is an essential preparatory procedure which performed in the introductory of the whole algorithm and consists of several steps. The first step is the tokenization that will split the input text into smaller parts, such as words. While, the second step includes removing punctuation. This step is very important because it will remove all types of punctuation, such as (:" !? ; _ - ?!{}/*//...[],) from the tokenized text. In

Table 60.1 List of regular expressions

No.	Regular expression	Results
1	[a-zA-Z]+	Remove English characters
2	[0-9]+	Remove English numbers
3	[0-9]+	Remove Arabic numbers
4	#\$% @^~(&*)+	Remove another



Fig. 60.2 The diacritics have to eliminate

the third step, the unwanted words, which are shown in Table 60.1, will be removed by using a list of regular expressions.

Finally, is the forth step which involves the normalization of Arabic words, where it is the most important step in preprocessing phase because of an Arabic language has many shapes for writing. To perform normalization, the proposed method applies three operations that will remove Diacritics, Tatwheel, and Latter. Beginning with the removal of diacritics from an Arabic word, and continuing in the conversion of an alphabetic word into another. An example of this process is displayed in Fig. 60.2. If these diacritics are not removed, then we will have many shapes for the same words, and hence, the vector becomes extremely large and more articles will be required for building, which will take a longer time.

Tatwheel in Arabic language is used to make a character looks longer than others. The main reason for using Tatwheel is to make the shape of word pretty. However, this represents a problem because the same words can be written in many ways. The final step in normalization will make a unique letter corresponds some letters. When all the above operations are completed, remove any word that consists of only two characters, as in Arabic language any word that consists of two characters does not make any sense. Finally, the Arabic stop word can be filtered from an article by removing any token that matches the word in the stop words list. The stop words list is a built-in NLTK library. In NLTK all words are sign words.

60.4.2 Features Extraction Phase

After removing stop words and the other parts that were mentioned in the preprocessing phase, what remains are pure and clear words such as (verbs, adjectives, adverbs, nouns). At this stage, the words with the argument's properties will be extracted. A special dictionary of words was used, extracted from the expert opinion in the Arabic language, containing the most important words dealing with arguments. Also, some Iraqi dialect words have been added that relate to arguments sources. The lexicon of arguments plays an important role in defining the beginning of causes related to arguments. An arguments lexicon plays an important role in determine the beginnings of the sentences which contain an argument. By knowing the words related to the arguments, the sentence regarding the arguments can be predicted in the texts of citizens' petitions. From the above, we can touch the importance and majority role that this stage plays in extracting the arguments from the texts of citizens' petitions.

60.4.3 Arguments Extraction Phase

In this phase, words related to arguments are identified and extracted, at this stage, the argument sentence in the text is determined and extracted which represents the citizens' petition. By depending on the help of some natural language processing principles, it is possible to know the end of the sentence that represents the argument and whose beginning was determined from the properties and which was represented in the previous stage. By defining the words that are and extracting the words related to the arguments, at this stage, the sentence that is the argument in the text is determined and extracted. With the help of some natural language processing principles, it is possible to know the end of the sentence that represents the argument and whose beginning was determined from the properties and which was represented in the previous stage.

A group of words that denote the argument have been used in the Arabic language. Table 60.2 shows examples of words related to the argument in the Arabic language. These are the words that most often appear in the argument we want to extract. The words in Table 60.2 were used as keywords to extract the argument from citizens' petitions. For this reason, the basis of the proposed work is lexicon-based, based on these words that formed the basis of the lexicon.

The words in Table 60.3 were approved based on the petitions collected from the sources concerned with the topic, as well as the assistance of Arabic language experts and those concerned with Arabic language affairs, especially in the field of

Table 60.2 Samples of Arabic argument tag words (translated to English)

That leads	That causes	To be	Considered as	Because
Whereas	To our need	For being	Then	It said

Table 60.3 Types and numbers of citizen's petitions

No.	Department	No. of petitions
1	Water Affairs	1050
2	Health Care	260
3	Sewage Treatment	1075
4	Municipality	500
5	Energy Affairs	1250
6	Education Affairs	300
7	Agricultural Affairs	125
8	Labor and Social Affairs	135
9	Youth and Sport Affairs	150
10	Compensation	75
11	Others	230
Total		5150

argument. New words can also be added to the proposed lexicon if citizens use words that are not found in the dictionary, especially if the dialects used in the petitions differ.

60.5 Experimental Results

Our dataset has been collected from the Department of Citizens' Petition in Baghdad Governorate in Iraq. More than 5000 Petitions/Petitions have been collected about several areas (Water Affairs, Health Care, Sewage Treatment, Municipality, Energy Affairs, Education Affairs, Agricultural Affairs, Labor and Social Affairs, Youth and Sport Affairs, Compensation and Others), Table 60.3 includes the total number of petitions for each department.

The real problem that we faced is that there are a large number of citizens' petitions that do not contain arguments, meaning that when the citizen wrote the petition, he neglected the arguments for several reasons, including:

- That the citizen does not have an argument for his petition.
- That the citizen does not know the argument formulation of his petition.
- That the citizen inadvertently neglected writing the argument for his petition.
- Other reasons.

The above points represent the most important reasons for citizens' neglect in writing the arguments in their applications to the government. Where 64% of citizens' applications collected from Baghdad governorate do not contain arguments, meaning that only 36% of applications contain arguments. Of the 5150 applications that were collected as shown in Table 60.4, there were only 1854 applications containing arguments and 3296 applications that did not contain arguments, the proportion of applications that did not contain large arguments for the above reasons. Of the 1854 applications containing arguments, the percentage of valid arguments was (91.5%), meaning (1698) applications. Table 60.4 shows the numbers and types of applications that contain arguments and the number that the system extracted from those petitions.

As we note, the number of applications containing arguments is relatively small compared to the number of applications. Table 60.4 shows the number of applications from which the arguments were extracted and their total was 1698 out of 1854, with a success rate of 91.5%.

60.6 Conclusions

The proposed system extracts arguments from citizens' petitions submitted to service foundations and others. The basis for the proposed system is the lexicon, which contains the basic words that play an important role in extracting arguments from the texts. The system was applied on real data taken from the Citizen Affairs Department

Table 60.4 Argument extraction from citizen's petitions

No.	Department	No. of petitions	No. of petitions that including argument	No. of automatic argument extraction
1	Water Affairs	1050	386	358
2	Health Care	260	89	76
3	Sewage Treatment	1075	393	375
4	Municipality	500	188	172
5	Energy Affairs	1250	458	447
6	Education Affairs	300	88	75
7	Agricultural Affairs	125	54	46
8	Labor and Social Affairs	135	63	52
9	Youth and Sport Affairs	150	58	41
10	Compensation	75	23	18
11	Others	230	54	38
Total		5150	1854	1698

in the service departments of the capital, Baghdad, Iraq. The system achieved a success rate of 91.5%. The most important reason influencing the data collected is that the citizen does not write the arguments on which his application is based, and as we mentioned for that there are several reasons. The dictionary that was adopted for the words of the arguments played a big role in extracting those arguments correctly. There is no doubt that the issue of extracting arguments from the texts is not an easy task with the different dialects in the texts of Iraqi citizens. As a future work, we suggest using deep learning to explore words that are more profound and effective in extracting arguments from Arabic texts.

References

1. Katsonis, M., Botros, A.: Digital government: a primer and professional perspectives. *Aust. J. Public Adm.* **74**(1), 42–52 (2015)
2. Samsor, A.M.: Challenges and Prospects of e-Government Implementation in Afghanistan. *International Trade, Politics and Development* (2020)
3. Astawa, I.P.M., Dewi, K.C.: E-government facilities analysis for public services in higher education. *J. Phys. Conf. Ser.* **953**(1), 012061 (2018)
4. Bertot, J.C., Jaeger, P.T., McClure, C.R.: Citizen-centered e-government services: benefits, costs, and research needs. In: DG.O, pp. 137–142 (2008)
5. Poudyal, P.: Automatic extraction and structure of arguments in legal documents. In: Gaggl, S.A., Thimm, M. (eds.), p. 19 (2016)
6. Florou, E., Konstantopoulos, S., Koukourikos, A., Karampiperis, P.: Argument extraction for supporting public policy formulation. In: Proceedings of the 7th Workshop on Language

- Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 49–54 (2013)
- 7. Sardianos, C., Katakis, I.M., Petasis, G., Karkaletsis, V.: Argument extraction from news. In: Proceedings of the 2nd Workshop on Argumentation Mining, pp. 56–66 (2015)
 - 8. Abutridy, J.A.A.: Text mining: principles and applications. *Rev. Fac. Ingen.* **7**, 57–62 (2000)
 - 9. Tao, D., Yang, P., Feng, H.: Utilization of text mining as a big data analysis tool for food science and nutrition. *Compr. Rev. Food Sci. Food Saf.* **19**(2), 875–894 (2020)
 - 10. Dang, S., Ahmad, P.H.: Text mining: techniques and its application. *Int. J. Eng. Technol. Innov.* **1**(4), 866–2348 (2014)
 - 11. Kadhim, N.J., Saleh, H.H., Attea, B.: Improving extractive multi-document text summarization through multi-objective optimization. *Iraqi J. Sci.* **59**(4B) (2018)
 - 12. Alatabi, H.A., Abbas, A.R.: Sentiment analysis in social media using machine learning techniques. *Iraqi J. Sci.* **61**(1) (2020)
 - 13. Mooney, R.J., Nahm, U.Y.: Text mining with information extraction. In: Daelemans, W., du Plessis, T., Snyman, C., Teck, L. (eds.) *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, Bloemfontein, Sept 2003*, pp. 141–160. Van Schaik Pub., South Africa (2005)
 - 14. Jasim, K., Sadiq, A.T., Abdullah, H.S.: A framework for detection and identification the components of arguments in Arabic legal texts. In: 2019 First International Conference of Computer and Applied Sciences (CAS), pp. 67–72. IEEE (2019)
 - 15. Jasim, K., Sadiq, A.T., Abdullah, H.S.: Unsupervised-based information extraction from unstructured Arabic legal documents. *Opcion* **35**, Especial No. 20 (2019)
 - 16. Mochales-Palau, R., Moens, M.: Study on sentence relations in the automatic detection of argumentation in legal cases. *Front. Artif. Intell. Appl.* **165**, 89 (2007)
 - 17. Moens, M.F., Boiy, E., Palau, R.M., Reed, C.: Automatic detection of arguments in legal texts. In: Proceedings of the 11th International Conference on Artificial Intelligence and Law, pp. 225–230 (2007)
 - 18. Poudyal, P., Gonçalves, T., Quaresma, P.: Using clustering techniques to identify arguments in legal documents. In: Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019), Montreal, QC, 21 June 2019
 - 19. Cabrio, E., Villata, S.: Five years of argument mining: a data-driven analysis. *IJCAI* **18**, 5427–5433 (2018)
 - 20. Moens, M.F.: Argumentation mining: how can a machine acquire common sense and world knowledge. *Argument Comput.* **9**(1), 1–14 (2018)
 - 21. Obaid, A.J., Sharma, S.: Data-mining based novel neural-networks-hierarchical attention structures for obtaining an optimal efficiency. In: Favorskaya, M.N., Peng, S.L., Simic, M., Alhadidi, B., Pal, S. (eds.) *Intelligent Computing Paradigm and Cutting-Edge Technologies. ICICCT 2020. Learning and Analytics in Intelligent Systems*, vol. 21. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-65407-8_36
 - 22. Arbabi, Z., et al.: *J. Phys. Conf. Ser.* **1530**, 012111 (2020)

Chapter 61

Multi-key Encryption Based on RSA and Block Segmentation



Rana JumaaSarih Al-Janabi and Ali Najam Mahawash Al-Jubouri

Abstract Security is an essential part of modern-day communications; this is mainly achieved using cryptography techniques, and one of its most important types is asymmetric encryption which is used to provide both confidentiality and authentication. One of those techniques is the RSA public key algorithm. In this paper, we propose an encryption structure that uses some characteristics of the block cipher techniques and the random ordering of blocks mixed with the traditional RSA algorithm to overcome the problem of increasing security at the expense of the time that is present when using large keys in RSA algorithm; small RSA keys are prone to factorization attacks, while large keys are very slow and inefficient. By using a fixed encryption block size, segmenting and encrypting each segment with a randomly chosen RSA key, we can increase complexity and allow the use of larger block sized without sacrificing speed.

61.1 Introduction

Encryption is a collection of tools and techniques used to ensure the confidential exchange of information over the network. Providing confidentiality and security of information, translating the data shape and type into an entirely different form the original keeping the data safe from unauthorized access [1]. Original data are known as plaintext where the newly transformed data are the ciphertext and the transforming procedure known as encryption.

Encryption is divided into two broad classes based on the feature [2]. Symmetric cryptography is one, and asymmetric cryptography is the other. In symmetric encryption (Shared key system), one key is used in both the encryption and decryption processes [1, 3] (such as DES, AES). In contrast, asymmetric encryption (Public key system), which is a revolution in cryptography [4], applies two different approaches to the keys used, such as a public key to encrypt the original text and a private key to

R. J. Al-Janabi (✉) · A. N. M. Al-Jubouri

Department of Computer Science, University of Al-Qadisiyah, Qadisiyah, Iraq

e-mail: rana.aljanaby@qu.edu.iq

decrypt the encrypted text (such as RSA and Al-Gamal). Some asymmetric encryption algorithms can be used in a digital signature process where the original data are encrypted with the private key. Anyone with the public key can verify the authenticity of the messages and to whom it belongs. This concept can also be used to ensure data confidentiality and correctness in communications [5].

The best known and most commonly used public key scheme at present is RSA. Three cryptologists invented it in 1978 by Ronald Rivest, Adi Shamir, and Leonard Adleman of MIT; RSA is used for privacy, for the authentication of digital records, for payment systems for electronic credit and debit cards, and business systems such as Web servers and browsers Web traffic safe [6]. It consists of three stages: primary key generation, encryption, and decryption stage; there are two different keys to the RSA cryptosystem, which are public and private keys. It is possible to disclose the public key, which is used for encrypting a plaintext. Still, the hidden key is used to decrypt the ciphertext [7]. RSA is a public key cipher system that uses number theory. Hence, its security depends on the difficulty of the analysis for large prime numbers, which is a mathematically known problem for which there is no solution. The private key is linked mathematically to the public key in public encryption. Data encryption and decryption are usually complicated problems with mathematical equations that need enormous computer resources to analyze and break. To break the RSA encryption, an attacker needs to apply computational number theory and face the practical difficulty of factoring large integers. In the RSA factoring challenge, they published a list of semiprimes (numbers with exactly two prime factors) When the challenge ended in 2007, only, RSA-576 and RSA-640 had been successfully factored [8]. This challenge verifies that the strength of RSA depends on the size of the keys used, So, when the key's value is big enough that it becomes much harder to know the common factors of the main number, RSA comes to be more protected [7].

