

Samarjeet Borah · Sambit Kumar Mishra ·
Brojo Kishore Mishra ·
Valentina Emilia Balas ·
Zdzislaw Polkowski *Editors*

Advances in Data Science and Management

Proceedings of ICDSM 2021

Lecture Notes on Data Engineering and Communications Technologies

Volume 86

Series Editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

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

More information about this series at <https://link.springer.com/bookseries/15362>

Samarjeet Borah · Sambit Kumar Mishra ·
Brojo Kishore Mishra · Valentina Emilia Balas ·
Zdzislaw Polkowski
Editors

Advances in Data Science and Management

Proceedings of ICDSM 2021



Springer

Editors

Samarjeet Borah
Sikkim Manipal University
Majhitar, Rangpo, Sikkim, India

Brojo Kishore Mishra
GIET University
Gunupur, Odisha, India

Zdzislaw Polkowski
Jan Wyzykowski University
Polkowice, Poland

Sambit Kumar Mishra
Gandhi Institute for Education
and Technology
Bhubaneswar, Odisha, India

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

ISSN 2367-4512

ISSN 2367-4520 (electronic)

Lecture Notes on Data Engineering and Communications Technologies

ISBN 978-981-16-5684-2

ISBN 978-981-16-5685-9 (eBook)

<https://doi.org/10.1007/978-981-16-5685-9>

© 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

Advances in data science and management is a collection of research and analysis works presented in the second international conference on data science and management conducted virtually during February 2021. Contributions of the volume are grouped into four different parts based on the application area.

The first part of the proceedings volume is on machine learning applications for data science. It consists of several state-of-the-work experimental research outcomes on various machine learning applications on data science. The first work is on metaheuristic approach which is being used for generalization of blocks linked with query response. The authors have observed that achieving maximum security measures the block chain technology provides the centralized authentication mechanism prioritizing the individual local block chains. In the next paper, an examination is conducted on a novel approach of big data because of its versatile opportunities and benefits. As currently, voluminous, vast varieties with high velocity data are being produced every day, efficient analysis and use of such data have a great importance for the research community. This follows a work on symmetrical performance assessment of large-scaled data using metaheuristic approach.

The part also contains some machine learning applications on future prediction analysis, gold price prediction, online product review monitoring, prediction in the finance sector and software quality prediction, forecasting of stock market, price of agricultural products, fuel consumption, etc. Few applications were also found on social science issues such as prediction of early childhood development, personality traits in Facebook users of Nigeria and privacy concerns in social networks.

Analysis of biomedical and biological data has been discussed in the second part. Basically, the machine learning is provisioned with the mechanism to learn from labeled data and of course being programmed explicitly. In fact, in many specified applications, supervised learning mechanism is applied because of its association with regression as well as classification techniques. Considering the proper immunizations and routine checkup related to health issues, it is essential to schedule the procedure where there may be chance of risks of accumulation of large biomedical datasets. In such situation, it has to be prioritized on the target groups and to minimize the high-level risks. Therefore, the predictive analysis along with machine learning

as well as multidimensional data mining can be applied to overcome this situation. Also, adoption of the deployed mechanisms along with computational intelligence and predictive analysis can identify the barriers preventing the universal coverage of immunization along with the delay response. Many times, some viral diseases do not show any specific symptoms by which it is really hard to predict manually. As per precautionary measures and to prevent such situations, supervised learning with proper prediction can be applied to observe the symptoms and maximize the rate of accuracy. Now also in the pandemic situation of corona virus (COVID-19), the timely decisions can help deceleration of the spreading of the symptoms at a greater extent. Accordingly, proper planning and prediction are most essential to control the situation. In general perspective, the machine learning along with predictive and behavioral analysis can be helpful toward planning for future policies linked to the asymptomatic symptoms as well as epidemiological data. This volume on data science and management contains several works on analysis of biomedical and biological data.

In normal situation, most of the people are suffering from bronchial asthma which is very sensitive to smoke, air pollution, etc. Though it is very hard to predict manually, but implementing with sensors linked with Internet of Things, the performance can be easily monitored. Of course in addition to this, deep learning can also be helpful to measure the parameters along with the sensitivity factors. Usually, deep learning helps the system to learn by natural happenings. Moreover, it learns to perform classification tasks directly from environment achieving optimum accuracy and beyond human-level performance. In the present situation, enablement of storage of biomedical data is highly essential during predictive modeling as well as enhancement of healthcare paradigms. So, to deal with these large-scaled biomedical data, meta-learning framework should be deployed associated with virtual machines. Also, in the present application, the computational intelligence along with big data analytics can be the great support in the patient health monitoring to enhance the performance even if linked with unstructured datasets exploring the applications and enabling the implementation of resources.

The third part deals with computer fraud and security data analysis. As the Internet is becoming the integral part of the modern society, data and system security issues started playing a major role in the same. New threats are being developed everyday, producing a huge scope of research in the domain. Attackers try to discover new type vulnerabilities in digital data and systems. Several counter measures are also developed to take care of the same. Primarily, intrusion detection system (IDS) is being used to monitor both the internal and external traffics of the network. Several research works on intrusion detection have been included in this volume highlighting the new trend in development. Again, data leakage may play a serious threat to medical data management system, insurance data management system, etc. Application of advanced technology like block chain in security mechanisms may provide more protection to the system.

A good number of research on analysis of image and stream data have been included in the fourth part of the proceedings volume. These include some works on agricultural data or image targeting classification of crop, crop type determination,

prediction of crop yield, leaf disease identification of tea plant, etc. Other works include fatigue detection of vehicle drivers, autonomous vehicles, video analysis and ranking noise restoration etc. One important work on detecting concentration level of a learner on e-learning platform is included in the volume. It is found to be relevant to the current status of the education system.

This volume includes mostly the recent research findings and reports, presented by researchers from various research laboratories, universities and institutions in the domain of data science. It is expected that the findings discussed in this volume will be useful to researchers, professionals and mankind as a whole. The editors of this volume extend deepest appreciations to all the authors, reviewers and specially Springer (the publisher) for making this volume possible.

Majhitar, India
Bhubaneswar, India
Gunupur, India
Arad, Romania
Lubin, Poland

Samarjeet Borah
Sambit Kumar Mishra
Brojo Kishore Mishra
Valentina Emilia Balas
Zdzislaw Polkowski

Contents

Machine Learning Applications for Data Science	
Linguistic Information for Decision-Making Using SVM	3
Ritesh Dash, Dillip Ku. Dash, and Radhe Shyam Panda	
Generalization of Blocks Linked with Query Response Using Meta-heuristic Approach	11
Anil Kumar Mishra, Jyoti Prakash Mishra, and Sambit Kumar Mishra	
Use of Big Data Analysis in Data Management Aspects	19
Suchismita Mishra, Srikanta Patnaik, and Bibhuti Bhusan Mishra	
An Efficient Procedure for Identifying the Similarity Between French and English Languages with Sequence Matcher Technique	29
M. Sree Ram Kiran Nag, G. Srinivas, K. Venkata Rao, Sairam Vakkalanka, and S. Nagendram	
Symmetrical Performance Assessment of Large-Scaled Data Using Meta-heuristic Approach	41
Jyoti Prakash Mishra, Zdzislaw Polkowski, Sambit Kumar Mishra, and Samarjeet Borah	
Examining Data Mining Classification Techniques for Predicting Early Childhood Development in Nigeria	51
Aimufua Ikponmwosa, Narasimha Rao Vajjhala, Sandip Rakshit, and Olumide Longe	
Significance of Machine Learning in Future Prediction Analysis	61
Soumya S. Mohapatra, Bunil Kumar Balabantray, Shiba Ch. Barik, Mousumi Acharya, and Ladu K. Sahoo	
Investigation of Enactment of a Lean Amenity in Joint Provision Centre—A Study of Amenity	75
S. M. Kaviya, R. Jayanthi, and M. Saravanan	

Online Product Review Monitoring System Using Machine Learning	83
Sandra Johnson, J. Madhumathi, R. Aishwarya, and V. Vedha Pavithra	
Gold Price Prediction Using an Evolutionary Extreme Learning Machine	93
Jyoti Prakash Mishra and Smruti Rekha Das	
Stock Market Evidence on Investor's Predispositions Impacting Portfolio Return	101
Sai Rashmi Patra and Shakti Ranjan Mohapatra	
Exploratory Review of Applications of Machine Learning in Finance Sector	119
Sandip Rakshit, Nyior Clement, and Narasimha Rao Vajjhala	
Prediction of Personality Traits in Facebook Users	127
Mamta Bhamare and K. Ashokkumar	
Software Quality Prediction Using Machine Learning	137
Aparna Mohapatra, Saumendra Pattnaik, Binod Kumar Pattanayak, Srikanta Patnaik, and Suprava Ranjan Laha	
Analysis of Stock Market and Its Forecasting	147
Sunil Wankhade, Adarsh Kaul, Sanjana Mohile, and Ruchira Kadamb	
Evaluation of the Technology Acceptance Model for Lean Six Sigma Approach—The Main Study	161
Slawomir Switek, Ludoslaw Drelichowski, and Zdzislaw Polkowski	
Recognition of Compound Characters from Degraded Kannada Documents	177
T. N. Sridevi and Lalitha Rangarajan	
Demand Forecasting and Design Thinking for a New Product Using Neural Networks and Generative Adversarial Networks	189
Shweta Upadhyaya, Prankul Kumar, and S. Ushasukhanya	
An Efficient Method of Predicting the Average Fuel Consumption in Automobiles Using Ensemble Stacking in Python	197
S. Anandamurugan, R. Deenadhayalan, B. Venkatesan, S. Sakthivel, and S. Rajesh	
An Empirical Study of Privacy Concern and Trust in the Decision to Revisit Personalized Social Networking Websites	211
Darshana Desai	
Distribution System Voltage Stability Index Determination with Nature-Inspired Meta-heuristic Cuckoo Search Algorithm	223
M. Sridhar Bhatlu, Satyajit Panigrahy, and Ashwani Kumar Chandel	

Effect of E-training on Employee Performance in IT Industry	233
Bidush Kumar Sahoo, Smruti Rekha Sahoo, Jyoti Prakash Mishra, and Binod Kumar Pattanayak	
Analysing Odisha Turmeric Price and Other Major Turmeric Producing States of India: A Longitudinal Approach	241
Bidush Kumar Sahoo, Rojalin Pani, Bikash Chandra Pattanaik, and Saumendra Pattnaik	
Document Classification Using Genetic Algorithm	253
Samarjeet Borah, Needhi Kumari Singh, Passang Uden Yolmo, Rahul Kumar, and Ranjit Panigrahi	
Analysis of Biomedical and Biological Data	
Prediction of Chronic Kidney Disease by Best Accuracy Using Supervised Classification Machine Learning Approach	265
Diddi Priyanka, Diddi Anusha, T. Anandhi, P. Indria, E. Brumancia, and R. M. Gomathi	
System for Full Immunization Coverage	273
Akanksha Pradhan, Rutuja Patil, Vrushali Alugade, and Varsha Patil	
Soft Computing Techniques to Identify the Symptoms for COVID-19	283
Sujogya Mishra, Aezened Mohmaed, Pradyumna Kumar Pattnaik, Kamalakanta Muduli, and Tunku Salha Tunku Ahmad	
Prediction of COVID-19 Cases in India Using Parametric Curve	295
Gopal Behera and Ashutosh Bhoi	
Nutritional Ingredients Analyzer for Food	303
A. Thilagavathy, Tadavarthi Rishi, Veeram Deepak Reddy, and Sudesh Nimmagadda	
Deep Learning Analysis for COVID-19 Using Neural Network Algorithms	313
V. Vijaya Baskar, V. G. Sivakumar, S. P. Vimal, and M. Vadivel	
Novel Approach to Monitor the Respiratory Rate for Asthma Patients	321
V. G. Sivakumar, S. P. Vimal, M. Vadivel, and V. Vijaya Baskar	
A Systematic Research on Identifying Mental Disorders in Social Networks Using Online Social Media Mining	329
S. Sai Jayanth, Shaik Nakarikanth Abja, and A. Mary Posonia	
Removal of Outliers and Missing Values in Diabetes Dataset Using Ensemble Method	335
M. D. Anto Praveena and B. Bharathi	

Literature Survey: Computational Models for Analyzing and Predicting the Spread of the Coronavirus Pandemic	343
Anubhav Soam, Kapeesh Kaul, and S. Ushasukhanya	
Regression Analysis and Prediction of the COVID-19	349
Santosini Bhutia and Bichitrananda Patra	
Deep Learning Neural Network-Based Pneumonia Classification	363
Anasua Banerjee and Minakhi Rout	
Fuzzy Time Series Model for Forecasting Agricultural Crop Production	371
K. Senthamarai Kannan, K. M. Karuppasamy, and R. Balasubramaniam	
Data Visualization on Breast Phantom Mammogram Images Using Kernel Performance of SVM	385
A. R. Venmathi and L. Vanitha	
Role of Fog-Assisted Internet of Things-Enabled System for Managing the Impact of COVID-19	397
Upendra Verma, Mayank Sohani, Samarjeet Borah, Kapil Kumar Nagwanshi, and Sunil Pathak	
Intelligent Big Data Analytics: A Perspective for IoHT and HealthCare	407
Preetishree Patnaik, Brojo Kishor Mishra, Vivek Jaglan, and Manoj Kumar Sahoo	
Biomedical Data Classification Using Meta-learning: An Experimental Investigation	419
Tapas Ranjan Baitharu, P. K. Bharti, and Subhendu Kumar Pani	
Computer Fraud and Security Data Analysis	
Advanced Blockchain Security for Medical Data	431
G. S. R. P. Prasanth, G. V. Harish, and B. Bharathi	
An IoT-Based Efficient Way of Monitoring Food Quality Management	441
Varsha Patil, Rajesh Kadu, Namrata Patel, and Kranti Bade	
Host-Specific Outlier Detection Using Process Relation Semantics with Graph Mining	449
Binayak Panda and Satya Narayan Tripathy	
Intrusion (Hybrid) Detection System for Cloud Computing Environments	463
G. Nagarajan, R. I. Minu, and T. Sasikala	
Intrusion Detection and Prevention Systems Using Snort	473
Shubham Sharma, Parma Nand, and Pankaj Sharma	

Contents	xiii
Performance Assessment of End-to-End Routing Protocols in Cognitive Radio Ad-Hoc Networks	487
Debabrata Dansana and Prafulla Kumar Behera	
Unstructured Log Analysis for System Anomaly Detection—A Study	497
Anukampa Behera, Chhabi Rani Panigrahi, and Bibudhendu Pati	
Malware Detection System Using API-Decision Tree	511
D. Anil Kumar, Susanta Kumar Das, and Manoj Kumar Sahoo	
ANFIS for Fraud Automobile Insurance Detection System	519
Gopikrishna Panda, Sunil Kumar Dhal, Rabinarayan Satpathy, and Subhendu Kumar Pani	
Image and Video Data Analysis	
Computer Vision-Based Alert System to Detect Fatigue in Vehicle Drivers	533
Jyotsna Rani Thota, B. J. Jaidhan, Mukkamala S. N. V. Jitendra, A. Shanmuk Srinivas, and A. S. Venkata Praneel	
Milestones in Autonomous Vehicle and Evaluation Using Computer Vision	545
J. Pavan Satish, Sai Harsha, T. Prem Jacob, A. Pravin, and G. Nagarajan	
Sentinel-2 Images-Based Intelligent Crop Type Determination	555
L. Sujihelen, N. Naga Praveen Kumar, P. Pramod Sai, and G. Nagarajan	
An Efficient Modeling Based on XGBoost and SVM Algorithms to Predict Crop Yield	565
G. S. Mallikarjuna Rao, Sujani Dangeti, and Shanmuk Srinivas Amiripalli	
Concentration Level of a Learner Using Facial Expressions on E-Learning Platform	575
S. Vijayakumar, Karlapudi Rahul Sai, and Bhumireddy Sohith Reddy	
A Non-negative Matrix Factorization for IVUS Image Classification Using Various Kernels of SVM	585
S. P. Vimal, M. Vadivel, V. Vijaya Baskar, and V. G. Sivakumar	
Tea Plant Leaf Disease Identification Using Hybrid Filter and Support Vector Machine Classifier Technique	591
S. Prabu, B. R. TapasBapu, S. Sridhar, and V. Nagaraju	
A Multidimensional Data Mining Approach for Video Analysis and Ranking System	603
Anjan Dutta, Vaibhav Sinha, Punyasha Chatterjee, Narayan C. Debnath, and Soumya Sen	

Impulse Noise Restoration Using Combined Fuzzy Logic and Adaptive Trimmed Median Filter	613
Priyaranjan Kumar, Aswini K. Samantaray, Anshuman Kashyap, and Chinmayee Biswal	
Author Index	621

Editors and Contributors

About the Editors

Dr. Samarjeet Borah is currently working as Professor in the Department of Computer Applications, SMIT, Sikkim Manipal University (SMU), Sikkim, India. Dr. Borah handles various academics, research and administrative activities such as curriculum development, Board of Studies, Doctoral Research Committee, IT Infrastructure Management etc. in Sikkim Manipal University. Dr. Borah is involved with various funded projects from AICTE (Government of India), DST-CSRI (Govt. of India) etc. in the capacity of Principal Investigator/Co-principal Investigator. He has organized various national and international conferences such as ISRO Sponsored Training Programme on Remote Sensing & GIS, NCWBCB 2014, NER-WNLP 2014, IC3-2016, IC3-2018, ICDSM-2019, ICAET-2020, IC3-2020 etc. Dr. Borah is involved with various book volumes and journals of repute for Springer, IEEE, Inderscience, IGI Global etc. in the capacity of Editor/Guest Editor/Reviewer. He is editor-in-chief of the book/proceedings series—*Research Notes on Computing and Communication Sciences*, Apple Academic Press, USA.

Dr. Sambit Kumar Mishra is having more than 22 years of teaching experience in different AICTE approved institutions in India. He obtained his Bachelor Degree in Engineering in Computer Engineering from Amravati University, Maharashtra, India in 1991, M.Tech. in Computer Science from Indian School of Mines, Dhanbad(Now IIT, Dhanbad), India in 1998 and Ph.D. in Computer Science and Engineering from Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India in 2015. He has more than 29 publications in different peer reviewed international journals and editorial board member of different peer reviewed indexed journals. Presently he is associated with Gandhi Institute for Education and Technology, Baniatangi, Bhubaneswar, Odisha, India.

Dr. Brojo Kishore Mishra is a Professor at Department of Computer Science and Engineering, School of Engineering and Technology, GIET University, Gunupur,

India. He received Ph.D. degree in field of Computer Science from Berhampur University, India in 2012. His main research areas include Data Mining, Machine Learning and Security. Dr. Mishra has published more than 30 papers mainly on peer-reviewed international journals indexed SCOPUS, ESCI, and SCI.

Dr. Valentina Emilia Balas is currently Full Professor in the Department of Automatics and Applied Software at the Faculty of Engineering, “Aurel Vlaicu” University of Arad, Romania. She holds a Ph.D. in Applied Electronics and Telecommunications from Polytechnic University of Timisoara. Dr. Balas is author of more than 350 research papers in refereed journals and International Conferences. Her research interests are in Intelligent Systems, Fuzzy Control, Soft Computing, Smart Sensors, Information Fusion, Modeling and Simulation. She is the Editor-in Chief to International Journal of Advanced Intelligence Paradigms (IJAIP) and International Journal of Computational Systems Engineering (IJCSysE), member in Editorial Board member of several national and international journals and is evaluator expert for national, international projects and Ph.D. Thesis. Dr. Balas is the director of Intelligent Systems Research Centre in Aurel Vlaicu University of Arad and Director of the Department of International Relations, Programs and Projects in the same university. Now she is working in a national project with EU funding support: BioCell-NanoART = Novel Bio-inspired Cellular Nano-Architectures—For Digital Integrated Circuits, 3M Euro from National Authority for Scientific Research and Innovation. She is a member of EUSFLAT, SIAM and a Senior Member IEEE, member in TC—Fuzzy Systems (IEEE CIS), chair of the TF 14 in TC—Emergent Technologies (IEEE CIS), member in TC—Soft Computing (IEEE SMCS).

Dr. Zdzislaw Polkowski is Professor of UJW at Faculty of Technical Sciences and Rector's Representative for International Cooperation and Erasmus+ Program at the Jan Wyzykowski University Polkowice. Since 2019 he is also adjunct Professor in Department of Business Intelligence in Management, Wroclaw University of Economics and Business, Poland. He holds a Ph.D. degree in Computer Science and Management from Wroclaw University of Technology, Post Graduate degree in Microcomputer Systems in Management from University of Economics in Wroclaw and Post Graduate degree IT in Education from Economics University in Katowice. He obtained his Engineering degree in Industrial Computer Systems from Technical University of Zielona Gora. He has published more than 55 papers in journals, 15 conference proceedings, including more than eight papers in journals indexed in the Web of Science. He served as a member of Technical Program Committee in many International conferences in Poland, India, China, Iran, Romania and Bulgaria. He is also the member of the Board of Studies and expert member of the doctoral research committee in many universities in India. He is also the member of the editorial board of several journals and served as a reviewer in a wide range of international journals. His area of interests includes IT in Business, IoT in Business and Education Technology. He has successfully completed a research project of worth 120,000 USD on Developing the innovative methodology of teaching Business Informatics

(www.dimbi.pl) funded by the European Commission. He also owns an IT SME consultancy company in Polkowice, Poland.

Contributors

Shaik Nakarikanth Abja Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Mousumi Acharya Department of Computer Science and Engineering, DRIEMS, Cuttack, Odisha, India

Tunku Salha Tunku Ahmad Faculty of Applied and Human Sciences, Universiti Malaysia Perlis, Kangar, Malaysia

R. Aishwarya Departments of Computer Science and Engineering, R.M.K. Engineering College, Kavaraipettai, India

Vrushali Alugade Department of Computer Engineering, SIES Graduate School of Technology, Navi Mumbai, India

Shanmuk Srinivas Amripalli Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam, Andhra Pradesh, India

S. Anandamurugan Kongu Engineering College, Erode, Tamil Nadu, India

T. Anandhi Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

D. Anil Kumar Berhampur University, Berhampur, Odisha, India;
National Institute of Science and Technology, Berhampur, Odisha, India

M. D. Anto Praveena Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Diddi Anusha Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

K. Ashokkumar Department of Computer Science and Engineering, Sathyabama Institute of Science & Technology, Chennai, India

Kranti Bade Computer Engineering Department, SIESGST, Mumbai University, Mumbai, India

Tapas Ranjan Baitharu SVU, Gajraula, Uttar Pradesh, India

Bunil Kumar Balabantray Department of Computer Science and Engineering, National Institute of Technology Meghalaya, Shillong, India

R. Balasubramaniam Department of Statistics, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu, India

Anasua Banerjee School of Computer Engineering, Kalinga Institute of Industrial Technology Deemed to be University, Bhubaneswar, India

Shiba Ch. Barik Department of Computer Science and Engineering, DRIEMS, Cuttack, Odisha, India

Anukampa Behera Department of Computer Science and Engineering, ITER, S'O'A (Deemed to be), Bhubaneswar, India;

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

Gopal Behera Government College of Engineering Kalahandi, Bhawanipatna, India

Prafulla Kumar Behera Department of Computer Science and Application, Utkal University, Bhubaneswar, Odisha, India

Mamta Bhamare Department of Computer Science and Engineering, Sathyabama Institute of Science & Technology, Chennai, India

B. Bharathi Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

P. K. Bharti SVU, Gajraula, Uttar Pradesh, India

M. Sridhar Bhatlu National Institute of Technology, Hamirpur, HP, India

Ashutosh Bhoi Government College of Engineering Kalahandi, Bhawanipatna, India

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

Chinmayee Biswal C V Raman College of Engineering, Bhubaneswar, India

Samarjeet Borah Sikkim Manipal Institute of Technology, Sikkim Manipal University, Majitar, Rangpo, East Sikkim, Sikkim, India

E. Brumancia Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Ashwani Kumar Chandel National Institute of Technology, Hamirpur, HP, India

Punyasha Chatterjee School of Mobile Computing and Communication, Jadavpur University, Kolkata, India

Nyior Clement American University of Nigeria, Yola, Adamawa, Nigeria

Sujani Dangeti Department of Computer Applications, Gayatri Vidya Parishad College of Engineering (Autonomous), Madhurawada, Visakhapatnam, Andhra Pradesh, India

Debabrata Dansana Department of Computer Science and Application, Utkal University, Bhubaneswar, Odisha, India

Smruti Rekha Das Department of Computer Science and Engineering, Gandhi Institute for Education and Technology, Bhubaneswar, India

Susanta Kumar Das Berhampur University, Berhampur, Odisha, India; National Institute of Science and Technology, Berhampur, Odisha, India

Dillip Ku. Dash Department of Mathematics, CCET, Bhilai, India

Ritesh Dash School of EEE, REVA University, Bengaluru, India

Narayan C. Debnath School of Computing and Information Technology, Eastern International University, Thu Dau Mot, Vietnam

R. Deenadhayalan Paavai Engineering College, Namakkal, Tamil Nadu, India

Darshana Desai Department of MCA, Indira College of Engineering and Management, Pune, India

Sunil Kumar Dhal Sri Sri University, Cuttack, Odisha, India

Ludoslaw Drelichowski WSG University, Bydgoszcz, Poland

Anjan Dutta Techno International NewTown, Kolkata, India

R. M. Gomathi Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

G. V. Harish Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Sai Harsha Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Aimufua Ikponmwosa American University of Nigeria, Yola, Adamawa, Nigeria

P. Indria Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Vivek Jaglan GEH University, Dehradun, Uttarakhand, India

B. J. Jaidhan Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam, AP, India

R. Jayanthi Sathyabama Institute of Science and Technology, Chennai, India

Mukkamala S. N. V. Jitendra Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam, AP, India

Sandra Johnson Departments of Computer Science and Engineering, R.M.K. Engineering College, Kavaraipettai, India

Ruchira Kadam Rajiv Gandhi Institute of Technology, Versova, Mumbai, India

Rajesh Kadu Computer Engineering Department, SIESGST, Mumbai University, Mumbai, India

K. M. Karuppasamy Department of Statistics, Manonmaniam Sundaranar University, Abishekappatti, Tirunelveli, Tamil Nadu, India

Anshuman Kashyap e-Infochips Private Limited, Ahmadabad, India

Adarsh Kaul Rajiv Gandhi Institute of Technology, Versova, Mumbai, India

Kapeesh Kaul Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

S. M. Kaviya Sathyabama Institute of Science and Technology, Chennai, India

Prankul Kumar Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

Priyaranjan Kumar Gandhi Institute for Education and Technology, Bhubaneswar, India

Rahul Kumar Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India

Suprava Ranjan Laha Department of Computer Science and Engineering, ITER, Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India

Olumide Longe American University of Nigeria, Yola, Adamawa, Nigeria

J. Madhumathi Departments of Computer Science and Engineering, R.M.K. Engineering College, Kavaraipettai, India

G. S. Mallikarjuna Rao Department of Computer Applications, Gayatri Vidya Parishad College of Engineering (Autonomous), Madhurawada, Visakhapatnam, Andhra Pradesh, India

A. Mary Posonia Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

R. I. Minu Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India; SRM Institute of Science and Technology, Kattankulatur, India

Anil Kumar Mishra Department of Computer Science and Engineering, Gandhi Engineering College, Affiliated to Biju Patnaik University of Technology, Rourkela, Odisha, Madanpur, Bhubaneswar, India

Bibhuti Bhushan Mishra Faculty of Management Studies (FMS), IBCS, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

Brojo Kishor Mishra GIET University, Gunupur, Odisha, India

Jyoti Prakash Mishra Department of Computer Science and Engineering, Gandhi Institute for Education and Technology, Affiliated to Biju Patnaik University of Technology, Rourkela, Odisha, Baniatangi, Bhubaneswar, India

Sambit Kumar Mishra Department of Computer Science and Engineering, Gandhi Institute for Education and Technology, Affiliated to Biju Patnaik University of Technology, Rourkela, Odisha, Baniatangi, Bhubaneswar, India

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

Sujogya Mishra Department of Mathematics, College of Engineering and Technology, Bhubaneswar, India

Aparna Mohapatra Department of Computer Science and Engineering, ITER, Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India

Shakti Ranjan Mohapatra Faculty of Management, Biju Patnaik University of Technology, Rourkela, Odisha, India

Soumya S. Mohapatra Department of Computer Science and Engineering, DRIEMS, Cuttack, Odisha, India

Sanjana Mohile Rajiv Gandhi Institute of Technology, Versova, Mumbai, India

Aezeden Mohmaed Department of Mechanical Engineering, Papua New Guinea University of Technology, Lae, Papua New Guinea

Kamalakanta Muduli Department of Mechanical Engineering, Papua New Guinea University of Technology, Lae, Papua New Guinea

N. Naga Praveen Kumar Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

G. Nagarajan Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

V. Nagaraju Department of Electronics and Communication Engineering, Rajalakshmi Institute of Technology, Chennai, Tamil Nadu, India

S. Nagendram Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Andhra Pradesh, India

Kapil Kumar Nagwanshi Department of Computer Science & Engineering, Amity School of Engineering & Technology, Amity University Rajasthan, Jaipur, India

Parma Nand Department of Computer Science Engineering, School of Engineering and Technologies, Sharda University, Greater-Noida, India

Sudesh Nimmagadda R.M.K. Engineering College, Gummidi poondi, India

Binayak Panda P.G. Department of Computer Science, Berhampur University, Berhampur, Odisha, India

Gopikrishna Panda Sri Sri University, Cuttack, Odisha, India

Radhe Shyam Panda SSGI, SSTC, Bhilai, India

Rojalin Pani Siksha ‘O’ Anusandhan University, Bhubaneswar, India

Subhendu Kumar Pani Krupajal Computer Academy, BPUT, Rourkela, Bhubaneswar, Odisha, India