The encryption can be strengthened by adding multiple keys or adding more structural complexity, rather than using bigger and slower keys. We propose using a combination of both, creating a new structure for encryption and decryption to encrypt large blocks employing multiple smaller segments each encrypted with a unique RSA key pair and using them in a random order, the order of which the keys are used is controlled by a PRNG to ensure the same order is used in both encryption and decryption.

61.2 Problem Statement

To discuss the solution, we must first explore the problem and the work done to try to solve it.

61.2.1 Overview

In this paper, we study the effect of using large numbers in RSA algorithm to increase security but at the expense of time and complexity. Although using larger than 256 bit keys in RSA can increase its security, but it also increases the time needed to encrypt and decrypt the data. Also, we propose a model using multiple keys to allow for stronger encryption without increasing the time needed to apply the transformation.

61.2.2 Related Works Discussion

Many scientists have put forward some ideas to improve the RSA algorithm. The algorithms used in each case are discussed along with a few recent and major modifications suggested.

Ayele and Sreenivasarao in 2013 proposed using two public key inputs; the algorithm is close to the standard RSA algorithm. It uses a pair of public keys and some mathematical relationships instead of sending a value directly. These two public keys are sent separately [9]; these additions make the algorithm harder to crack since an attacker needs to factorize two numbers instead of one, but the time needed for transformation is doubled.

Patidar and Bhartiya published a paper in 2013 introducing a new algorithm principle that uses three prime numbers; they also propose using offline storage of key parameters. RSA key pairs are stored in a database that is identical in all networks. All parameters which are used in the RSA algorithm are stored before starting the algorithm. This is used for speeding the application of the RSA algorithm during data exchange over the network [10]. Using three primes increase the complexity and the time required to do the transformations, but they combat this by relying on the use of predefined offline datasets to speed the process of transformation. This implementation can only be used in a closed network system and cannot be used in public systems on the Internet without the predefinition of such offline datasets which is not possible, limiting the use case of this implementation.

Minni et al. worked on a paper in 2013 eliminating the distribution of n which is the large number whose factor if found compromises the RSA algorithm [11]. This implementation still suffers from the drawback of time increase with key size increase.

61.3 Proposal

The new structure splits the block to smaller segments then encrypts each of them using the traditional RSA steps explained below.

61.3.1 Standard RSA Steps

RSA uses two keys one for encryption known as the public key and the other for decryption known as the private key. We can summarize the steps to use the algorithm as follow:

1. Two prime numbers are chosen and kept secret p, q .
2. Compute $n = pq$. n is used as the modulus for both the public and private keys. Its length in bits is the key length.
3. Compute $\lambda(n) = \text{lcm}(p - 1, q - 1)$. This is kept secret.
4. Choose an integer e such that $1 < e < \lambda(n)$ and $\gcd(e, \lambda(n)) = 1$; (e and $\lambda(n)$ are coprime).
5. Calculate d as the multiplicative inverse of e modulo $\lambda(n)$. $d \equiv e^{-1} \pmod{\lambda(n)}$.
6. (e, n) pair is released as the public key, and the rest are kept secret.
7. $(d, n, \lambda(n))$ are used as a private key.
8. Encryption is handled with $m^e \equiv c \pmod{n}$.
9. Where decryption is $c^d \equiv (m^e)^d \equiv m \pmod{n}$.

Refer to the algorithm's original search address for more detailed information on the basic RSA algorithm, including how it operates from all angles, the way to construct it, and deal with data [12–14].

61.3.2 Proposed Structure

After closely reviewing previous studies, it is clear that there are issues with complexity and time. As complexity grows, time increases. The proposal discussed in this paper does not alter the basic structure of the RSA algorithm, but we propose mixing the structure of RSA with block cipher characteristics. Where encryption and decryption block is set to a fixed length with 2048 bits, then the 2048 bit block is segmented; the segment can be of bit sizes 64, 128, 256, 512, etc. then generate the same number of RSA key pairs to the segment numbers. Each segment is encrypted with a randomly selected key; to control the random generation and preserve the order on decryption, we use a random generator with seed. The seed is used later in decryption to reproduce the same random sequence used in encryption.

The proposed structure adds to the complexity of the algorithm, and improves the speed, using a large RSA key can yield stronger encryption, but at the expense of time, using the proposed structure solves the issue, we can increase the complexity and strength of encryption by implementing the randomness of choosing different RSA keys for each segment; the attacker will have a harder time breaking the encryption; even if the attacker can use mathematics to factor the primes used in the keys, he cannot order in which they are used without obtaining the seed key. This improved complexity allows us to use smaller segment sizes with the RSA key pairs to allow for faster encryption times. See Fig. 61.1. This can be explained with steps.

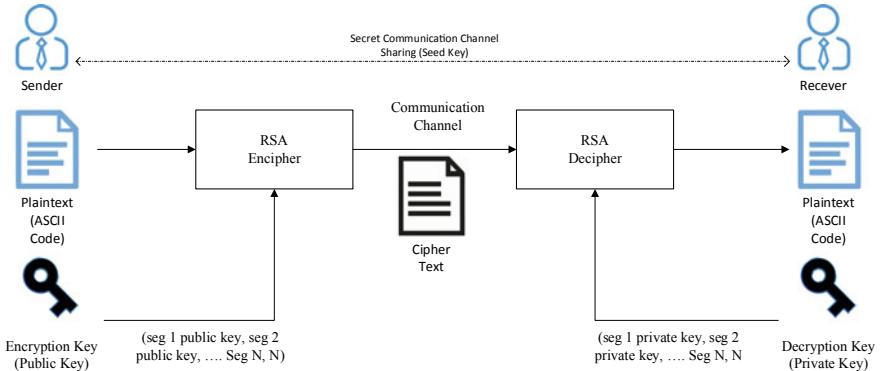


Fig. 61.1 Showing the communication between two sides using the proposed structure

Public and private key pairs are generated using the following algorithm:

Algorithm 1: Proposed Key Generation

Input: Segment size
Output: Public key and private key

- 1 START
- 2 Set block size to be 2048
- 3 Calculate number of segments $N = \text{block size}/\text{segment size}$
- 4 Create empty list of size N to hold public keys, PUB_KEYS[N]
- 5 Create empty list of size N to hold private keys, PRI_KEYS[N]
- 6 For i in range $(1, N)$ Do
 - 7 Public key, private key = Generate new RSA key pair of size (Segment Size)
 - 8 PUB_KEYS[i] = PublicKey
 - 9 PRI_KEYS[i] = PrivateKey
- 10 Next i
- 11 Return public key as (PUB_KEYS[i], segment size, N)
- 13 Return private key as (PRI_KEYS[i], segment size, N)
- 14 Generate random seed key (Shared with other party secretly)
- 15 END

Encryption process is illustrated in Algorithm 2; the text is cut to block; each block is processed separately; a single block is segmented, and each segment is encrypted with different key for more complexity.

Algorithm 2: Encryption

Input: PublicKey, Bits To Encrypt, Seed Key
Output: Encrypted Bits

- 1 START
- 2 Declare **Encrypted Bits** = empty
- 3 From **PublicKey** extract **List of Public RSA Keys** and number of Segments N
- 4 **Generator** = Initialize sequence generator with (**Seed Key**)
- 5 While **Bits To Encrypt** not empty
 - 6 **Block** = cut a block of 2048 bits from **Bits To Encrypt**
 - 7 Add padding if needed
 - 8 **Encrypted Block** = Initialize empty block 2048 bits
 - 9 **List of Segments** = Divide **Block** to the desired number of segments (N)
 - 10 For each segment in the **List of Segments**
 - 11 **Sequence** = **Generator**, Get Number (0, N – 1)
 - 12 Encrypt segment with RSA using **List of Public RSA Keys** [**Sequence**]
 - 13 Append to result to **Encrypted Block**
 - 14 Next segment
 - 15 Append **Encrypted Block** to **Encrypted Bits**
- 16 Go to step 5
- 17 Output **Encrypted Bits**
- 18 END

Decryption process is similar to encryption, the difference being we use private keys instead of public keys. This can be illustrated in Algorithm 3.

Algorithm 3: Decryption

Input: PrivateKey, Bits Top Decrypt, Seed Key
Output: Decrypted Bits

- 1 START
- 2 Declare **Decrypted Bits** = empty
- 3 From **PrivateKey** extract **List of Private RSA Keys** and number of Segments N
- 4 **Generator** = Initialize sequence generator with (**Seed Key**)
- 5 While **Bits To Decrypt** not empty
 - 6 **Block** = cut a block of 2048 bits from **Bits To Decrypt**
 - 7 **Decrypted Block** = Initialize empty block 2048 bits
 - 8 **List of Segments** = Divide **Block** to the desired number of segments (N)
 - 9 For each segment in the **List of Segments**
 - 10 **Sequence** = **Generator**, Get Number (0, N – 1)

(continued)

(continued)

- 11 Decrypt segment with RSA using **List of Private RSA Keys [Sequence]**
- 12 Append to result to **Decrypted Block**
- 13 Next segment
- 14 Remove padding if found
- 15 Append **Decrypted Block** to **Decrypted Bits**
- 16 Go to step 5
- 17 Output **Decrypted Bits**
- 18 END

61.4 Results and Discussion

The testing is done on a system with the following specifications: CPU: 4700HQ, RAM 16 GB DDR3, OS Windows 10 Home v2004. The application is coded and tested in C#.Net. All the testing is run on a single thread without using the parallel process to speed the encryption or decryption.

We tested the traditional RSA algorithm as a reference with key sizes (64, 128, 256, 512, 1024, 2048) and tested the proposed structure with two segmentation each is 1024 bit, 4 segmentations each is 512 bit, 8 segmentations each is 256 bit, and finally 16 segmentations each is 128 bit. Speed tests are performed on a file of size 10,174,700 bytes (Approximately 10 MB).

Speed results can be seen in Table 61.1.

Looking at the results of Table 61.1, we can see that using the proposed algorithm in all of its possible segmentations are faster than the corresponding original algorithm; the proposed algorithm uses a fixed 2048 bit block size regardless of the underlying segmentation used; all the segmentation options are faster than the traditional RSA with a key size of 2048 bits.

61.5 Conclusion

In this paper, we propose that the original algorithm RSA be improved by using several public keys and private keys for encryption and decryption processes to make it more complex without affecting encryption and decryption speed. The proposed structure uses a fixed block size of 2048 bit and encrypts each of its segment with a randomly chosen RSA key to increase the security rather than using slow and large traditional RSA keys. This allows the use of larger block sized for encryption and decryption without the need for larger keys. Making the use of public key encryption for purposes other than key exchange is possible.

Table 61.1 Encryption and decryption performance results

Algorithm	Encryption speed	Decryption speed	Encryption time	Decryption time
RSA 64	3564.7	713.4	3262	14,261
RSA 128	3462.9	333.7	3134	30,487
RSA 256	2635.6	119.1	3985	85,362
RSA 512	1630.0	36.7	6341	277,089
RSA 1024	951.1	10.5	10,781	964,368
RSA 2048	515.6	2.8	19,809	3,583,957
Proposed (2 segments)	894.7	10.3	11,461	984,037
Proposed (4 segments)	1526.9	35.0	6769	290,050
Proposed (8 segments)	2575.5	121.1	4078	83,955
Proposed (16 segments)	3434.5	316.6	3160	32,134

References

1. Barakat, M., Eder, C., Hanke, T.: An introduction to cryptography. Timo Hanke RWTH Aachen Univ. 1–145 (2018)
2. Kahate, A.: Cryptography and Network Security. Tata McGraw-Hill Education (2013)
3. Simmons, G.J.: Symmetric and asymmetric encryption. ACM Comput. Surv. **11**, 305–330 (1979)
4. Kumar, Y., Munjal, R., Sharma, H.: Comparison of symmetric and asymmetric cryptography with existing vulnerabilities and countermeasures. Int. J. Comput. Sci. Manag. Stud. **11**, 60–63 (2011)
5. Jamgekar, R.S., Joshi, G.S.: File encryption and decryption using secure RSA. Int. J. Emerg. Sci. Eng. **1**, 11–14 (2013)
6. Boneh, D.: Twenty years of attacks on the RSA cryptosystem. Not. Am. Math. Soc. **46**, 203–213 (1999)
7. Obaid, T.S.: Study a: public key in RSA algorithm. Eur. J. Eng. Technol. Res. **5**, 395–398 (2020)
8. RSA Factoring Challenge—Wikipedia: https://en.wikipedia.org/wiki/RSA_Factoring_Challenge. Accessed 28 Mar 2021
9. Ayele, A.A., Sreenivasarao, V.: A modified RSA encryption technique based on multiple public keys. Int. J. Innov. Res. Comput. Commun. Eng. **1**, 859–864 (2013)
10. Patidar, R., Bhartiya, R.: Modified RSA cryptosystem based on offline storage and prime number. In: 2013 IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–6. IEEE (2013)
11. Minni, R., Sultania, K., Mishra, S., Vincent, D.R.: An algorithm to enhance security in RSA. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–4. IEEE (2013)
12. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM. **21**, 120–126 (1978)

13. Lavanya, K., Obaid, A.J., Sumaiya Thaseen, I., Abhishek, K., Saboo, K., Paturkar, R.: Terrain mapping of LandSat8 images using MNF and classifying soil properties using ensemble modelling. *Int. J. Nonlinear Anal. Appl.* **11**(Special Issue), 527–541 (2020). <https://doi.org/10.22075/ijnaa.2020.4750>
14. Sharma, S., Obaid, A.J.: Mathematical modelling, analysis and design of fuzzy logic controller for the control of ventilation systems using MATLAB fuzzy logic toolbox. *J. Interdiscip. Math.* **23**(4), 843–849 (2020). <https://doi.org/10.1080/09720502.2020.1727611>

Chapter 62

Chaotic Pseudo Random Number Generator (cPRNG) Using One-Dimensional Logistic Map



Ayan Mukherjee, Pradeep Kumar Mallick, and Debahuti Mishra

Abstract In this paper, a method of generating non-periodic pseudo random numbers has been proposed based on a chaotic PRNG using one-dimensional logistic map. PRNGs generate deterministic sequences of numbers that appear random; such a sequence is reproducible given the state of the generator is known. The proposed cPRNG is based on the one-dimensional logistic map; by adjusting the value of the control parameter lambda, a sequence of random numbers is generated. The sequence so generated is then put through a test of randomness. In this paper, the Wald–Wolfowitz runs test has been used as a test of randomness.