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

Ranjit Panigrahi Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India

Satyajit Panigrahy National Institute of Technology, Hamirpur, HP, India

Namrata Patel Computer Engineering Department, SIESGST, Mumbai University, Mumbai, India

Sunil Pathak Department of Computer Science & Engineering, Amity School of Engineering & Technology, Amity University Rajasthan, Jaipur, India

Bibudhendra Pati Department of Computer Science, Rama Devi Women’s University, Bhubaneswar, India

Rutuja Patil Department of Computer Engineering, SIES Graduate School of Technology, Navi Mumbai, India

Varsha Patil Department of Computer Engineering, SIES Graduate School of Technology, Mumbai University, Mumbai, India

Preetishree Patnaik GIET University, Gunupur, Odisha, India

Srikanta Patnaik Department of Computer Science and Engineering, ITER, Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India

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

Sai Rashmi Patra College of Engineering and Technology, Bhubaneswar, India

Bikash Chandra Pattanaik Gandhi Institute for Education and Technology, Bhubaneswar affiliated to Biju Patnaik University of Technology, Rourkela, India

Binod Kumar Pattanayak Department of Computer Science and Engineering, ITER, Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India

Pradyumna Kumar Pattnaik Department of Mathematics, College of Engineering and Technology, Bhubaneswar, India

Saumendra Pattnaik Department of Computer Science and Engineering, ITER, Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India

J. Pavan Satish Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Zdzislaw Polkowski Wroclaw University of Economics and Business, Wroclaw, Poland

S. Prabu Department of Electronics and Communication Engineering, Mahendra Institute of Technology, Namakkal, India

Akanksha Pradhan Department of Computer Engineering, SIES Graduate School of Technology, Navi Mumbai, India

P. Pramod Sai Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

G. S. R. P. Prasanth Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

A. Pravin Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

T. Prem Jacob Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Diddi Priyanka Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

S. Rajesh Paavai Engineering College, Namakkal, Tamil Nadu, India

Sandip Rakshit American University of Nigeria, Yola, Adamawa, Nigeria

Lalitha Rangarajan Department of Studies in Computer Science, University of Mysore, Mysuru, Karnataka, India

Bhumireddy Sohith Reddy Computer Science and Engineering, R.M.K. Engineering College, Chennai, India

Veeram Deepak Reddy R.M.K. Engineering College, Gummidipoondi, India

Tadarvarthi Rishi R.M.K. Engineering College, Gummidipoondi, India

Minakhi Rout School of Computer Engineering, Kalinga Institute of Industrial Technology Deemed to be University, Bhubaneswar, India

Bidush Kumar Sahoo Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;

Gandhi Institute for Education and Technology, Bhubaneswar affiliated to Biju Patnaik University of Technology, Rourkela, India

Ladu K. Sahoo Department of Computer Science and Engineering, DRIEMS, Cuttack, Odisha, India

Manoj Kumar Sahoo National Institute of Science and Technology (Autonomous), Berhampur, Odisha, India;

Biju Patnaik University of Technology Rourkela, Rourkela, Odisha, India;

Berhampur University, Berhampur, Odisha, India

Smruti Rekha Sahoo College of Engineering and Technology, Bhubaneswar, India

S. Sai Jayanth Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Karlapudi Rahul Sai Computer Science and Engineering, R.M.K. Engineering College, Chennai, India

S. Sakthivel Paavai Engineering College, Namakkal, Tamil Nadu, India

Aswini K. Samantaray National Institute of Technology Goa, Ponda, India

M. Saravanan Sathyabama Institute of Science and Technology, Chennai, India

T. Sasikala Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Rabinarayan Satpathy Sri Sri University, Cuttack, Odisha, India

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

K. Senthamarai Kannan Department of Statistics, Manonmaniam Sundaranar University, Abishekappatti, Tirunelveli, Tamil Nadu, India

A. Shanmuk Srinivas Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam, AP, India

Pankaj Sharma Department of Computer Science Engineering, School of Engineering and Technologies, Sharda University, Greater-Noida, India

Shubham Sharma Department of Computer Science Engineering, School of Engineering and Technologies, Sharda University, Greater-Noida, India

Needhi Kumari Singh Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India

Vaibhav Sinha Techno International NewTown, Kolkata, India

V. G. Sivakumar Department of Electronics and Communication Engineering, Vidya Jyothi Institute of Technology, Hyderabad, Telangana, India

Anubhav Soam Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

Mayank Sohani SVKM's NMIMS University, MPSTME Shirpur Campus, Shirpur, Maharashtra, India

M. Sree Ram Kiran Nag Department of Computer Science and Engineering, Andhra University, Vizag, Andhra Pradesh, India

T. N. Sridevi Department of Studies in Computer Science, University of Mysore, Mysuru, Karnataka, India

S. Sridhar Easwari Engineering College, Chennai, Tamil Nadu, India

G. Srinivas Department of Computer Science and Engineering, Gitam University, Vizag, Andhra Pradesh, India

L. Sujihelen Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

Slawomir Switek WSB University, Dabrowa Gornicza, Poland

B. R. TapasBapu Department of Electronics and Communication Engineering, S. A. Engineering College, Chennai, Tamil Nadu, India

A. Thilagavathy R.M.K. Engineering College, Gummidipoondi, India

Jyotsna Rani Thota Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam, AP, India

Satya Narayan Tripathy P.G. Department of Computer Science, Berhampur University, Berhampur, Odisha, India

Shweta Upadhyaya Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

S. Ushasukhanya Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

M. Vadivel Department of Electronics and Communication Engineering, Vidyajyothi Institute of Technology, Hyderabad, Telangana, India

Narasimha Rao Vajjhala University of New York Tirana, Tirana, Albania

Sairam Vakkalanka Department of Computer Science and Engineering, Gitam University, Vizag, Andhra Pradesh, India

L. Vanitha Department of Electronics and Communication Engineering, Prathyusha Engineering College, Chennai, India

V. Vedha Pavithra Departments of Computer Science and Engineering, R.M.K. Engineering College, Kavaraipettai, India

A. S. Venkata Praneel Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam, AP, India

K. Venkata Rao Department of Computer Science and Engineering, Andhra University, Vizag, Andhra Pradesh, India

B. Venkatesan Paavai Engineering College, Namakkal, Tamil Nadu, India

A. R. Venmathi Department of Electronics and Communication Engineering, Kings Engineering College, Chennai, India

Upendra Verma SVKM's NMIMS University, MPSTME Shirpur Campus, Shirpur, Maharashtra, India

V. Vijaya Baskar Department of Electronics and Communication Engineering, School of EEE, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India

S. Vijayakumar Department of CSE, R.M.K. Engineering College, Chennai, India

S. P. Vimal Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu, India

Sunil Wankhade Rajiv Gandhi Institute of Technology, Versova, Mumbai, India

Passang Uden Yolmo Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India

Machine Learning Applications for Data Science

Linguistic Information for Decision-Making Using SVM



Ritesh Dash, Dillip Ku. Dash, and Radhe Shyam Panda

Abstract Extracting parameters by using linguistic variable which is capable of modelling an electrical signal for power system analysis on the different fault condition has been presented in this paper. Different pattern selection, grammar formulation and representing the signal parameter in terms of linguistic variable are the primitive work carried out in this sponsored research. Half cycle signal data from the fault location has been identified as the researchable area. Using discrete wavelet transformation, the parameters were extracted from the original signal, and by using linguistic variable with the help of machine learning, the classification for the different types of fault has been investigated in detail with MATLAB/Simulink software and Python. The proposed work in this research paper has been tested with different types of fault data for checking the robustness of the controller and its logic.

Keywords SVM · Linguistic variables · Gradient descent

1 Introduction

Electrical power system is highly scattered and nonlinear, and therefore, fault in the transmission line is an usual issue. Out of the different types of electrical transmission line fault, the most common type of fault is line to ground fault, and the most severe type of fault is dead short circuit. Proper identification of electrical fault and its isolation from grid safety point of view is a very critical task [1]. All these mentioned issues can be critically addressed by installing a proper relay which could prevent these problems and lead to minimal amount of loss in the transmission line. In these days, microprocessor-based relays were also used in transmission

R. Dash (✉)
School of EEE, REVA University, Bengaluru, India

D. Ku. Dash
Department of Mathematics, CCET, Bhilai, India

R. S. Panda
SSGI, SSTC, Bhilai, India

line. Microprocessor relays provide better security in terms of short time interval calculation and also filter out the noise signal more precisely which makes the relays self-monitoring. In contrast to the above features, microprocessor-based relays were not able to provide speed in terms of signal classification and identification. Therefore, relay having high security will provide slower response as compared to a relay with less security. Highly secured relay is also triggered unintentionally leading to mall operation [2].

Fault recognition can be increased by limiting the analysis of the signal up to half cycle and by changing the implementation strategy. In this research, half cycle analysis has been considered for visualizing and classifying the different quality of fault occurring in an overhead transmission line.

2 Problem Formulation

Problem formulation means designing the mathematical equation for the case under investigation. In SVM, the problem consists of some mathematical objective function and constraints. As observed, linear SVM operates in input space, and nonlinear SVM operates on feature space [3]. However, nonlinear SVM can be solved by applying Kernel function to linear SVM as per Eq. 1

$$x_1^T x_2 \rightarrow K(x_1 x_2) \quad (1)$$

where x_1 and x_2 denote input data point and K represents the Kernel function.

Let us consider there are “ n ” training data set collected during a fault condition, i.e. line to ground fault condition within zone –1 presented in a matrix format “ A ” with text and labelled by +1 or –1. A diagonal matrix θ consists of all elements with all diagonal elements as +1 or –1. Linear Kernel over matrix A has been applied to transfer the data from nonlinear points to linear points. As mentioned earlier, the accuracy of the algorithm is best described by the robustness of the hypothesis [4]. Here, it has been assumed that the hypothesis is a sigmoid type of function bounded by $0 \leq h_\theta(x) \leq 1$. From linear regression

$$h_\theta(x) = \theta^T x \quad (2)$$

Or

$$h_\theta(x) = g(\theta^T x) \quad (3)$$

where g represents the sigmoid function.

Or

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \quad (4)$$

So the cost function for the proposed SVM becomes

$$J\theta = -\log h\theta x, \quad \text{if } y = 1 - \log(1 - h\theta x), \quad \text{if } y = 0 \quad (5)$$

Or

$$J(\theta) = -y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x)) \quad (6)$$

So for “ n ” number of data sets, the cost function becomes

$$J\theta = -1mi - 1ny(i) \log h\theta xi + (1 - yi \log)(1 - h\theta xi) \quad (7)$$

Here $h_\theta(x)$ represents the predicted value and can be represented as a single cost function.

After proper classification, now it is required to apply regression analysis to predict the value of K_p and K_i in the PI controller of the STATCOM for achieving the co-ordinated control action; therefore, the objective function for the regression analysis is shown in Eq. 8.

$$\min_{(\theta, y) \in R^{n+1+l}} y + \frac{1}{2} \|w\|^2 \quad (8)$$

$$\text{s.t. } \theta(Aw - y) + y \geq 1$$

This can be solved by applying the Lagrangian multiplier.

So applying the Lagrangian multiplier and its dual form the equation can be further modified as

$$\min 0 \leq u \leq Rl12uTQu - ITu \quad (9)$$

where u represents the duality of the variable.

After the formulation of a support vector machine problem, it is required to find out the proper solver to solve the optimization problem. The method to use the solver for a particular optimization problem depends upon many factors which can be summarized as below.

Optimization of a particular problem depends upon its properties such as convex or non-convex optimization problems. Generally, the support vector machine algorithm is applied to the convex type of problem, and this is to avoid minima and local minima while solving the optimization problem. It is also important to check whether the problem is a primal or dual, and accordingly constraints have to be identified if present [5].

In addition to the above properties, it is also required to identify whether accuracy is important or not. If accuracy is important, then time factor is also needed or not.

So based on the above two factors, generally the solver has to be identified and applied to find the optimized value. The type of SVM problem that is convex

or non-converts is determined while designing the problem. As mentioned above, solving a convex problem is easier as compared to finding a solution for the non-convex problem, and it is because in convex problem all the minimum is identified to be a global minimum. Again convex optimization problem can be further classified into two categories such as unconstrained convex optimization problem and constrained convex optimization problem. Therefore, it is always a practice to convert unconstrained convex optimization problem into a constrained convex optimization problem by identifying whether it is an equality constraint or inequality constraint or hybrid constraint.

The gradient method for the gradient descent method is generally preferred to solve the unconstrained convex optimization problem. Generally, the unconstrained convex optimization problem is transferred into a constrained convex optimization problem. Large-scale optimization problems are generally solved by using the iterative method. If the complexity in the problem is larger, then finding the solution will take maximum time to get the minima. Therefore, for initiating the iteration process, it is required to find a stationary point in free space, and the stationary point has to be identified in such a manner that the iteration mostly leads to a global minimum or optimized value. First-order optimization problems are very simple in the form of complexity; however, the solution takes a lot of time to find the optimized value, and also the convergence is very slow. In contradiction, the second-order optimization problem is quite complex in computation; however, convergent is very fast, and it is very little time to solve the problem. To find the optimization for a problem, select the stationary point arbitrary and apply a gradient to find the next close to the selected point. The procedure must be repeated until convergence occurs to a stationary point.

The gradient descent method required two other parameters apart from stationary point, i.e. direction and state size. The direction determines in which direction we must search, and the step size determines the length that we can go in that particular direction. Mathematically,

$$x^{(k+1)} = x^k + t_k X \Delta(x_0) \quad (10)$$

Here,

$x^{(k+1)}$ = next stationary point, i.e. $k + 1$

x^k = initial stationary point, i.e. at $t = 0$

t_k = step size

$\Delta(x_0)$ = direction of movement.

As we are applying gradient descent, so we are moving in a negative direction. Therefore, Eq. 10 can be modified as

$$x^{(k+1)} = x^k - \underline{\underline{\text{def}}}(t_k X \Delta(x_0)) \quad (11)$$

In Eq. 11, the step size “ t_k ” determines the convergence of the SVM optimization problem. Therefore, the step size must be identified in such a manner that the new step point is a minima point, or it minimizes the function at that point.

$$tk = \min x \geq 0 f xk - \underline{\underline{def}} tkX \Delta x0 \quad (12)$$

3 Result Analysis

Table 1 shows the test result for a different percentage of training data set. It is found that root means square error remains the same for 60, 65, 70, 75, 80 and 85% of the training data. In this model, the best *R*-squared value is 0.94 for 60% of the training data (Tables 2, 3, 4, 5 and 6).

Again the mean absolute error is about 1% for test sample 60% data and 85% data. And also the time taken by model no. 6 for the training of the data is 26.066 s as compared to 19.977 s for model no. 1. So the overall evaluation of the model is that model no. 1, 6 or 2 is best fitted for the training data. In the present model, all three models have been evaluated for finding the accuracy of the model (Fig. 1).

Table 1 Error evaluation.
60% of the training data set

RMSE	0.03
<i>R</i> -squared	0.94
MSE	0.00
MAE	0.01
Prediction speed	–600,000 obs/s
Training time	19.977 s

Table 2 65% of the training data set

RMSE	0.01
<i>R</i> -squared	0.91
MSE	0.02
MAE	0.02
Prediction speed	–410,000 obs/s
Training time	24.501 s

Table 3 70% of the training data set

RMSE	0.05
<i>R</i> -squared	0.80
MSE	0.00
MAE	0.02
Prediction speed	–630,000 obs/s
Training time	21.929 s

Table 4 75% of the training data set

RMSE	0.03
<i>R</i> -squared	0.89
MSE	0.00
MAE	0.00
Prediction speed	–380,000 obs/s
Training time	21.0166 s

Table 5 80% of the training data set

RMSE	0.05
<i>R</i> -squared	0.77
MSE	0.00
MAE	0.02
Prediction speed	–760,000 obs/s
Training time	1.9795 s

Table 6 85% of the training data set

RMSE	0.01
<i>R</i> -squared	0.93
MSE	0.00
MAE	0.01
Prediction speed	–540,000 obs/s
Training time	26.066 s

4 Conclusion

Predicting the fault performance using linguistic variable in association with support vector machine has been presented in this paper. In this research work, 1700 samples were considered for the proposed algorithm. From the sample space, 80% of the data has been utilized for training purpose, and 20% of the data has been utilized for testing purpose. The model and its algorithm are predicting the performance with 87% of accuracy.

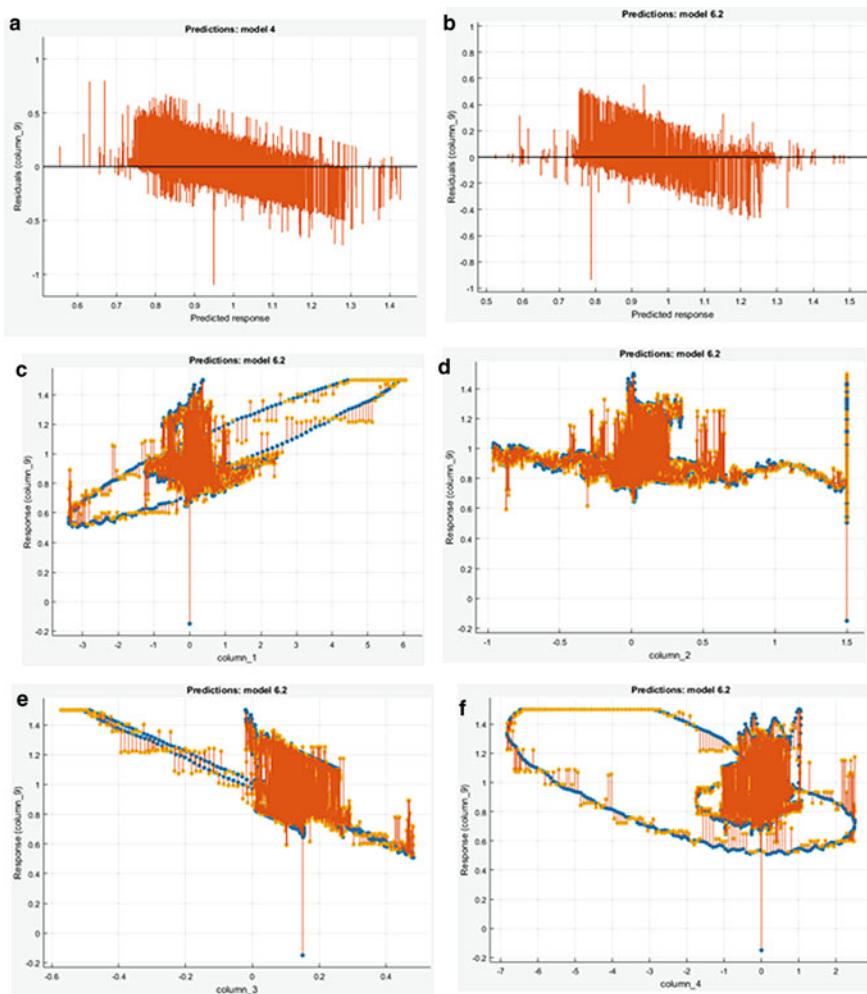


Fig. 1 Prediction of **a** 60% of the training data set. Prediction for model no. 1. **b** 65% of the training data set. Prediction for model no. 2. **c** 70% of the training data set. Prediction for model no. 3. **d** 75% of the training data set. Prediction for model no. 4. **e** 80% of the training data set. Prediction for model no. 5. **f** 85% of the training data set. Prediction for model no. 6

Acknowledgements This project is supported by TEQIP-III of CCSVTU, Bhilai, under collaborative research project sanctioned under CCSVTU Project Ref. No. CRP-II/79.

References

1. Andonovski G, Music G, Blazic S, Skrjanc I (2018) Evolving model identification for process monitoring and prediction of non-linear systems. *Eng Appl Artif Intell* 68:214–221
2. Angelov P, Gu X (2018) Deep rule-based classifier with human-level performance and characteristics. *Inf Sci* 463:196–213
3. Bezerra C, Costa B, Guedes LA, Angelov P (2016) An evolving approach to unsupervised and real-time fault detection in industrial processes. *Expert Syst Appl* 63:134–144
4. Bretas AS, Moreto M, Salim RH, Pires LO (2006) A novel high impedance fault location for distribution systems considering distributed generation. In: IEEE/PES transmission and distribution conference, pp 433–439
5. Dash R, Dash DK, Biswal GC (2021) Classification of crop based on macronutrients and weather data using machine learning techniques. *Results Eng* 9

Generalization of Blocks Linked with Query Response Using Meta-heuristic Approach



Anil Kumar Mishra, Jyoti Prakash Mishra, and Sambit Kumar Mishra

Abstract The term blockchain can be practically viewed as a database as storage and retrieval of information can be performed within blocks. But database cannot be termed as blockchain. In the simplest definition, it is the time stamped series of data along with heterogeneity which is managed by a cluster of systems with security implementing the cryptographic applications. Practically, it uniformly stores the information in blocks, where each block contains the hashed information from the previous block to provide cryptographic security. The coded symbol representing the data links to the desired blocks and process to address the previous blocks. This implementation mechanism is absolutely decentralized supporting peer-to-peer storage of data to all the access points. Accordingly, it captures the prerequisite and relevant information logically structured in blocks. Each block is attached to its preceding blocks by applying the cryptographic generic function where the details of the initial block are required to be focussed. Based upon the non-recursive processing of data, it cannot repeat the data already validated in the block. In general, there are some basic aspects of this technology to ascertain decentralization, maintain transparency as well as immutability and scalability in the system. In case of decentralization, rather having faith on single entity, it is essential to focus on other linked entities and to maintain integrity of data during the application. After that as soon as the data is associated with the blockchain, the information can be stored permanently, and the particular block will be associated with the succeeding blocks to update the

A. K. Mishra

Department of Computer Science and Engineering, Gandhi Engineering College, Affiliated to Biju Patnaik University of Technology, Rourkela, Odisha, Madanpur, Bhubaneswar, India

J. P. Mishra

Gandhi Institute for Education and Technology, Affiliated to Biju Patnaik University of Technology, Rourkela, Odisha, Baniatangi, Bhubaneswar, India

e-mail: jpmishra@gietsbsr.com

S. K. Mishra (✉)

Department of Computer Science and Engineering, Gandhi Institute for Education and Technology, Affiliated to Biju Patnaik University of Technology, Rourkela, Odisha, Baniatangi, Bhubaneswar, India

e-mail: sambitmishra@gietsbsr.com

information. In this paper, it is intended to generalize the response of queries linked with blocks implementing abstraction as well as meta-heuristic approach.

Keywords Block · Meta-heuristic · Query index · Protocol · Streaming · Generalization · Aggregation

1 Introduction

Considering the chaining mechanism, it is understood that the new block usually maintains the procedure from fetching and transact the data focussing on the list of unused data of specific address. The indexing technique in this case maintains all the block information performing the search expansion and responsible towards evaluation of queries. It is also clear that the blockchain technology supports the processing of queries efficiently and maintains the integrity of data. Somehow, the implementation mechanism in some specific applications is bit challengeable.

Many times the protocols linked with this technique seeks the identification of space with objects during processing. As such the mechanism with sequential approach is not so effective to obtain optimality during query processing. In such case, the cryptographic approach can be adopted during propitiating the application with traceable data. Also it will be essential to concentrate on the scale of system implementation as well as support of query system. Sometimes linking with coded keys as well as data generation tends to location of the queries and processed data. So the main challenge here is the frequent updates of the data values. As the authenticated data value captures the entire relevancy during processing, managing the complete transaction of data is really a great challenge. The technology linked with these transactions will initiate from the initial block and updates the position up to final block. Categorizing on multiple databases with heterogeneity and focussing on the specified range of queries, it is observed that the optimized cost of queries depends on the recursive traversal of queries within the range as well as relational operators applied to them. Accordingly, the resultant queries transact linking with the blocks within the specified timestamp.

2 Review of Literature

Casino et al. [1] in their work have focussed on blockchain technology and discussed various impacts associated with decentralization, immutability and transparency capabilities. They have also in their work discussed the developmental issues linked with crypto currencies.

Mammadzada et al. [2] during their research have cited the examples of crypto currencies and observed the significance of permitting individual entity with data

from sensors during communication as well as enabling the crypto currencies towards blockchain applications.

Leible et al. [3] in their study observed the dissimilarities of blockchains linked with scarcity as well as market principles. In this regard, they have also impacted towards principle and proper implementation of blockchain techniques.

Makridakis and Christodoulou [4] in their work focussed the growth of the blockchain visualizing the limitations and implications towards the future of the Internet. Also they prioritized on the disruptive changes linked towards blockchain along with the general characteristics and unique value of blockchain.

Anwar [5] in their work have signified the distributed technology approach linked with data and applications with blockchain. They have also observed the implementation mechanisms of distributed framework as well as synchronization of transactions. They have also found similarities with data in blocks linked chronologically during cryptographically applications.

In [6], many applications linked with blockchain have been discussed. While focussing on timestamps, it is observed that specific data can be accumulated and can be reliably processed concurrently by applying crypto currencies.

Deshpande et al. [7] in their work have prioritized the application of blockchains. As per their survey, the enhancement as well as accessibility of blockchain applications is the major factors towards designing and implementing new redundant interfaces.

In [8], the specific set of rules like ocean protocol is being responsible towards incentivizing the re-use of data implementing the blockchain technology in the system which can be used to link the data with smart real type applications.

Dannen [9] in their work have focussed on smart contracts linked to counteract vulnerabilities. During their research, they have come through the possibilities of implementation of commercial service providers focussing on runtime verification with security.

Vasilev and Stoyanova [10] in their research focussed on financial stability linked to investment of software towards supply chain management. As per the study, it is focussed that the same can be implemented towards transmission of real market information on the supply chain. Basically, the main intention is to provide mechanisms towards information sharing in supply chains with upstream partners.

3 Application of Particle Swarm Optimization Towards Query Response as Well as Streaming the Blocks

In general, the particle swarm optimization is otherwise termed as meta-heuristic approach and is linked with autonomous components. In this case, the particle searches for optimality in the process of application. Somehow, the implementation of the algorithm is based on the particles in the form of data packets. Each data packet will be associated with other relevant information along with the position of

the corresponding particle while initiating the search mechanism. Accordingly, the implementation of specific tasks, as well as clusters can be efficiently carried out implementing the technique.

Algorithm 1: Listing the query response and streaming the blocks

- Step 1: Obtain the query sets and fix the recursion parameter, recn . Initially, $\text{recn} = \text{null}$
- Step 2: Link the query sets with other databases updating the recursion parameter, recn
- Step 3: Obtain the join index value as well as point of split of queries
- Step 4: Perform sorting operation to the queries based on join index value as well as query performance index
- Step 5: Assign the recursion parameter, recn and join index value to blocks
- Step 6: Obtain the updated performance index of queries along with join index values
- Step 7: Simulate the blocks containing updated recn and join index values to obtain optimality.

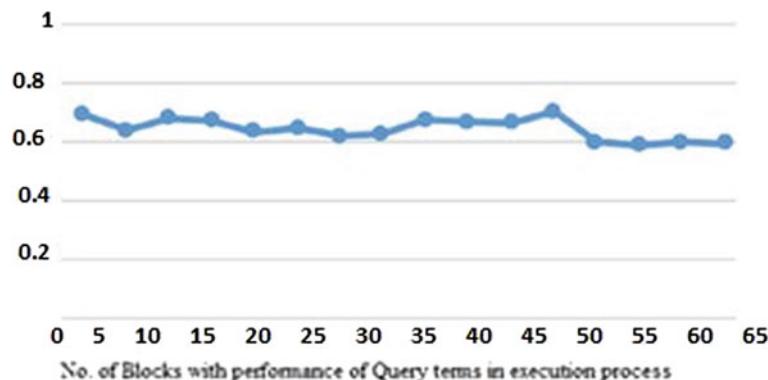
Algorithm 2: Linking blocks towards generalization based on abstraction

- Step 1: Obtain the size of blocks
- Step 2: Update the query performance index linked with each block and evaluate the query execution time during query transact
- Step 3: Find the query specifying nodes with processing elements assigned with each block
- Step 4: Obtain the optimality of each block
- Step 5: Based on query terms and recursion parameter of each block, obtain the query transact response and simulate each block
- Step 6: Generate the updated query evaluation performance index of each block and search for other unconnected blocks to obtain optimized result
- Step 7: Reconfigure again all the linked blocks and apply the aggregation mechanisms.

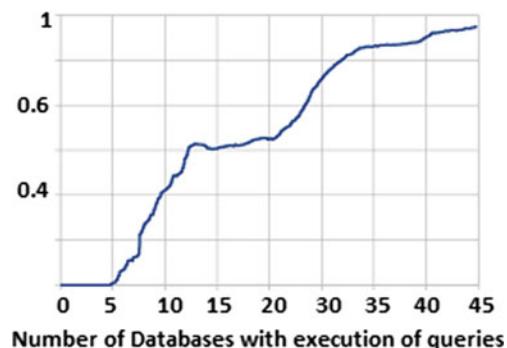
4 Experimental Analysis

Considering Fig. 1 as well as Table 1, it is clearly understood that during process of execution, the query terms are directly dependent upon the number of blocks associated with the system. The performance of query terms during process of execution is based on its retrieval mechanism. Finally, the obtained data are fed to MATLAB 13B for further analysis and result generation (Fig. 2 and Table 2).

As cited in Fig. 3 and Table 3, the consistency of query execution plans is maintained up to certain level. Considering fixed size queries, the efficiency of query plans is enhanced with linkage of more servers in the system.