62.1 Introduction

A random number sequence can be defined as a sequence of numbers or symbols which cannot be predicted reasonably well than by a random chance. The use of random numbers can be seen virtually in every field; random numbers are used in computer simulation, finance, cryptography, gambling, data modeling, statistics, and gaming to name a few. There exists two classes of random number generators, hardware random number generators (HRNGs) and pseudo random number generators (PRNGs) [1]. Hardware random number generators rely on extra piece of hardware that measures a natural phenomenon which is seemingly random to generate random numbers. These generators lack speed and rely on extra hardware. The results produced by such generators cannot be easily reproduced. Pseudo random number generators (PRNGs) on the other hand produce random numbers that

A. Mukherjee · P. K. Mallick (✉)

Department of Computer Science and Engineering, Kalinga Institute of Industrial Technology
(Deemed to be) University, Bhubaneswar, Odisha, India
e-mail: pradeep.mallickfcs@kiit.ac.in

D. Mishra

Department of Computer Science and Engineering, Siksha ‘O’ Anusadhan (Deemed to be)
University, Bhubaneswar, Odisha, India
e-mail: debahutimishra@soa.ac.in

appear to be random but given the state of the generator are deterministic and reproducible. PRNGs are considerably faster in comparison to HRNGs, and the results of a PRNG can be easily reproduced [2]. There exists various algorithms that generate sequences of pseudo random numbers. The linear congruential generator (LCG) which generates the sequence using a discontinuous piecewise linear equation is the most well-known PRNG. It is also one of the oldest PRN algorithms.

Chaos theory on the other hand is the study of chaotic system; these are dynamic systems which exhibit a hyper-sensitive dependence to initial conditions. Chaotic systems are deterministic, yet predicting their long-term behavior is virtually impossible [3]. Chaotic systems are also non-periodic. These properties make them a suitable candidate for PRN generation. A chaotic system generates a sequence of non-periodic random numbers which can then be transformed to obtain a sequence of random numbers as per the user's need. The logistic map is one of the simplest examples of a chaotic system. The map was popularized in a 1976 paper by the biologist Robert May, [4] in part as a discrete-time demographic model analogous to the logistic equation written down by Pierre François Verhulst [5]. The logistic map is a polynomial mapping of degree 2 that exhibits chaotic behavior.

The rest of the paper is organized as follows; the Section 62.2 discusses the literature survey, the details of chaotic systems and logistic map are discussed in Section 62.3, the algorithm and implementation along with results obtained from this study has been given in Section 62.4, and finally, Section 62.5 concludes the paper.

62.2 Literature Review

Several researchers have employed various techniques to design random number generators. There has been intensive work carried out on random number generation since the 1970s. In the design and analyze RNG content and in attempt of finding unknown parameters, researchers have undertaken different systems, processes, and phenomena. An automated methodology of producing HRNs was presented by Ray et al. [6], for arbitrary distributions using the inverse cumulative distribution function (ICDF). The ICDF is evaluated via piecewise polynomial approximation with a hierarchical segmentation scheme that involves uniform segments and segments with size varying by powers of two which can adapt to local function nonlinearities [7].

Gu and Zhang [8] proposed a uniform random number generator using leap ahead LFSR architecture. This work introduced novel URNG using leap ahead LFSR architecture that could produce m-bits RN/cycle using only one LFSR. A normal LFSR can produce only one random bit per cycle. In most applications, multi-bits are required to form a random number, multi-LFSRs architecture is used to implement a URNG [7]. Jonathan et al. [9] proposed a new architecture using cellular automata. CA-based PRNGs are well suited for implementation on field programmable gate arrays (FPGAs) [7]. Dabal and Pelka [10] presented an FPGA implementation of chaotic pseudo random bit generators. To ensure protection against unauthorized access,

modern communication systems require the use of advanced methods of information protection; key generation is a fundamental problem of cryptography. Chaotic digital systems are gaining the attention of cryptographers in the recent years. Chaotic signals can be used to carry information [7].

62.3 Chaotic Systems and Logistic Map

Chaotic systems are non-linear deterministic dynamic systems that exhibit a high sensitivity to initial conditions [3]. Minute variations in initial conditions that might occur as a result of errors in measurements or due approximation error in computation can produce highly diverging results for such systems, thus making it impossible to predict its long-term behavior in general, even though such systems are of deterministic nature, which means that the anticipated outcomes exhibit a distinct evolution that can be completely determined by its initial conditions, without any element of randomness involved. In other words, chaotic systems are unpredictable despite being deterministic in nature [3]. For the evolving variable, there may exist sequences of values that repeat itself in a chaotic system; this induces periodicity and imposes a periodic behavior from that point on in the sequence; such sequences, however, are repelling instead of being attractive. This implies that if the evolving variable is not contained in the sequence, the same will diverge from it rather than entering the system. Thus, it has been observed that for most of the values of initial condition taken in a chaotic system, the evolution of the variable is non-periodic. The unpredictability, non-periodicity, and hyper-sensitive dependence on initial condition of chaotic systems can be used to generate sequences of pseudo random numbers by taking a parameter from the initial conditions as the seed, varying which sequences of pseudo random numbers can be generated. The one-dimensional logistic map [3] is a polynomial mapping of degree 2 that exhibits chaotic behavior. Mathematically, the map is described as given in Eq. (62.1).

$$x_{n+1} = \lambda x_n (1 - x_n) \quad (62.1)$$

where λ is the control parameter and $0 < x_n < 1$

A bifurcation diagram is used to investigate the dynamics of nonlinear system. It is a visual representation of the succession of period-doubling produced as the value of the control parameter is increased. Here, the control parameter lambda is plotted on the horizontal axis, and the set of values of the function that is asymptotically visited from the initial conditions is taken on vertical axis. Figure 62.1 illustrates the bifurcation diagram of the map. The bifurcation diagram illustrates the forking of the periods of stable orbits from 1 to 2 to 4 to 8 etc. Each of these bifurcation points is a period doubling bifurcation. The ratio of the lengths of successive intervals between values of r for which bifurcation occurs converges to the first Feigenbaum constant [3].

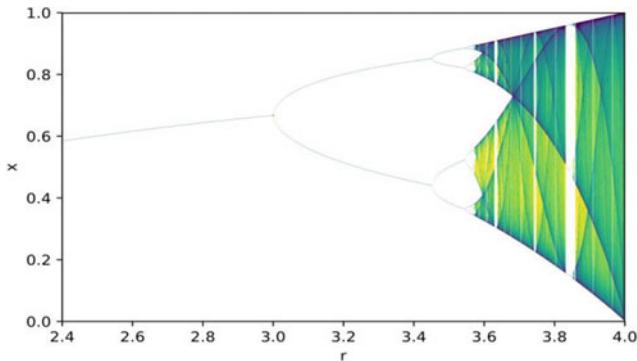


Fig. 62.1 Bifurcation diagram of one-dimensional logistic map

62.4 Algorithm and Implementation

Figure 62.2 illustrates the algorithm for random number generation. A seed is initially taken and fed to the logistic map equation. The results so obtained are then transformed, and the values so obtained are stored. The same has been implemented in Python; three variables are declared, namely x —the seed, r —the control parameter, and size which defines the number of numbers to be generated. A list is then created to store the sequence. The initial value of x is set and then fed to the equation, the values so obtained is transformed using Eq. (62.2).

$$x_n = (x_{n-1} \times a) \bmod m \quad (62.2)$$

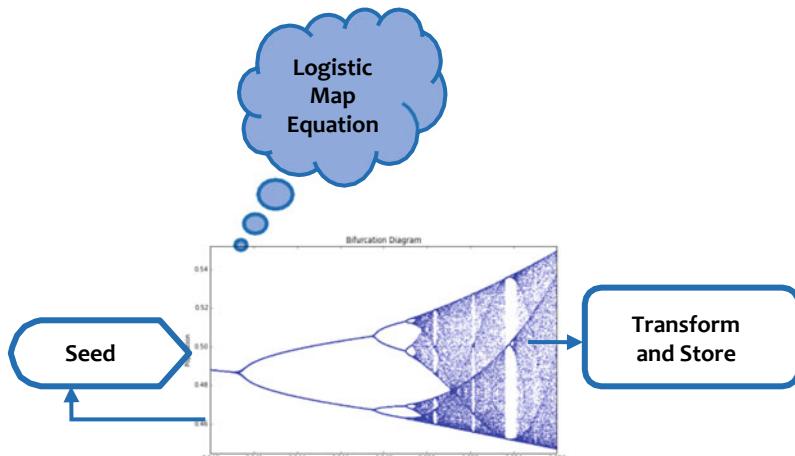


Fig. 62.2 cPRNG algorithm

[999, 224, 85, 824, 750, 897, 472, 629, 606, 480, 149, 905, 183, 502, 924, 277, 676, 908, 575, 232, 690, 8, 679, 886, 894, 1, 72, 610, 429, 190, 601, 969, 118, 531, 737, 281, 654, 871, 204, 67, 741, 850, 117, 462, 24, 0, 327, 727, 199, 962, 303, 908, 740, 30, 447, 289, 836, 226, 766, 326, 759, 754, 197, 199, 800, 350, 373, 228, 360, 813, 83, 2, 700, 424, 238, 832, 897, 532, 394, 566, 692, 104, 931, 84, 857, 262, 603, 314, 338, 28, 429, 502, 476, 555, 174, 333, 45, 5, 223, 928, 944, 513, 444, 36, 790, 806, 219, 315, 706, 166, 148, 834, 485, 612, 694, 29, 533, 100, 111, 492, 334, 132, 34, 3, 5, 381, 858, 336, 711, 610, 792, 566, 46, 220, 664, 967, 608, 303, 276, 232, 979, 624, 895, 963, 868, 630, 429, 947, 793, 872, 384, 595, 474, 536, 14, 152, 760, 60, 441, 826, 989, 893, 976, 281, 528, 310, 735, 706, 913, 997, 372, 852, 493, 190, 3, 31, 68, 564, 425, 544, 251, 265, 553, 198, 384, 97, 701, 265, 607, 860, 726, 101, 503, 232, 233, 720, 792, 871, 392, 150, 92, 4, 407, 108, 659, 468, 86, 572, 495, 638, 170, 586, 525, 343, 110, 923, 796, 12, 793, 69, 94, 394, 724, 850, 317, 807, 902, 832, 854, 664, 951, 32, 907, 578, 688, 520, 452, 165, 468, 800, 122, 724, 436, 822, 529, 515, 782, 233, 214, 444, 837, 225, 701, 824, 526, 226, 0, 496, 6, 102, 131, 478, 56, 6, 208, 459, 838, 793, 784, 540, 823, 443, 538, 830, 175, 526, 982, 925, 3, 84, 1, 506, 192, 599, 939, 908, 858, 478, 162, 25, 517, 478, 839, 16, 0, 640, 379, 596, 381, 218, 561, 166, 784, 649, 51, 60, 2, 811, 734, 685, 806, 241, 263, 252, 389, 581, 900, 742, 749, 903, 665, 130, 860, 174, 773, 260, 586, 859, 173, 296, 65, 23, 6, 88, 721, 833, 34, 550, 568, 835, 236, 783, 631, 262, 981, 159, 996, 387, 594, 378, 544, 784, 901, 61, 368, 4, 414, 90, 69, 0, 798, 64, 144, 312, 862, 793, 89, 338, 672, 646, 37, 132, 672, 87, 400, 888, 679, 480, 344, 586, 466, 866, 412, 386, 438, 0, 86, 642, 842, 452, 650, 142, 429, 822, 247, 564, 850, 279, 390, 262, 205, 498, 240, 802, 424, 232, 778, 964, 336, 572, 84, 8, 937, 525, 864, 897, 308, 244, 823, 726, 717, 845, 232, 557, 824, 568, 805, 819, 285, 798, 122, 705, 581, 833, 656, 672, 2, 1, 685, 184, 111, 516, 434, 113, 156, 854, 989, 82, 329, 482, 954, 228, 310, 204, 672, 252, 159, 58, 828, 709, 297, 197, 92, 2, 678, 868, 789, 31, 315, 510, 247, 798, 435, 779, 600, 254, 618, 711, 476, 748, 739, 291, 210, 86, 567, 386, 970, 736, 73, 4, 138, 27, 977, 920, 287, 499, 642, 87, 367, 924, 208, 64, 440, 280, 342, 392, 747, 287, 709, 399, 868, 517, 346, 910, 345, 495, 285, 184, 720, 24, 429, 636, 676, 697, 444, 393, 98, 920, 170, 579, 398, 279, 541, 697, 551, 790, 306, 625, 525, 2, 64, 559, 242, 954, 46, 745, 809, 466, 4, 314, 803, 936, 229, 395, 176, 279, 812, 686, 4, 532, 714, 528, 817, 153, 922, 640, 606, 946, 931, 466, 290, 712, 183, 594, 87, 37, 204, 171, 670, 961, 118, 766, 303, 59, 857, 192, 385, 531, 671, 515, 830, 9, 73, 891, 832, 17, 221, 980, 842, 985, 908, 556, 184, 987, 642, 931, 807, 913, 800, 676, 599, 488, 130, 757, 282, 106, 828, 5, 37, 785, 812, 480, 289, 345, 372, 838, 56, 962, 140, 950, 6, 797, 795, 199, 128, 324, 392, 733, 160, 695, 439, 15, 796, 351, 541, 909, 356, 88, 563, 552, 795, 213, 448, 572, 651, 922, 235, 199, 858, 394, 566, 752, 508, 299, 808, 661, 24, 548, 37, 3, 1, 203, 538, 405, 400, 196, 49, 125, 922, 820, 852, 894, 836, 420, 996, 358, 676, 132, 843, 858, 452, 368, 544, 380, 165, 74, 3, 737, 293, 610, 654, 626, 128, 422, 799, 206, 267, 807, 489, 916, 651, 694, 896, 609, 747, 992, 359, 896, 430, 978, 92, 72, 3, 590, 685, 876, 172, 452, 904, 0, 748, 744, 989, 872, 71, 384, 823, 183, 782, 64, 1, 794, 556, 236, 790, 745, 640, 528, 22, 6, 727, 471, 632, 746, 82, 490, 689, 326, 464, 811, 233, 596, 72, 338, 115, 580, 899, 398, 516, 669, 449, 696, 14, 827, 121, 112, 70, 898, 614, 668, 602, 122, 793, 421, 353, 54, 69, 400, 982, 988, 845, 441, 191, 995, 789, 652, 37, 61, 386, 572, 652, 816, 419, 630, 958, 775, 842, 346, 376, 428, 551, 574, 169, 266, 872, 623, 349, 492, 345, 645, 132, 698, 18, 733, 256, 986, 351, 896, 893, 330, 911, 376, 612, 650, 899, 582, 682, 322, 724, 184, 120, 448, 395, 248, 548, 500, 14, 238, 354, 769, 117, 400, 207, 901, 654, 930, 728, 539, 973, 814, 169, 969, 350, 534, 744, 966, 140, 119, 704, 278, 399, 654, 722, 95, 289, 494, 714, 160, 9, 359, 684, 56, 790, 764, 306, 758, 268, 81, 778, 283, 721, 192, 749, 502, 328, 84, 25, 521, 181, 530, 203, 363, 345, 434, 424, 723, 551, 280, 842, 94, 768, 436, 790, 362, 882, 658, 621, 976, 679, 429, 196, 230, 894, 756, 541, 322, 495, 40, 38, 649, 298, 266, 747, 636, 380, 583, 942, 231, 110, 630, 807, 73, 596, 636, 780, 393, 261, 371, 793, 6, 825, 155, 607, 465, 28, 49, 998, 484, 321, 972, 136, 380, 75, 633, 342, 581, 810, 51, 286, 852, 626, 815, 935, 457, 242, 955, 677, 631, 57, 1, 124, 146, 302, 284, 132, 11, 116, 45, 877, 551, 292, 446, 484, 385, 896, 725, 139, 928, 943, 867, 921, 979, 540, 650, 53, 5, 168, 693, 536, 140, 484, 752, 5, 985, 158, 985]