**Fig. 1** Number of blocks with performance of query terms**Table 1** Performance evaluation of blocks with query term execution

S. No.	No. of blocks	No. of query terms per block	Cost of query terms during execution process (ms)
1	29	55	0.64
2	37	67	0.61
3	67	82	0.63
4	79	91	0.67

Fig. 2 Number of databases with execution of queries**Table 2** Association of databases with query execution

S. No.	No. of associated databases	Cost of execution of queries (ms)
1	19	0.51
2	29	0.82
3	37	0.91
4	40	0.98

Fig. 3 Number of servers with query execution plans

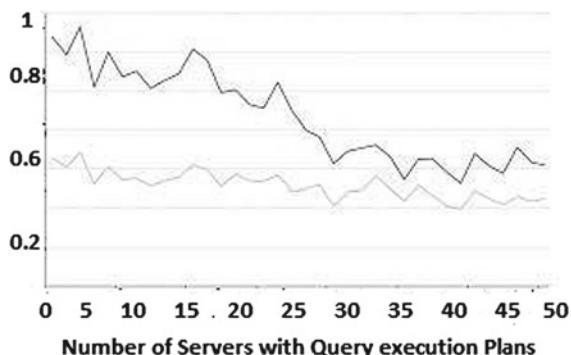


Table 3 Association of servers with query execution plans

S. No.	No. of database servers	Queries associated with each database	Cost of query execution plans
1	20	90	0.81
2	30	90	0.64
3	40	90	0.56
4	50	90	0.46

5 Discussion and Future Direction

While considering the storage of data as well as query processing mechanisms, it is essential to maintain integrity with all the associated blocks, and also, it is required to enhance the efficiency of queries updating the query evaluation performance index parameter. The rules applicable in this case during linking the entities, and particularly, the strong entities should be authenticated with proper identification of keys in the application. With the varying size of queries, the experimental results should focus towards application achieving the effectiveness and obtaining optimal solution in practical situations.

6 Conclusion

Practically, the blockchain technology is provisioned with appropriate solution to enhance the device to device communication security issues and as such it is a great challenge in present situation. The application in such cases should address the issues related to security measures towards internet of things. Also the blockchain technology is deployed to concentrate on the issue linked with decentralized systems. To authenticate and judge each cluster within the blocks as well as to stream each block, the meta-heuristic technique in this case is applied. It is also observed that achieving

maximum security measures the blockchain technology provides the centralized authentication mechanism prioritizing the individual local blockchains.

References

1. Casino F, Dasaklis TK, Patsakis C (2018) A systematic literature review of blockchain-based applications: current status, classification and open issues. *Telemat Informat*
2. Mammadzada K, Iqbal M, Milani F, García-Bañuelos L, Matulevičius R (2020) Blockchain oracles: a framework for blockchain-based applications. In: Book proceedings: business process management: blockchain and robotic process automation forum, BPM 2020 blockchain and RPA forum, Seville, Spain, 13–18 Sept 2020
3. Leible S, Schlager S, Schubotz M (2019) A review on blockchain technology and blockchain projects fostering open science. *Front Blockchain*
4. Makridakis S, Christodoulou K (2019) Blockchain: current challenges and future prospects/applications. *Future Internet*
5. Anwar H (2019) Blockchain vs. distributed ledger technology. Available online at: <https://bit.ly/2SFTRZ0>
6. Werbach K (2018) Trust, but verify why the blockchain needs the law. *Berkeley Technol Law J* 33:487–550. <https://doi.org/10.15779/Z38H41JM9N>, CrossRefFullText|GoogleScholar
7. Deshpande A, Stewart K, Lepetit L, Gunashekhar S (2017) Distributed ledger technologies/blockchain: challenges, opportunities and the prospects for standards. Technical report. The British Standards Institution (BSI)
8. Ocean Protocol (2019) The ocean protocol—a decentralized data exchange protocol to unlock data for AI. Available online at: <https://oceanprotocol.com/tech-whitepaper.pdf>
9. Dannen C (2017) Introducing ethereum and solidity: foundations of cryptocurrency and blockchain programming for beginners, 1st edn. Apress, Berkeley, CA. <https://doi.org/10.1007/978-1-4842-2535-6>
10. Vasilev J, Stoyanova M (2019) Information sharing with upstream partners of supply chains. In: International multidisciplinary scientific geoconference: SGEM, pp 329–336

Use of Big Data Analysis in Data Management Aspects



Suchismita Mishra, Srikanta Patnaik, and Bibhuti Bhushan Mishra

Abstract The concept of big data is spread all over all the sectors of industries. As it is concerning about the most important concept, the data, it need to be managed properly and effectively. By managing the data effectively, it will improve the decision making quality for better business. With the geometrical growth of data, appropriate solutions need to be studied and designed appropriately to handle and extract valuables and knowledge from targeted datasets. The strategic management level needs easy access of valuable insights from such varying and rapidly changing data, which ranges from daily transactions on customer interactions and social network. These values can be extracted using different data mining methods like association rule mining and cluster analysis. This value is provided with big data analytics, which is represented as an application of advanced analytics techniques of big data. The end point use of such finding is supply chain management, risk management fields of different domains.

Keywords Big data analysis · Supply chain management · Risk management · Quality management · Fraud detection · Experimental analysis

1 Introduction

In the present era, we are completely surrounded by data. Even it is impossible to think a day without data. The concept of data is nothing new. It is the organized way of representation about any object. The object could be any organization, event, person, goods, or service.

S. Mishra · S. Patnaik

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

e-mail: srikantapatnaik@soa.ac.in

B. B. Mishra (✉)

Faculty of Management Studies (FMS), IBCS, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

e-mail: bibhutibhusamishra@soa.ac.in

With the advancements in technologies and use of Internet, the use of data become extremely essential and wide. This advancement has come with issues and challenges to store and manage the data. Due to the use of computer and Internet, the amount of data volume get huge and the speed of data generation become rapid. This gave rise to the concept of big data management. The heap of data we can find in every sector existing with management issues. This paper focuses on some of the issues in managing the big data.

2 Characteristics of Big Data

The concept of big data is the representation of data whose scale, diversity, distribution along with timeliness to fulfill the requirement of the use of new technical issues and tools in order to enable insights that discover new sources of business values. Three main features those characterize big data are: volume, variety, and velocity, popularly known as the three V's. The volume of the data means its size showing how enormous it is. Velocity refers to the rate with which data is getting changed, or how innovatively it can be created. Finally, the third V, the variety includes the different forms and types of data, along with the different kinds of uses and ways of analyzing the data [9].

Data volume is the primary issue of big data. Big data is quantified by size in terms of TBs or PBs. It also measured by the number of records, transactions, tables, or files present in the dataset. One of the most versatile characteristics of big data due to which it is a combination of its various and vast sources. The data is gathered from a large and different variety of sources including, data logs, click streams, and social media.

The sources of data consist of unstructured data, such as text and human language, and semi-structured data, such as extensible markup language (XML) or rich site summary (RSS) feeds and structured data such as databases. There is also data, which is hard to categorize because it comes from audio, video, and other devices known as mixed data. Also, multi-dimensional data can be drawn from a data warehouse to add historical data.

3 Big Data Storage and Management

The most important concern at the time of big data adoption is safety storage of data. In the traditional methods of structured data, storage and retrieval include relational databases, data marts, and data warehouses. Generally, the data stored in repositories got collected from operational databases by various ETL tools (extract, transform, load tools). ETL tool extracts data from external data source; afterward, it transforms them to proper operational format as per requirement. Then, finally, load it into the repositories. Likewise, in the further stages, the data get cleaned, transformed into

appropriate formats, and cataloged to be available for mining and various online analysis operations [3].

The big data environment is also known as magnetic, agile, and deep (MAD) analysis skills. It can also be seen different from many aspects of a traditional enterprise data warehouse (EDW) environment. The first difference we can see that the traditional EDW approaches do not exhibit compatibility with incorporation of new data sources unless they are made available in cleaned and integrated format as the existing one. Due to this messy nature of data, big data environments should be magnetic and attracting all types of data sources, irrespective of data quality [5].

Along with growing numbers of data sources day by day, and sophisticated data analysis requirements, big data storage must allow analysts to effectively produce as well as adapt data rapidly. This characteristic has given rise to an agile database, having logical and physical contents in the flexibility to adapt in sync with rapidly evolving data [11]. As most of the current data analysis methods use complex statistical methods, the data analysts should efficiently handle enormous datasets by drill up and drill down approaches. Also, the big data repositories need to serve as a sophisticated algorithmic runtime engine [5].

The non-relational database, for example, not only SQL (NoSQL), has been developed for the purpose of storage and management of unstructured data. The main aim of NoSQL databases are massive scaling, data model flexibility and simplified application development, and deployment. The NoSQL databases have a clear separation between data management and data storage. The databases focus on the high-performing scalable data storage. It also allows data management tasks to be handled in the application layer.

Simultaneously, the in-memory databases manage the data in server's executing memory by eliminating the disk input/output (I/O) and enabling real-time responses from the data source.

Again, the in-memory databases are gaining popularity due to its advanced analytics on big data, specifically in terms of speedup the access in scoring of analytic models for analytical tasks. This gives rise to scalability factor of big data and speed for discovery analytics [17].

For all such purposes, a special framework named Hadoop used to perform big data analytics and provide reliability, scalability, and manageability by implementation of the MapReduce paradigm on the databases. This Hadoop framework consists of two major components which includes

- (1) the HDFS for the big data storage and
- (2) MapReduce for big data analytics [9].

The HDFS storage function provides a redundant and reliable distributed file system. It also provides methods for optimize large files by splitting a single file into blocks and distributed across cluster nodes.

In addition, the data is protected among the nodes by a replication mechanism by ensuring the availability and reliability despite any node failures [3] (Figs. 1 and 2).



Fig. 1 Flow of big data analysis

Fig. 2 Different aspects of big data



4 Supply Chain and Performance Management

In SCM, the concept of big data analytics is used in forecasting demand changes matching with supply accordingly. This can increase benefits in manufacturing, retail, and logistics industries. The stock data analysis will lead the utilization of geospatial data on deliveries. It will also help organizations to automate and replenishment decisions as well as reduce times and optimize costs and process interruptions.

In addition, alternate pricing scenarios can execute instantly, which may result in a reduction in inventories and increase the profit margins [4]. Big data can identify the root causes of cost increment and provide methods for better planning and forecasting [17]. In other fields where big data analytics can be proved to be valuable is performance management. Here, the government sectors and healthcare industries can easily find benefits. With the gradually increasing need of improvement in productivity, human resource performance information can be easily monitored and forecasted by applying predictive analytics tools. This allows departments to be linked with their strategic objectives along with the service. Also, the availability of big data and performance information and its accessibility to operational managers, the use of predictive KPIs, balanced scorecards, and dash boards within the organization

can introduce operational benefits by implementing the monitoring of performance system. It also improves transparency, goal setting, and planning and management functions [4].

5 Quality Management and Improvement

Being specific in terms of manufacturing, energy, and utilities and telecommunications industries, big data vastly get used for quality management in order to maximize profitability and minimize costs by improving the quality of goods and services. This will help in reduction in scrap rates by identifying any interruption in the production process before they happen can save significant expenditures [4]. The big data analytics can result in manufacturing improvements also [17]. It also leads the real-time data analyses by monitoring the machine logs, enabling managers to make swifter decisions for quality management.

Big data analytics also allows real-time monitoring of network demand to forecast the bandwidth in order, response to customer behavior. Also, healthcare and IT systems can improve the efficiency and quality by communicating and integrating patient data across different departments with preserving privacy controls [4]. With the incrementing use of electronic health records, and the advancements in analytics tools, an opportunity arises to mine the available de-identified patient information for assessing the quality of healthcare [22].

Big data can be used for better understanding of changes in the location, frequency, and intensity of weather and climate. This can benefit citizens and businesses that rely upon weather, such as farmers, as well as tourism and transportation companies. Also, with new sensors and analysis techniques for developing long-term climate models and nearer weather forecasts, weather-related natural disasters can be predicted, and preventive or adaptive measures can be taken beforehand [22].

6 Risk Management and Fraud Detection in Big Data

In industries like investment banking or retail banking, insurance sector is beneficial from the point of big data analytics in the concern of risk management. The evaluation of risk is a critical aspect in financial services sector. Big data analytics may prove helpful in churning out the investments produced by analysis of the likelihood of gains versus the likelihood of losses in case of the internal and external big data analysis, the complete and dynamic appraisal of risk exposures [4]. The big data can be useful for organizations enabling the quantification of risk factors [17]. This works as an aid in risk mitigation due to comprehensive view of different types of risk and existing interrelations is provided to decision makers [4].

From the point of view of fraud detection, specifically in the fields like government, banking, and insurance industries, this big data approach is vastly used in detection

and prevention of fraudulent cases [17]. Data analytics is in common practice in automated fraud detection mechanism. But most organizations are looking towards harnessing the existing potentials of big data to improve their system efficiency. Also, big data allows the systems to match all electronic data across multiple sources, including both public and private sector data, and perform faster analytics result [4].

The customer intelligence is also get used to model general customer behaviors and detect suspicious or diversified activities through the mechanism of accurately flagging of outlier occurrences. By providing systems enabled with big data facilities, is about to prevailing fraudulent patterns, which may allow these systems to learn the new patterns of frauds and act accordingly. At the same time, SNAs are used to identification of the networks after collaborating fraudsters and discover evidence of fraud activities [4]. In this way, the big data tools, techniques, and governance processes increase the prevention and recovery of fraudulent transactions by dramatically increasing the speed of identification and detection of compliance patterns within all available datasets [22].

7 Experimental Analysis

Data mining is one of the techniques popularly used for big data analysis. In data mining, there are many different techniques are available, like association analysis, cluster analysis, etc. Again in each, there are many algorithms are available for getting the result.

Choosing the association analysis technique, there is the famous Apriori algorithm to churn out the future relationships existing among products. But the traditional Apriori algorithm is not so efficient to accommodate dynamically growing data optimally.

In this paper, a new algorithm is proposed to improve the efficiency of the association analysis in a transactional dataset. In our proposed algorithm, Apriori generation method is used to find out the frequency of the subsets. Here, we have used telecommunication dataset for testing purpose.

7.1 Proposed Algorithm

Input-itemset I = {i₁, i₂, ... in}

Transaction database TD

Minsup, minconf

Output—frequent itemset F

Algorithm:

Step-1

 Initialize itemset, I – 1, minsup, minconf, F

Step-2

For $n = 2$; $F_n - 1 \neq \emptyset$; $n++$
 $C_n = \text{aprioriGen}(F_{n-1})$, for all transactions $t \in TD$
 $C_t = \text{subset}(C_n, t)$
 For all candidate set $c_n \in C_n$ increment c_n
 $C_n = \{c_n \in C_n : c_n \geq \text{minsup}\}$

Step-3
 Un F_n for all $n = 1, 2, \dots, n$
 To generate Rules
 Input—transactional dataset TD , attribute set C , consequence set S , minsup, minconf
 Output—rule set R .
 Algorithm:
 Step-1
 Initialize $R(C)$, minsup, minconf
 Step-2
 Take input the TD
 Reduce S to adjust with minsup and minconf
 Step-3
 For all $S_i \in S$ where $i = 1, 2, \dots, n$
 If $S_i \in R(S)$
 Remove S_i from TD
 $S = S - \{S_i\}$
 Step-4
 Using apriori find the frequencies of itemset I
 Step-5
 Generate rule $F_i - 1$

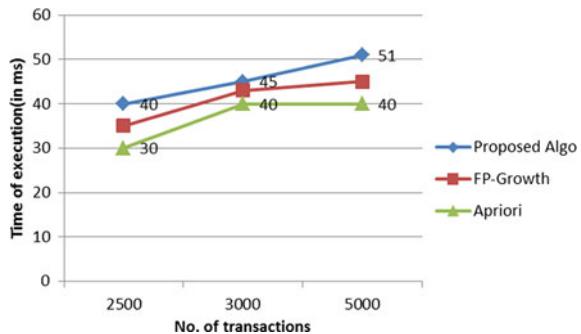
In this algorithm, the rule set R provides the targeted association relation between attribute set and consequence set. The attribute set C consists of properties or characteristics of the products. The consequence set S consists of expected combination of products with C . This algorithm provides improved result in comparison to traditional rule generation algorithm.

7.2 Comparison Analysis

To determine the efficiency of the new proposed algorithm, let's look at the execution performance. We took the execution time and the number of transactions to depict the analysis. Here, we take two existing methods, namely Apriori and FP-Growth to compare with the proposed algorithm. In the graph, the performance of each algorithm is represented (Fig. 3).

In order to determine the efficiency of the proposed method, it is necessary to compare its execution with the existing methods. We also considered the number of database passes to measure the performance. When the minimum support varies from high support (1000, 1%) to low support (40, 0.04%), the run time of Apriori is

Fig. 3 Output of the comparison



58% versus 24% of that of new proposed algorithm, and is 26% versus 15% of that of existing Apriori algorithm. For the Cereals database, when the minimum support varies from high support (2000, 25%) to low support (600, 7.5%), the run time of FP-Growth is 13% versus 23% of that of proposed method, and is 18% versus 1.8% of that of Apriori. The performances of Apriori-based algorithms mainly depend on the number of candidate itemsets and the efficiency of calculating intersection. The dataset are collected from Extended Bakery Datasets, the Repository (<https://wiki.cse.calpoly.edu/datasets/wiki/apriori>). As per the observation, the methods of generating candidate itemsets in Apriori and FP-Growth are nearly the same. In Table 1, we found the number of candidate set generated basing upon the support value.

In the performed experiment, we compared Apriori, one of the most popular association rule mining algorithm, and FP-growth algorithm, an efficient association rule mining algorithm, with our proposed algorithm. First, the proposed algorithm has been tested by using the Cereals dataset, Cosmetics dataset and Wine dataset. Then by using the same cereals dataset, cosmetics dataset and wine dataset, the said algorithm has been compared with Apriori and FP-growth. The comparison has done between the number of transactions get executed in time duration taken in terms of milliseconds. The analysis has been carried out using the weka7.2 software. The performance is determined by calculating number of transactions are processed in milliseconds. From this study, we derive that the proposed algorithm performs better than existing algorithms.

Table 1 The numbers of candidate itemsets for mining dataset

Algorithm methods	Cereals dataset (Nos.)	Cosmetics dataset (Nos.)	Wine dataset (Nos.)
<i>Min_sup (Standard)</i>	2500	3000	5000
Apriori	595	683	118
FP-growth	555	623	105
Proposed method	535	587	985

8 Conclusion

In this research, we have tried to examine a novel approach of big data because of its versatile opportunities and benefits. In this information era, our life is emerged in huge voluminous, vast varieties with high velocity data are being produced every day. For numerous benefits, big data analytics is being applied to leverage business changes, enhancing decision making.

Nowadays many new technologies are getting applied in various fields. The big data approach can reveal several benefits and innovative ideas, which shall result in remarkably improvement in concern field with a bright future plan if approached properly. Simultaneously, this big data approach is very difficult to implement and maintain. Its requirements include proper storage of data with security, data management, system, process integration and analyzing, etc. With all these problems were being faced with traditional data management. The big data has exponentially increased such difficulties due to drastically addition of volume, velocity, and varieties of data and sources which need to be dealt with intelligently. Therefore, further research work can focus to provide a perfect roadmap or framework in the direction of big data management which definitely encompass the previously stated difficulties.

The big data analytics has a great significance in this state of data overflow and can provide unforeseen insights. This may prove to be beneficial for decision makers in various sectors. With proper exploiting data and application, the big data analytics can provide potential to a strong basis for advancements on the scientific, technological, and humanitarian levels.

References

1. Adams MN (2010) Perspectives on data mining. *Int J Mark Res* 52(1):11–19
2. Asur S, Huberman BA (2010) Predicting the future with social media. In: ACM international conference on web intelligence and intelligent agent technology, vol 1, pp 492–499
3. Bakshi K (2012) Considerations for big data: architecture and approaches. In: Proceedings of the IEEE aerospace conference, pp 1–7
4. Cebr (2012) Data equity, unlocking the value of big data. In: SAS reports, pp 1–44
5. Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C (2009) MAD skills: new analysis practices for big data. *Proc ACM VLDB Endow* 2(2):1481–1492
6. Cuzzocrea A, Song I, Davis KC (2011) Analytics over large-scale multidimensional data: the big data revolution! In: Proceedings of the ACM international workshop on data warehousing and OLAP, pp 101–104
7. Economist Intelligence Unit (2012) The deciding factor: big data & decision making. In: Capgemini reports, pp 1–24
8. Elgendi N (2013) Big data analytics in support of the decision making process. MSc thesis, German University in Cairo, p 164
9. EMC (2012) Data science and big data analytics. In: EMC education services, pp 1–508
10. He Y, Lee R, Huai Y, Shao Z, Jain N, Zhang X, Xu Z (2011) RCFFile: a fast and space efficient data placement structure in MapReduce-based warehouse systems. In: IEEE international conference on data engineering (ICDE), pp 1199–1208

11. Herodotou H, Lim H, Luo G, Borisov N, Dong L, Cetin FB, Babu S (2011) Starfish: a self-tuning system for big data analytics. In: Proceedings of the conference on innovative data systems research, pp 261–272
12. Kubick WR (2012) Big data, information and meaning. In: Clinical trial insights, pp 26–28
13. Lee R, Luo T, Huai Y, Wang F, He Y, Zhang X (2011) Ysmart: yet another SQL-to-MapReduce translator. In: IEEE international conference on distributed computing systems (ICDCS), pp 25–36
14. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. In: McKinsey global institute reports, pp 1–156
15. Mouthami K, Devi KN, Bhaskaran VM (2013) Sentiment analysis and classification based on textual reviews. In: International conference on information communication and embedded systems (ICICES), pp 271–276
16. Plattner H, Zeier A (2011) In-memory data management: an inflection point for enterprise applications. Springer, Heidelberg
17. Russom P (2011) Big data analytics. In: TDWI best practices report, pp 1–40
18. Sanchez D, Martin-Bautista MJ, Blanco I, Torre C (2008) Text knowledge mining: an alternatives to text data mining. In: IEEE international conference on data mining workshops, pp 664–672
19. Serrat O (2009) Social network analysis. *Knowl Netw Solut* 28:1–4
20. Shen Z, Wei J, Sundaresan N, Ma KL (2012) Visual analysis of massive web session data. In: Large data analysis and visualization (LDAV), pp 65–72
21. Song Z, Kusiak A (2009) Optimizing product configurations with a data mining approach. *Int J Prod Res* 47(7):1733–1751
22. TechAmerica (2012) Demystifying big data: a practical guide to transforming the business of government. In: TechAmerica reports, pp 1–40
23. Van der Valk T, Gijsbers G (2010) The use of social network analysis in innovation studies: mapping actors and technologies. *Innov Manag Policy Pract* 12(1):5–17
24. Zeng D, Hsinchun C, Lusch R, Li SH (2010) Social media analytics and intelligence. *IEEE Intell Syst* 25(6):13–16
25. Zhang L, Stoffel A, Behrisch M, Mittelstadt S, Schreck T, Pompl R, Weber S, Last H, Keim D (2012) Visual analytics for the big data era—a comparative review of state-of-the-art commercial systems. In: IEEE conference on visual analytics science and technology (VAST), pp 173–182

An Efficient Procedure for Identifying the Similarity Between French and English Languages with Sequence Matcher Technique



M. Sree Ram Kiran Nag, G. Srinivas, K. Venkata Rao, Sairam Vakkalanka, and S. Nagendram

Abstract Through this research article, we have compared multi-lingual document which are used in plagiarism detection, bilingual lexicon extraction and NLP application. The comparison of multi-lingual text is done using a parallel corpus which is a collection of aligned sentences and sentences which are translation of each other. Here, we have used sequence matcher in order to retrieve the similarity of sentences, lexicons, and words in a multi-lingual content. The technique sequence matcher is found to perform better over the other existing techniques.

Keywords NLP · Plagiarism · Sequence matcher

1 Introduction

In artificial intelligence, natural language processing is one of the significant areas. NLP plays a major role in carrying the ability to make machines comprehend human language [1]. Machines can realize and extract patterns from text data as NLP applies various techniques like text similarity and encoding.

Text similarity is used to find out the nearness between two chunks of text by its meaning which is one of the important methods of NLP. “Bag of Words,” “TF-IDF,” and “word2vec” are the techniques are being used to encode the text data [2, 3]. The

M. Sree Ram Kiran Nag (✉) · K. Venkata Rao

Department of Computer Science and Engineering, Andhra University, Vizag, Andhra Pradesh, India

G. Srinivas · S. Vakkalanka

Department of Computer Science and Engineering, Gitam University, Vizag, Andhra Pradesh, India

e-mail: sgorla@gitam.edu

S. Vakkalanka

e-mail: svakkala@gitam.edu

S. Nagendram

Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Greenfields, Vaddeswaram, Andhra Pradesh 522502, India

process of encoding allows to find the similarity between sentences from the database. The major steps involved in the text similarity with NLP are: 1. Preprocessing of Text, 2. Feature Extraction, 3. Vector similarity, and 4. Decision function.

2 Literature Survey

Fung and Cheung noisy-parallel, comparable, and quasi-comparable non-parallel corpora are the three levels proposed by the authors. Many parallel sentences in the same order in the texts in noisy-parallel corpora. Texts have topic-related documents and are not translations of each other in comparable corpora. The bilingual documents that are not necessarily related to each other contained by quasi-comparable corpora. The researchers are more interested because of extraction of parallel texts can be done with comparable corpora [4].

Langlois et al. compared documents which are not necessarily translations of each other for categorizing which are mined from dissimilar sources and written in dissimilar languages. They collected data in French, Arabic, and English. With the available hyperlinks for Wikipedia and Euro news two corpora are built [5, 6].

Rao et al. two techniques were implemented one is Levenshtein distance and the second is sequence matcher in this work to check the similarity between math keywords [7]. The experimental study presented based on the time of retrieval and accuracy. True positives and true negatives are used to find the exact retrieval of wanted and unwanted math formulae.

3 Proposed Methodology

As long as the sequence elements are comparable, for matchup of such duos of successions, the more flexible class is sequence matcher [8, 9]. The fundamental idea with the proposed method is to find the best matching math formulae. Sequence matcher retrieves more related and very less number of unwanted math keywords. 0 and 1 are the values used to compares two strings with Sequence Matcher. After comparison, if the ration is greater than 0.8, then the word will be considered as a key word and will be appended to dataset [10, 11].

4 Methodology

1. Take a French document dataset 1.1. Load the dataset and read 1.2. Remove the special characters from document 1.3. Remove French stop words 1.4. Convert the document to list of words 1.5. Find French synonyms to each of the French word in list using Natural Language Processing 1.6. Then find the corresponding English synonyms from each French word 1.7. Prepare the final list which has all the English synonyms of French list	1. Take an English document dataset 1.1. Load the dataset and read 1.2. Remove the special characters from document 1.3. Remove all English stop words 1.4. Convert the paragraph to list of words 1.5. Find the English synonyms for each English word using Natural Language Processing 1.6. Prepare the final list which has all English synonyms
2. Load the two final lists of step 1 and step 2	
3. Apply sequence matcher model to find SIMILARITY between two documents	
3.1. Take a value as threshold for each model while calculating similarity 3.2. Compare the list 1 with list 2 and remove all the words which are matched with the words in English list 2 3.3. For the left over words check semantic closeness and if it satisfies threshold remove the words. Like perform for all until we get final left over list	
4. To find total similarity check between documents define a formula	
$100 - \left(\frac{\text{finalLeftOverEnglishListLength}}{\text{OriginalEnglishLength}} \right) * 100$	

5 Flow Chart

See Fig. 1.

6 Results and Discussions

Accuracy is the metric used to find out the similarity between the languages French and English. As shown in Tables 1, 2, 3, 4, 5, 6 and 7. The accuracy is measured by considering French and English documents in one-to-one, one-to-many, and many-to-many mapping patterns, and the results are produced with the help of the accuracy metric, and it is clear from the tables that with the presented sequence matcher-based language similarity produced the accuracy values in the range between 56 and 81%. The total experimentation is done on nearly 100 French and English documents. The process of finding the similarity was performed with help of algorithm shown in Fig. 1. The purpose of using sequence matcher technique for language similarity is

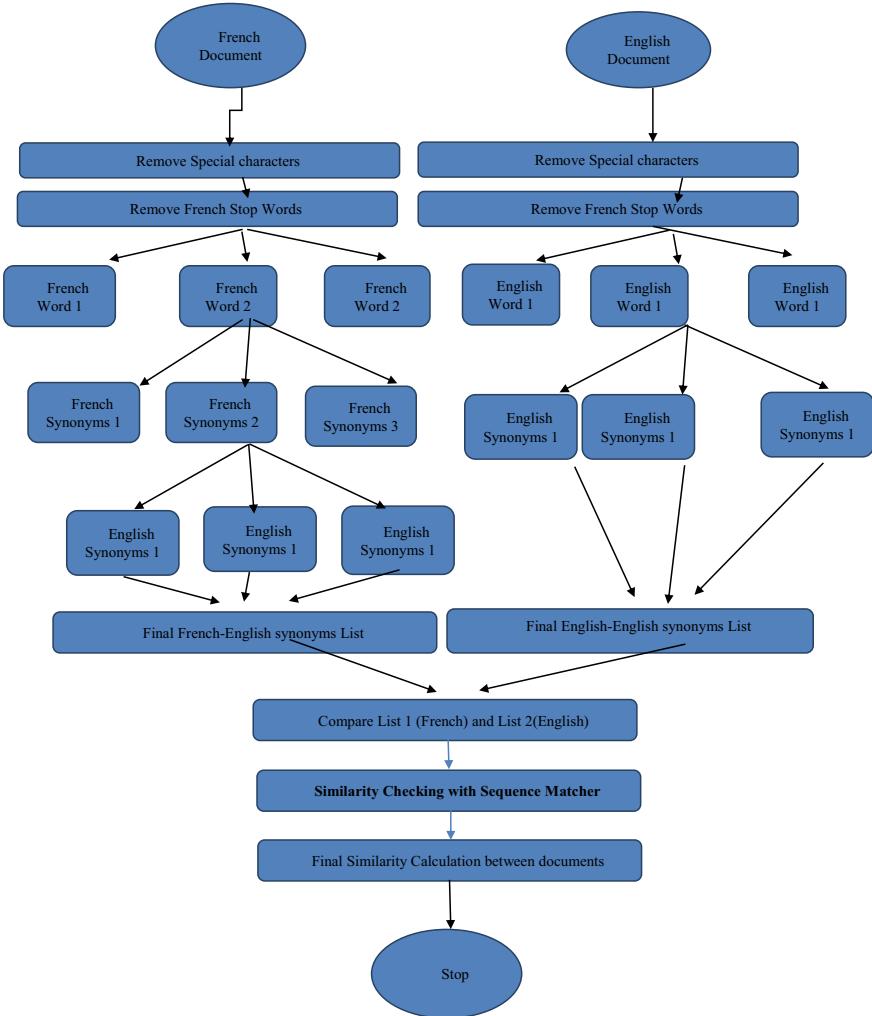


Fig. 1 Flow diagram of the proposed sequence matcher algorithm

that sequence matcher is a familiar technique for finding the similarities between the strings (Table 8).