Fig. 62.3 1000 three-digit random numbers generated using the proposed methods for set parameters

Here, a is the multiplier which is used to transform the values to integers from fractional values, and mod m is the value control parameter which can be adjusted to determine the number of digits in the final value. In this paper, the value of a is set as 10^{16} , and the values of m taken is [100, 1000, 10000] to generate 2-digit, 3-digit, and 4-digit random numbers. An example of the same is displayed in Fig. 62.3, a sequence of 1000 three-digit random numbers is generated with $x = 0.0000001$, $r = 3.915$, $\text{size} = 1000$, $a = 10^{16}$ & $m = 1000$.

62.4.1 Observations

Several sequences of random numbers are generated by adjusting the parameters of the algorithm. The results are then visually represented using a density distribution graph and a scatter plot (x_{n+1} vs. x_n).

Observation 1: A sequence of a thousand two-digit random numbers is generated with the following parameters:

$x = 0.00001$, $r = 3.915$, $a = 10^{16}$, and $m = 1$. Figure 62.4 illustrates the density distribution of this sequence, and Fig. 62.5 shows the scatter plot of x_{n+1} versus x_n .

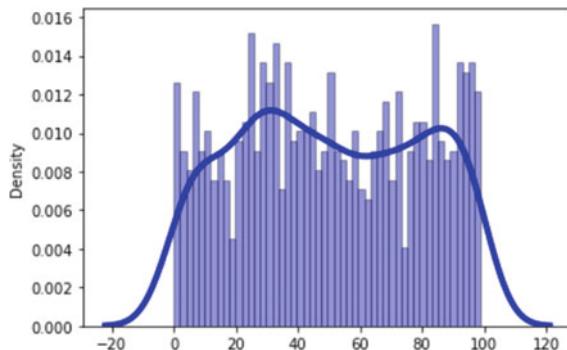


Fig. 62.4 Density distribution of sequence generated in Observation 1

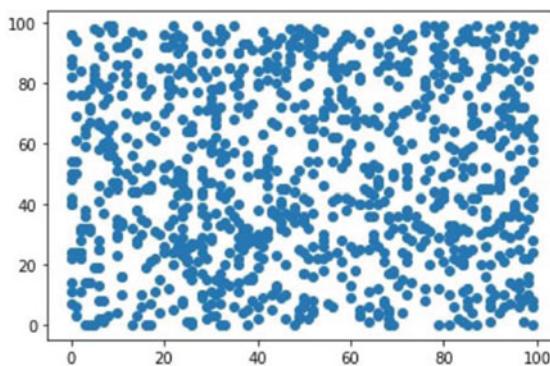


Fig. 62.5 Scatter plot of sequence generated in Observation 1

Observation 2: A sequence of a thousand two-digit random numbers is generated with the following parameters:

$x = 0.00002$, $r = 3.915$, $a = 10^{16}$, and $m = 100$. Figure 62.6 illustrates the density distribution of this sequence, and Fig. 62.7 shows the scatter plot of x_{n+1} versus x_n .

Observation 3: A sequence of a thousand three-digit random numbers is generated with the following parameters:

$x = 0.000007$, $r = 3.915$, $a = 10^{16}$, and $m = 1000$. Figure 62.8 illustrates the density distribution of this sequence, and Fig. 62.9 gives the scatter plot of x_{n+1} versus x_n .

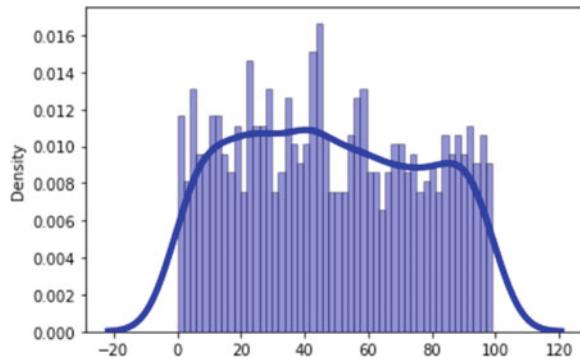


Fig. 62.6 Density distribution of sequence generated in Observation 2

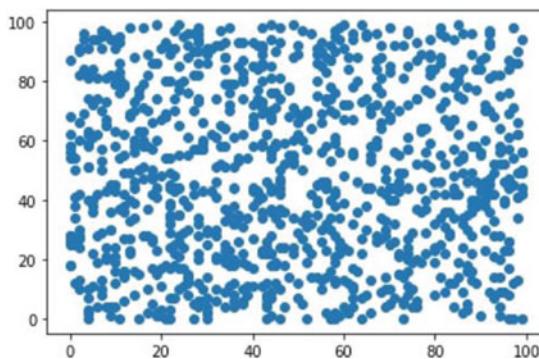


Fig. 62.7 Scatter plot of sequence generated in Observation 2

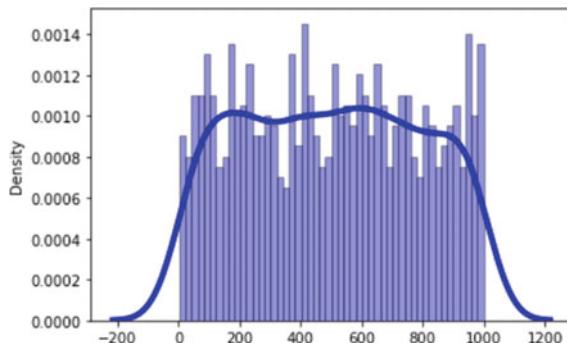


Fig. 62.8 Density distribution of sequence generated in Observation 3

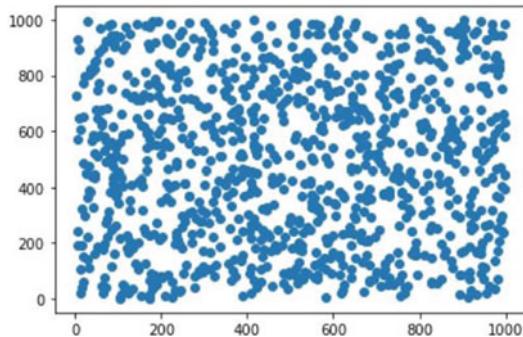


Fig. 62.9 Scatter plot of sequence generated in Observation 3

Observation 4: A sequence of a thousand 3-digit random numbers is generated with the following parameters:

$x = 0.000003$, $r = 3.915$, $a = 10^{16}$ and $m = 1000$. Figure 62.10 illustrates the density distribution of this sequence, and Fig. 62.11 the scatter plot of x_{n+1} versus x_n .

Observation 5: A sequence of a thousand four-digit random numbers is generated with the following parameters:

$X = 0.000006$, $r = 3.915$, $a = 10^{16}$, and $m = 10,000$. Figure 62.12 illustrates the density distribution of this sequence, and Fig. 62.13 shows the scatter plot of x_{n+1} versus x_n .

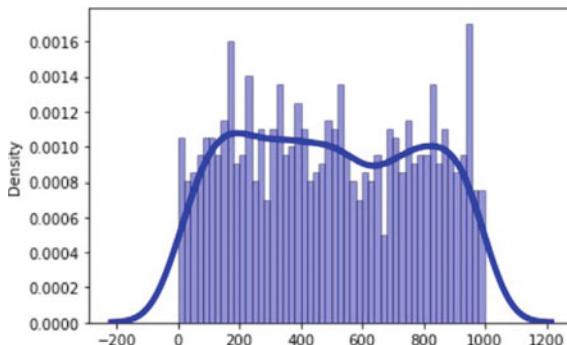


Fig. 62.10 Density distribution of sequence generated in Observation 4

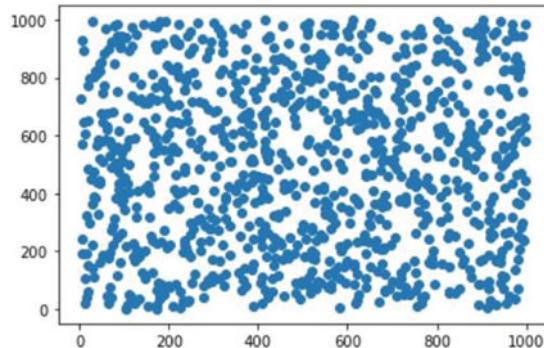


Fig. 62.11 Scatter plot of sequence generated in Observation 4

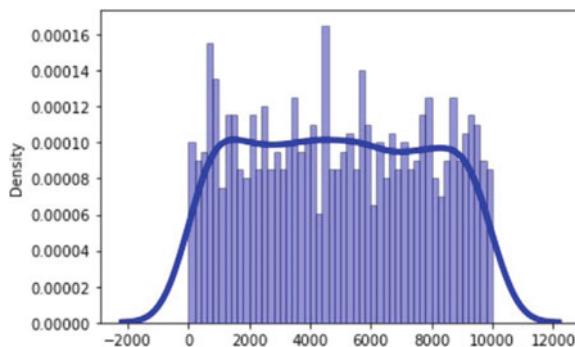


Fig. 62.12 Density distribution of sequence generated in Observation 5

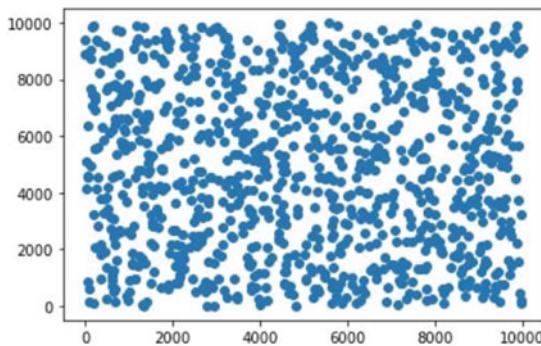


Fig. 62.13 Scatter plot of sequence generated in Observation 5

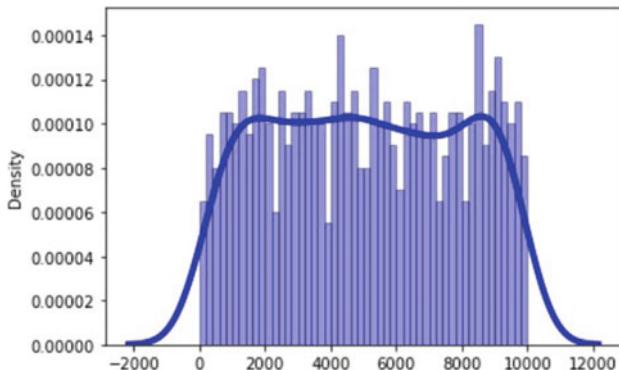


Fig. 62.14 Density distribution of sequence generated in Observation 6

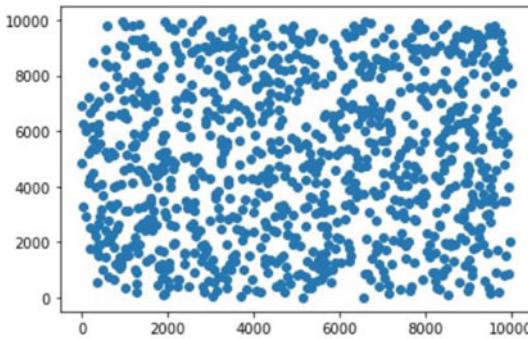


Fig. 62.15 Scatter plot of sequence generated in Observation 6

Observation 6: A sequence of a thousand four-digit random numbers is generated with the following parameters:

$x = 0.000008$, $r = 3.915$, $a = 10^{16}$, and $m = 1000$. Figure 62.14 illustrates the density distribution of this sequence, and Fig. 62.15 the scatter plot of x_{n+1} versus x_n .