7 Conclusion

In this paper, we have discussed the language similarity between French and English documents. Here, in order to retrieve the similarity of sentences and words in a

Table 1 Accuracy measure with sequence matcher in between French–English one–one documents

Samples	Datasets	Sequence matcher accuracy
Sample 1	French doc + English doc	78.03
	French doc re-write + English doc	71.94
	French doc + English doc re-write	73.86
	French doc re-write + English doc re-write	73.19
Sample 2	French doc + English doc	74.23
	French doc re-write + English doc	70.02
	French doc + English doc re-write	73.31
	French doc re-write + English doc re-write	74.71
Sample 3	French doc + English doc	80.15
	French doc re-write + English doc	76.66
	French doc + English doc re-write	78.37
	French doc re-write + English doc re-write	80.11
Sample 4	French doc + English doc	74.29
	French doc re-write + English doc	73.12
	French doc + English doc re-write	73.46
	French doc re-write + English doc re-write	78.2
Sample 5	French doc + English doc	76.21
	French doc re-write + English doc	70.31
	French doc + English doc re-write	78.01
	French doc re-write + English doc re-write	77.76
Sample 6	French doc + English doc	71.46

(continued)

Table 1 (continued)

Samples	Datasets	Sequence matcher accuracy
	French doc re-write + English doc	69.63
	French doc + English doc re-write	71.64
	French doc re-write + English doc re-write	73.18
Sample 7	French doc + English doc	70.24
	French doc re-write + English doc	66.43
	French doc + English doc re-write	70.76
	French doc re-write + English doc re-write	73.08
Sample 8	French doc + English doc	71.3
	French doc re-write + English doc	71.3
	French doc + English doc re-write	72.97
	French doc re-write + English doc re-write	76.75

multi-lingual content a string similarity technique sequence matcher is used. The performance of the proposed sequence matcher-based language similarity technique is measured in terms of accuracy. More techniques need to be identified for finding the similarity between the languages for the improvement in accuracy.

Table 2 Accuracy measure with sequence matcher in between French–English one-one documents

Samples	Datasets	Sequence matcher accuracy
Sample 9	French doc + English doc	75.87
	French doc re-write + English doc	73.42
	French doc + English doc re-write	73.31
	French doc re-write + English doc re-write	77.81
Sample 10	French doc + English doc	73.9
	French doc re-write + English doc	67.155
	French doc + English doc re-write	72.65
	French doc re-write + English doc re-write	75.52
Sample 11	French doc + English doc	76.61
	French doc re-write + English doc	72.88
	French doc + English doc re-write	75.53
	French doc re-write + English doc re-write	79.51
Sample 12	French doc + English doc	75.21
	French doc re-write + English doc	72.046
	French doc + English doc re-write	74.05
	French doc re-write + English doc re-write	75.94
Sample 13	French doc + English doc	75.91
	French doc re-write + English doc	71.64
	French doc + English doc re-write	77.59
	French doc re-write + English doc re-write	78.99
Sample 14	French doc + English doc	73.94

(continued)

Table 2 (continued)

Samples	Datasets	Sequence matcher accuracy
	French doc re-write + English doc	71.00
	French doc + English doc re-write	72.56
	French doc re-write + English doc re-write	74.695

Table 3 Accuracy measure with sequence matcher in between French–English one-one documents

Samples	Datasets	Sequence matcher accuracy
Sample 15	French doc + English doc	78.16
	French doc re-write + English doc	71.83
	French doc + English doc re-write	76.19
	French doc re-write + English doc re-write	75.35
Sample 16	French doc + English doc	71.42
	French doc re-write + English doc	72.85
	French doc + English doc re-write	71.25
	French doc re-write + English doc re-write	77.5
Sample 17	French doc + English doc	78.72
	French doc re-write + English doc	71.8
	French doc + English doc re-write	74.14
	French doc re-write + English doc re-write	76.58

Table 4 Accuracy measure with sequence matcher in between French–English one–many documents

No of French samples	No of English samples	Datasets	Accuracy
1	1	French doc + English doc	78.23
1	2		60.28
1	3		67.01
1	4		60.58
1	5		54.95
1	1	French doc re-write + English doc	72.94
1	2		61.21
1	3		67.01
1	4		59.11
1	5		53.61
1	1	French doc + English doc re-write	74.86
1	2		66.95
1	3		70.93
1	4		65.86
1	5		62.22
1	1	French doc re-write + English doc re-write	73.79
1	2		63.91
1	3		69.37
1	4		62.13
1	5		61.97
2	1	French doc + English doc	67.05
2	2		75.23
2	3		65.97
2	4		61.17
2	5		56.56
2	1	French doc re-write + English doc	68.23
2	2		71.02
2	3		68.04
2	4		62.05
2	5		55.76
2	1	French doc + English doc re-write	66.31
2	2		73.91
2	3		68.43
2	4		64.266
2	5		60.49

(continued)

Table 4 (continued)

No of French samples	No of English samples	Datasets	Accuracy
2	1	French doc re-write + English doc re-write	66.84
2	2		75.21
2	3		72.18
2	4		66.13
2	5		62.71

Table 5 Accuracy measure with sequence matcher in between French–English many–many documents

French pair samples	English pair samples	Accuracy	French pair samples	English pair samples	Accuracy
[1 6]	[1 6]	77.53	[3 10]	[1 9]	74.03
[1 6]	[1 8]	69.85	[3 10]	[3 6]	76.33
[1 6]	[4 9]	71.83	[3 10]	[2 6]	69.31
[1 6]	[3 8]	73.99	[3 10]	[4 7]	69.56
[1 6]	[2 8]	69.57	[3 10]	[2 8]	69.18
[1 6]	[5 6]	65.83	[3 10]	[2 9]	74.52
[1 6]	[1 8]	69.85	[3 10]	[5 10]	66.21
[1 6]	[4 7]	67.22	[3 10]	[2 7]	70.97
[1 6]	[5 9]	68.32	[3 10]	[2 10]	69.49
[1 6]	[1 9]	74.03	[3 10]	[4 8]	69.78

Table 6 Accuracy measure with sequence matcher in between French–English many–many documents

French pair samples	English pair samples	Accuracy	French pair samples	English pair samples	Accuracy
[5 7]	[3 9]	78.88	[5 6]	[2 6]	76.11
[5 7]	[2 9]	74.85	[5 6]	[1 9]	74.57
[5 7]	[2 8]	72.22	[5 6]	[5 7]	69.71
[5 7]	[1 6]	75.87	[5 6]	[3 7]	75.30
[5 7]	[2 6]	76.11	[5 6]	[1 6]	71.92
[5 7]	[5 7]	70.91	[5 6]	[1 7]	70.36
[5 7]	[5 10]	70.43	[5 6]	[3 9]	78.88
[5 7]	[1 7]	69.50	[5 6]	[3 10]	79.77
[5 7]	[5 9]	68.32	[5 6]	[5 10]	73.69
[5 7]	[1 10]	73.54	[5 6]	[1 10]	76.54

Table 7 Accuracy measure with sequence matcher in between French–English many–many documents

French pair samples	English pair samples	Accuracy	French pair samples	English pair samples	Accuracy
[1 10]	[1 10]	75.98	[4 6]	[3 6]	74.75
[1 10]	[4 10]	75.63	[4 6]	[3 8]	72.94
[1 10]	[4 7]	73.85	[4 6]	[1 7]	68.39
[1 10]	[1 8]	68.95	[4 6]	[4 6]	72.49
[1 10]	[4 9]	76.42	[4 6]	[2 8]	70.54
[1 10]	[1 9]	73.94	[4 6]	[1 8]	70.06
[1 10]	[1 6]	73.54	[4 6]	[5 10]	74.96
[1 10]	[3 7]	72.52	[4 6]	[5 6]	75.80
[1 10]	[5 8]	69.83	[4 6]	[2 6]	71.59
[1 10]	[5 9]	71.81	[4 6]	[2 8]	70.54

Table 8 The French and English documents for finding the language similarity

French document	English document
UN DOLLAR ET 80 CENTS. C'était tout. Et soixante cents étaient centimes	ONE DOLLAR AND EIGHTY-SEVEN CENTS. That was all. And sixty cents of it was in pennies

References

1. Stalls BG, Knight K (1998) Translating names and technical terms in Arabic text. In: Proceedings of the 1998 COLING-ACL, Montreal
2. Evans DK (2005) Identifying similarity in text: multi-lingual analysis for summarization. Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy, Columbia University
3. Langlois D, Saad M, Smaili K (2018) Alignment of comparable documents: comparison of similarity measures on French–English–Arabic data. Nat Lang Eng 24(5):677–694
4. Cheung P, Fung P (2004) Sentence alignment in parallel, comparable, and quasi-comparable corpora, pp 30–33
5. Rao GA, Srinivas G, Rao KV, Prasad Reddy PVGD (2018) Characteristic mining of mathematical formulas from document—a comparative study on sequence matcher and Levenshtein Distance procedure. Int J Comput Sci Eng 6(4):400–403
6. Rao GA, Srinivas G, Rao KV, Prasad Reddy PVGD (2018) A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulas from documents. IJSC 8(4):1728–1732
7. Rao K et al (2019) An experimental study with tensor flow for characteristic mining of mathematical formulae from a document. EAI Endorsed Trans Scalable Inf Syst 6:e6
8. Brahmaji Rao KN, Srinivas G, Prasad Reddy PVGD, Surendra T (2019) A heuristic ranking of different characteristic mining based mathematical formulae retrieval models. IJEAT 9(1)
9. Ballesteros L, Croft B (1996) Dictionary methods for cross-lingual information retrieval. In: Wagner RR, Thoma H (eds) Database and expert systems applications. DEXA 1996. Lecture notes in computer science, vol 1134. Springer, Berlin, Heidelberg. <https://doi.org/10.1007/BFb0034731>

10. Saad M, Langlois D, Smaïli K (2013) Extracting comparable articles from Wikipedia and measuring their comparabilities. *Procedia Soc Behav Sci* 95:40–47
11. Brahmaji Rao KN, Srinivas G, Prasad Reddy PVGD, Tarakeswara Rao B (2020) Non-negative matrix factorization procedure for characteristic mining of mathematical formulae from documents. In: Communication software and networks. Lecture notes in networks and systems, vol 134. Springer, pp 539–551

Symmetrical Performance Assessment of Large-Scaled Data Using Meta-heuristic Approach



Jyoti Prakash Mishra, Zdzislaw Polkowski, Sambit Kumar Mishra, and Samarjeet Borah

Abstract Considering the present situation of enhancement of computing resources, the processing as well as flow of large scaled data can be deployed towards several data intensive applications to prioritize the scalability as well as consistency of data even if in virtual platform. In such situation, the technologies can be provisioned with specific paradigms to minimize the cost linked with computing resources. The provision in virtual platform can also be able to manage the specific infrastructure obtaining the security and privacy with the applications. The association of physical components as well as the storage components in this situation can have support on resource abstraction with control while generating applications in virtual platforms. In specific term, it can be understood that the management as well as support in virtual platform should be provisioned with portability as well as interoperability requirements. During the applications in virtual machines, it has also been observed that the specified techniques somehow generate massive heterogeneous large scaled data which sometimes a big challenge as many virtual platforms are typically acquainted with homogeneous linked data along with resources. Considering the performance of virtual machines, it is observed that the heterogeneity of the databases linked with virtual machines can be sequenced and the processes associated with the services can be implemented easily through these virtual machines. Based on all the situations, in this work, the meta-heuristic techniques, i.e., ant colony optimization and

J. P. Mishra (✉) · S. K. Mishra

Gandhi Institute for Education and Technology, Affiliated to Biju Patnaik University of Technology, Rourkela, Baniatangi, Bhubaneswar, India
e-mail: jpmishra@gietsbsr.com

S. K. Mishra

e-mail: sambitmishra@gietsbsr.com

Z. Polkowski

Wroclaw University of Economics and Business, Wroclaw, Poland
e-mail: zdzislaw.polkowski@ue.wroc.pl

S. Borah

Sikkim Manipal Institute of Technology, Majitar, East Sikkim, India
e-mail: samarjeet.b@smit.smu.edu.in

firefly approach have been applied to these large-scaled data towards symmetrical performance assessment.

Keywords Virtual machine · Heterogeneity · Meta-heuristic · Pheromone · Light absorption coefficient

1 Introduction

In general, large-scaled data can be easily provisioned towards accessibility of computational resources provided associated with virtual machines and virtual computing facilities. Particularly during conceptualization, this can be better towards intensive computations. While evaluating the performance of individual processing elements in virtual platforms, it is very much essential to concentrate on heterogeneity of data on processing elements along with the associated complexities. In such situation, it is also required to focus on the waiting time of processes along with the performance indices. Purposefully in this situation, the heterogeneous environment is chosen as it may be associated with various categories of servers with variant hardware configurations. Accordingly, the heterogeneity mechanism can reflect the performance of large-scaled data.

2 Review of Literature

Yang et al. [1] in their work prioritized on queuing system comprised of schedule queue, computation queue along with transmission queue to characterize the service process in a multimedia focussing on resource optimization problems. In fact they focussed on cloud service performance linked to fault recovery.

Chiang and Ouyang [2] during their research discussed on energy proportional system based on queuing mechanism to enhance the performance efficiency. In their study, they considered some factors lined to the service towards multi-server system to optimize the server configuration and to maximize the computation n in virtual platforms.

Cao et al. [3] in their work discussed on different heterogeneous multi-core servers with variant sizes associated with queuing system to optimize power allocation and load distribution in a virtual platforms.

Paola et al. [4] in their work discussed the technological aspects along with the computing requirements towards incorporating more number of computing and storage resources.

Ye et al. [5] in their work have focussed on different data centres linked to virtual systems provisioned with global storage services along with scalability and built in activities. In such situation, they also focussed on solutions inclusive of servers, networking equipment as well as storage systems.

Chan and Seung [6] during their studies have projected on control over the computing resources linked to virtual platforms. As such the storage in virtual platform is primitive applications as due to data intensive, it can enhance storage capacity requirement along with usage of data.

Ghani-Ur et al. [7] in their work have focussed on storage allocation provisioned with cost factors and performance issues in virtual platforms incorporating multiple storage. They specified the systems provisioned in virtual platform to meet specific requirements in virtual platforms.

Vasilev and Cristescu [8] in their work focussed on determination of open-source applications and the replacement towards specialized security applications. According to the study, it is observed that the open-source components can be selected as on the identification of security features which is mostly based on determination of performance indicator.

Vasilev and Kehayova-Stoycheva [9] during their studies focussed on specific dimensions lined to Internet usage and clustered procedures. Basically, the clustering procedure is associated with specific activities linked to web services. More focus has been given towards hierachal clustering towards linkage with applied metrics.

Vasilev and Atanasova [10] in their work prioritized towards obtaining the dependencies among the complacency from education and other factors. Accordingly, they have implemented descriptive statistics, cross-tabulations as well as neural networks towards analysis of the dataset. Basically, they focussed on statistical methods towards obtaining significant dependencies along with artificial intelligence mechanisms towards extraction of knowledge from the dataset.

3 Conceptualization of Meta-heuristic Approach

As it is understood that heuristic is a problem-solving mechanism used to discover the suitable paths and generate better solutions within a limited time frame. Of course, the heuristic values are closely associated with solution itself. The meta-heuristic in such scenario is an optimal range approach to obtain heuristic provisioning better solution towards optimization problems.

In this manuscript, specifically two meta-heuristic approaches, i.e., ant colony optimization along with firefly algorithms have been incorporated to assess the performance of large-scaled data.

3.1 *Implementation of Ant Colony Optimization*

Basically, implementation using ant colony optimization is done traversing complete linked paths along with instantiated decision variables and pheromone values. Usually, each pheromone value permits the method of distribution linked to components of solutions and further can be updated during search implementation. Very

often the unnecessary and insufficient mechanisms enhance the needs of storage as well as computing capabilities, and in such situation, the retrieval mechanism is really challenging. So, considering the large-scaled heterogeneous data, it is really essential to implement ant colony optimization technique as large-scaled attributes may be unknown with the focussed datasets. To maintain the accuracy in performance measurement, it is essential to focus on dimension of data as the redundancies within large-scaled data are required to be minimized through appropriate normalization mechanisms.

3.2 Steps for Implementation of Ant Colony Optimization Algorithm

Step 1: Initialize the ants, parametric values linked to the storage system for specified time periods
 Step 2: Initialize the pheromone trails with the absorption parameter
 Step 3: Iterate till the condition is satisfied,
 while (condition is not achieved) do
 identify the position of ant and link to initial position
 Step 4: Reinitiate the iteration for individual ant
 for each ant do
 determine the heuristic values and select the next level node
 link to next level node and apply the transition rule and update the pheromones
 Step 5: Apply the similar procedure to each ant and regenerate the solution focussing on fitness parameters
 Step 6: Update the pheromones
 Step 7: Based on moves of ants, determine the differential values of pheromone levels and update the trail.

3.3 Algorithm

Step 1: Define the maximum number of iterations, e.g., 500
 Step 2: Determine the size of population, no. of ants, e.g., 500
 Step 3: Initialize the pheromone level, i.e., 10
 Step 4: Define the exponential parameter of pheromone
 Step 5: Define the pheromone evaporation rate
 Step 6: Position the ants and obtain the optimal solution
 Step 7: Update the pheromone and record the iteration information.

Generally, many vital facts linked to databases in majority are uncommon during the time of optimization of queries. In such situation, the optimization process associated with parametric queries determines the relevant query execution plans with

optimality and also obtains the feasible run time constraints to enhance the effectiveness of newly submitted query plans. Though it is known that the query plans during optimization process can be valued through cost parameters, but the query plans can be consistent towards parametric values based on optimization criteria as well as predicate selectivity as reflected in Fig. 1. In general, the parametric cost linked to query plans based on the pheromone level can be treated as independent and individual entity, and also the optimizer estimates the query plans based on its retrieval mechanism. Accordingly, the data can be processed through MATLAB 13B for further analysis and result generation (Table 1).

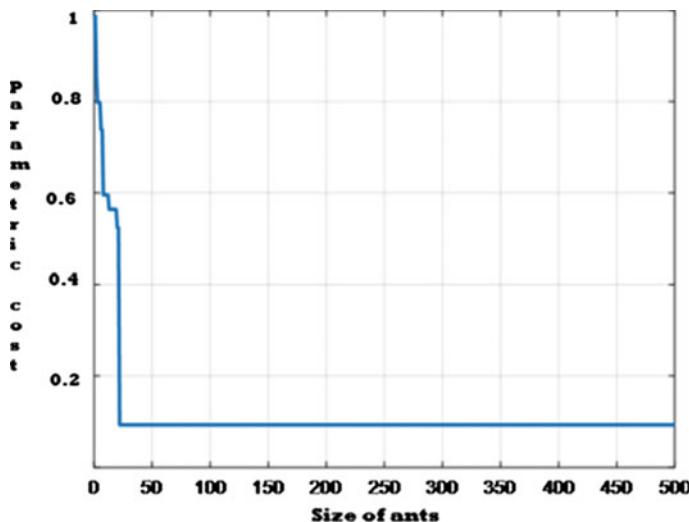


Fig. 1 Size of ants versus parametric cost based on pheromone level

Table 1 Evaluation of parametric cost based on pheromone level

S. No.	Size of ants	Parametric cost (based on pheromone level) (ms)
1	50	0.15
2	100	0.15
3	200	0.15
4	300	0.15
5	400	0.15
6	450	0.15
7	500	0.15

4 Implementation of Firefly Techniques

The mechanisms associated with firefly techniques along with its variants can be applied towards the basic optimization criteria focussing and obtaining the feasible solutions. Practically, the solution in such scenario can be obtained following the expansion mechanisms on queries associated with large-scaled data. Also the optimality linked to firefly techniques can be observed obtaining the optimal query response along with determination of the dimension of queries in the system. As allocation of resources is based on specified scheduling mechanisms closely associated with factors of time and associated cost, constraints can be technically categorized sequencing the tasks implementing the meta-heuristic techniques.

4.1 Steps to Formulate the Objective Function with Associated Mechanisms

Step 1: Obtain the initial population of fireflies, parametric values, x_i , where $i = 1$ to q

Step 2: Determine the light intensity value, I linked with operations on parametric values, $O(x)$, the value of light intensity should be directly proportional to the operations on parametric values, $O(x)$

Step 3: Define the parametric constraint with absorption measures

Step 4: Iterate with each firefly

```

while (t < population_init)
    for i = 1: n (all n fireflies)
        for j = 1: i (n fireflies)
            if (I(j) > I(i)), shift ith firefly to jth firefly
            end if
            end for j
        end for i
    end while

```

Step 5: Obtain the new feasible solutions and update light intensity parameter

Step 6: Schedule the fireflies and obtain the recent optimal result

Step 7: Determine the analysis and check the optimality.

4.2 Algorithm

Step 1: Define the number of decision parameters, $newpar$

Step 2: Define the lower bound and upper bound of decision parameters

Step 3: Obtain the number of iterations along with size of fireflies

Step 4: Obtain the initial base parameter of attraction coefficient value along

with light absorption coefficient value

Step 5: Define the mutation coefficient value along with damping parameter

Step 6: Compare the parametric values associated with lower bound and upper bound decision parameters

Step 7: Initiate the process to obtain most optimal cost values

```
for j = 1:itmax
```

```
newpar = regen(firefly,tpar,1);
```

```
for i = 1:newpar
```

```
newpar(i).Cost = null;
```

```
for j = 1:itmax
```

```
if newpar(j).Cost < newpar(i).Cost
```

```
tij = normalize(newpar(i).Position-newpar(j).Position)/distn;
```

```
t1 = t0*exp(-q*tij^n);
```

```
end
```

```
end
```

```
end
```

Basically, there are some mechanisms associated with firefly technique to observe the performance mainly linked to light absorption coefficient, randomization, and enhancement of coefficient values. The selection of parameters in such situation can be analysed to ensure the optimality within the system. The performance indication and cost analysis can also be defined to observe the attractiveness of fireflies with others. Also accordingly, the position of fireflies can be updated. As shown in Fig. 2, the parametric cost based on attraction coefficient and light absorption coefficients are proportionate with size of fireflies. It is seen that the parametric cost associated with the light absorption is treated as independent while proportionating with attraction coefficient parameters. Practically, the data are processed through MATLAB 13B for further analysis and result generation (Table 2).

Fig. 2 Size of fireflies versus parametric cost based on attraction and light absorption coefficient values

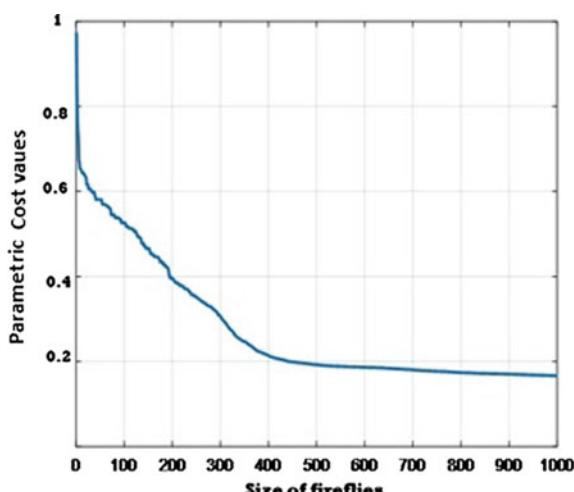


Table 2 Parametric cost of fireflies based on attraction coefficient and light absorption coefficient values

S. No.	Size of fireflies	Parametric cost based on attraction coefficient and light absorption coefficient values (ms)
1	300	0.34
2	400	0.22
3	500	0.20
4	600	0.19
5	700	0.18
6	800	0.17
7	900	0.16

5 Discussion and Future Direction

During analysis, it is observed that the performance of large-scaled data in heterogeneous environments is quite sensitive. Considering the response time as well as average waiting time, it is seen that the accuracy during performance analysis can be more implementing the queuing mechanisms. It is required to prioritize on the primary data schedule along with the execution queue. The implementation of simulation mechanisms in such situation should be provisioned with high order accuracy for individual processing elements as well as heterogeneous systems.

6 Conclusion

Access of large-scaled data in general in virtual platform sometimes become challenging as the many times the computational resources are managed centrally. In this work, as it is provisioned with heterogeneous computing resources, again it is essential to prioritize the virtual technology. To optimize the operations and to obtain the optimal solution with more accuracy, it is required to expedite the virtual mechanism at a great extent.

References

1. Yang B, Tan F, Dai Y-S (2013) Performance evaluation of cloud service considering fault recovery. *J Supercomput* 65(1):426–444
2. Chiang YJ, Ouyang Y-C (2014) Profit optimization in SLA-aware cloud services with a finite capacity queuing model. *Math Probl Eng* 2014. Article ID 534510, 11 pages
3. Cao J, Li K, Stojmenovic I (2014) Optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and data centers. *IEEE Trans Comput* 63(1):45–58

4. Paola P, Roberto C, Alberto B, Lorenzo P (2020) Amazon, Google and Microsoft solutions for IoT: architectures and a performance comparison. *IEEE Access* 8:5455–5470
5. Ye T, Peng X, Hai J (2019) Secure data sharing and search for cloud-edge-collaborative storage. *IEEE Access* 7:15963–15972
6. Chan KP, Seung JB (2020) Blockchain of finite-lifetime blocks with applications to edge-based IoT. *IEEE Internet Things J* 7(9):120–133
7. Ghani-Ur R, Anwar G, Muhammad Z, Syed Husnain AN, Dhananjay S (2019) IPS: incentive and punishment scheme for omitting selfishness in the internet of vehicles (IoV). *IEEE Access* 7:109026–109037
8. Vasilev J, Cristescu MP (2021) Specialized applications used in the mobile application security implementation process. In: *Organizations and performance in a complex world*, pp 39–49
9. Vasilev J, Kehayova-Stoycheva M (2018) Measuring the problem use of internet—internal structures and dependences. *Econ Comput Sci* (1)
10. Vasilev J, Atanasova T (2015) Parallel testing of hypotheses with statistical and artificial intelligence methods: a study on measuring the complacency from education. *Comput Sci Appl* 2(5):206–211

Examining Data Mining Classification Techniques for Predicting Early Childhood Development in Nigeria



Aimufua Ikponmwosa, Narasimha Rao Vajjhala^{ID}, Sandip Rakshit^{ID}, and Olumide Longe^{ID}

Abstract Early childhood is a critical part of a child's development as it involves physical, cognitive, and psychological development. In the educational domain, especially early childhood education, there is rich data available that we could leverage to determine the development stage of a child and hidden patterns of a child's learning ability or disability. This study investigates which data mining classification technique will be most suitable in building a predictive model that can identify the social, cognitive, and emotional stages of a child. The authors compared J48, Naïve Bayes, random forest, support vector machines (SVM), and k-nearest neighbors (KNN) classifiers using performance measures like Kappa statistics, receiver operating characteristic (ROC), root-mean-square error (RMSE), and mean absolute error (MAE) using a data mining analytical tool called WEKA. The authors also compared the accuracy measures like true positive (TP) rate, false positive (FP) rate, precision, recall, and *F*-measure. The results indicate that the J48 classifier has a better classification accuracy and prediction rating over other tested algorithms using the early childhood dataset.

Keywords Data mining · Childhood development · Classification · Nigeria · Prediction · Accuracy · Model · Random forest · Classifier · Support vector machine

A. Ikponmwosa · S. Rakshit · O. Longe
American University of Nigeria, Yola, Adamawa, Nigeria
e-mail: ikponmwosa.aimufua@aun.edu.ng

S. Rakshit
e-mail: sandip.rakshit@aun.edu.ng

O. Longe
e-mail: olumide.longe@aun.edu.ng

N. R. Vajjhala (✉)
University of New York Tirana, Kodra e Diellit, Tirana, Albania
e-mail: narasimharao@unyt.edu.al

1 Introduction

Over 200 million children under the age of 5 in low and middle-income countries, particularly in Africa and Asia, do not attain full developmental potential because of various factors, including poverty, poor nutrition, and other factors [1]. Early childhood development is considered as a predictor of adult health and productivity [2]. Research indicates that investment in early childhood development can help countries in improving human development, human capital formation, economic growth, and social progress [1, 2]. Also, the exposure of a child to an environment that is not so conducive for upbringing during the first few years of their life lowers the child's intelligence quotient (IQ). This can also lead to low academic achievement, increased anti-social behavior, and reduces economic productivity in adulthood [2].

2 Review of Literature

2.1 *Data Mining*

The advent of smartphones over the last decade, coupled with other technological advances, has led to large volumes of data. Health, manufacturing, and other leading industries use large data repositories to help design business strategies and analyze unstructured and structured data to gain useful knowledge. Data mining involves analyzing large-scale observation datasets and identifying previously unknown relationships and summarizing the data in a novel manner. Data mining algorithms include both predictive and descriptive algorithms [3]. Data mining algorithms have been successfully applied in several domains, including biomedical research, and decision making at various management levels [4].

Data mining has several application types, including classification, estimation, prediction, correlation analysis, and visualization. According to Kumar and Khatri [4], data mining comprises the analysis of large data to discover trends and meaningful information that can be converted into a form of intelligence. Data mining is mostly used to interpret knowledge and discover hidden patterns from various domains of expertise. Ming-Syan et al. [5] stated that the difference between data mining and traditional data analysis is the ability to mine information and discover knowledge and the premise of no clear assumption. Data mining uses automated data analysis techniques to uncover previously undetected relationships among data items [6].