62.4.2 Tests for Randomness

A test for randomness is done to check if the given distribution of data is truly random or if there exists some pattern or trend in the same. In this paper, Wald–Wolfowitz runs test has been used to test for randomness. It is a statistical procedure that examines whether a sequence of data is occurring randomly from a specific distribution. That is, run test is used for examining whether or not a set of observations constitutes a random sample from an infinite population. The runs test results are summarized in

Table 62.1 Runs test results for various observations

Set of observations	z —test statistics	p —value	Inference
Observation 1	-0.5619514869490162	0.5741490727971621	Random sequence
Observation 2	-1.8002975944469373	0.0718136605525433	Random sequence
Observation 3	-0.16366341767699444	0.8699961176213894	Random sequence
Observation 4	-0.16366341767699444	0.8699961176213894	Random sequence
Observation 5	-0.09463204468147698	0.9246070960693767	Random sequence
Observation 6	-0.09463204468147698	0.9246070960693767	Random sequence

Table 62.1; the z —test static value and the p —value for each observation is computed and inference is made. $\alpha = 0.5$.

62.5 Conclusion

Random number generation is an important area in computer science and cryptography. The literature on this topic is quiet rich and vast. Several studies have been published on both HRNGs and PRNGs since 1970s. In this paper, we have generated non-periodic sequences of random numbers using a chaotic system. The test results indicate the sequences generated are truly random. There exists several chaotic systems in the nature; in this paper, the one-dimensional logistic map has been taken; however, in this paper, we have proposed a generalized algorithm wherein other chaotic systems can be used to solve a wide spectrum of problems. The algorithm discussed in this paper is of simple nature with a time complexity of $O(n)$. The non-periodic nature of the sequences generated makes the algorithm a suitable candidate for generating large sequences of pseudo random numbers.

References

1. The Numerical Algorithms Group.: G05—Random number generators. NAG Library Manual, Mark 23. Retrieved 02 Sept 2012
2. Wang, Y.: Statistical properties of pseudo random sequences and experiments with PHP and Debian OpenSSL. In: Computer Security—ESORICS 2014. Lecture Notes in Computer Science. 8712, pp. 454–471. Springer LNCS, Heidelberg (2014). https://doi.org/10.1007/978-3-319-11203-9_26. ISBN 978-3-319-11202-2
3. Boeing, G.: Chaos theory and the logistic map. Retrieved 17 May 2020
4. May, R.M.: Simple mathematical models with very complicated dynamics. Nature. **261**(5560), 459–467 (1976). <https://doi.org/10.1038/261459a0>
5. Weisstein, E.W.: Logistic Equation. *MathWorld*
6. Cheung, R.C.C., Villasenor, J.D., Luk, W.: Hardware generation of arbitrary random number. Int. J. Sci. Res. **4**(1), (2015)
7. Dabal, P., Pelka, R., Gayoso, C.A., Rabini, M., Moreira, J.: A review on implementation of random number generation based on FPGA (2015)

8. Gu, X.-C., Zhang, M.-X.: Uniform random number generator using Leap-Ahead LFSR architecture. In: International Conference on Computer and Communications Security (2009)
9. Jonathan, M.C., Cerda, J.C., Martinez, C.D., David, H.K.: Hoe 44th IEEE Southeastern Symposium on System Theory University of North Florida, Jacksonville, FL Mar 11–13 (2012)
10. Dabal, P., Pelka, R.: FPGA implementation of chaotic pseudo-random bit generators, MIXDES 2012. In: 19th International Conference on Mixed Design of Integrated Circuits and Systems, 24–26 May 2012, Warsaw, Poland

Chapter 63

Consumer Buying Behavior and Bottom of Pyramid (BoP): The Diffusion of Marketing Strategy



Artta Bandhu Jena and Lopamudra Pradhan

Abstract BoP may be regarded as common term in economics and business. There are few arguments to define the low-income group by experts. Few researchers classify by taking the annual salary less than USD 1500 and others define less than USD 700/USD 400. But, the wide definition of BoP group refers to Karnani which emphasizes 2.5 billion people live with less than \$2 a day in world. Poverty is the main economic, social and moral problem. Poverty alleviation is an urgent challenge. Governments and international organizations i.e., World Bank and UNO have been trying to address this challenge. BoP offers the opportunities to create the value for companies (CSR) and poor people also. Simply, design the products and selling them is a marketing strategy to achieve the business goal. Success of business would depend on knowing the BoP intimately. Against this backdrop, the present study would discuss the consumer buying behavior which includes thorough analysis of buying motive, buying habit, buying attitude and post-purchase behavior of consumers in BoP retail market.

63.1 Introduction

Low-income markets show the prodigious opportunity to world's wealthiest companies to seek the fortunes and bring prosperity to aspiring poor people. Distribution of wealth and capacity to generate income would be captured in the form the economic pyramid in the world. More than 4 billion live at BoP in less than \$2 per day. Consumer behavior is the study of how people buy, what they buy, when they buy and why they buy. The study of consumer behavior deals with decision-making process and physical activity, individuals engage in when evaluating, acquiring and pattern in consuming the goods and services. The behavior pattern of consumer has

A. B. Jena (✉)

Department of Business Management, Fakir Mohan University, Vyasa Vihar, Balasore, Odisha, India

L. Pradhan

Department of Economic, Nilgiri College, Balasore, Odisha, India

been a major change in retail sector. The term, “BoP” is first coined by Management Guru, CK Prahalad along with Stuart Hall of Cornell University. Their study on “The Fortune at the BoP”, provided the first articulation of how companies can profitably serve the huge untapped market of more than 4 billion people living in less than \$2 per day. However, success of business giants like P&G, Unilever, ICICI Bank and Cavinkare see the increased number of firms now evincing interest the BoP concept in internationally. Prahalasd and Hart [1] published a paper in Strategy + Business Magazine that introduced the idea of BoP, and poor people have the vast untapped business opportunities. If big companies serve the poor people, they can eradicate poverty and also earn the profit. The concept of BoP suggests, the best way to meet the wants of poor people can be through this profit driven market-based approach. The BoP represents the market made up by the poorest people of region.

63.2 Need of the Study

The study consumer behavior is rooted in modern marketing concept. To operationalize the BoP concept, marketers solve the consumption problem of consumers in BoP market. But, business will not help consumer to solve the consumption problems unless marketers study the baying behavior of consumers and make to understand the buying process and the factors affecting it. The idea of retell business is retail in detail, emergence of giant-sized retail business in the names of malls and polices of government allow FDI in retail business and also change the life style and preferences of consumers in globally more particularly in BoP market. For solving the problem of consumers and marketers in BoP market. There should be regular research in area of consumer buying behavior. In present globalized economic scenario, the consumers have ample opportunities to select any shop for purchasing good quality products.

63.3 Review of Literature

Prahalad and Hart [1] stated on the matter of MNCs on BoP: “for corporations that have distribution and brand presence throughout the developing world, such as Coca-Cola Company, the BoP offers a vast untapped market for such products as water and nutritionals.” Sarin and Venugopal (2003) studied on BoP market and highlighted that it is considered that marketing strategies of large organizations may influence the quality of life to people of BoP. Prahalad and Hart [1] and Prahalad [2] made the studies that poor people are innovative entrepreneur and their abilities, if honed, can be of great use for multinational business organizations to reduce the overall cost and enhance their profitability. Prahalad (2005) pointed out that lack of resources make them more risk averse. His study conducted to develop strategies to make brand positioning in BoP market. Karnani [3] studied that poverty will be decreased only by considering BoP as producers and not merely as customers.

To include BoP into value chain, marketers will lead to increase in real income of BoP that will enhance the self-esteem, social cohesion and empowerment of people. Prahalad (2007) suggested that poor people can choose small packages if they are dissatisfied because of switching cost. Kauffman (2007) argued, MNCs offer products in traditional markets which are usually not suitable for BoP market. The reason is that local conditions make the provisions of such products are challenging. Hammond et al. [4] studied that BoP is to be estimated 4 billion people in world who are poor and they have the limited/no access to consume the products. People in socio-economic group earn US\$ 1 to US\$ 8 per day in purchasing power parity in globally. Davidson (2009) studied that when large organizations succeed in working in BoP, they often may not take all credit for their business success. Karnani [5] highlights that poor people are not always rational in economic decision-making. The poor people may incur their income on unhealthy products instead of purchasing more food for their families/children education. Western MNCs are accustomed to innovative approaches for products offered to BOP market. Davidson (2009) opined that BoP model is treated as a popular strategy to provide poor women as an opportunity to earn income by distributing goods and services door-to-door recent years. It was explored the recent example of BoP entrepreneurship: the CARE Bangladesh Rural Sales Program (BRSP). Alur [6] studied that BoP as market or producers poses a set of challenge to private sector, if addressed appropriately will transform the current urban horizon by unleashing the true potential which lies in BoP. Rabino [7] made a study that despite disagreements in BoP definition, there is the idea that BoP market with millions of potential consumers is the profitable market.

63.4 Research Gap

The goal of a business invasion of BoP largely depends upon its strategy. This may not achieve the goal if something is wrong in marketing strategy i.e., marketing-mix. In doing so, it would create an emerging area of research where the information about the subsistence consumers profile and their purchasing behavior has to be known. Consumer behavior is deeply influenced by various factors i.e., culture, subculture, social class, income, age, gender, etc. Therefore, marketers should analyze these factors for BoP marketing. This is the main research gap. Hence, present study would try to bridge the gap. There are few previous studies towards BoP marketing in rural Odisha. So, the present study is a new one in BoP marketing and intention to find a solution.

63.4.1 Scope of the Study

The primary data was collected from 4 districts of Odisha of Khordha, Cuttack, Balasore, Puri and Dhenkanal and also regarded sampling units for the purpose of

study. 597 samples have been collected randomly from consumers from the sampling areas through a well-structured questionnaire.

63.4.2 Objective of the Study

The present study has the following objectives (i) to analyze the attitude of consumers and buying behavior towards products, (ii) to measure the perception differences on market efficiency of BoP across age and gender and (iii) to study about the satisfaction of the customers on BoP marketing practices.

63.4.3 Hypothesis of Study

The following hypotheses are set and tested during the study. Hypothesis-1: Motivation on BoP is significantly different across age group of consumers, Hypothesis-2: Cultivating customer peer groups in BoP sale is significantly different across the age groups of consumers and Hypothesis-3: Cultivating customer peer groups in BoP sale is similar irrespective of the gender types of consumers.

63.5 Research Methodology

The present study is based on empirical nature. The primary data has been collected from 597 respondents based on stratified random sampling method. Further, secondary data has also been collected from various sources. Then, primary data has been tabulated and analyzed through SPSS software to get the inference.

63.6 Analysis and Interpretation of Data

The age of respondents has been divided into five categories i.e., below 20, 21–35, 36–50, 51–65, above 65 years. Similarly, the income has been categorized i.e., less than Rs. 10, 000, Rs. 10,000–Rs.15,000, Rs. 15,000–Rs. 20,000, Rs. 20,000–Rs. 25,000 and Rs. 25,000 and above. Similarly, in relation to their education qualification, it is categorized as no education, below HSC, technical, graduation and any other. Further, from companion point of view, it is divided such as friends, family, colleagues, alone and relatives.

Table 63.1 KMO and Bartlett's Test on Motivation

	KMO measure of sampling adequacy	0.957
Bartlett's test of sphericity	Approx. Chi-square	17.513
	Df	10
	Sig	0.064

Source SPSS Output Source

A. Factor Analysis on Motivation

Table 63.1 shows KMO and Bartlett's test on "Motivation" for purchase on BoP model. The KMO measures the adequacy which shows the proportion of variance on "Motivation" that may be influenced by underlying by underlying 5 factors. Bartlett's test measures correlation matrix that variables are associated for structure detection as the value is 0.957. Thus, factor analysis is useful with input variables on "Motivation" on BoP retail.

It is observed from Table 63.2 that the factors response on "Motivation" in BoP market, the initial value of 5 factor components are one and extraction values are greater than 0.5 which means all factors are fit to data and will be explored to measure the variance to know most significant factor. All variable factors of "Motivation" are fit for further measurement.

Table 63.2 Communalities on motivation

		Initial	Extraction
A1	Physically infra-Attractiveness has an impact on motivation	1.000	0.676
A2	Do you think products available in retail are of best using?	1.000	0.751
A3	Do you think that of perfect brand match?	1.000	0.646
A4	Do you think of target customer-match?	1.000	0.923
A5	Do you think of perfect Price-product match?	1.000	0.848

Extracted Method Principal Component Analysis

Source SPSS Output

Table 63.3 Total variance explained on motivation

Components	Initial Eigen value			Total of extraction squared loading		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	1.128	22.552	22.552	1.128	22.552	22.552
2	1.118	22.353	44.905	1.118	22.353	44.905
3	0.971	19.420	64.325			
4	0.937	18.737	83.062			
5	0.847	16.938	100.000			

Source SPSS Output

It is seen from Table 63.3, total variance of individual factors (five) and extraction value of sum of squares loading values 3 factors were found in initial Eigen values which are positive and finite value. The process adopted in principal component analysis with loading values of component scores on responses from migrated customers which are presented in last column of Table 63.3. The results are found in extraction value for those 2 factors which are significant among selected factors in initial Eigen value. It has been concluded that in measuring through principal component analysis, only 2 factors out of five 5 factors are significant in total variance of 45% which shows a loss of 55% data and needs for further analysis. The extraction value of loadings are most equal with Eigen value of the components which also indicate of no loss of data due to responses as presented from the responses on Likert scale.

Table 63.4 highlights result of component matrix of each components variance on “Motivation” where in 1st column more value on “A3” i.e., 0.709 and followed by “A5” i.e., 0.862 are two the most significant cause for better option for “Motivation”. Similarly, results are also found in 2nd columns for these two significant variables on A3 and A5. It is concluded that the effectiveness of motivation is found more significantly within two columns i.e., A3 and A5. The customers who have not changed their attitude toward purchasing from the retail store when stores cannot fulfill the expectation of customers more particularly on perfect brand match and perfect price-product match in BoP marketing.

Table 63.4 Component matrix on motivation

	Component	
	1	2
A1	Physically infra-Attractiveness has an impact on motivation	-0.096
A2	Do you think products available in retail are of best using?	0.436
A3	Do you think that of perfect brand match?	0.709
A4	Do you think of target customer-match?	0.331
A5	Do you think of perfect Price-product match?	0.862

Extracted Method Principal Component Analysis a. 2 components extracted

Source SPSS Output

Table 63.5 KMO and bartlett's test on pricing mechanism

	KMO measure of sampling adequacy	0.648
Bartlett's test of sphericity	Approx. Chi-square	266.975
	Df	28
	Sig	0.000

Source SPSS Output

B. Pricing Mechanism

Table 63.5 shows KMO and Bartlett's test on "Pricing Mechanism" for purchase on BoP model. The KMO measure of adequacy highlights proportion of variances on "Pricing Mechanism" which may be influenced by underlying 8 factors. Bartlett's test of sphericity measures the correction matrix, and variables are associated for structure detection as the value is 0.648. Thus, factor analysis is useful with input variables on "Pricing Mechanism" on BoP retail.

Table 63.6 indicates the factors responses on "Pricing Mechanism" in BoP marketing. The initial of all eight factor components are one, and extraction values indicate greater than 0.5 which means all factors are fit to data to measure the variances to know the significant factor. Thus, all eight factor variables of "Pricing Mechanism" are fit for measurement.

It is noticed from Table 63.7, total variance of individual factors (eight) and extraction values of sum of square loading value of 3 factors are measured in initial

Table 63.6 Communalities on pricing mechanism

		Initial	Extraction
B1	Market pricing' innovation	1.000	0.602
B2	Price-performance-process matrix' incorporate local know-how.'	1.000	0.618
B3	Innovative distribution and communication strategies	1.000	0.863
B4	Effective and prolong activity	1.000	0.977
B5	Use purchase parity base	1.000	0.706
B6	Use high price sensitivity mechanism	1.000	0.898
B7	Connected efficiently-provided communities	1.000	0.879
B8	Market pricing' innovation	1.000	0.766

Extracted Method Principal Component Analysis

Source SPSS Output

Table 63.7 Total variance explained on pricing mechanism

Components	Initial Eigen value			Total of extraction squared loading		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.851	23.137	23.137	1.851	23.137	23.137
2	1.202	15.026	38.163	1.202	15.026	38.163
2	1.055	13.189	51.352	1.055	13.189	51.352
4	0.971	12.139	63.491			
5	0.822	10.272	73.763			
6	0.763	9.537	83.300			
7	0.688	8.603	91.903			
8	0.648	8.097	100.000			

Extracted Method Principal Component Analysis

Source SPSS Output

Eigen values which are positive and finite values. The process adopted in principal component analysis with loading values of the component scores on responses from migrated customers, which are presented in the last column of Table 63.7. The results are found in extraction value for those 2 factors which are significant among selected factors in initial Eigen value. It is concluded that in measuring through principal component analysis, only 3 factors out of 8 factors are significant with total variance of 51% that shows a loss of 49% of data and needs for further analysis. Further, the extraction value of loading is most equal with Eigen value of the components which also indicate of no loss of data due to responses as presented from the responses on Likert scale.

Component matrix in Table 63.8 highlights the results of the each components Variance on “Pricing Mechanism”. The customers made their exercise on selecting the options on the eight variables. Out of eight factors, only five factors are found those have more significant positive value in the columns of components. Here, the more value is marked B3, B5 and B6, and these are the three most significant causes for better option on “Pricing Mechanism”. So, the “Effectiveness of Pricing Mechanism” is highlighted more significantly on B3 (Innovative Distribution and Communication Strategies), B5 (Use purchase parity base) and B6 (Use high price sensitivity mechanism). The other five factor variables are to be solved for more customer satisfaction and enhancing its efficiency in BoP marketing.

C. Product Characteristics

Table 63.9 shows KMO and Bartlett’s Test on “Product Characteristics” for purchase on BoP model. The KMO measure of adequacy shows proportion of variance on “Product Characteristics” that may be affected by underlying 18 factors. Bartlett’s test of sphericity measures correlation matrix, variables are associated for structure

Table 63.8 Component matrix on pricing mechanism

		Component		
		1	2	3
B1	Market pricing' innovation	0.433	-0.638	0.081
B2	Price-performance-process matrix' incorporate local know-how.'	-0.200	0.167	0.742
B3	Innovative distribution and communication strategies	0.638	-0.170	0.164
B4	Effective and prolong activity	0.350	-0.576	0.151
B5	Use purchase parity base	0.604	0.355	-0.120
B6	Use high price sensitivity mechanism	0.606	0.418	-0.235
B7	Connected efficiently provided communities	0.575	0.184	0.117
B8	Market pricing' innovation	0.171	0.267	0.605

Extracted Method Principal Component Analysis

a.3 components extracted

Source SPSS Output

Table 63.9 KMO and Bartlett's test on product characteristics

	KMO measure of sampling adequacy	0.951
Bartlett's test of sphericity	Approx. Chi-Square	664.628
	Df	153
	Sig	0.000

Source SPSS Output

detection as the value is 0.951. Factor analysis is useful with input variables of perception on “Product Characteristics” on BoP retail.

Table 63.10 indicates the factors responses on “Product Characteristics” in BoP marketing where initial values of all 5 factor components are one, and extraction values indicate more than 0.5. All factors are fit to data to measure the variance to know the most significant factor. All six factor variables of “Product Characteristics” are fit for further measurement.

Table 63.11 shows, total variance of individual factors (eighteen) and extraction values of sum of squares loading value of 7 factors are found in initial Eigen values which are positive and finite values. The process adopted in principal component

Table 63.10 Communalities on product characteristics

		Initial	Extraction
C1	Availing of personalized products	1.000	0.793
C2	Makes wide advertisement for creating awareness on product	1.000	0.614
C3	Feeling of product development	1.000	0.804
C4	Improve access to retail products	1.000	0.887
C5	Creating buying power(low cost product ranges)	1.000	0.675
C6	Shaping product aspiration to urban life styles	1.000	0.673
C7	Any time product availability	1.000	0.769
C8	Bundle Pack for cheaper cost	1.000	0.645
C9	Refill pack availability	1.000	0.791
C10	Well maintained supply chain	1.000	0.995
C11	Experiencing products in malls for new products	1.000	0.832
C12	Localized product availability	1.000	0.748
C13	Pay-per-use” system(laundry system)	1.000	0.981
C14	Well maintained supply chain	1.000	0.899
C15	Meet the needs of consumers in a different area	1.000	0.708
C16	Greenfield-market expansion	1.000	0.952
C17	Product selling devotion to social missions	1.000	0.885
C18	Nurture the creation of a new market	1.000	0.796

Extracted Method Principal Component Analysis

Source: SPSS Output

Table 63.11 Total variance explained on product characteristics

Components	Initial Eigen values			Total of extraction squared loadings		
	Total	% of Variance	Cumulative%	Total	% of Variance	Cumulative %
1	1.837	10.206	10.206	1.837	10.206	10.206
2	1.603	8.904	19.110	1.603	8.904	19.110
3	1.470	8.167	27.277	1.470	8.167	27.277
4	1.368	7.601	34.878	1.368	7.601	34.878
5	1.149	6.383	41.261	1.149	6.383	41.261
6	1.135	6.307	47.568	1.135	6.307	47.568
7	1.085	6.026	53.594	1.085	6.026	53.594
8	0.986	5.480	59.074			
9	0.962	5.345	64.418			
10	0.897	4.983	69.401			
11	0.853	4.741	74.142			
12	0.783	4.350	78.492			
13	0.762	4.236	82.728			
14	0.712	3.953	86.682			
15	0.677	3.759	90.440			
16	0.654	3.634	94.074			
17	0.573	3.181	97.255			
18	0.494	2.745	100.000			

Extracted Method Principal Component Analysis

Source: SPSS Output

analysis with loading values of the component scores on responses from migrated customers which are presented in the last column of the above table. The results are fund in extraction value for those 7 factors which are significant among selected factors in initial Eigen value. It is concluded that in measuring through principal component analysis, only 7 factors out of 18 factors are significant with total variance of 54% which indicates a loss of 46% data and needs for further analysis. Further, the extraction value of loading are most equal with the Eigen value of the components which also indicate of no loss of data due to responses as presented from the responses on Likert scale.

Component matrix in Table 63.12 notices the results of the each components Variance on “Product Characteristics”. The customers made their exercise on selecting the options on the eighteen variables. Out of eighteen factors, only seven factors are found those have more significant positive value in column of components. Here, the more value is marked on C2, C4, C5, C8, C9, C13 and C18 and these are the seven most significant causes for better option on “Product Characteristics”. So, it is concluded that the Effectiveness of Empowerment is highlighted more significantly on C2 (Makes wide advertisement for creating awareness on product), C4 (Improve

Table 63.12 Component matrix on product characteristics

		Components						
		1	2	3	4	5	6	7
C1	Availing of personalized products	0.200	0.133	-0.007	0.145	-0.084	0.372	0.410
C2	Makes wide advertisement for creating awareness on product	0.681	0.180	-0.041	-0.159	-0.033	0.210	-0.214
C3	Feeling of product development	0.581	0.202	-0.163	-0.233	0.077	-0.125	-0.154
C4	Improve access to retail products	0.734	0.004	0.093	0.159	0.092	-0.056	0.053
C5	Creating buying power(low cost product ranges)	0.879	0.403	0.279	-0.055	0.194	-0.053	0.534
C6	Shaping product aspiration to urban life styles	-0.089	0.546	-0.553	0.172	0.150	-0.047	-0.082
C7	Any time product availability	-0.196	0.150	-0.538	0.241	-0.382	-0.123	0.004
C8	Bundle Pack for cheaper cost	0.843	-0.054	0.359	0.550	-0.232	0.107	-0.353
C9	Refill pack availability	0.706	0.233	0.470	0.493	0.046	-0.142	-0.198
C10	Well maintained supply chain	0.225	-0.626	-0.124	0.034	-0.053	0.030	0.180
C11	Experiencing products in malls for new products	0.098	-0.501	-0.351	0.168	0.222	0.134	0.229
C12	Localized product availability	-0.113	0.108	0.431	-0.552	-0.062	0.171	0.016
C13	Pay-per-use" system(laundry system)	0.776	0.053	0.034	-0.112	0.075	0.455	-0.158
C14	Well maintained supply chain	0.000	0.257	-0.086	0.192	0.294	0.632	0.052
C15	Meet the needs of consumers in a different area	0.212	0.161	0.181	0.370	-0.007	-0.211	0.471
C16	Greenfield-market expansion	0.060	0.147	-0.085	0.088	-0.611	0.367	0.063
C17	Product selling devotion to social missions	0.049	-0.398	0.192	0.079	-0.212	0.170	0.090

(continued)

Table 63.12 (continued)

		Components						
		1	2	3	4	5	6	7
C18	Nurture the creation of a new market	0.848	-0.207	-0.004	0.269	0.546	0.139	-0.203

Extracted Method Principal Component Analysis

a. 7 components extracted

Source SPSS Output

access to retail products), C5 (Creating buying power (low cost product ranges)), C8 (Bundle Pack for cheaper cost), C9 (Refill pack availability), C13("Pay-per-use" system(laundry system)) and C18 (Nurture the creation of a new market). The other eleven factor variables are to be solved for more customer satisfaction and enhancing its efficiency in BoP marketing.

D. Cultivating Customer Peer Groups BoP Sale (in five-point scale)

Table 63.13 shows KMO and Bartlett's Test on "Cultivating Customer Peer Groups" for purchase on BoP model. The KMO measure of adequacy shows the proportion of variance on "Cultivating Customer Peer Groups" that may be influenced by underlying 10 factors. Bartlett's test of sphericity measures the correlation matrix that variables are related for structure detection as the value is 0.689 that factor analysis is useful with input on "Cultivating Customer Peer Groups" in BoP retail.

Table 63.14 indicates the factors responses on "Cultivating Customer Peer Groups" in BoP marketing where initial values of all 10 factor components are one, and extraction values show greater than 0.5 which means all factors are to fit to data and can be explored to measure the variance to find out the significant factor. All 10 factor variables on "Cultivating Customer Peer Groups" are fit for further measurement.

Table 63.15 shows, total variances of individual factors (ten) and extraction values of sum of squares loading values 3 factors are found in initial Eigen values which are positive and finite value. The process adopted in Principal Component analysis with loading value of component scores on responses from migrated customers which are presented in last column of Table 63.15. The results are found in extraction value for those 3 factors which are significant among selected factors in initial Eigen value. It

Table 63.13 KMO and Bartlett's test on cultivating customer peer groups

	KMO measure of sampling adequacy	0.689
Bartlett's test of sphericity	Approx. Chi-square	814.884
	Df	45
	Sig	0.000

Source SPSS Output

Table 63.14 Communalities on cultivating customer peer groups

		Initial	Extraction
D1	Create a sewing circle	1.000	0.821
D2	Create an investment club,	1.000	0.973
D3	Create a customer peer group	1.000	0.636
D4	Create high-touch benefits of an enabling service,	1.000	0.656
D5	Adopt new behaviors and mind-sets orientation	1.000	0.701
D6	Create capacity for recharging	1.000	0.683
D7	Adding more functionality, and ability	1.000	0.697
D8	Create cultural competence for product consumption	1.000	0.602
D9	Create Bundle pack for use	1.000	0.713
D10	Create a hygienic environment around the villages	1.000	0.844

Extracted Method Principal Component Analysis

Source SPSS Output

Table 63.15 Total variance explained on cultivating customer peer groups

Components	Initial Eigen value			Total of extraction squared loading		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	2.528	25.281	2.281	2.528	25.281	25.281
2	1.415	14.153	39.435	1.415	14.153	39.435
3	1.182	11.820	51.255	1.182	11.820	51.255
4	0.978	9.778	61.033			
5	0.881	8.805	69.838			
6	0.848	8.476	78.314			
7	0.636	6.360	84.674			
8	0.580	5.801	90.474			
9	0.548	5.484	95.959			
10	0.404	4.041	100.000			

Extracted Method Principal Component Analysis

Source SPSS Output

is concluded that in measuring through principal component analysis, only 3 factors out of 10 factors are significant with total variance of 51% which highlights a loss of 49% data and needs for further analysis. Further, the extraction value of loading are most equal with Eigen value of the components which also indicate of no loss of data due to the responses as presented from the responses on Likert scale.