2.2 Applications of Data Mining

There are several domains in which data mining can be applied, a domain like social media analytics, medical data mining, educational data mining, business intelligence, etc. In the medical and healthcare domain, there are issues. Healthcare domain is a large domain, especially as the healthcare information systems generate a huge amount of data regarding the patient's medical history, current ailment, and diagnoses. With this huge data being generated, the medical domain can watch out for repeated patterns and predict the outcome of such a result, leading to improved quality relating to the overall quality of patient services, early prediction and diagnosis of diseases [7]. Data mining techniques can be employed again in the medical field for clinical test analysis and its relationship pathology [8].

Komi et al. [9] conducted a study on applying data mining methods in diabetes prediction. The authors emphasized using data mining methods in shedding more light on the prediction of diabetes mellitus. In their report, the authors comparatively analyzed five different mining techniques using MATLAB tools, including Gaussian mixture modeling (GMM), extreme learning machines (ELM), support vector machines (SVM), logistic regression, and artificial neural networks (ANN), to propose a technique that will be most effective in the early prediction of diabetes. This technique model was trained and validated against a test dataset. The experiment's effect proved that provided with the diabetes dataset ANN technique provided the highest accuracy.

Kumar and Khatri [4] compared different classification techniques and their prediction accuracy for a specific dataset (chronic kidney disease). Kumar and Khatri [4] used WEKA as a data mining tool to analyze five classification techniques, namely J48, Naïve Bayes, random forest, SVM, and KNN, using performance measures like TP rate, FP rate, and precision. Olukunle and Ehikioya [10] proposed using a fast association rule mining (ARM) algorithm to be used on medical image dataset. They established that ARM aims at discovering strong, interesting patterns between items in a vast dataset. Olukunle and Ehikioya [10] further proposed the use of the frequency pattern (FP) growth algorithm, a standard ARM algorithm that is efficient for mining a large dataset from frequency pattern. Olukunle and Ehikioya [10] suggested that his approach will have a compactable parallel data representation scheme for input and output structure. Further, Olukunle and Ehikioya [10] experimented with showing that FP growth has the desirable features in handling large medical data images. In their conclusion, Olukunle and Ehikioya [10] indicated that it is necessary to mine medical images because of the vast information available for knowledge support and to apply the ARM technique because it is simple and explanatory. As the education data mining gets larger into early childhood education by so doing having different characteristics, different data mining techniques will have their predictive efficiency. Nigeria is the most populous African country. However, there is limited literature predicting early childhood development challenges using various data mining classification techniques, even though there is significant data available in this domain. In this study, the authors comprehensively studied and

compared other data classification techniques and their prediction accuracy for early childhood datasets.

Shouman et al. [11] used decision tree algorithm techniques in diagnosing heart disease. The authors investigate applying a range of techniques to a different type of decision tree seeking better performance in heart disease diagnosis. Shouman et al. [11] proposed a model that enhances the decision tree accuracy by integrating a multiple classifier voting technique with different types of discretization methods and different decision tree types. The research seeks to improve diagnosis accuracy by applying the multi-interval discretization method, multiple classifier voting, and reduced error pruning to the decision tree. The experiment was conducted using Microsoft Visual Studio 2019, and it involved systematically testing different discretization techniques, multiple classifiers, voting techniques, and various decision tree types in the diagnosis. Shouman et al. [11] concluded by saying that by applying multi-interval equal frequency with nine voting gain ratios, and the accuracy of the decision tree will be improved by a percentage of 84.1%.

Kumar and Pal [12] investigated engineering students' performance improvement using some data mining techniques. A 17-attribute dataset was created to be run by WEKA open-source analytical tool to enable the user to apply classification and regression on the resulting dataset over a tenfold cross-validation, thereby estimating the predictive model's accuracy. The authors used the Iterative Dichotomizer 3 (ID3), C4.5, and Classification and Regression Tree (CART) algorithms in the classification model to test the predictive model. In their study, Shouman et al. [11] found that the C4.5 technique had the highest predictive accuracy in identifying students who were more likely to fail than other methods.

Comendador et al. [13] examined the students' history of accessing a university learning management system (LMS) data, applying some data mining techniques to build a model that will predict the user's learning behavior. The study's objective was to categorize the typical online behavior of distance education and identify influencers for learning outcomes. In their study, the authors used a dataset from the University history to access the Polytechnic University of the Philippines (PUP) and choose a dataset of 248 student records. Comendador et al. [13] conducted an experiment using WEKA and understudied reduced error punning tree (REPTree), CART, and J48 tree algorithm to evaluate the appropriate classification algorithm that can be utilized for predicting student finals based on usage data in the LMS. Comendador et al. [13] further applied discretization and tested the algorithms on the provided dataset using tenfold cross-validation in WEKA. After obtaining the final attribute, Comendador et al. [13] further applied a three-feature selection technique CH, GR, and IG in the classification of student performance in an E-Learning environment. The final result showed that the score obtained from participation in the online activity was the most valuable influencer to completing the program [13].

de Paula Santos et al. [14] proposed an evaluation model using educational data mining techniques to analyze students' responses during an institutional teaching evaluation. The mining data process begins to identify the categories of analysis that students find most important in the teaching practices and identify the semantics orientation of student response to the instructor, whether positive or negative. de

Paula Santos et al. [14] further categorized the model into five stages and compared it to existing Higher Technological Education, in which the statistical models and data mining were not used. They aim to promote EDM use in particular sentiment analysis, to identify which teaching practice is good and not considered from the student perspective. Their research's relevance is student-centered, making students cease from being mere spectators and becoming the protagonist in reconstructing the teaching model and roles within the institution.

3 Methodology

In trying to understudy the educational data mining (EDM) of early childhood, we came up with this approach on:

- Which classification technique (J48, random forest, Naïve Bayes, SVM, and KNN) will be most appropriate in developing a predictive model for early childhood education system in Nigeria?
- What are the relevant factors that influence early childhood development in Nigeria?

Classification can be classified as a learning technique that organizes items in collection to target categories which aims to accurately predict the large dataset into classes. Kesavaraj and Sukumaran [15] indicated in their research that classification can be used to predict group membership of same instance and can be used to classify item into a set of classes. According to Umadevi and Marseline [16], classification techniques are classified into two groups, the supervised learning and the unsupervised learning. In the supervised learning, the classified data are grouped into classes based on insight of different classes, while the unsupervised classification data are not predicted by the user.

Clustering technique in data mining is the process of grouping a collection of same datasets into classes of similarity in order that data of same object type are grouped differently from object of another cluster. The number of similarities between the object of a cluster is calculated by the use of similarity function. Clustering is very useful for document organization, it helps in data recovery technology, and it also increases the efficiency of a database system [17]. In analyzing the clustering technique, the first step involves computing “proximity indices” between particular groups in relating to interest area. Having known the proximity indices, a clustering algorithm can then be applied to group with similarity in object data. There are factors for selecting that leads to choosing an appropriate clustering method which includes the nature of the data (continuous or nominal) and the size of the data matrix [18]. In this phase, we obtained early childhood learning data from Word of Faith College (nursery section) in Benin City, Nigeria and created an online questionnaire (Google form) which the links to the questionnaire was sent to Church Day care

centers, social media platforms, students, staff, and members of American University of Nigeria with the help of the Student Government Association (SGA). A survey instrument was used to collect the data containing eighteen (18) questions.

4 Analysis and Findings

Descriptive statistics involve summarizing the figures of data collected. Descriptive analysis includes numbers in average and/or percentage, graph, and tables [19]. For this research thesis, the target population are individuals who have children or give care to children within the early childhood development stage.

Table 1 shows that majority of the respondent were mothers and they made up of 42.2% of the whole 279 population of respondent but 9 responses in general was incomplete, while Father's has a total of 6.6%. From our WEKA analysis, we overlaid the environment in which the child is growing up with the type of responses given, and we could deduce a factor that matter to the overall development of early childhood. From Table 2, we can notice that children in the averagely conducive environment and that of the very conducive environment have higher and better number of positive responses than that of the children living in the not so convenient environment.

In Table 3, five major classifiers, namely J48, random forest, Naïve Bayes, SVM, and KNN were considered for use in this thesis research using WEKA. Various performance measures relating to the aims and objective of the thesis have been measured in Table 3.

Figure 1 shows the level of accuracy for all classifiers. It shows that J48 and random forest have better reading of accuracy, while K-nearest neighbor reveals the lowest. We can deduce that J48 has slightly better accuracy than random forest algorithm.

Table 1 Descriptive analysis of collected data

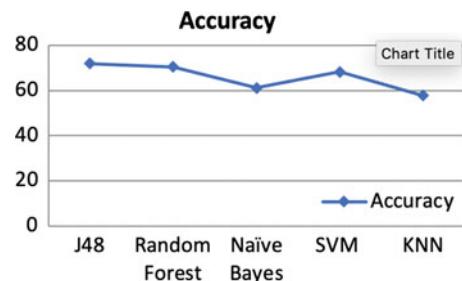
Category	Percentage %
Mothers	42.2
Class teacher	39.6
School administrator	1.9
Fathers	6.6
Care giver	3
Other (brothers, sisters, relatives)	6.7

Table 2 Environmental condition of children

Count	Environment of child
78	Very conducive (color code red)
154	Averagely conducive (color code blue)
38	Not so conducive (turquoise blue)

Table 3 Experimental result of all algorithms

Algorithm	Accuracy	ROC	Kappa statistics	RMSE	Mean absolute error	Time to build model (s)
J48	71.85	0.727	0.476	0.3994	0.247	0.06
Random forest	70.37	0.811	0.451	0.3668	0.274	0.19
Naïve Bayes	61.11	0.767	0.377	0.4509	0.266	0.01
SVM	68.14	0.733	0.425	0.3892	0.301	0.24
KNN	57.77	0.619	0.226	0.5072	0.286	0

Fig. 1 Classification accuracy value of algorithm

In Fig. 2, ROC curves reveal good result for random forest and Naïve Bayes, a little fair outcome for SVM and J48, and a decreased performance for KNN. However, the time it took to build models is less in the case of KNN, SVM, and J48 compared over random forest and Naïve Bayes.

The *F*-measure represents the combining of precision and recall. We can then put together that a classifier that has high precision and low recall is adequate and most accurate as shown in Fig. 3.

In Table 4, J48 revealed a significantly high figure in the true positive (TP) rate, the precision, recall, and the *F*-measure which indicates a good performance of the algorithm and a low figure in the false positive (FP) rate which indicates that the algorithm can handle amount of false positive attribute/numbers and reduce the outcome of false result.

5 Conclusion

This study uses data mining classification techniques to predict and provide a better understanding of early childhood to develop to understand the teaching methods to meet the children's needs. A predictive model was developed that could be used to collect, process, and review hidden information. This information can help in predicting the challenges that children face as part of early childhood development. Previous studies in the education domain had mainly focused on higher education, so this predictive model should help educators and policymakers better understand the

Fig. 2 ROC versus time to build

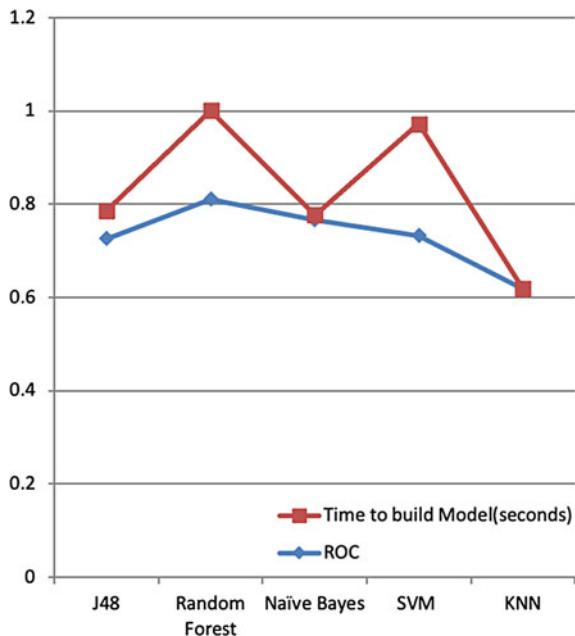


Fig. 3 Result of K -value versus error

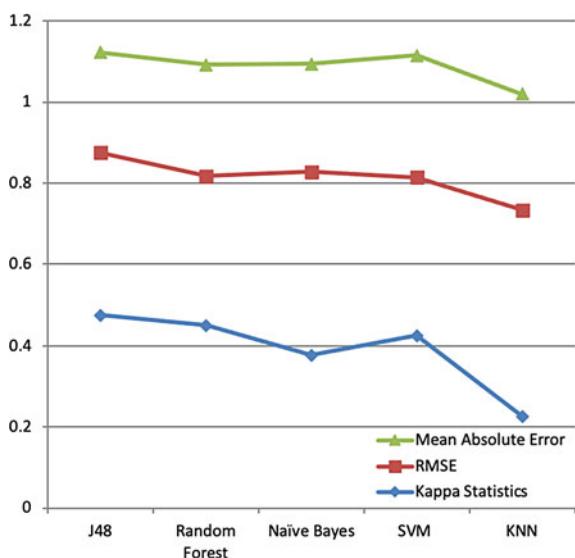


Table 4 Major accuracy measure value

Algorithm	TP rate	FP rate	Precision	Recall	F-measure
J48	0.719	0.280	0.719	0.719	0.711
Random forest	0.704	0.292	0.702	0.704	0.694
Naïve Bayes	0.611	0.235	0.644	0.611	0.613
SVM	0.681	0.289	0.678	0.681	0.677
KNN	0.578	0.390	0.566	0.578	0.567

challenges encountered during early childhood development. Several classification techniques along with the key performance indicators were analyzed in this study in the context of early childhood development.

References

1. Baker-Henningham H (2014) The role of early childhood education programmes in the promotion of child and adolescent mental health in low- and middle-income countries. *Int J Epidemiol* 43(2):407–433
2. Masterov D (2007) The productivity argument for investing in young children. *Rev Agric Econ* 29:446–493
3. Wright MOD, Masten AS (2005) Resilience processes in development. In: Goldstein S, Brooks RB (eds) *Handbook of resilience in children*. Springer US, Boston, MA, pp 17–37
4. Kumar N, Khatri S (2017) Implementing WEKA for medical data classification and early disease prediction. In: 2017 3rd international conference on computational intelligence & communication technology (CICT). IEEE, Ghaziabad
5. Ming-Syan C, Jiawei H, Yu PS (1996) Data mining: an overview from a database perspective. *IEEE Trans Knowl Data Eng* 8(6):866–883
6. Sahu H, Shrma S, Gondhalakar S (2011) A brief overview on data mining survey. *Int J Comput Technol Electron Eng (IJCTEE)* 1(3):189–207
7. Aigbovo O (2019) Trend and pattern of economic and financial crimes statutes in Nigeria. *J Financ Crime* 26(4):969–977
8. Podgorelec V, Hericko M, Rozman I (2005) Improving mining of medical data by outliers prediction. In: 18th IEEE symposium on computer-based medical systems (CBMS’05). IEEE, Dublin
9. Komi M et al (2017) Application of data mining methods in diabetes prediction. In: 2017 2nd international conference on image, vision and computing (ICIVC). IEEE, Chengdu
10. Olukunle A, Ehikioya S (2002) A fast algorithm for mining association rules in medical image data. In: IEEE CCECE2002. Canadian conference on electrical and computer engineering. Conference proceedings (Cat. No. 02CH37373). IEEE, Winnipeg, Manitoba
11. Shouman M, Turner T, Stocker R (2011) Using decision tree for diagnosing heart disease patients, vol 121, pp 23–30
12. Kumar S, Pal S (2012) Data mining: a prediction for performance improvement of engineering students using classification. *World Comput Sci Inf Technol J* 2:51–56
13. Comendador BEV, Rabago LW, Tanguilig BT (2016) An educational model based on knowledge discovery in databases (KDD) to predict learner’s behavior using classification techniques. In: 2016 IEEE international conference on signal processing, communications and computing (ICSPCC). IEEE, Hong Kong

14. de Paula Santos F, Lechugo CP, Silveira-Mackenzie IF (2016) “Speak well” or “complain” about your teacher: a contribution of education data mining in the evaluation of teaching practices. In: 2016 international symposium on computers in education (SIEE). IEEE, Salamanca
15. Kesavaraj G, Sukumaran S (2013) A study on classification techniques in data mining. In: 2013 fourth international conference on computing, communications and networking technologies (ICCCNT). Tiruchengode, IEEE, pp 1–7
16. Umadevi S, Marseline KSJ (2017) A survey on data mining classification algorithms. In: 2017 international conference on signal processing and communication (ICSPC). IEEE, Coimbatore
17. Patel D, Modi R, Sarvakar K (2014) A comparative study of clustering data mining: techniques and research challenges, vol iii, pp 67–70
18. Antonenko PD, Toy S, Niederhauser DS (2012) Using cluster analysis for data mining in educational technology research. *Educ Technol Res Dev* 60(3):383–398
19. Agresti A (2018) Statistical methods for the social sciences, 5th edn. Pearson Inc., Boston, MA

Significance of Machine Learning in Future Prediction Analysis



Soumya S. Mohapatra, Bunil Kumar Balabantray, Shiba Ch. Barik,
Mousumi Acharya, and Ladu K. Sahoo

Abstract The current century has astonished us in transiting through the gravest epochs of human tragedy. COVID-19 is a wakeup call for the humanity. This present scenario challenges the demand of a technology that can analyze the entire world's fatality rate as well as the situation in our country India and forecast the next few years so that the human lives can take necessary steps to overcome from the prolonged pain. So our objective is to show how machine learning has added its great impact in future forecasting and analyzing the challenges regarding the current pandemic. Machine learning is a subset of artificial intelligence where we are acknowledging the different figured data from the repository and further the information technology system helps to predict the spread of the disease in future by implementing different models and algorithms. Our research methodology mostly emphasizes on few standard forecasting models like linear regression (LR), support vector machine (SVM), time series analysis (we used the Holt's linear model) and ARIMA model and SARIMA model. Predictions are done based upon the confirmed cases, and it has been found that time series model is providing the best result.

Keywords COVID-19 · Machine learning · Artificial intelligence · Linear regression · ARIMA · SARIMA · Support vector machine

1 Introduction

COVID-19 was originally identified in one of the prominent cities (Wuhan) of China in the early of December 2019 and was named as COVID-19 by the World health Organization (WHO). It has been recorded officially that there are 17,396,943 con-

S. S. Mohapatra (✉) · S. Ch. Barik · M. Acharya · L. K. Sahoo
Department of Computer Science and Engineering, DRIEMS, Cuttack, Odisha, India

B. K. Balabantray
Department of Computer Science and Engineering, National Institute of Technology Meghalaya, Shillong 793003, India
e-mail: bunil@nitm.ac.in

firmed cases and 675,060 deaths till today (August 1, 2020) by WHO [1]. People who are affected with the COVID-19 used to show the symptoms of mild-to-severe illness and but the most strenuous difficulties with this virus are that the person may remain asymptomatic although being affected for so many days which lead to multiorgan dysfunction and cause untimely death. To overcome the rapid infection, the respective government has taken lots of step to avail social distancing, and several researches have been carried out by the researchers and medical practitioner to help the human lives on behalf of the society [2].

Among the different researching technology, today's world mostly relies on a technology called as machine learning which is a subset of artificial intelligence. It is the branch of computer science which makes the machine to learn and self-reformation. Its significance mostly used to build the mathematical model which works on the sample data to predict about the future. The sample data is represented through matrix and iterated through different optimization technique to predict the result [3]. The technology is categorized into different types which include supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. The significant research area of machine learning involved in future prediction and forecasting which has played a leading role through various application areas like temperature forecasting, disease prediction, market value future prediction, etc. Particularly our study is based on predicting the infected cases of COVID-19 through the autoprediction analysis by which early recognition can be done based on the present scenario and will be helpful in saving the humanity [4]. To bestow with the human confrontation, our goal is to evolve a forecasting system for COVID-19. The forecasting is done on three wavering entities like (1) number of confirmed cases, (2) number of deaths, and (3) the number of recoveries. The forecasting problem is considered as a regression problem [3], so the experimentation is done through some of the supervised machine learning models like linear regression (LR) model, support vector machine (SVM), Holt's linear model, AR model, ARIMA model and SARIMA model for time series forecasting [5]. By performing the linear regression, we are taking the COVID dataset from the data scientist community, Kaggle, where the training datasets (80%) and test datasets (20%) as well as we are training the model, further finding the predicted value by re-enforcing the trained model with the test dataset. Also, we are applying the time series model with the trained dataset to find out the forecasted values. For each model, we are evaluating the root mean square error (RMSE) and finally comparing the RMSE score for each model. Model having less RMSE score is considered to be the best model [4].

2 Input Data and Methodologies

2.1 Dataset

The objective of our study is to forecast the future COVID cases based on our present confirmed cases, deaths, and the number of recoveries. The current dataset is collected from the data repository of data science community, Kaggle named as (covid19data.csv). It contains the abstract in the tabular format regarding the confirmed cases, deaths, and recoveries. Overall data is summarized on daily basis. The various time series data samples are illustrated in Table 1 respectively.

2.2 Supervised Machine Learning Model

In supervised ML, we used to evaluate the accuracy of the predicted values based on the input as the labeled dataset [3]. For example, if we are taking the image of a cat as our trained dataset, then if any new images of cats are given to the model, it should recognize each one as cat, i.e., the model should work on the training algorithm and predict the label correctly. There are two major areas where supervised learning is needful. Those are classification and regression problem statements. The classification problem used to predict the disjunction values through the algorithm by recognizing the input values as a group or class whereas the regression problem inspects the unceasing values. It used to map the value of ‘Y’ with the given value of ‘X’, i.e., $Y = f(X)$. In our proposed system, we are performing the predictive analysis with the help of the supervised algorithm like: (1) linear regression (LR), (2) support vector machine (SVM), and (3) time series analysis for future forecasting of COVID-19 [4, 6].

Table 1 Sample data

Observation date	Province/state	Country/region	Last update	Confirmed	Details	Recovered
01/22/2020	Anhui	Mainland China	01/22/2020 17:00	1	0	0
01/22/2020	Beijing	Mainland China	01/22/2020 17:00	14	0	0
01/22/2020	Chongqing	Mainland China	01/22/2020 17:00	6	0	0
01/22/2020	Fujian	Mainland China	01/22/2020 17:00	1	0	0
01/22/2020	Gansu	Mainland China	01/22/2020 17:00	0	0	0

2.2.1 Linear Regression

Regression analysis is used to demonstrate the impact of independent variable upon the dependent variable. It helps to predict the continual values and shows the relationship between the dependent and independent variable. Linear regression is the simplest statistical method which used to decipher in machine learning for predictive analysis. Linear regression shows the linear relationship between the independent variable (X -axis) and dependent variable (Y -axis) [3]. The linear regression equation can be represented through the mathematical statement as:

$$Y = b_0 + b_1 X + \epsilon \quad (1)$$

where Y = target variable, X = predictor variable, b_0 = Y -intercept, b_1 = slope, ϵ = error term.

The error term uses to represent the variability between X and Y . The equivalent form of the above equation can be written as:

$$Y = b_0 + b_1 X \quad (2)$$

The objective of the machine learning algorithm is to find the best fit line variable which infers the difference between the actual value and the predicted value should be minimum which can be represented as:

$$\min 1/n \sum_{i=1}^n (Y(\text{Predi}) - Y(Y_i))^2 \quad (3)$$

Hence,

$$g = 1/n \sum_{i=1}^n (Y(\text{Predi}) - Y(Y_i))^2 \quad (4)$$

where g is called as the cost function. $Y(\text{Predi})$ is the predicted value and $Y(Y_i)$ is the actual value, ‘ n ’ used to represent the total number of data points.

2.2.2 Support Vector Machine (SVM)

Support vector machine is a renowned algorithm provided by machine learning which is mostly used for classification between different features. It is used to differentiate the n -dimensional spaces into multiple classes where new data points can be put according to the features matching with the respective class. The decision boundary or best fit line in the n -dimensional space is called as the hyperplane. With respect to machine learning, the input vector (x) is mapped to the n -dimensional space called as feature space (z) with the help of nonlinear mapping techniques, after which linear regression is applied to it [3]. With respect to a multivariate training dataset (x_n)

having number of observations (N) with y_n as a set of observed retaliations, the linear function can be delineated as:

$$f(x) = x' \beta + b \quad (5)$$

The objective is to find the minimal norms of values of $f(x)$ with $(\beta' \beta)$. So the problem will fit in minimization function as

$$J(\beta) = 1/2(\beta' \beta) \quad (6)$$

With a special condition of the values of all residuals as not more than ϵ as in the following equation

$$\forall n |y_n - x'_n \beta + b| \leq \epsilon \quad (7)$$

2.2.3 Time Series Analysis (Holt's Linear Model)

Here, we have considered the Holt's linear model which used to give progressively more weights to the recent points. In the Holt's model, we keep

$$F_{t+1} = a_t + b_t \quad (8)$$

where F_{t+1} = Forecast for the period $t + 1$.

a_t = level representing the smoothed value up to and including the last data.

b_t = it is the slope of the line that we are fitting at point t .

So a_t is the constant and when we add slope to it, we get the forecast value of the next period. So forecast for the next period will be level at the end of the previous period or the present period. Both a_t and b_t are updated for every data point using exponential smoothing [5]. So a_t is represented by:

$$a_t = \alpha D_t + (1 - \alpha)(a_{t-1} - b_{t-1}) \quad (9)$$

where α = weight, D_t = Demand during the last period $a_{t-1} - b_{t-1}$ = Forecast for the period t . Now, b_t is the slope which has two components. One is the component that comes out of the level, and the other is the component that comes out of the slope.

$$b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \quad (10)$$

where β = exponential smoothing constant. $a_t - a_{t-1}$ = Slope $(1 - \beta)b_{t-1}$ = comes from the actual component that was computed. So in Holt's model, we are redefining the slope at every point up to and including the last point. Here for every data point, we used to find out the level and a slope and we add them so that we get the forecast for the next period [5].

2.2.4 Autoregressive Model

It is a time series model where the forecasting is based on the past values called as lags. Here distributions depend on only on time rather than location on time. Models that depend on it are called stationary models. In general, it can be represented as

$$Y_t = \omega + \phi Y_{t-1} + \epsilon_t \quad (11)$$

where Y_t = target, Y_{t-1} = lagged target, ϵ_t = error term = coefficient, ω = intercept A time series that is a linear function of p past values with error is called autoregressive process of order p (AR - p) [7] and represented as:

$$Y_t = \omega + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t \quad (12)$$

2.2.5 Moving Average Model (MA Model)

Here forecast series based solely on the past errors in the series called as the error lags. In MA model to know what is the order of residuals is to be taken, we plot the acf (autocorrelation) function with the number of lags and we see for what order of lags there is nonzero acf and use that order. In this way we find the coefficient with respect to the lags. Hence, a time series that is a linear function of q past error values with current error is called as a moving average process of order q that is MA(q) [7].

$$Y_t = \omega + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (13)$$

2.2.6 Autoregressive Integrated Moving Average Model (ARIMA Model)

In reality, most economic variables are non-stationary which cannot be handled by the AR model, MA model and ARMA model. Hence, they have to go through a transformation process called differencing before they become stationary. This process is called integration which is done through ARIMA model [7]. ARIMA (p, d, q) specifies the number of lags of dependent variable (p), how many times the variable is differenced to become stationary (d), the number of lags of the error term (q). Differenced variable

$$\Delta Y_t = Y_t - Y_{t-1} \quad (14)$$

It is the model where at least once the time series is differenced in order to make the stationary data and further the AR and MA terms get combined. So the mathematical statement becomes:

$$Y_t = \omega + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} \\ + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \quad (15)$$

2.2.7 Seasonal Autoregressive Integrated Moving Average Model (SARIMA Model)

Data might contain seasonal periodic component in addition to correlation with recent lags. It repeats for every observation. So SARIMA models the seasonal element in univariate data [5]. To have a better understanding of it, we will take

$$Y_t = \Delta^d X_t \text{ (Here we are taking } X_t \text{ for ARIMA process)} \quad (16)$$

If Y_t is ARIMA then $\Delta^d X_t$ is ARIMA with difference and order of difference d and

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad (17)$$

It can be written as

$$\phi(B)Y_t = \theta(B)Z_t \quad (18)$$

where B = backshift operator,

$$\theta(B) = 1 + \theta_1(B) + \cdots + \theta_q(B)^q$$

$$\phi(B) = 1 - \phi_1(B) - \phi_2(B) - \cdots - \phi_p(B)^p.$$

Along with the trends of ARIMA model, SARIMA is configured with four seasonal elements which can be demonstrated as $S(p, d, q, P, D, Q, s)$ where the variables p = order of non-seasonal AR term, d = order of non-seasonal difference, q = order of non-seasonal MR terms, P = seasonal autoregressive order, D = seasonal difference order, Q = seasonal moving average order, s = number of time steps for a single seasonal period. Hence, $S(p, d, q, P, D, Q, s)$ has the form

$$\phi_p(B^s)\phi_p(B)(1 - B^s)^D(1 - B^d)X_t = \Theta_Q(B^s)\theta_q(B)Z_t \quad (19)$$

$$\text{where } \theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$$

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

$$\phi_p B^s = 1 - \phi_1 B^s - \phi_2 B^{2s} - \cdots - \phi_p B^{ps}.$$

Seasonal differencing can be represented as:

$$D = 1 \triangledown S X_t = (1 - BS)X_t; \quad D = 2 \triangledown^2 S X_t = (1 - BS)^2 X_t \quad (20)$$

2.3 Performance Metrics

In our current analysis, we are evaluating the performance of each model by implementing one of the measures of machine learning called as root mean square error (RMSE) for each model.

2.3.1 Root Mean Square Error (RMSE)

In the view of determining the accuracy of the model, we are calculating the statistical RMSE that has occurred between the predicted values and test values. The predicted errors are computed in terms of distance residuals. It is used for determining the exactness to differentiate the forecasting error of different models for a specific dataset [4]. The mathematical statement for RMSE can be represented as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{Predicted}_i - \text{actual}_i)^2} \quad (21)$$

3 Proposed Model and Methodologies

Our study is enhanced with some scientific research by taking the help of machine learning technology which can be helpful in speeding up the research for early detection of the positive cases, death rate, and recoveries in the forthcoming days by taking the data sample of past few days. The input content in the dataset is the time series data containing the information about the past confirmed cases, death cases, and number of recoveries in Tables 2, 3 and 4, respectively.

From the above analysis, we can see the death rate is quite low in the 14th week as it was expected to rise looking at the trend of the infection of death trend of previous weeks [2]. The death cases are reducing in a consistent manner since 14th week to 19th week. Further there is a spike for two consecutive weeks. Again for further analysis keeping the view of above analysis, our dataset has been taken for preprocessing from the repository of Kaggle on the date of (01/01/2020) till (01/07/2020). Then it has been spitted into two sets as training set (80%) and test set (20%). Training set is used to train the model, whereas test set is used for prediction through several models provided by machine learning models. Further the learning model has been accessed based on the performance metric like RMSE [4]. The recommended system has been depicted in Fig. 3.

4 Experimental Results and Discussion

Till today we all are unaware of this deadly virus. So our attempt is to predict the number of infected including the expected death cases for the upcoming months through the trained machine learning models.

Table 2 Periodic sample of confirmed cases on datewise

Observation date	Province/state	Country/region	Last update	Confirmed
2020-01-22	Anhui	Mainland China	01/22/2020 17:00	1.0
2020-01-22	Beijing	Mainland China	01/22/2020 17:00	14.0
2020-01-22	Chongqing	Mainland China	01/22/2020 17:00	6.0

Table 3 Periodic sample of death cases on datewise

Observation date	Province/state	Country/region	Last update	Death
2020-01-24	Guangxi	Mainland China	01/24/2020 17:00	0
2020-01-24	Shanghai	Mainland China	01/24/2020 17:00	0
2020-01-24	Jiangxi	Mainland China	01/24/2020 17:00	0

Table 4 Periodic sample of recoveries on datewise

Observation date	Province/state	Country/region	Last update	Recovered
2020-01-22	Anhui	Mainland China	01/22/2020 17:00	0.0
2020-01-22	Beijing	Mainland China	01/22/2020 17:00	0.0
2020-01-22	Chongqing	Mainland China	01/22/2020 17:00	0.0

4.1 New Positive Cases Future Prediction

In our study, we are training the models to predict the newly infected cases per month. We are giving the train data of confirmed cases and predicting the future values of newly infected cases. All the models are summarized above according to their RMSE values. From that assessment, LR model and SARIMA model both are showing the lowest root mean square error among the others. But we have considered SARIMA model as the good forecasting model with respect to stability. If we compared with other models, ARIMA is showing poor performance.

It is observed from the LR model by seeing the graph that the predicted values are not at all linear to the actual values [3]. Future predictions of confirmed cases for other models are shown above. Figure 4 shows the performance of different model to predict the confirmed cases by different graph plots graphical and showing that the confirmed cases are increasing (Figs. 1, 2, 3, 4, 5; Table 5).

Weekly Growth of different types of Cases in World

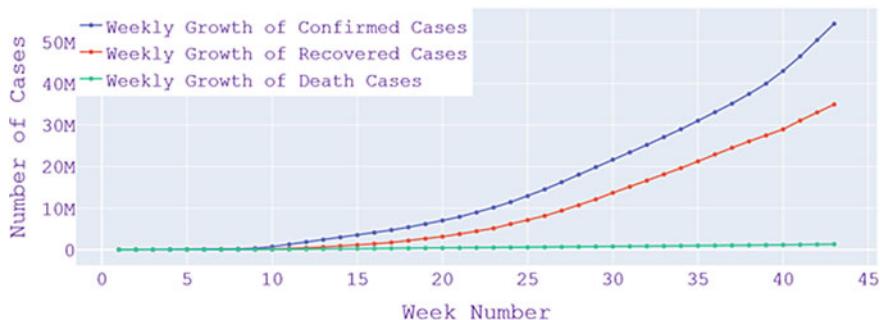


Fig. 1 Growth of different no. of cases worldwide on weekly basis

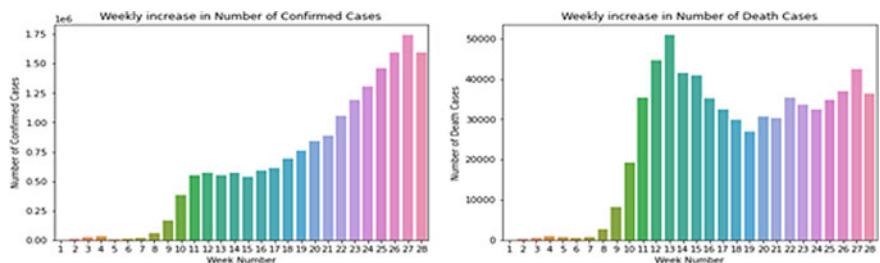


Fig. 2 Analysis of growth of confirmed cases and death cases worldwide on weekly basis

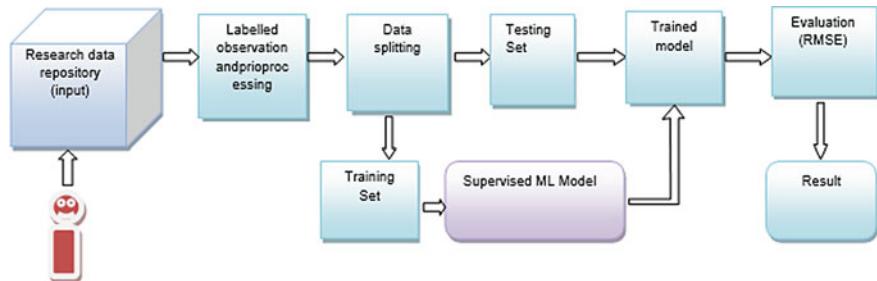


Fig. 3 Proposed model workflow

4.2 Mortality Rate and Recovery Rate Prediction

Mortality rate = (No. of death cases/No. of confirmed cases) * 100

Recovery rate = (No. of recovered cases/No. of confirmed cases) * 100.

From the below pattern, it is clearly visible that for a long period of time mortality rate is showing a positive sign [3].

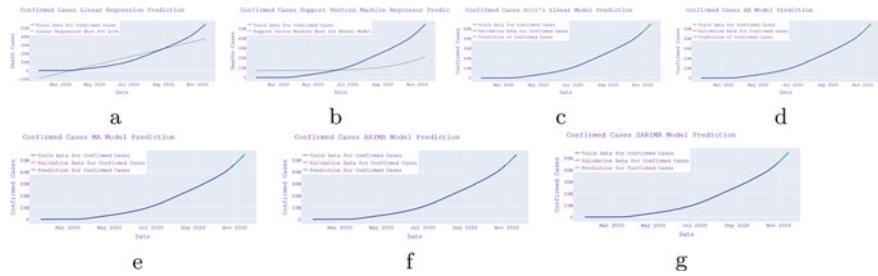


Fig. 4 Monthly prediction of confirmed cases using different models: **a** linear regression model, **b** SVM regression model, **c** Holt's linear model, **d** AR model, **e** MA model, **f** ARIMA model, and **g** SARIMA model

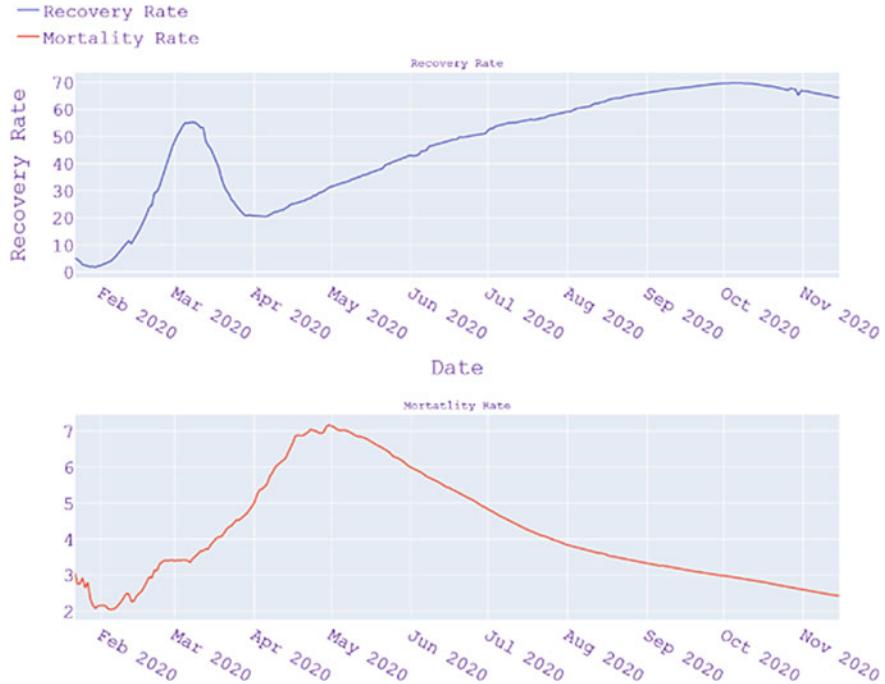


Fig. 5 Mortality rate and recovery rate prediction monthly (till the month of August)

4.3 Death Cases Future Prediction

From the above analysis, we can conclude by analyzing the patterns that there is an improvement rally in case of recovery rate and falling of death cases. As we know, number of closed cases = number of recoveries + number of deaths. So, number of deaths = number of closed cases – number of recoveries [3]. By considering the

Table 5 Performance measurements (prediction and RMSE) of different models for confirmed cases

Dates	Linear regression	SVM	Holt's linear model	AR model	MA model	ARIMA model	SARIMA model
02/08/2020	12,039,092.18	11,846,525.19	17,710,466.16	18,067,087.89	18,100,185.36	18,072,381.20	18,192,747.21
03/08/2020	12,117,043.08	12,093,700.57	17,937,665.98	18,325,345.57	18,367,119.70	18,322,203.55	18,464,381.17
04/08/2020	12,194,993.97	12,346,025.24	18,164,865.80	18,595,027.29	18,635,454.93	18,591,857.08	18,750,271.43
05/08/2020	12,272,944.87	12,603,579.22	18,392,065.62	18,873,991.28	18,905,191.05	18,880,992.27	19,051,677.72
06/08/2020	12,350,895.76	12,866,443.39	18,619,265.44	19,152,541.44	19,176,328.05	19,175,562.27	19,349,446.69
RMSE	70,383.8455	425,612.0205	425,612.0205	425,612.0205	286,175.2579	15,815,137.21	264,666.9359

above formula, we are going to predict the no. of deaths in future which is represented in plots by taking different models and compare it with the actual data.

Although from the graph of ARIMA model and SARIMA model, it is not prominent about the death cases but if we analyze the RMSE values then ARIMA and SARIMA are almost closer but SARIMA is showing the best result. Second best result is shown by ARIMA. We are calculating the RMSE for each model by taking the train input death cases. According to the errors, we can visualize SARIMA and ARIMA models are nearly equal as well as showing the best result. Holt's linear model and AR model are showing equal values as well as considered as the worst.

5 Conclusion

Machine learning is one of the most emerging technologies that has aim to construct intelligent machines that extract decisions from data and applications of statistical techniques. Due to the global threat of COVID-19 pandemic, various researches have been taken up arms to alleviate its impact [8]. We have used ML-based prediction system in this paper for predicting the risk of COVID-19 outbreak. A comprehensive review has done on daywise confirmed and death cases for the analysis of the future



Fig. 6 Monthly prediction of death cases using different models: **a** LR model, **b** SVM, **c** Holt's linear model, **d** AR model, **e** MA model, **f** ARIMA model, and **g** SARIMA model

Table 6 Performance measurements (prediction and RMSE) of different models for death cases

Dates	Linear regression	SVM	Holt's linear model	AR model	MA model	ARIMA model	SARIMA model
02/08/2020	609,452.63	1,074,512.66	676,874.94	690,570.75	690,704.35	690,010.56	692,328.73
03/08/2020	613,287.18	1,101,043.62	681,818.43	696,636.94	696,252.95	695,570.13	697,831.36
04/08/2020	617,121.73	1,128,127.28	686,761.92	703,170.05	703,234.45	701,979.24	705,083.00
05/08/2020	620,956.27	1,155,772.24	691,705.41	709,970.15	709,548.84	709,140.74	712,326.75
06/08/2020	624,790.82	1,183,987.17	696,648.90	716,705.37	715,896.13	716,428.39	719,505.87
RMSE	70,383,845.557	4953,595234	286,175,257950	286,175,257950	9184,325647	2383,052262	1681,268083

prediction of COVID-19 cases. By evaluating the results among different model, we concluded that SARIMA model is providing the best result although it is difficult to render the exact value where it shows the confirmed cases and deaths are increasing as compared to the current situation. LR model also performs well up to some extent. As per the forecasted result, death cases will rise and recovery rate will rise in the initial few weeks of August. Out of all the all models built, SVM produces poor results in all scenarios [9]. Finally, we concluded that model prediction for the current scenario is correct and it can be used to understand the upcoming situation. Our future work will emphasize on using the updated dataset and doing real-time prediction of any pandemic situation [8]. Figure 6 signifies the performance of different models to forecast the death rate by graphical representation and showing that the death cases are increasing significantly (Table 6).

References

1. World Health Organization. <https://covid19.who.int/>
2. John Hopkins University & Medicine, Corona Virus Resource Centre. <https://coronavirus.jhu.edu>
3. Rustam F, Rishi AA, Mehmood A, Ullah S, On BW, Aslem W, Choi GS. Covid-19 future forecasting using supervised machine learning model. IEEE. <https://doi.org/10.1109/ACCESS.2020.2997311>
4. Chakraborty T, Ghosh I. Real time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis
5. Time series analysis, Wikipedia. <https://en.wikipedia.org/wiki/Timeseries>
6. Jahhan EJ (1974) Time series analysis. IEEE Trans Autom Control AC-19(6)
7. The mathematical structure of ARIMA model. http://people.duke.edu/rnau/Mathematical_structure_of_ARIMA_models--Robert_Nau.pdf
8. Jamshidi M, Lalbaksh A, Tall J. Artificial intelligence and COVID-19: “deep learning approach for diagnosis and treatment”. <https://doi.org/10.1109/ACCESS.2020.3001973IEEE>
9. A Pearson’s correlation coefficient based decision tree and its parallel implementation. scholar.google.com
10. Nadella P, Swaminathan A, Subramanian SV (2020) Forecasting efforts from prior epidemics and COVID-19 predictions. Eur J Epidemiol. <https://link.springer.com/article/10.1007/s10654-020-00661-0>

Investigation of Enactment of a Lean Amenity in Joint Provision Centre—A Study of Amenity



S. M. Kaviya, R. Jayanthi, and M. Saravanan

Abstract The primary objective of our research is to design a cell utility for lean provider management the usage of cloud storage carrier. Users can evaluation the stored data to the cloud with protected and safe cloud garage and simple to make use of cloud-based lean service control. Here we concentrate on lean awareness in step with the development technology. This paper builds some extent by means of point investigation of the markers for the procedure of lean execution in a mutual administration center of attention. The lean method might make bigger the stage of elegance in the shared provider heart, and an investigation of dependability and prohibit is applied to verify if this purpose is come to. The strategy considers the advancement of measurements along the employees of the organization examined. The guidelines that reflected the existing circumstance of the procedures are characterized, in mild of the application module such as hard work, techniques, fabrics, machines, measure, and the board. From those markers, five have not been referenced ahead of in the examined writing. The aftereffect of the application demonstrates that even a process that is a piece of an adult shared carrier center may introduce a couple of improvement openings and be nearer to its degree of brilliance through the joint usage of lean and shared service middle.

Keywords Capacity · Excellence in products and services · Lean place of work · Shared carrier middle (SSC) · Stability

1 Introduction

Administrators are persistently gaining ground toward for higher results available in the market, searching for progressively valuable, capable, nimble errands. Their methods by and large fuse come with the advancement of systems and diminishing

S. M. Kaviya (✉) · R. Jayanthi · M. Saravanan
Sathyabama Institute of Science and Technology, Chennai, India

M. Saravanan
e-mail: saravanan.cse@sathyabama.ac.in

of prices. Back place of work services, which can be actions that help associations, speak to around 25–30% of an organization's price range [1–4]. Therefore, to extend the potency of a trade, it is important to be sure that such supporting methodologies are similarly gainful and deft. Setting up a shared organization network (SSC) has been commonly gotten in associations as a technique for supporting strategies and lessening costs helpfully [5–9]. The lean assistance or lean office framework is in like manner noteworthy at the present time, and it was proposed to assist chiefs with making their strategies continuously deft and compelling, regardless of the way that this subject has no longer been extensively discussed in the composition. The lean workplace means applying of lean rules to managerial actions to enhance progress and scale back managerial waste [10–14]. The advantages of consolidate larger adaptability and snappier response to exhibit alterations and it has been carried out with victory in private and public sectors.

2 Literature Survey

Protik Basu, Indranil Ghosh, "Structural Equation Modelling-Based Empirical Analysis of Technical Issues for Lean Manufacturing Implementation in the Indian Context," 2018. This paper method to examine and join the precise problems for tough use of lean manufacturing. A large composing investigation is done to collect an extensive review of all the knowledge explicit presentations fundamental for lean execution, joined with a near thorough once-over of the significant number of focal issues gathered from its powerful use. A crucial model is along those traces conceptualized, which is observationally affirmed matter to the data from the Indian collecting division. The basic type showed is depended upon to be used by means of the lean grasp for appropriate affiliation and the administrators of the tilt utilization procedure. This work is one in all without a doubt the primary asks going to have a survey-based take a look at examination of an in each way that truly matters exhaustive once-over of specific knowledge components and points of interest of lean manufacturing usage in the Indian atmosphere.

Chen Lixia and Meng Bo, "How to Make 5S as a Culture in Chinese Enterprises," 2008. In over 20 years, a ton of endeavors in Chinese Mainland have achieved 5S that is the reason for lean introduction in delicate of its bizarre accomplishment in a lot of countries, alternatively a significant number of them omitted to snatch their distinctive targets since they benefit from 5S absolute best as an device and hope to benefit from it briefly time without combine with inside of sight tradition. The paper brings up fundamental flawed assumptions and blunders of Chinese undertakings in actualizing 5S by the use of exam in assembling ventures, which ship in regards to the unhappiness of 5S management, and proposes steps to complete 5S program effectively to be explicit learn to make 5S as a convention. Besides, the paper proposes elementary components for ventures in Chinese Mainland to make 5S as a convention, which can be critical for them to prevail or flop in 5S management. We believe that

Chinese undertakings might succeed in 5S management and provides a company social premise to the association of lean era in China.

T. Žilić and V. Čošić, "Implementing Shared Service Center in Telecom Environment as More Efficient and More Cost Effective Business Model," 2016. Looking for constantly precious and even more fiscally shrewd strategy is a consistent factor for manager in IT trade, specifically in affiliations like immense telecom providers which might be population. This paper is regulating a kind of new tactics, execution of Technology Shared Service Center in bizarre in comparison to other telecom providers in Europe. To have any such device, and stunningly extra to fuse it within two or three national affiliations, is an enormous authoritative trade. Prompting telecom development is an additional inspiration for transferring to new tactics. Along these strains, utilization of the shared service center in telecom, or any IT business, could be both dynamic and creative test. Other than the difficulties went going through while on the similar time executing Technology Shared Service Centers, this document is portraying the elemental focal points of this style with amazing spotlight on telecom situation.

Malak Baslyman; Daniel Amyot; Yasser Alshalabi, "Lean Healthcare Processes: Effective Technology Integration and Comprehensive Decision Support Using Requirements Engineering Methods," 2019. Medicinal products and services face a lot of difficulties in conveying higher assistance quality and pleasurable, quickly, and evolving wishes. Lean administration attracts near, in most cases embraced in social insurance, give techniques and apparatuses to process growth. Be that as it will, the main lure of the lean methodology in human services and products is a skinny spotlight on continual qualities and must have that incessantly bars the ones of parental figures, among others. Guardians in clinics are an uncommon more or less employees, no longer quite the same as those of most different associations. They are critically inundated in a novel environment in which they deal with end-clients (patients) with a large number of components to imagine and different basic strategies to pursue. It is fundamental to take into consideration their needs and targets, however, authoritative targets and qualities, in any exchange the executives procedure, in particular if adjustments come with new innovation. We propose a lean-AbPI type that joins the main ideas of lean administration with the activity-based process integration means (AbPI). AbPI gives a few reconciliation choices of new procedures into present ones while breaking down the impact of the progressions on spouse/consumer wishes, hierarchical targets, and execution locations. The usage of the lean-AbPI fashion is proven through a genuine medical institution contextual investigation. The outcomes display that utilizing the style, an exhaustive investigation is given that activates improved choice assist and utilization similar strategies of reasoning recognizability.

3 Related Work

3.1 *Lean Service/Lean Office and SSC*

Lean brooding about is in line with the Toyota Production System, which used to be as soon as evolved in a manufacturing environment inside the automobile business. The core goal of the tilt manufacturing is a continuing search for improvements, increasing the efficiency of processes via turning in merchandise or products and services in a waste-free means. The primary goal of this implies is to eliminate waste and processes that do not upload value to the product or purchaser. A service corporate may also be outlined as one that has intangible, perishable products and services and products, and principally products and services and merchandise performed within the presence of the client. Damrath comprises the function of variety of products and assistance, that is, in keeping with the Damrath, the human factor is a distinctive inside the procedure, which makes normalization of products and assistance tricky. In this state of affairs, the tilt assistance can assist the id of scrap, by the use of making use of lean production ideas to carrier operations.

4 Existing System

The existing concept implements the lean management system will represent an accountable and difficult job, the belief of which require considerable financial sources in addition to human and mechanical resources. The present set of rules, Black–Scholes pricing style is in large part used by possibility buyers who buy choices which might be priced beneath the formulation calculated worth and sell choices that are priced higher than the Black–Scholes calculated value. The handiest problem is not showing the correct information about lean carrier.

5 Proposed System

The lean method would possibly make bigger the degree of magnificence in the shared service heart, and an investigation of dependability and prohibit is applied to confirm if this objective will be reached. The development of measurements alongside the workers of the group tested. The present set of rules, Black–Scholes pricing style is in large part used by possibility buyers who buy choices which might be priced beneath the formulation calculated worth, and sell choices that are priced higher than the Black–Scholes calculated value. The handiest problem is not showing the correct information about Lean carrier.

6 Modules

6.1 *Login and Registration*

In this module, we design to extend login and signup monitor. Android used xml to develop classical displays in our utility. The modules describe signup website contains email identity or consumer identify, password and conform password those quite details will need to be stored in database. Login display contains email identity or username and password when the buyer to login the app it should be retrieve the info to the database and blend in response to person enter if its match user title and password to allow within the app differently alert and display a message to the person.

6.2 *Production Analytics*

In this module, the manufacture product is uploaded within the non-public cloud storage. The uploaded product will probably be IN, OUT, and DAMAGE to the gross sales management analytics of production knowledge can also be get and store knowledge in cloud server.

6.3 *Historical Analytics*

The ancient data of all merchandise can be seen and analyzed. The information has been segregated in the personal cloud which may also be retrieved with seek research. Historical knowledge is processed to evaluate the tilt management.

6.4 *Quality Waste Analytics*

The quality waste management is among the fastest growing waste streams, which includes a broad and increasing spectrum of products. Accurate estimation of e-waste era is tricky, basically because of loss of fine quality information referred to market and socio-economic dynamics.



Fig. 1 System architecture

6.5 Database Creation

User electronic mail identification or user name and password had been stored after registration. Android used SQLite database for storing and fetching user software main points.

7 System Architecture

A generalized system architecture for the discussed scheme has been given in Fig. 1. It is a modularized architecture and consists of various components, such as, user, analytics module, data storage, wastage management, interface etc. The analytics module will basically be using historical data.

8 Future Enhancement

The long run contains AI solution to handle composition and execution of an Internet sale framework. This framework displays an online exhibit of sophistication of wised items they need to promote or offer.

9 Conclusion

In view of the dialogs did all the way through the evolution of this research, it was noticed that the broke down fashion of markers might be utilized for SSCs that seek to execute a lean technique, paying little appreciate to the specific concept of the procedure, gave that its estimation group used to be accurately adjusted. A correlation with the writing regarding the means of markers demonstrated that seven of the investigated guidelines had been referred to within the writing in regards to the matter, regardless of whether no longer implemented to SSC. Subsequently, discoveries of this paper add to both practice and hypothesis. To get started with, in hypothetical phrases, in view of the aftereffects of the stability and limit learn about, it is used to be obvious that even an increase SSC composition with a style perceived in evaluating investigation had a large number of development openings when broke down from lean ideas. For the situation study displayed, Company A demonstrated that no longer precisely 50% of its markers have been seen as competent, and an enormous volume of the tips demonstrated flimsiness, with territory 1 because the region with essentially the most noticeably terrible exhibition, with part of the flimsy markers and 80% inadequate ones. In this feeling, this paper raised sagacious effects referring to significance of soundness and limit in the case of a lean utilization in administrations. At the top of the day, the writing has been widely underlining the pertinence of strength and prohibit with recognize to lean execution, with distinctive thoughtfulness regarding fabricating stipulations. Be that as it will, an identical investigation in management associations have been so much lesser, which subvert any pageant on such matter. This paper connects this hollow via tending to this factor inside the SSC, giving observational proof to have the same opinion that, paying little thoughts to this sort of affiliation, soundness, and limit were essential viewpoints for a lean execution, despite the fact that they may not be at a similar level.

References

1. Li G, Field JM, Davis MM (2017) Structuring lean procedures with advanced assistance high quality: an application in cash related administrations. Qual Manag J 24(1):6–20
2. Ramos LJT (2005) Serviços Compartilhados como Forma de Estruturação Organizacional. Ace's postulation. Department of Administrator, Bahia Federal University, Salvador
3. Shinetal C (2008) Expanded without cellular DNA focuses in patients with obstructive rest apnea. Psychiatry Clin Neurosci 62(6):721–727
4. Su N, Akkiraju R, Nayak N, Goodwin R (2009) Shared services transformation: conceptualization and valuation from the perspective of real options. Decis Sci 40(3):1–36
5. Saravanan M, Mathew S (2014) Structural clustering approach: to enhancing performance of automation reference tool GRASP. Int J Appl Eng Res 9(22):15723–15731. ISSN 0973-4562
6. Saravanan M, Mythili (2014) A novel approach of horizontal aggregation in SQL for data mining. Int J Eng Trends Technol (IJETT) 9(1)
7. Saravanan M (2015) A novel approach for ant colony optimization algorithm in artificial immune network system. 10(1). ISSN 0973-6077

8. Saravanan M, Jyothi VL (2015) Clustering of knowledge sets by way of using fuzzy algorithm. ARPN 10(4). ISSN 1819-660
9. Khan MY, Md. Sabeelur Rahman K, Albert Mayan J, Saravanan M (2015) Improvising group work using clustering and sequential pattern mining. EJSR Eur J Sci Res 133(4):496–500
10. Saravanan M, Nithya (2015) Recapitulation of coding to enable data integrity protection in cloud storage. Glob J Pure Appl Math (GJPAM) 11(6):4169–4175
11. Sanchana VB, Renuga S, Saravanan M (2016) A novel approach for environment friendly data handling in cloud setting. ARPN J Eng Appl Sci 11(17)
12. Jailin Reshma A, Jenushma James J, Kavya M, Saravanan M (2016) An evaluation of persona recognition all for offline handwriting. ARPN J Eng Appl Sci 11(15)
13. Nagarajan G, Minu RI (2018) Wireless soil monitoring sensor for sprinkler irrigation automation system. Wireless Pers Commun 98(2):1835–1851
14. Nagarajan G, Minu RI (2015) Fuzzy ontology based multi-modal semantic information retrieval. Procedia Comput Sci 48:101–106

Online Product Review Monitoring System Using Machine Learning



Sandra Johnson, J. Madhumathi, R. Aishwarya, and V. Vedha Pavithra

Abstract In this fast moving world, having major proportion of people depending on online Web sites for purchase of their day-to-day necessities, their only scope of trust is the reviews available in the about the particular product. These reviews can either be authentic made by loyal buyers or even it may be forged for marketing purposes. According to survey, almost sixty percent of the reviews present in Amazon are fake and nearly fifteen percent of companies as sellers, wage individuals for creating fake reviews. In order to settle the imbalance and forgery, a study has been made to identify the fake reviews and filter it out from the user's sight, such that they cannot be manipulated as well as the particular company's reputation will not be at stake. This system majorly works on the basis of identifying fake reviews by user extracting the multiple review of single user from reviewers

Keywords Reviews · Machine learning · Fake · Signed inference algorithm · Genetic algorithm

1 Introduction

Reliability is a common and major concept which that everyone has come across as an evaluation constraint in most scenarios, in case on online shopping the only source that buyers rely on is the reviews made by users and the star rating of the

S. Johnson (✉) · J. Madhumathi · R. Aishwarya · V. Vedha Pavithra

Departments of Computer Science and Engineering, R.M.K. Engineering College, Kavaraipettai, India

e-mail: sjn.cse@rmkec.ac.in

J. Madhumathi

e-mail: madh16218.cs@rmkec.ac.in

R. Aishwarya

e-mail: aishle17401.cs@rmkec.ac.in

V. Vedha Pavithra

e-mail: vedhle17409.cs@rmkec.ac.in

product. More than star rating, the reviews are what make up the buying or denying decisions of users. The user rating and reviews serve as a major factor in driving sales, though these reviews sound and are projected as progressive improvement for purchaser benefit, the research and statistical output says the otherwise.