Component matrix highlights the results of the each components variance on “Cultivating Customer Peer Groups”. The customers made their exercise on selecting the options on the ten variables. Out of ten factors, only 3 factors have been found those have more significant positive value in column of components. Here, more

Table 63.16 Component matrix on cultivating customer peer groups

		Component		
		1	2	3
D1	Create a sewing circle	0.044	0.413	-0.385
D2	Create an investment club	0.734	0.029	0.183
D3	Create a customer peer group	0.402	-0.671	-0.156
D4	Create high-touch benefits of an enabling service,	0.653	-0.220	0.284
D5	Adopt new behaviors and mind-sets orientation	0.624	-0.275	-0.192
D6	Create capacity for recharging	0.013	0.289	-0.632
D7	Adding more functionality, and ability	0.103	0.695	0.063
D8	Create cultural competence for product consumption	0.671	0.068	0.383
D9	Create Bundle pack for use	0.255	0.272	0.523
D10	Create a hygienic environment around the villages	0.696	0.158	0.187

Extracted Method: Principal Component Analysis

a.3-components extracted

Source SPSS Output

value is marked on D2, D8and D10 and, these are the three most significant causes for better option on “Cultivating Customer Peer Groups”. So, it is concluded that the Effectiveness of Cultivating Customer Peer Groups is highlighted more significantly on D2 (Create an investment club), D8 (Create cultural competence for product consumption) and D10 (Create a hygienic environment around the villages). The other seven factor variables are to be solved for more customer satisfaction and enhancing its efficiency in BoP marketing.

E. Measurement of Satisfaction based on Life Style and Culture

Table 63.17 shows KMO and Bartlett's Test on “Life Style and Culture” for purchase on BoP model. KMO measure of adequacy indicates the proportion of variance in variables on “Life Style and Culture” that may be influenced by underlying 10 factors. Bartlett's test of sphericity measures the correlation matrix that variables are associated for structure detection as the value is 0.852. It highlights that factor analysis is useful with input variables on “Life Style and Culture” on BoP retail.

Table 63.18 shows the factors responses on “Life Style and Culture” in BoP marketing where initial values of all 10 factor components are one and extraction

Table 63.17 KMO and Bartlett's test on life style and culture

	KMO measures of sampling adequacy	0.852
Bartlett's test of sphericity	Approx. Chi-square	1844.786
	Df	45
	Sig	0.000

Source SPSS Output

Table 63.18 Communalities on life style and culture

		Initial	Extraction
E1	Consumer behave as per their need and desire	1.000	0.853
E2	Arrangement of process and activities of people on occasion	1.000	0.846
E3	searching for, selecting special products	1.000	0.820
E4	Purchasing right quality product	1.000	0.712
E5	Using brand preferences	1.000	0.877
E6	Evaluating product and packaging	1.000	0.994
E7	Evaluate on disposing of products and services	1.000	0.863
E8	Retail choice	1.000	0.771
E9	Purchase timing	1.000	0.747
E10	Purchase amount	1.000	0.786

Extraction Method Principal Component Analysis

Source SPSS Output

values show more than 0.5 which means all factors are best fit to data and could be explored in measuring the variances to know the significant factor. Thus, all 6 factors on “Life Style and Culture” are best fit for further measurement.

Table 63.19 shows that total variance of individual factors (ten) and extraction values of sum of square loading values 4 factors are seen in initial Eigen values which are positive and finite values. The process adopted here in principal component analysis with loading value of the component scores on responses from migrated customers which has been presented in last column of Table 63.19. The results are found in extraction value for those 2 factors which are significant among selected

Table 63.19 Total variance explained on life style and culture

Component	Initial Eigen values			Total of extraction squared loadings		
	Total	% of variance	Cumulative %	Total	% of \$Variance	Cumulative %
1	2.744	27.440	27.440	2.744	27.440	27.440
2	1.713	17.133	44.573	1.713	17.133	44.573
3	1.332	13.320	57.893	1.332	13.320	57.893
4	1.179	11.792	69.685	1.179	11.792	69.685
5	0.992	9.922	79.607			
6	0.788	7.882	87.489			
7	0.494	4.939	92.428			
8	0.305	3.052	95.480			
9	0.240	2.402	97.882			
10	0.212	2.118	100.000			

Extraction Method Principal Component Analysis

Source SPSS Output

Table 63.20 Component matrix on life style and culture

		Components			
		1	2	3	4
E1	Consumer behave as per their need and desire	0.603	0.525	-0.221	0.407
E2	Arrangement of process and activities of people on occasion	0.677	0.048	0.224	0.579
E3	searching for, selecting special products	0.549	-0.390	0.586	0.151
E4	Purchasing right quality product	-0.283	0.625	0.447	-0.207
E5	Using brand preferences	0.813	0.217	0.216	-0.143
E6	Evaluating product and packaging	-0.251	0.142	-0.192	0.272
E7	Evaluate on disposing of products and services	0.646	0.432	-0.135	0.490
E8	Retail choice	0.705	-0.078	0.286	0.430
E9	Purchase timing	0.826	-0.365	0.552	0.182
E10	Purchase amount	-0.157	0.707	0.449	-0.245

Extraction Method Principal Component Analysis.

Source SPSS Output

factors in initial Eigen value. It is concluded that in measuring through principal component analysis, only 4 factors out of 10 factors are significant with the variance of 70% which highlights a loss of 30% of data and needs for further analysis. Further, the extraction value of loading is most equal with the Eigen value of the components which also indicate of no loss of data due to the responses as presented from the responses on Likert scale.

Component matrix in Table 63.20 highlights the results of the each components Variance on “Life Style and Culture”. The customers made their exercise on selecting the options on the twelve variables. Out of twelve factors, only five factors are found those have more significant positive value in column of components. Here, more value is marked on E1, E2, E7 and E9, and these are the five most significant causes for better option on “Life Style and Culture”. So, the Effectiveness of Life Style and Culture is highlighted more significantly on E1 (Consumer behave as per their need and desire), E2 (Arrangement of process and activities of people on occasion), E7 (Evaluate on disposing of products and services) and E9 (Purchase timing). The other six factor variables are to be solved for more customer satisfaction and enhancing its efficiency in BoP marketing.

It has been found from Table 63.21 that determining the relative importance of the significant predictor i.e., factor of “total variables on changing business practices of retail malls”, is having a high standardized positive coefficient Beta value. Further, the standard error reveals very low and insignificant values, which can be taken as a significant satisfaction is based on applicability relation management practices in companies. The co-efficient table reported that out of five factors, all factors are mostly significant except two factors, these are C i.e., Pricing Mechanism and E i.e.

Table 63.21 Co-efficients of factors on satisfaction on BoP marketing practices

Model		Unstandardized coefficients		Standardized coefficients		
		B	Std. Error	Beta	t	Sig
1	(Constant)	4.169	0.198	0.000	21.002	0.000
	A	0.055	0.029	0.079	1.924	0.055
	B	-0.003	0.028	-0.004	-0.097	0.923
	C	0.028	0.024	0.048	1.165	0.245
	D	-0.106	0.036	-0.119	-2.919	0.004
	E	0.197	0.015	0.032	3.354	0.001

a. Dependent Variable Satisfaction on BoP marketing practices

Source SPSS Output

Cultivating customer peer groups BoP sale as these two factors revealed negative Beta values.

But among all the significant factors which are showing the positive Beta values, factor E i.e., “Measurement of satisfaction based on life style and culture” is most significant and more elastic as it revealed high B-values, i.e., 0.197 across other factors. Further t-value supports to this. So, these seven variables are more elastic, and Beta (unstandardized) values indicate less value with the change of satisfaction on business practices of retail malls they prefer mostly to purchase for all time.

63.7 Testing of Hypothesis

The following hypotheses have been tested in the present study.

Hypothesis-1: *Motivation on BoP is significantly different across age group of consumers.*

The motivation includes following factors;

1. Physically infra-Attractiveness has an impact on motivation
2. Do you think products available in retail are of best using?
3. Do you think that of perfect brand match?
4. Do you think of target customer-match?
5. Do you think of perfect Price-product match?

63.8 One Way ANOVA: Age

Here, the descriptive analysis in Table 63.22 indicates the results of “Motivation” in BoP marketing with respect to the age of the customers. Five statements on this factor i.e., “Motivation” are incorporated here which were asked to the customers of BoP

Table 63.22 Descriptive analysis of motivation on BoP

		N	Mean	Std. Deviation	Std. Error
A3	Below 20	133	2.195	1.378	0.119
	21–35	254	2.000	1.285	0.080
	36–50	130	1.992	1.197	0.105
	51–65	40	2.050	1.036	0.163
	65 >	40	1.500	0.847	0.133
	Total	597	2.011	1.255	0.051
A5	Below 20	133	3.165	1.457	0.126
	21–35	254	3.606	1.316	0.082
	36–50	130	3.161	1.487	0.130
	51–65	40	2.900	1.549	0.244
	65 >	40	3.475	1.501	0.237
	Total	597	3.355	1.431	0.058

Sources Primary Data and SPSS Output

marketing. The lower Mean Score values indicate much better positive towards the attainment of more effective results on business strategy maintained by BoP retails. It is marked that, customers more than 65 age group in factor A3 show lower Mean score i.e., 1.5000 and in A5 it is 2.900. So, it is concluded that customers of more than 65 age group in A3 (Do you think that of perfect brand match) and customers of 51–65 age group in A5 (Do you think of perfect Price-product match) are more influenced by the “motivation” practices in BoP marketing irrespective to their age.

Table 63.23 reports the results of ANOVA to test the changes on “motivation” by the retail malls in BoP marketing across the age group of the customers. It shows a significant F-statistic of 0.819, indicating the significance value to be 0.514 in factor A3 whereas F statistic is 4.310 and the significance value of 0.002 in A5. Hence, a strong variation has been marked in A3 and in A5 with respect to change in “motivation” across the age groups of consumers. So, the hypothesis taken here was accepted.

Table 63.23 ANOVA on motivation

		Total of squares	Df	Mean square	F	Sig
A3	Between groups	15.108	4	3.777	2.420	0.047
	Within groups	923.810	592	1.560		
	Total	938.918	596			
A5	Between groups	34.543	4	8.636	4.310	0.002
	Within groups	1186.174	592	2.004		
	Total	1220.717	596			

Sources Primary Data and SPSS Output

Hypothesis 2: Cultivating customer peer groups in BoP sale is significantly different across the age groups of consumers.

(1) Cultivating customer peer groups in BoP sale (in five point scale) It includes following factors; Create a sewing circle (2) Create an investment club (3) Create a customer peer group (4) Create high-touch benefits of an enabling service, (5) Adopt new behaviors and mind-sets orientation	(6) Create capacity for recharging (7) Adding more functionality, and ability (8) Create cultural competence for product consumption (9) Create Bundle pack for use (10) Create a hygienic environment around the villages
--	--

63.9 One Way ANOVA: Age

Here, the descriptive analysis in Table 63.24 indicates the results of “cultivating customer peer groups in BoP sale” in BoP marketing with respect to the age of the

Table 63.24 Descriptive analysis on cultivating customer peer groups in BoP sales

		N	Mean	Std. Deviation	Std. Error
E2	Below 20	133	2.398	1.386	0.120
	21–35	254	2.503	1.667	0.104
	36–50	130	3.453	1.570	0.137
	51–65	40	3.475	1.395	0.220
	65 >	40	2.650	1.561	0.246
	Total	597	2.762	1.620	0.066
E8	Below 20	133	4.075	1.449	0.125
	21–35	254	4.283	1.161	0.072
	36–50	130	2.846	1.681	0.147
	51–65	40	2.000	1.300	0.205
	65 >	40	3.125	1.571	0.248
	Total	597	3.693	1.573	0.064
E10	Below 20	133	4.218	0.890	0.077
	21–35	254	4.122	1.016	0.063
	36–50	130	2.915	1.403	0.123
	51–65	40	2.725	1.198	0.189
	65 >	40	3.400	1.127	0.178
	Total	597	3.738	1.247	0.051

Source Primary Data and SPSS Output

Table 63.25 ANOVA on cultivating customer peer groups in BoP sale

		Sum of squares	Df	Mean square	F	Sig
E2	Between groups	117.551	4	29.388	12.026	0.000
	Within groups	1446.674	592	2.444		
	Total	1564.224	596			
E8	Between groups	328.768	4	82.192	42.454	0.000
	Within groups	1146.137	592	1.936		
	Total	1474.905	596			
E10	Between groups	201.699	4	50.425	41.144	0.000
	Within groups	725.537	592	1.226		
	Total	927.236	596			

Source SPSS Output

customers. Ten statements on this factor i.e. “cultivating customer peer groups in BoP sale” have been incorporated here, which were asked to the customers of BoP marketing. The lower mean score values indicate much better positive towards the attainment of more effective results on business strategy maintained by BoP retails. It is marked that the customers of 51–65 age group in factor E8 shows lower mean score i.e. 2.000 than customers of all other factors. So, it is concluded that the customers of 51–65 age group in factor E8 (Create cultural competence for product consumption) are more affected by the “cultivating customer peer groups in BoP sale” practices in BoP marketing with respect to their age in relation to customers of other two significant factor.

Table 63.25 presents the results of ANOVA to measure changes on “cultivating customer peer groups in BoP sale” by retail malls in BoP marketing across the age group of customers. It shows a significant F statistic of E2, E8 and E10 is 12.026, 42.454 and 41.144, respectively, with significance value of 0.000 in all factors. Hence, strong differences arise in all the above factors with respect to change in “cultivating customer peer groups in BoP sale” across the age groups of consumers. *So, the hypothesis taken here was also accepted.*

Hypothesis-3: *Cultivating customer peer groups in BoP sale is similar irrespective of the gender types of consumers.*

Table 63.26 presents the results of ANOVA to test the changes on “cultivating customer peer groups in BoP sale” by the retail malls in BoP marketing across the gender groups of the customers. It shows a significant F-statistic of E2, E8 and E10 is 5.635, 6.123 and 6.252, respectively, with significance value of 0.018, 0.014 and 0.013, respectively. Hence, a strong variation has been marked in all the above factors with respect to change in “cultivating customer peer groups in BoP sale” across the gender groups of consumers. *Hence, the hypothesis taken-3 here was rejected.*

Table 63.26 ANOVA on cultivating customer peer groups in BoP sale

		Sum of squares	Df	Mean square	F	Sig
E2	Between groups	14.675	1	14.675	5.635	0.018
	Within groups	1549.550	595	2.604		
	Total	1564.224	596			
E8	Between groups	15.023	1	15.023	6.123	0.014
	Within groups	1459.881	595	2.454		
	Total	1474.905	596			
E10	Between groups	9.641	1	9.641	6.252	0.013
	Within groups	917.595	595	1.542		
	Total	927.236	596			

Sources Primary Data and: SPSS Output

63.10 Summary of Major Findings and Recommendations

The main finding of the study obtained on the basis of analysis and interpretation of collected primary data have been given in the following paragraphs.