The scholars found into being that average consumer ratings associated off-color with the tallies from consumer reports when compared, having alteration in average user rating between twosomes of yields larger than one star, the item with the upper user rating was appraised more favorably by consumer reports only around two-thirds of values provided. Additionally, for instance when comparing a laptop provided with average rating of four out of five stars along with one more laptop which has average rating of three out of five stars, the first laptop would only be objectively better 65% and not the expected 100% at most of the situation. It makes a far difference in quality of product; this applies even more apt for reviews than star rating.

There are many reasons as to why the reviews are manipulated; the reasons would either be in favor of the marketer or competitor. For the marketer, in order to have positive reputation in market, as it serves as a most needed asset for driving sales, they would generate positive reviews. On the other hand, for competitors to increase their sales and to plummet the success of their rival companies, they would write negative reviews about the product by their rivals. In either of these cases, customers are those who suffer the most.

To overcome this problem, a new and novel online product review monitoring system to detect the net spam is proposed. The system serves the purpose by evaluating the reviews and then labeling them to be fake or genuine, in a way that fake reviews are hidden from the sight of users, protecting the customers from being manipulate from spending their money on unworthy products. The system uses signed inference algorithm and genetic algorithm to achieve this evaluation and labeling of reviews.

2 Literature Review

Liang et al. [1] explained a mechanism to detect spam reviews by taking both feature and relationship of the reviewer in consideration, as at the present time, most of the customers can gain copious information and help for making decisions in purchasing products and service from online review assets, through reviews in social media. It, on the other hand, stimulates some manufactures in appointing spammers for writing fake criticisms as well on some target products. The concept of how to perceive fake review as well as review creator is seeking the consideration of marketers and ecommerce. A fresh and unique graph model with multiedge, having for each node demonstrating a referee and apiece edge embodies a reviewer's inter-relationship on each one of special product. Reviewers' unreliability score is the feature based on which combing is done, and an unendorsed continuous working out framework is proposed. Moreover, it is by far the first set of rules where both the reviewer's

possibilities and inter-relationships between them are considered together. Experimental results show that the method is effective in detecting spam reviewers with a satisfied precision, but requires additional implementation requirements such as dataset powered up with detailed attributes.

Lin et al. [2] discussed on spotting of the bogus reviews from the provided sequential reviews in online social sites, as distinguishing and identifying spam in review is way more significant for present e-commerce bids. However, the displayed order of review has been abandoned by the previous research works. It considers the problem on fake review uncovering in review categorization, and it is crucial for instigating online anti-opinion spam implementation. The characteristics of fake reviews are first needed to be analyzed. Then based on valuation contents and critics behavior, a six period profound details are suggested for fake reviews highlighting, following that the process devises an administered elucidation followed by threshold-based solution for spotting the fake reviews at earliest possible situation. The trial results indicate that undertaken methodology is capable of identifying the fake reviews methodical with higher accuracy and recollection, but the limitations are that it requires training datasets, as well as the model works only on sequential data.

Istiaq Ahsan et al. [3] discussed about the issues in online marketing such that, entire of the e-commerce has begun to get mammoth as the days passes by, even if it does not by every passing minute. Online evaluations have a crucial role in the online marketing arena, as well as, it has proved itself for being promising in terms of judgment constructing from the eyes of shoppers. The shoppers' only scope of trust was the reviews. Moreover, these are precisely profound and substantial facts according to the customer, that would make certain the genuineness of user-generated gratified discussion groups, reviews, blogs, media, blogs, and so on is unpredictably perceptible. The limitations were that, it allowed spam recognition in only those contents with word count above 150, and shorter reviews were not taken into consideration.

Rajamohana et al. [4] have performed a survey on techniques that can be used in spam detection of reviews. Having our commonplace activities getting majorly impacted by Internets influence, e-commerce is facing the rapid development zones in the Internet era. The survey that has a detailed survey is done using various mechanism learning practices for sensing spam and sincere reviews, but the limitations are that they provide multiple approaches but not specify the efficient and suitable one.

Zhang et al. [5] proposed a collective hyping mechanism for identifying the forged reviews in online activities as ever since the advent of online shopping bogus reviews always misinform consumers shopping online. He proposes a new Concerted Promotion Hyping Recognition solution. But the limitation is that it cannot identify the individual spammers in social media as it focuses on masses.

Jia et al. [6] articulated a LDA-based system for detecting spam in online reviews as it is obligatory for latent consume to contribute a conclusion based on reviews been made by users online. The disadvantage is that it requires more evaluation time, as the system checks the reviews. Artificial intelligence is the ability to process information properly in a complex environment. The criteria of properness are not predefined and hence not available beforehand [7, 8].

Shehnepoor et al. [9] portray the influential nature of online social media in information transmission among people. The limitation is that it just focuses on classification of fake reviews and it does not work on users [10].

3 Proposed Work

In simple terms, this project “online product review monitoring system” considers the delinquent of perceiving fake commentators, thereby as a result identifying fake reviews in online review datasets obtained. The datasets of online review predominantly comprise of customers or reviewers as users, a set of products made available such as mobile phones, laptops, and finally the reviews. Individually each criticism is transcribed from a particular user to a particular product and contains a star rating, often an integer from 1 to 5. As such, a Bayesian network is used to represent the review dataset. The network is displayed in such a way that, user nodes are associated with the merchandise nodes, having the links signify the “reviewed” connections and a review rating attained for each review. The manipulators, goods, and assessment object in the analysis grid are grouped into certain classes, such that, two classes for each object type: yields are what is more good or bad quality, users are either truthful or deceit, and finally the reviews as are either genuine or forged.

The different phases and internal flows within the system are portrayed in the figure, and the system functioning begins with registration and login if in case of user level access and only login if the user has admin level access. Initially the admin logs into the online product review monitoring system and adds products along with its description and images. Later when user logs in and searches for products, list of products registered by admin is displayed to the user, from which user can select their desired products after going through the description, images, and reviews made available for the product by admin. All these are functioning that take place in the surficial level in open eyes.

Internally, after reviews are made by user for products, the reviews are evaluated and labeled to be of deceit or authentic by the system using esteemed algorithms. These algorithms label a review to be fake on three bases, if a review is made by user before buying or product, if the reviews originate from same user numerous times, and if the contents of review does not match with the metadata of particular product for which review was formed.

After the reviews are labeled, only the reviews categorized as genuine are listed at the user side, the rest are made available only under admin level access login. The admin has additional functionalities such that, the admin can remove a product from list and also delete the reviews marked as fake. The admin is also provided with user management rights, such that the admin can monitor the functioning of users based on the extracted user id.

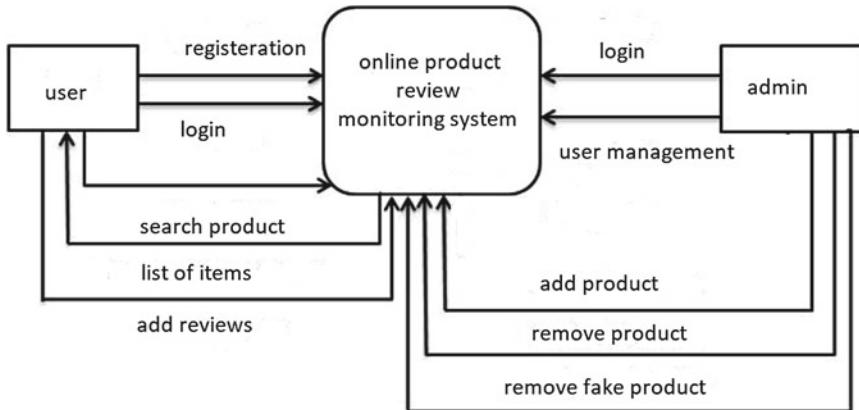


Fig. 1 System architecture

4 System Architecture

From Fig. 1, this system takes products and reviews as input for the user and admin, the provided data are sequentially processed by signed inference algorithm and genetic algorithm, and these algorithms work on processing the reviews and products by labeling them and on how to display the processed data. The signed inference algorithm is majorly about labeling, and the genetic algorithm is inclined toward display mechanism of the labeled contents.

The output of the system would hide the fake reviews from customers, as well as the system provides an interactive display of feedback to user, by including a friend option in system, where in case for any product searched by customer, contains reviews made by their friends from the application, their reviews will be displayed on top, provided with higher priority, whereas to admin all fake and honest reviews will be displayed but in a classified manner, where fake reviews are categorized to find out if the fake reviews are created by particular users repeatedly, subsequently tracing out the spammers as well.

5 Methodology

The system depicts the real-world online shopping problem, and the product review which, on the one hand, serves as scope of trust for customers but, on the other hand, also as a weak spot, which the manipulators target to change the rate of profit in their favor. This system, online product review monitoring system to detect the net spam, uses two algorithms signed inference algorithm and the genetic algorithm. Both these algorithms together as a single entity is implemented in multiple functionalities and in mapping of modules. As the system as whole works in evaluation, identification

and labeling of reviews, through multiple submodules of process, the execution of same algorithm exists in multiple submodules. To ideate and display this real-world working, a basic online shopping Web site is developed and incorporated.

From Fig. 2, the signed inference algorithm has its algorithms pseudocode involved all the way along in program in numerous functionalities, and the algorithm serves many purposes such that, it helps in concealing the fake and bogus reviews from users, it evaluates if the product has been brought by the user for the particular review, if the particular review is the first review made by the esteemed user for that particular product, and finally the algorithm check for the relevance of metadata of product for with review is made along with the review itself.

The genetic algorithm is implied for this system in two aspects in low key rating and for user review sorting purposes. The algorithm is applied to the low key rating functionality of the system in a way that, the reviews with lower star rating provided that less than 2 would lead to product being removed from system automatically without intervention from admin. This functionality included acts as an added advantage to the system, as it automatically removed product of lower star rating level, than being removed manually by the admin.

Then, the algorithm is applied for user review sorting functionality in order to sort the way reviews are displayed to the user, and the algorithm works in a manner that all the reviews with similar values are displayed for a given product and also the friend functionality executes on the basis of this algorithm, if any reviews for the selected product are from friends of the user, then these reviews are displayed with higher priority above the reviews from other users.

6 Results and Discussion

Thus, the reviews are classified into fake and genuine by considering the user id, product id, number of products bought by the user in a particular time, review content, and product metadata.

In Table 1, the user details are displayed to the admin. In Table 2, the reviews given by the user for particular products they bought are displayed. In Table 3, the reviews which are evaluated as fake are displayed to the admin.

The accuracy of fake review detection is tested with 100 product reviews containing both fake and genuine reviews. The hybrid signed inference algorithm with the genetic algorithm has identified fake reviews with 96% accuracy. Out of 100 product reviews, 25 product reviews were fake and the rest are from genuine reviews. A total of 24 fake reviews out of 25 fake reviews in the given dataset is identified correctly with 96% accuracy.

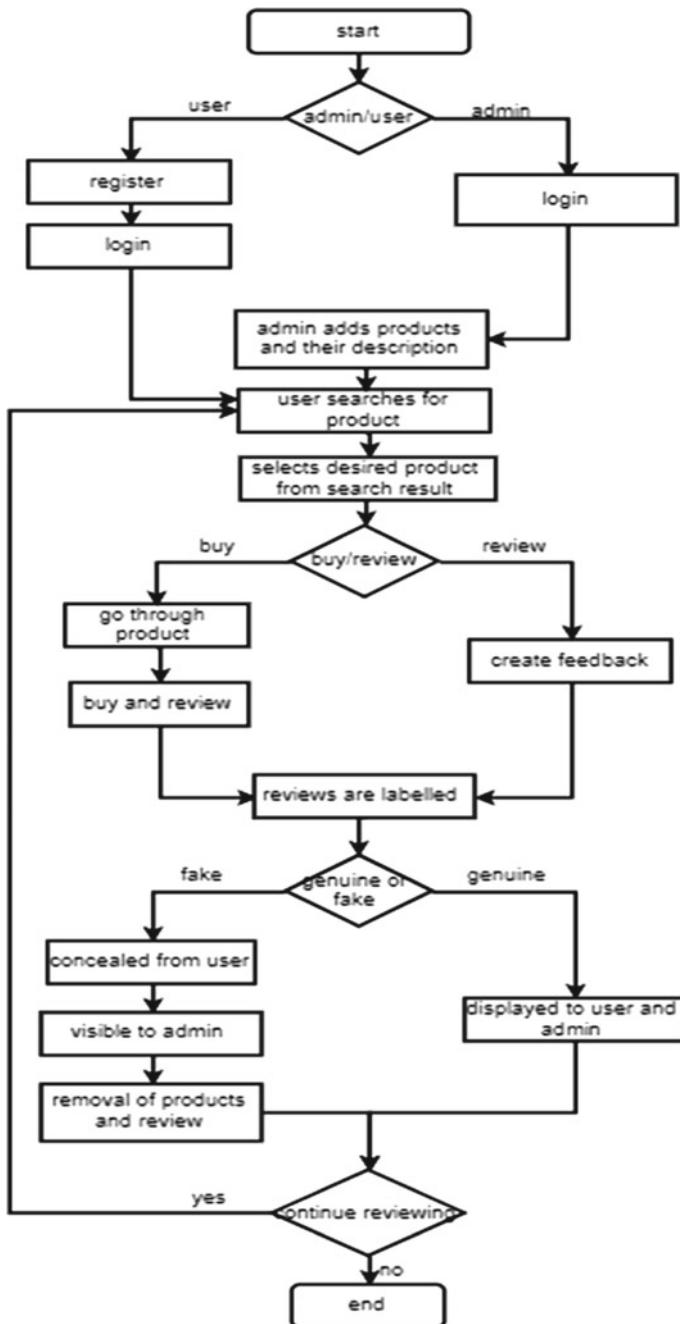


Fig. 2 Methodology

Table 1 User details

User name	Mail id	Number	Location	Gender	Block
Anu	venkyt1991@gmail.com	9,790,721,621	Chennai	Female	Block
Kiran	venkyt1991@gmail.com	9,790,721,621	Chennai	Male	Block
pavithra	pavithrat1994@mail.com	9,790,721,621	Kallakurichi	Female	Block
Harini	harini@gmail.com	9,234,321,567	Villupuram	Female	Block
kubra	kubra@gmail.com	9,234,321,533	Hyderabad	Female	Block
Indira	indira@gmail.com	9,234,321,123	villupuram	Female	Block
Indu	indu@gmail.com	9,234,321,221	Chennai	Female	Block
Jen	jen@gmail.com	9,234,321,567	Chennai	Male	Block

Table 2 Product reviews

Product id	User id	User name	Ratings	Feedback
6	3	Pavithra	5	Good
4	3	Pavithra	3	nice
5	1	Anu	5	Very nice
6	1	Anu	5	Very nice
2	1	Anu	5	Good
1	1	Anu	5	Supper
1	2	Kiran	4	Not bad
1	2	Kiran	5	Good
1	2	Kiran	5	Very good

Table 3 Fake reviews

Product id	User id	User name	Ratings	Feedback
5	1	Anu	5	Very nice
6	1	Anu	5	Very nice
2	1	Anu	5	Good
1	1	Anu	5	Supper
3	1	Anu	1	Laptop is good
8	5	Kubra	5	Amazing movie
7	1	Anu	4	Good
Null	Null	Null	5	Good
Null	Null	Null	2	Khdgfkad

7 Conclusion

Thus, this is used to identify the genuine reviews from all other reviews given by the customers for a particular product by concealing the fake reviews from the users using machine learning. This helps the customer to identify quality product and helps the online shopping Web site to achieve customer satisfaction. Our work on fake review identification using a hybrid signed inference algorithm with the genetic algorithm has given 96% accuracy which is an improvement over the other works. This work can be further enhanced with other machine learning algorithms and evaluated using different datasets collected from e-commerce Web sites.

References

1. Liang D, Liu X, Shen H (2014) Detecting spam reviewers by combing reviewer feature and relationship. In: International conference on informative and cybernetics for computational social systems (ICCSS)
2. Lin T, Zhu T, Wu H, Zhang J, Wang X, Zhou A (2014) Towards online anti-opinion spam: spotting fake reviews from the review sequence. In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014)
3. Istiaq Ahsan MN, Nahian T, Kafi AA, Hossain MI, Shah FM (2016) Review spam detection using active learning. In: IEEE 7th annual information technology, electronics and mobile communication conference (IEMCON)
4. Rajamohana SP, Umamaheswari K, Dharani M, Vedackshya R (2017) A survey on online review spam detection techniques. In: IEEE international conference on innovations in green energy and healthcare technologies (ICIGEHT'17)
5. Zhang Q, Wu J, Zhang P, Long G, Zhang C (2017) Collective hyping detection system for identifying online spam activities. IEEE Intell Syst 32(5)
6. Jia S, Zhang X, Wang X, Liu Y (2018) Fake reviews detection based on LDA. In: 4th IEEE international conference on information management
7. Sundaram BV (2009) Review of software architectural styles for artificial intelligence systems. Int J MC Square Sci Res 1(1):96–112
8. Hemanth Kumar G, Ramesh GP (2019) Reducing power feasting and extend network life time of IOT devices through localization. Int J Adv Sci Technol 28(12):297–305
9. Shehne poor S, Salehi M, Farahbakhsh R, Crespi N (2017) NetSpam: a network-based spam detection framework for reviews in online social media. IEEE Trans Inf Forensics Sec 12(7)
10. Swarnalatha A, Manikandan M (2020) Intravascular ultrasound image classification using wavelet energy features and random forest classifier. In: Intelligent computing in engineering. Springer, Singapore, 803–810

Gold Price Prediction Using an Evolutionary Extreme Learning Machine



Jyoti Prakash Mishra and Smruti Rekha Das

Abstract Prediction of gold price has always been most fascinated due to its nonlinearity and dynamic time series behavior, which is constrained by so many influencing factors such as economic data, monetary policy, supply and demand, inflation, and currency movements. Immemorial gold is always having the highest degree of monetary rewards and has been termed as oldest precious metal used in global currency. After understanding the hidden pattern behind the prediction of various things, which needs very fast decisions to make the computational cost of the market, researchers have proposed many statistical and machine learning models for gold price prediction. In this study, an evolutionary extreme learning machine (ELM) is designed for future gold price prediction, where two evolutionary estimation paradigms are suggested such as particle swarm optimization (PSO) and differential evolution (DE) during the training stage to optimize the weights of the network. The performance of the prediction model is measured through mean square error (MSE) and evaluated on GOLD/USD collected with six-year period of time. Through this study, a better prediction model can be designed, which will help the gold investor in taking decision for the best time of investing money in the gold market.

Keywords Gold price prediction · Extreme learning machine (ELM) · Particle swarm optimization (PSO) · Differential evolution (DE)

1 Introduction

In recent years, it is noticed that there is an outstanding growth by investing money in gold by the investor due to the instability, which is exist in flat money value and the increase trend of the price of gold. Hence, time series prediction of future gold price

J. P. Mishra · S. R. Das (✉)

Department of Computer Science and Engineering, Gandhi Institute for Education and Technology, Bhubaneswar, India
e-mail: smrutirekhadas@gietbbsr.com

J. P. Mishra
e-mail: jpmishra@gietbbsr.com

becomes a most important tools for the investors, which will help the investors to take decisions in [1] investing their money in selling or buying transaction of golds. Based on the hypothesis, there is a complex relationship exist between current value and the previous value. On that basis, prediction methods can be categorized into three types such as [2] objective forecasting method (prediction methods based on the quantity), subjective forecasting method (prediction methods based on the qualitative), and prophecy (educated guessing). This study focuses on objective forecasting method. Various prediction methods have been designed till now for financial as well as commodity market, and it is found that the main [3] categories are classical methods, hybrid approaches, and artificial intelligence. Though classical methods can solve many linear applications and artificial neural network (ANN) can solve nonlinear problems exist in time series data, there is some drawback that exists between the above two methods which can be solved by the hybrid methods. Hence, this study emphasizes over hybridized prediction model.

After getting the huge popularity of neural network-based methods for market prediction, this study considers extreme learning machine (ELM) as the prediction model, optimizing the network with two evolutionary based methods such as particle swarm optimization (PSO) and differential evolution (DE). The performance of the prediction model is evaluated using GOLD/USD dataset.

This paper is organized in the following ways: Sect. 2 covers the methodology adopted for this experimentation. The dataset description and result analysis is arranged in Sect. 3 and Sect. 4, respectively. Finally, the paper concludes in Sect. 5.

2 Methodology Adopted

The methods that have adopted for this experimentation is described in this section.

2.1 *Extreme Learning Machines (ELMs)*

The working principle of ELM is based on the concept of neural network, where there is a connection between each nodes of the input layer with each other node of the hidden layer. The weights and biases are chosen randomly, but unlike ANN, in ELM, the output weights are generated analytically. The random selection of input weights and biases is the main cause behind creating the non-optimal solution, which needs a greater number of neurons than the required neurons for conventional [4, 5] learning algorithm, which needs tuning at each iteration. The essential part of ELM, in which it differs from single layer feed neural network (SLFN) is that its output weights are calculated analytically.

2.2 Particle Swarm Optimization (PSO) Algorithm

This population-based PSO algorithm designed by following the social activities of the biological communities. In PSO, a swarm of particles is kept on and trying to search the global best solution, where moving of each individual is continued [6–9] in the direction of the best particle of the entire population, which occurs at each iteration. Let x_i indicates the position i in the search space. P indicates the discrete time steps, and v_i denotes the velocity at the time instant P . The position of the particle is changed by adding velocity to the current position which is given in (1).

$$x_i(P + 1) = x_i(P) + v_i(P + 1) \quad (1)$$

This is the equation for the position. The particle velocity for the gbest, PSO is calculated using (2).

$$\begin{aligned} v_i[P + 1] = & w \times v_i[P] + C_1 \times r_1(P\text{best}_i - x_i[P]) \\ & + C_2 \times r_2(g\text{best} - x_i[P]) \end{aligned} \quad (2)$$

Here, w is the particle inertia, and C_1 and C_2 are the positive acceleration constant. The quantities r_1 and r_2 are positive random numbers.

2.3 Differential Evolution (DE)

DE is an evolutionary algorithm. It works through a simple cycle of stages, such as (i) initialization of the parameters, (ii) mutation with different vectors, (iii) crossover, and (iv) selection. The operation of DE is very similar to genetic algorithm (GA) where it is having the same phases [10–12], but the sequence of phases is different. Here, in DE, at every generation, mutation occurs first and then recombination takes place. It is like multipoint crossover in GA. Selection is the third operator in DE, whereas it is the first operator in GA to perform comparison among the solutions in the same generations. Here, in DE through selection, it pushes the better solution to the next generation by passing crossover and mutation, and it is performed between parent and offspring.

3 Dataset Description

The commodity market data, such as GOLD/USD dataset within the range of Dec 3, 2015, to Dec 3, 2020, period of time are considered for prediction in this study. The available features for GOLD/USD dataset are price, open price, high price, low price, volume, and change of price. The gold data available for predictions are in 1

Troy Ounce Unit. This experimentation has worked out for four time horizons such as 1 day, 3 days, 7 days, and 1 month in advance. A total of 70% of data is selected for training and 30% is selected for testing.

4 Result Analysis

The simulation is carried out for GOLD/USD dataset using ELM, ELM-DE, and ELM-PSO prediction model. ELM is a neural network-based model and in the last decades, it is observed that neural network-based methods have shown excellent performance in prediction in the financial market as well as commodity market. Though standard ELM performs well in the prediction, but through optimizing the weight and bias, a better prediction can be generated. Hence, for optimization, two standard evolutionary optimization methods such as PSO and DE are considered here in this study. Though the simulated graph of all the prediction models such as ELM, ELM-DE, and ELM-GA evaluating over GOLD/USD are generated, due to lack of space, only the simulated result of ELM-DE is given in this study. The result of the rest of the prediction model can be observed through the MSE result during training given in Table 1. The actual versus predicted simulated graph of GOLD/USD dataset, considering four time horizons such as 1 day, 3 days, 7 days, and 1 month in advance using ELM-DE is shown in (a), (c), (e), and (g) of Fig. 1, respectively. Simultaneously, the error convergence graph for the same time horizon using GOLD/USD dataset is depicted in (b), (d), (f), and (h) of Fig. 1, respectively.

From the actual versus predicted graph, it can be visualized that the predicted graph is very close to the actual one and also if it will be compared among the four time horizons, then it is observed that in 1 day ahead predicted outcome, predicted graph is closer to the actual one than the rest days ahead prediction. Similarly, from the error convergence graph, it can be observed that 3 days ahead is converging faster than the rest days, but here, one thing can be visualized that after 3 days, 1 month is converging faster. The converging speed of error is coming next for one day and 7 days, respectively.

From the MSE result during training is given in Table 1, it can be observed that ELM optimized with two evolutionary optimizing techniques such as PSO and DE is showing better results than base model of ELM. It can also be stated that ELM-DE is generating minimum error than ELM-PSO. Comparing the generated error among

Table 1 MSE calculation of ELM, ELM-DE, and ELM-PSO using GOLD/USD dataset

Days ahead	ELM	ELM-DE	ELM-PSO
1 day	5.6216	2.33183	2.90278
3 days	6.9674	4.23981	5.3335
7 days	21.8941	12.8007	14.8531
1 month	23.0471	13.5418	30.1187

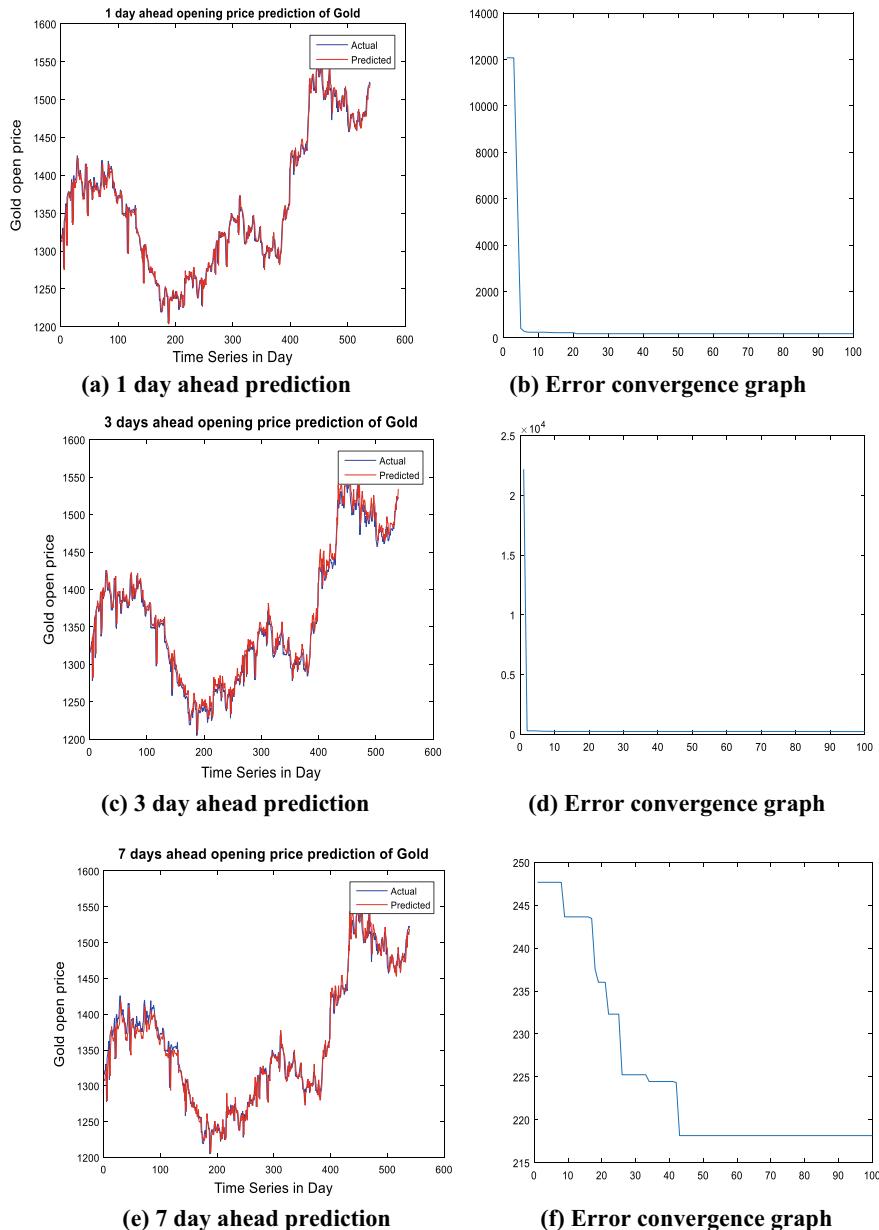


Fig. 1 Simulation graph for ELM-DE for **a** 1 day, **c** 3 days, **e** 5 days, and **g** 1 month ahead prediction and error convergence graph of evolutionary ELM-DE **b** 1 day, **d** 3 days, **f** 5 days, and **h** 1 month ahead prediction

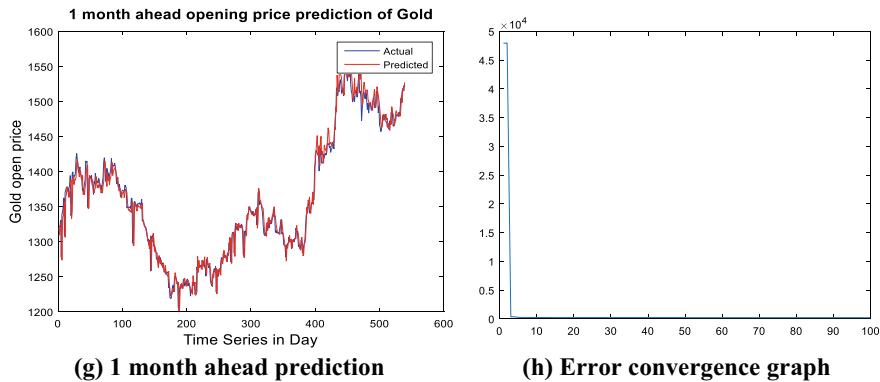


Fig. 1 (continued)

the experimented days, it is noticed that 1 day ahead is showing a minimum error and the error rate increases with the increasing days.