Out of 597 respondents, 254 respondents are found from in the age group of 21–35 followed by 133 respondents in below 20 years age group. 254 respondents are found from in the age group of 21–35 followed by 133 respondents in below 20 age group. Maximum of 195 respondents are married in age group of 21–35 followed by 81 customers in 36–50 age group. Maximum 111 customers earn less than Rs.10,000 in 21–35 age group followed by 92 customers are earning Rs.10,000–15,000 in the same age group as compared to other consumers. Maximum of 116 respondents are qualifying graduation in the age group of 21–35. Most preferred customers come alone followed by coming with their family members.

Two factors out of five have been highlighted more significantly on the analysis of “Motivation”. The two significant factors are A3 i.e. Do you think that of perfect brand match (0.709) followed by A5 i.e. Do you think of perfect Price-product match (0.862). Three factors out of eight are identified significant in the study of “Pricing Mechanism”. The three significant factors are B3 (Innovative Distribution and Communication Strategies), B5 (Use purchase parity base) and B6 (Use high price sensitivity mechanism). Eight factors out of eighteen are marked significant in the study of “Product Characteristics”. The eight significant factors are C2 (Makes wide advertisement for creating awareness on product), C4 (Improve access to retail products), C5 (Creating buying power (low cost product ranges)), C8 (Bundle Pack for cheaper cost), C9 (Refill pack availability), C13 (Pay-per-use” system (laundry system)) and C18 (Nurture the creation of a new market). Three factors out of ten are marked significant in the study of “Cultivating Customer Peer Groups BoP Sale”. The three significant factors are D2 (Create an investment club), D8 (Create cultural competence for product consumption) and D10 (Create a hygienic

environment around the villages). Four variables out of ten are marked as significant in the study of “Life Style and Culture” in BoP i.e. E1 (Consumer behave as per their need and desire), E2 (Arrangement of process and activities of people on occasion), E7 (Evaluate on disposing of products and services) and E 9 (Purchase timing). The hypothesis taken-1 and 2 here was accepted, and the hypothesis taken-3 here was rejected. The marketers should create a customer peer group for its sustainable development. Products should be produced according the suitability of the poor consumers. The marketer also should adopt the low-enough cost approach which will be affordable to the consumers and profitable to it.

References

1. Prahalad, C.K., Hart, S.L.: The fortune at the bottom of the pyramid. *Strategy+Bus Mag.* (26), 273 (2002)
2. Prahalad, C.K.: *The Fortune at the Bottom of the Pyramid: Eradicating Poverty through Profits*. Wharton School Publishing, NJ (2004)
3. Karnani, A.: The mirage of marketing to the bottom of the pyramid: how the private sector can help alleviate poverty. *Calif. Manage. Rev.* **49**(4), 90–111 (2007)
4. Hammond, A.L., Kramer, W.J., Katz, R.S., Tran, J.T., Walker, C.: *The Next four Billion: Market size and Business Strategy at the Base of the Pyramid*. World Resources Institute and International Finance Corporation, Washington DC (2007)
5. Karnani, A.: The bottom of the pyramid strategy for reducing poverty : a failed promise. UN/Desa working paper, pp.1–14. Available at: http://unclef.com/esa/desa/papers/2009/wp80_2009.pdf (2009)
6. Alur, S.: Retailers and new product acceptance in India’s base of pyramid (BoP) markets: propositions for research. *Int. J. Retail Distrib. Manag.* **41**(3), 189–200 (2013)
7. Rabino, S.: The bottom of the pyramid: an integrative approach. *Int. J. Emerg. Mark.* **10**(1), 2–15 (2015)
8. Ansari, S., Munir, K., Gregg, T.: Impact at the bottom of the pyramid: the role of social capital in capability development and community empowerment. *J. Manage. Stud.* **49**(4), 813–842 (2012)
9. Barki, E., Parente, J.: Consumer behavior of the base of the pyramid market in Brazil. *Greener. Manag. Int.* **56**, 11–23 (2010)
10. Ireland, J.: Lessons for successful BOP Marketing from Caracas’ Slums. *J. Consum. Mark.* **25**(7), 430–438 (2008)
11. Kolk, A., Rivera-Santos, M., Rufin, C.: Reviewing a decade of research on the “base/bottom of the pyramid” (BOP) concept. *Bus. Soc.* **20**(10), 1–40 (2013)
12. London, T.: The base of the pyramid perspective: A new approach to poverty alleviation. In: G. T. Solomon (ed.) *Academy of Management Best Paper Proceedings* (2008)
13. London, T.: Making better investments at the base of the pyramid. *Harv. Bus. Rev.* **87**(5), 106–114 (2009)
14. Prahalad, C.K., Hammond, A.: Serving the world s poor profitably. *Harv. Bus. Rev.* **80**(9), 48–59 (2002)
15. Rangan, K., Chu, M., Petkoski, D.: Segmenting the base of the pyramid. *Harv. Bus. Rev.* **89**(June), 113–117 (2011)
16. Subrahmanyam, S., Gomez-Arias, J.T.: Integrated approach to understanding consumer behavior at the bottom of the pyramid. *J. Consum. Mark.* **25**(7), 402–412 (2008)
17. Sánchez, C.M., Schmid, A.S.: Base of the pyramid success: a relational view. *S. Asian J. Global Bus. Res.* **2**, 59–81 (2013)

18. Singh, R., Bakshi, M., Mishra, P.: Corporate social responsibility: linking bottom of the pyramid to market development? *J. Bus. Ethics* **131**(2), 361–373 (2014)
19. Wood, V.R., Pitta, D.A., Franzak, F.J.: Successful marketing by multinational firms to the bottom of the pyramid: connecting share of heart, global umbrella brands and responsible marketing. *J. Consum. Mark.* **25**(7), 419–429 (2008)
20. Muhammad, Y.: Creating A World without Poverty : Social Business and the Future of Capitalism. Public Affairs Publication, New York (2009)

Websites

21. www.imf.org/external/pubs
22. <https://www.turnitin.com/viewGale.asp?oid>
23. <http://www.ece.ucsb.edu/~roy/classnotes>
24. www.iimahd.ernet.in/publications/data/2007
25. www.update.un.org
26. <http://www.un.org/esa coordination/Mirage.BOP>

Author Index

A

- Abdul Aleem, 181
Abhishek Das, 221
Aditya Kunar, 613
Ahmed T. Sadiq, 677
Aishwarya, R., 639
Ajit Kumar Behera, 29
Ali Najam Mahawash Al-Jubour, 687
Aman Kaushal, 243
Amrut Patro, 15
Amulya Raj, 181
Aniket Kumar, 191
Ankita Sahu, 377
Ankit Pasayat, 603
Anukampa Behera, 107
Arta Bandhu Jena, 441, 709
Asma Mohiuddin, 507
Atta-ur-Rahman, 1
Ayan Mukherjee, 697
Ayush Raj, 181
Ayusman Mishra, 119

B

- Barnali Sahu, 95
Bhabani Shankar Mishra, 415
Bhabani Shankar Prasad Mishra, 483, 495
Bhavesh Kumar Behera, 119
Bibek K. Nayak, 107
Bibudhendu Pati, 389
Bichitrana Patra, 467
Bidush Kumar Sahoo, 285
Bijayalaxmi Panda, 389
Biranchi Narayan Nayak, 345
Bishal, J. S., 315
Bishwa Ranjan Das, 433

Biswajit Swain, 211

Brijendra Pratap Singh, 191

C

- Chandani Raj, 95
Chhabi Rani Panigrahi, 389
Ch. Sanjeev Kumar Dash, 29, 141

D

- Danish Raza, 171
Dayal Kumar Behera, 575
Debadutta Nayak, 441
Debahuti Mishra, 191, 697
Debasmita Mohapatra, 243
Debolina Mahapatra, 357
Dhiaa Musleh, 1
Dibya Ranjan Das Adhikary, 131
Dibyasundar Das, 211
Dilip Kumar Bagal, 403
Dillip Ranjan Nayak, 403
Diptendu Sinha Roy, 15
Dipti Pratima Minz, 265

E

- Esaú Villatoro-Tello, 345

G

- Gayatri Mohapatra, 307
Gayatri Nayak, 231

H

- Harini, K. R., 517

Harshit Raj Sinha, 253
 Hima Bindu Maringanti, 433
 Hiranmay Dey, 95

I

Ishani Sengupta, 495
 Itismita Pradhan, 367

J

Jayashree Ratnam, 315
 Jay Bijay Arjun Das, 337
 Julie Palei, 575

K

Kalaiarasi, G., 627
 Kalaivani, J., 649
 Kedar Nath Das, 537, 551
 Kulamala Vinod Kumar, 527

L

Lakshmanan, L., 639
 Lambodar Jena, 265
 Lirika Singh, 315
 Lopamudra Pradhan, 709
 Lubesh Kumar Behera, 273
 Lukman Audah, 667

M

Madhuri Rao, 527
 Maël Fabien, 345
 Maheshwari, M., 627, 639
 Manadeepa Sahoo, 253
 Manaswini Pradhan, 585
 Manesh Kumar Behera, 323
 Manish Rout, 315
 Manosmita Swain, 203
 Manvik B. Nanda, 483
 Md Iqbal, 507
 Mihir Narayan Mohanty, 39, 53, 61, 221, 337
 Minu, R. I., 603, 613
 Mohammed Aftab A. Khan, 1
 Mohanty, J. R., 141
 Mustafa Maad Hamdi, 667

N

Nagarajan, G., 603, 613, 639
 Nandita Nupur, 151

Narendra Kumar Kamila, 527
 Neelamadhab Padhy, 403
 Nehad M. Ibrahim, 1
 Niharika Mohanty, 585
 Nikhil Kumar, 507
 Nikunj Agarwal, 159
 Niladri Sekhar Dash, 433
 Nilesh Nath, 107
 Nishant Niraj, 191

O

Oindrila Das, 367

P

Pachipala Yellamma, 517
 Pallavi Nanda, 171
 Parag Bhattacharjee, 159
 Paresh Baidya, 151
 Parimal Kumar Giri, 71, 467
 Petr Motlicek, 345
 Prabhat Kumar Sahu, 159
 Prabira Kumar Sethy, 567, 575
 Pradeep Kumar Mallick, 403, 495, 585, 697
 Prakash Kumar Sarangi, 285
 Prathima Devadas, 627
 Pratik Dutta, 273
 Pratyush Kumar Sinha, 119
 Premananda Sahu, 285
 Pritikrishna Biswal, 297
 Priyabrat Sahoo, 151
 Pubali Chatterjee, 297
 Pulak Sahoo, 141
 Pushkar Kishore, 231

R

Raghda Salam Al Mahdawi, 657
 Rahul Raj Sahoo, 181
 Rajeev Das, 551
 Rana JumaaSarih Al-Janabi, 687
 Ranjit Kumar Behera, 15
 Rashmi Ranjan Mahakud, 467
 Rashmita khilar, 39, 61
 Renu Sharma, 323, 561
 Riddhi Chatterjee, 357, 367
 Ridhy Pratim Hira, 231
 Rishabh Bhatia, 613
 Ritesh Kumar, 231
 Ritweek Das, 307
 Rohit Kumar, 231
 Ronak Singhania, 649
 Ronit Sahu, 131

S

- Sagar S. De, 71
Sahil Anjum, 507
Sami Abduljabbar Rashid, 667
Sampa Sahoo, 527
Samuka Mohanty, 203
Sanjeeb Kumar Kar, 307
Santi Kumari Behera, 575
Sarat Chandra Nayak, 29
Sarmistha Satrusallya, 53
Sasmita Behera, 243
Saswat Subhadarshan, 107
Satchidananda Dehuri, 29, 71, 141, 467
Satya Prakash Biswal, 345
Satya Ranjan Dash, 345, 357, 367
Saumendra Kumar Mohapatra, 221, 337
Saumya Ranjan Lenka, 561
Sawsan D. Mahmood, 657
Saykat Dutta, 537
Selvi, M., 627
Sghaier Chabani, 1
Shaimaa K. Ahmed, 657
Shantipriya Parida, 345, 357
Shlok Garg, 649
Shrabani Sahu, 243
Shubhrajit Parida, 203
Sidhant Kumar Dash, 567
Simran Sahoo, 297
Sitikantha Mallik, 415
Smriti Nayak, 357
Sobhit Panda, 323
Somnath Mandal, 273
Sonali Goel, 323, 561
Soumya S. Acharya, 151
Srikanta Kumar Mohapatra, 285
Sri Srinivasa Raju, M., 537
Sthita Pragyna Panda, 131
Stuti Snata Ray, 307
Subetha, T., 39, 61

Subhashree Mishra, 507

- Subhra Swetanisha, 575
Subrat Kar, 297
Suchintan Mishra, 253
Sudip Mondal, 171
Sujata Dash, 1
Suji Helen, L., 639
Sumit Pal, 203
Suneeta Mohanty, 415
Supriya Agrawal, 95
Sura Sabah Rasheed, 677
Suryakant Prusty, 211
Sushmidha, S., 517
Swadhin Kumar Barisal, 231
Swati Samantaray, 377

T

- Thangavel, M., 119
Trilok Nath Pandey, 467
Tushar Mitra, 253
Tushar Sharma, 211

U

- Utkarsh Jha, 273

V

- Venkat Narsimam Tenneti, 191
Vijayaraghavan, D., 517
Vishal Anand, 483
Vishnuvardhan, S., 603
Vithya Ganesan, 517

Y

- Yogitha, R., 627, 639