5 Conclusion

An evolutionary gold price prediction model is introduced, which is evaluated on GOLD/USD dataset. Since the last decades, prediction of gold price has always been more enthralled, which needs a perfect prediction to generate a great wealth. Hence, this study has considered three prediction model such as ELM, ELM-DE, and ELM-PSO. Both PSO and DE are two widely accepted optimization techniques, performing excellent in optimizing prediction model. In this experimental work, ELM-DE is noticed to be a better prediction model than the rest experimented model. In addition to this, one day ahead is observed to be given a minimum error with maximum accuracy than the rest of the day considered here for experimentation. In future, this work can be extended for a different gold price dataset, considering the influencing factors over the fluctuating data.

References

1. Hussein, Shamsul FM, Shah MBN, Jalal MRA, Abdullah SS (2011) Gold price prediction using radial basis function neural network. In: 2011 fourth international conference on modeling, simulation and applied optimization. IEEE, pp 1–11
2. Zhang G, Eddy PB, Hu MY (1998) Forecasting with artificial neural networks: The state of the art. *Int J Forecast* 14(1):35–62
3. Farahani K, Mahsa, Mehralian S (2013) Comparison between artificial neural network and neuro-fuzzy for gold price prediction. In: 2013 13th Iranian conference on fuzzy systems (IFSC). IEEE, pp 1–5

4. Zhu Q, Qin AK, Suganthan PN, Huang GB (2005) Evolutionary extreme learning machine. *Pattern Recognition* 38(10):1759–1763
5. Huang G, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst Man Cybern Part B (Cybern)* 42(2):513–529
6. Chakravarty S, Dash PK (2012) A PSO based integrated functional link net and interval type-2 fuzzy logic system for predicting stock market indices. *Appl Soft Comput* 12(2):931–941
7. Pulido M, Melin P, Castillo O (2014) Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange. *Inf Sci* 280:188–204
8. Abdual-Salam ME, Abdul-Kader HM, Abdel-Wahed WF (2010) Comparative study between differential evolution and particle swarm optimization algorithms in training of feed-forward neural network for stock price prediction. In: *Informatics and systems (INFOS)*, 2010. The 7th international conference on IEEE, 1–8
9. Briza AC, Naval PC (2011) Stock trading system based on the multi-objective particle swarm optimization of technical indicators on end-of-day market data. *Appl Soft Comput* 11(1):1191–1201
10. Das S, Suganthan PN (2011) Differential evolution: a survey of the state-of-the-art. *IEEE Trans Evoluti Comput* 15(1):4–31
11. Hegerty B, Hung C, Kasprak K (2009) A comparative study on differential evolution and genetic algorithms for some combinatorial problems. In: *Proceedings of 8th Mexican international conference on artificial intelligence*, 9–13.
12. Hachicha N, Jarboui B, Siarry P (2011) A fuzzy logic control using a differential evolution algorithm aimed at modelling the financial market dynamics. *Inf Sci* 181(1):79–91

Stock Market Evidence on Investor's Predispositions Impacting Portfolio Return



Sai Rashmi Patra and Shakti Ranjan Mohapatra

Abstract The behavioral finance studies the human psychology behind investments in securities. This article investigates the investor's predisposition toward heuristics over rational decision making in stock market. It also explores the impact of such intuitive-based decisions on the investment performance. The sample consisted of 410 registered individual investors of National Stock Exchange, India. The study is a novel one as it examines the compounding impact of heuristics of investors on investment behavior and prospects of earning desired return. The study finds that the investors emphasize more on heuristics than on rational decision-making process due to time and resource constraints. Collection of detailed information about stocks followed by making meticulous analysis of alternatives prior to investment as envisaged in the Mintzberg model is followed sparingly. The research concludes that every time the investors go for heuristics, they generate less return on their investment. Multivariate analyses have proved the hypotheses. The implication of this study lies in educating the individual investors about the shortcomings of heuristics in investment decisions and its negative impact on the return on investment. Further, investment advisors would stand to benefit of this research.

Keywords Heuristics · Rational investment decision · Investors · Stock market · Investment performance

1 Introduction

Neoclassical economics makes three fundamental assumptions about people. These are (i) people have rational preferences, (ii) they aim at maximizing utility, and (iii) they make well-informed independent decisions, Weintraub [1]. It holds that people make rational economic decisions every time as rationality is fundamental to their

S. R. Patra
College of Engineering and Technology, Bhubaneswar, India

S. R. Mohapatra (✉)
Faculty of Management, Biju Patnaik University of Technology, Rourkela, Odisha, India

financial decisions and socioeconomic activities. Neumann and Morgenstern [2] proposed in their expected utility theory that investors make consistent and independent decisions among various alternatives. Ackert and Deaves [3] write that Maurice Allais contradicted the expected utility theory popularly known as Allais Paradox. Kahneman and Tversky [4] came up with their famous prospect theory. They studied human psychology and confirmed that individuals' decisions are guided more by their intuitions and emotions. Bernstein [5] observed repeated patterns of irrationality and inconsistency in investment decisions of people. According to Shefrin [6], behavioral finance recognizes human psychology which explores human behavior under risky and uncertain circumstances which is different from traditional economic assumptions. Subash [7] pointed out that human traits like emotions, preferences, likes or dislikes, passion, love or hatred, beliefs, judgments, or heuristics influence them to take an easy path and end up in a poor decision. Parikh [8] states that "people have mind and heart, but they do not always make decisions out of their mind. When decisions are made from the heart, they are emotional and may not be rational." Baker and Nofsinger [9] highlighted that cognitive errors, psychological biases, and fundamental heuristics collectively influence an individual to take an easy path for an investment decision. Dutta et al. [10] observe that if the market is performing well but the retail investors are performing poorly, it would construe that investors do not behave rationally. Chandra and Kumar [11] studied the behavior of individual investors and observed that they are not free from intuition-based judgments in the stock market scenario. Mintzberg et al. [12] had proposed a three-step process of decision making for arriving at a rational decision. This study attempts to explore to what extent the investors make use of the model and if they achieve desired return on their investment.

The next sections of the article are as follows; Sect. 2 is devoted to literature review and hypothesis development. Section 3 discusses the research methods. In Sect. 4, an analysis of the study is made, and Sect. 5 concludes with discussions on the results and policy implications of the research study.

2 Survey of Literature and Hypothesis Development

In this section, the following paragraphs present the available research studies on rational decision-making process, behavioral attributes of individuals, and investment performance as outcome of such decisions of individual investors.

2.1 Rational Decision-Making Process

All investors aim at optimal decisions for a good return. Standard finance assumes that people gather full information, analyze the same, and then make rational decisions. Individual investors think rationally and invest based on estimations, Kubilay et al.

[13]. They analyze and use different models of standard finance to judge if expected return commensurates with the risk while making investment decisions, Arora and Kumari [14]. Mintzberg et al. [12] have postulated a rational decision model that involves (i) identifying problem or goal setting, (ii) developing alternatives, and (iii) selecting the best alternative. The model assumes that investors should adhere to the three-step process for an optimal investment solution. The decision makers need to gather information and evaluate competing substitutes from different possible situations before selecting a choice, Oliveira [15]. Uzonwanne [16] presented a seven-step model of rational decision making involving (i) identifying the problem; (ii) identifying the solution scenario; (iii) carrying out a gap analysis; (iv) gathering facts, options and alternatives; (v) analyzing option outcomes; (vi) selecting best possible options; and (vii) implementing decision for solution and evaluating final outcome. The latter is an extension of the Mintzberg model. But, in reality, the investors might not be following either due to lack of time and resource constraints. There is every possibility that they might look for shortcuts, popularly termed as heuristics in psychology for their investment decisions. In the process, it is likely that they commit cognitive and emotional errors leading them to earn low return.

2.2 *Heuristics in Decision Making*

Heuristics are behavioral shortcuts that people adopt in their daily activities including financial decisions. According to Kahneman and Tversky [4], people use heuristics as rule of thumb to make simple decision. They concluded that heuristics often lead to systematic error, and decision maker is subjected to adverse outcomes. As and when investors are tempted to apply heuristics, they usually make irrational decisions, Bazerman [17]. Ritter [18] reaffirms that investors use rule of thumb to deal with uncertainties in market. Sometimes, heuristics prove to be beneficial, particularly when the investors do not have sufficient time in hand, and when they have to make quick decisions, Waweru [19]. Researchers like Bikhchandani et al. [20], Pompian [21], Shefrin [22], Baker and Nofsinger [9], have investigated different psychological biases concerning the investment decisions and market outcomes. Dangol and Manandhar [23] have pointed out that there is a significant relationship between irrationality in investment decision making and heuristic biases.

In this paper, all five types of heuristics are studied to show how they impact investment decision-making process, and to examine if the investors follow the Mintzberg model of rational decision making to enhance the quality of decision and the expected return.

2.3 Representativeness Bias and Investment Decision

Kahneman and Tversky [4] explain the representativeness as the extent to which an event represents its population. Ritter [18] observes that due to representativeness, people give importance to the recently acquired knowledge. Pompian [21] confirms that investors show excessive dependence on stereotypes and generalize an outcome based on too small samples. Quite often, the investors update their beliefs based on simple classification of data and ignore complex data. Due to too much reliance on stereotypes, individuals make inappropriate predictions. They make wrong forecasts and arrive at dubious conclusions, Shefrin [6]. Researchers like Chen [24] and Waweru et al. [19] confirm that representativeness bias and investment decisions are negatively correlated. Arthur [25] and Yaowen [26] have concluded that the decision-making quality gets adversely affected by the representativeness bias. The study of Toma [27], however, finds that the return on individuals' investment has gone up due to representativeness bias. In view of opposite findings, the current study proposes to test the following hypothesis:

H₁: Representativeness has no impact on the individual investor's decision-making behavior.

2.4 Availability Bias and Investment Decision

Investors generally rely on information readily available to them and accordingly form their opinion about a stock, Kahneman and Tversky [4]. Brahmana et al. [28] observe that investors give great importance to events occurring frequently and ignore the need for a thorough analysis. Consequently, they fail to construct an optimal portfolio. But, Ikram et al. [29] observed an increase of returns as individual investors depended mostly on available information at the Islamabad Stock Exchange. However, Waweru et al. [19] confirmed that there was negative relationship of availability heuristic with that of the investment decisions of financial institutions trading at the Nairobi Stock Exchange. Since studies have divergent conclusions as regards impact of availability bias on investment behavior and the returns, the researchers develop the following hypothesis for testing in the article:

H₂: Availability heuristic has least impact on the individual investor's decision.

2.5 Anchoring Bias and Investment Decision

Investors make decisions based on data that come to their notice, and their decisions revolve around it. It is most likely that they often fall into the trap of anchoring bias. Kahneman and Tversky [4] observe that people tend to make investment decisions based on some known anchors. They look at the purchase price when they intend to sell a stock. They define a range for a share price based on the past trends. Thus, they

under-react to changes, provided it is within the range. Waweru et al. [16] find that the anchoring bias has positive effect on investment decisions. Luong and Ha [30] describe the positive influence of anchoring bias on the individual investor's decisions at the Ho Chi Minh Stock Exchange. But, Murithi [31] confirms that anchoring bias has negative effect on investment decisions. Investors afflicted with anchoring bias make irrational decisions leading to poor investment return. As such, the following hypothesis shall be tested in the present study:

H₃: Anchoring heuristic has no impact on the individual investor's decision.

2.6 *Gambler's Fallacy Bias and Investment Decision*

Gambler's fallacy refers to a situation where investors believe that they have adequate information to predict the point of reversal of stock prices in the near future. Shefrin [22] observes that financial strategists are more prone to gambler's fallacy. They often inappropriately predict the reversals of events. O'Neill and Puza [32] state that investment with the hope of reversal of trend is an irrational behavior. Waweru et al. [19] suggest that gambler's fallacy arises in the stock market when people dispose their stock as it has gone up in many past consecutive trading sessions expecting an imminent fall. The investors forget that it is the fundamental strength of the company and not the price of the previous trading session that will take the stock to the higher level, Singh [33]. Singh [33] also states that investors who hold or buy a stock visualize a very low probability of further decline as the price has fallen in past trading sessions. They forget that stock goes down due to its fundamental weakness. The current study tests the following hypothesis:

H₄: Gambler's fallacy does not impact the individual investor's decision making.

2.7 *Overconfidence Bias and Investment Decision*

Overconfidence bias is defined as having a strong faith in one's own intuition and reasoning. Investors believe that their cognitive abilities are outstanding. They often overestimate their strategies. But at the same time, they underestimate the risks associated with the investment. Investors take the happening of an event for granted, which may or may not happen because of changing economic scenario. According to Pompian [21], investors who evaluate themselves more than what they are, and attribute positive results to their strategies and skill, suffer from overconfidence bias. Moore and Healy [34] observe that individuals affected by overconfidence bias show overestimation, over-placement, and over-precision. Investors rely more on their skills and competency, consider themselves better than others, and become exceedingly confident of their judgment. These investors ignore the risk factors in the investment, Odean, [35]. Das and Mohapatra [36] have observed that every individual has self-confidence, but sometimes they show overconfidence and ignore to the

contrary opinion as they are overconfident about their prediction. Baker and Yi [37] have observed that overconfidence bias impacts decision making. Investors underestimate risk and overestimate expected profit. The studies made by Odean [35], Park et al. [38] pointed out that overconfidence makes the investors trade excessively. As a result, they get a low return on their investment. Kengatharan and Kengatharan [39], Waweru et al. [19] argue that overconfidence affected the decisions of individual/institutional investors. DeBondt and Thaler [40], Gervais [41], Kansal and Singh [42] find that the overconfidence bias leads the investors to commit judgmental errors. Madaan and Singh [43] have, however, concluded that overconfidence has significant positive impact on investment decisions. Based on the above studies, following hypothesis is tested in the present study:

H₅: Overconfidence heuristic does not impact the individual investor's decisions.

2.8 *Heuristic Factors Impacting Investment Decisions and Return on Investment*

Investment return or performance motivates investors to stay invested in stock market. But, their investment decisions mostly drive them to make mistakes leading to either earn less return or lose substantial value of their investment. Various research findings have documented that investors tend to avoid rational analysis when they are confronted with uncertainties and time constraints. Simon [44] introduced a concept called bounded rationality which states that investors are not completely irrational. They make some analysis to ascertain if their investment would have the prospects of earning a good return for them. But, there are contradictory findings as regards the correlation of application of heuristics in investment and earning a desired return. Anderson, Henker, and Owen [45] found that overconfident individual investors made higher amount of transactions and consequently earned greater returns than individuals with fewer transactions. Kim and Nofsinger [46] claimed that stocks lying long with the individual investor's possession yielded a negative return, whereas the stocks in possession for a short duration yielded a positive return. Abdin et al. [47] have confirmed that availability and overconfidence biases do not impact investment performance.

The study involves three aspects related to investment such as (i) rationality (ii) preference of individuals for heuristics, and (iii) return on investment as a measure of outcome of the decisions. The present study is expected to establish a relationship of heuristics, investment decision behavior, and the consequential impact on generating expected return. The following hypothesis is formulated for the purpose.

H₆: The heuristic variables have significant adverse impact on the portfolio performance.

3 Objectives and Methods

3.1 Objectives

Based on the theoretical inputs as presented above, the objectives of this research paper are:

- (i) To investigate the predisposition of individual investors toward heuristics over rational decision-making process, and
- (ii) To study the impact of heuristics on the investment decision making and return on investment.

3.2 Research Methods and Design

The research study uses a survey method through collection of primary data. The population for the study was over forty million registered individual investors who were actively trading through authorized broking houses at National Stock Exchange of India as on 30.6.2019. According to the Taro Yamane formula (as referred by Babajide and Adetiloye, [48], the valid sample size should be 399.98 (rounded to 400) at an expected standard error level of 0.05. The researchers used Surveymonkey.com facility and got responses from 421 individuals residing in 12 commercial cities of four sides of India. Responses of 410 individuals complete in all respect were retained for the analysis.

The survey instrument, a closed-ended questionnaire, consisted of three parts. The first part of the questionnaire was meant for collection of demographic and investment data. The second part consisted of 16 questions on heuristic items. It also contained seven questions to assess if the investors make use of the rational model for their investment decisions. The questionnaire on heuristics was developed in the line of the research conducted by Kengatharan and Kengatharan [39] and Waweru et al. [19] and rationality items in the line of studies undertaken by Kumar and Goyal [13], Arthur [25] and Shah and Oppenheimer [49]. The final part included three items to make an evaluation of the investment performance as an outcome of their decisions. The items were picked up from the studies of Kengatharan and Kengatharan [39]. Part-2 and Part-3 used five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Responses were put to descriptive and multivariate analysis using SPSS—22.

4 Analysis and Interpretations

4.1 Reliability

The reliability and validity of the questionnaire were checked through the Cronbach's Alpha coefficient. The test result shows the alpha coefficient to be more than 0.75 for each of the heuristic variables, rationality items, and investment performance questions. Thus, the items included in the questionnaire are reliable.

4.2 Demographic and Investment Profile Analysis

The compiled data reveals that 71.2% of the respondents are male, 75.8% are within the age group of 35–55 years. More than 86% of respondents were graduates and above, 65.1% belonged to salaried class. More than 71% had annual income within the range of INR 500,000 to 1,500,000. It concludes that majority of investors are financially comfortable. It further displays the investment experience and corpus. About 70% of respondents had an investment experience of over five years, the average experience being 6.78 years. As large sections of the respondents are experienced, their opinions seemed to reflect their true investment behavior.

4.3 Rational Investment Decision Making Process of Respondents

Table 1 shows the mean responses of the decision making process under the three heads, such as goal setting, information seeking, and evaluating alternatives for investment and the corresponding t-values and significance.

The responses on goal setting and problem identification relate to investment plans and objective of potential gain. The mean score response is 2.83. The next three questions broadly studied the information-seeking behavior like fundamental strength, performance indicators, and growth parameters of the target companies. The

Table 1 Rationality in investment decision making ($N = 410$)

Sl no	Statements	Mean	SD	T value	Sig
1	Goal setting/identification (q_1, q_2)	2.83	0.87	-5.61	0.00
2	Information-seeking behavior (q_3-q_5)	2.92	0.82	-5.81	0.00
3	Evaluating alternatives for investment (q_6, q_7)	2.91	0.79	-4.88	0.00
	Composite mean score	2.89			

Source Survey data, (significant level: 5%)

mean score for this behavior is 2.92. The last two items on evaluating alternatives for investment gets a mean score of 2.91. The composite mean score is 2.89. The mean scores below 3.00 indicate that respondents do not take the investment decision too seriously. The *t*-values show the significance of the mean difference. It concludes that they have not followed the three-step model meticulously before an investment decision.

4.4 Investment Decision Making Under the Influence of Heuristic Variables

Table 2 presents the mean scores of responses as regards the investment behavior of the respondents under five broad heuristic variables.

The availability heuristic receives the highest grand mean of 4.23, followed by the overconfidence heuristic with a grand mean of 4.21. The mean score level lying within 4–5 implies the great influence of these two biases on the investor's decision making. The biases like representativeness, anchoring, and gambler's fallacy occupy the third, fourth, and fifth positions with mean values of 3.94, 3.90, and 3.68, respectively. These three heuristics also impact the investment decision of the respondents.

Close scrutiny of the item-wise responses on all the biases displays the following main behavioral issues of investors:

- (i) The respondents consider availability of ready information as strong motivator to buy local stocks (Mean = 4.55, SD = 0.91, and *t* = 34.44). Besides, they prefer to use information from close friends and relatives as a reliable reference for their investment decisions (Mean = 4.16, SD = 0.96, and *t* = 24.68). Both these responses point out the availability disposition of investors.
- (ii) The responses on overconfidence variable show that the respondents claim to have appropriate skill and knowledge to identify good stocks for their investment (Mean = 4.45, SD = 0.97, and *t* = 30.21). They can predict the stock price movement after simple analysis (Mean = 4.27, SD = 0.99, and *t* = 25.91). They also believe that their stock price will rise in bull market (Mean

Table 2 Mean score data for heuristics variables (*N* = 410)

Sl. no	Heuristics	Mean	SD	Rank
1	Representativeness (<i>Q1</i> to 3)	3.94	0.952	3
2	Availability (<i>Q4</i> to 6)	4.23	0.812	1
3	Anchoring (<i>Q7</i> to 9)	3.90	0.944	4
4	Gambler's fallacy (<i>Q10</i> and 11)	3.18	0.812	5
5	Overconfidence (<i>Q12</i> to 16)	4.21	0.824	2
	Grand average	3.89		

Source (Survey data)

$= 4.24$, $SD = 0.93$, and $t = 26.96$). The respondents also attribute successful transactions to their own skill (Mean = 4.15, SD = 1.02, and $t = 25$) which shows their overconfidence in taking decisions.

- (iii) The respondents exhibit representativeness in their investment decisions as they prefer to buy hot stocks and avoid stocks that have performed poorly in recent past (Mean = 4.18, SD = 1.01, and $t = 23.6$).
- (iv) The respondents are found anchoring on the current price because of good performance of the stock (Mean = 4.17, SD = 0.89, and $t = 27.21$).
- (v) The respondents fall into gambler's fallacy as they use their past experience for buying stocks (Mean = 3.58, SD = 1.07, and $t = 10.94$). They also feel that they are able to predict trend reversal looking at the recent stock price movements (Mean = 3.28, SD = 1.04, and $t = 5.49$).

4.5 Mutual Exclusivity Test Between Heuristics and Rationality

The mean scores of heuristic-induced decisions and rational decisions of the respondents are compared to test the mutually exclusiveness of these two variables. The composite mean value of rational decision-making variable is 2.89. The mean scores of each heuristic variable ranges between 3.18 and 4.23. The composite mean below 3.00 indicates the low level of importance attached to the rational decision-making model. In contrast, scores above 3.00 signifies the high level of importance attached to heuristics in decision making. This proves that the investors are inclined to take decisions based on heuristics. In order to explore if there is any correlation between each of the heuristics with each component of rational decision-making process, the r-matrix is prepared and presented in Table 3.

The data in the table depicts the degree of positive correlation between individual items of heuristics and components of rational decision-making model. It shows adoption of heuristics in decision-making process.

Table 3 R-matrix: correlation between heuristics and rational decision-making components

Serial no	Heuristics	Rational decision-making process		
		Goal setting	Information seeking	Evaluating alternatives
1	Representativeness	0.14	0.11	0.18
2	Availability	0.16	0.19	0.13
3	Anchoring	0.12	0.14	0.17
4	Gambler's fallacy	0.16	0.12	0.14
5	Overconfidence	0.14	0.11	0.16

Source Survey data, (Significant level: 5%)

Table 4 Response on investment performance ($N = 410$)

Sl. no	Statements	1	2	3	4	5	Mean
1	My last year stock investment return meets my expectation	113 (27.6)	150 (36.7)	59 (14.4)	52 (12.6)	36 (8.7)	2.38
2	My investment return is more than the average market index return	99 (24.1)	148 (36.1)	55 (13.3)	59 (14.3)	49 (12)	2.62
3	I am satisfied with my investment decisions	72 (17.6)	79 (19.2)	130 (31.8)	57 (13.8)	71 (17.4)	2.93
		Grand mean of investment performance					2.64

Source Survey data (Figures in parenthesis indicate percentage)

4.6 Portfolio Performance

Table 4 exhibits the details of the responses on the three items concerning the portfolio performance. These question items were similar to the ones used by Kengatharan and Kengatharan [39]. The mean values on first two items reflect that majority of respondents (64.3%) do not agree to have met their expectation. About 60% of respondents have not earned better than the average index return. More than 36% of respondents do not seem to have been satisfied with their investment decisions, whereas 31.2% of respondents report satisfaction of their investment decisions.

The average score below 3.0 for each question implies that the investors have not met desired return due to some latent mistakes in their decision-making process.

4.7 Relatedness of Heuristic Variables with Portfolio Performance

. The relationship of each heuristic variable with its other components and lastly with that of the investment performance variable is studied through correlation coefficients as presented in Table 5.

Table 5 shows the positive correlation among all the heuristic variables. But, each heuristic variable shows a negative correlation with the performance. For example, the representativeness bias and portfolio performance is negatively correlated (with r value = -0.156). The r value between availability bias and portfolio performance is -0.226. Similarly, other heuristic variables are also negatively correlated. It implies that

Table 5 Correlation between the heuristic variables and portfolio Performance

Variables	Mean	1	2	3	4	5	6
Representativeness	3.94	1					
Availability	4.23	0.345	1				
Anchoring	3.90	0.321	0.376	1			
Gambler's fallacy	3.68	0.242	0.237	0.352	1		
Overconfidence	4.11	0.434	0.384	0.386	0.379	1	
Portfolio Performance	2.76	(–)0.156	(–)0.226	(–)0.122	(–)0.112	(–)0.217	1

Source Survey data

application of heuristics has negative impact on the investment decisions resulting in poor performance of their portfolio. As the bias increases, the quality of investment decision gets affected, and consequently, the rate of return on investment falls. The positive correlation among the heuristic variables indicates that as and when the investors adopt one heuristic, it also induces them to go for another heuristic. The negative r values of each heuristic with the portfolio performance support the hypothesis (H_6).

4.8 Multicollinearity Test and Linear Regression Equation

The degree of influence of heuristics on investment performance is studied through multiple regression analysis. As a precondition to regression analysis, the study tests multicollinearity among the independent variables. The minimum acceptable tolerance value should be 0.11, and the variance inflation factor (VIF) values should remain below 5, Sharma and Firoz, [50]. In this study, each of these independent variables are found to be more than 0.455 and their VIF values in the range of 1.315 and 1.634. It concludes that there does not exist any multicollinearity issue.

Tables 6, 7, and 8 present the model summary, ANOVA, and the regression coefficient for the regression equation, respectively. The impact of behavioral biases on investment performance is calculated through a linear regression model. In the following equation, investment performance (IP) is taken as dependent variable, and five heuristic biases are independent variables.

Table 6 Model summary

Multiple R	0.696
R^2	0.484
Adjusted R^2	0.480
Standard error	3.540
Observations	410

Table 7 Analysis of variance

	Df	SS	MS	F	Sig
Regression	5	14,576	2915	216.57	0.000
Residual	404	5437	13.46		
Total	409	20,013			

Table 8 Regression coefficient for investment decision making

	Unstandardized coefficient		Beta	t-stat	p-value
	B	SE			
Constant	5.676	1.826		3.108	0.000
representative (Rp)	0.235	0.106	-0.161	2.519	0.016
Availability (Av)	0.332	0.146	-0.186	2.774	0.029
Anchoring (An)	0.220	0.124	-0.121	2.776	0.012
Gambler's fallacy (Gf)	0.169	0.112	-0.102	2.511	0.010
Overconfidence (Ov)	0.227	0.120	-0.227	2.892	0.011

$$\begin{aligned} \text{IP} = & \alpha + \beta_1(\text{Representativeness}) + \beta_2(\text{Availability}) \\ & + \beta_3(\text{Anchoring}) + \beta_4(\text{Gamblers fallacy}) + \beta_5(\text{Overconfidence}) + e \end{aligned}$$

This equation explains the contribution of each independent variable to the dependent variable.

The results of Table 6 show that 48.4% change in the dependent variable is due to independent variables. The variation in the investor's decision making is predicted due to heuristic biases. Rest 51.6% change in the dependent variable is due to other variables which are not included in the study. The model is fit for the prediction of the investor's decision making. Table 7 shows the values of different statistics of the analysis of variance

By putting the values in the regression equation, we get,

$$\begin{aligned} \text{IP} = & 5.676 - 0.161(\text{Rp}) - 0.186(\text{Av}) \\ & - 0.121(\text{An}) - 0.102(\text{Gf}) - 0.227(\text{Ov}) \end{aligned}$$

The above equation indicates that the independent predictors with negative coefficients adversely impact the return on investment. It can also be stated that for every unit application of heuristics like representativeness, availability, anchoring, gambler's fallacy, and overconfidence in decision making, the investment performance will drop by 0.161, 0.186, 0.121, 0.102, or 0.227, respectively. The *t*-values (Table 8) in respect of these heuristics are individually more than the benchmark of 2.4, and the *p* values are less than 0.05 at significance level of 95%. This confirms that the heuristic variables have significant negative impact on the decision-making behavior of the individuals. As a result, the performance of the investment gets

Table 9 Results of tests of hypothesis

Hypothesis statements	Results	Study result/finding also coincides with
H ₁ : Representativeness has least impact on the individual investor's decision making	Not supported	Arthur et al. [25]
H ₂ : Availability heuristic has least impact on the individual investor's decision	Not supported	Clark [51]
H ₃ : Anchoring heuristic has least impact on the individual investor's decision	Not supported	Yaowen et al. [26]
H ₄ : Gambler's fallacy does not impact the individual investor's decision making	Not supported	Loweis et al. [52]
H ₅ : Overconfidence heuristic does not impact the individual investor's decisions	Not supported	Waweru et al. [19]
H ₆ : The heuristic variables have significant adverse impact on the portfolio performance	Supported	Kengatharan et al. [39]

adversely affected. The results do not support the hypothesis that the heuristic variables have least impact on the rationality of the investors and the investment performance. The research findings conclude that the investors prefer heuristics to rational analysis, and consequently, their investment does not perform as expected.

Thus, all the hypotheses (H₁, H₂, H₃, H₄, and H₅) about heuristic variables with no adverse influence on decision-making behavior are rejected. The hypothesis that the investment decisions have significant adverse impact on return of the portfolio of the investors is accepted.

5 Summary Results and Tests of Hypothesis

Table 9 presents the summary results of hypothesis tested above. It is found that all the hypotheses have not been supported except the H₆ which proves the adverse impact of heuristics on the portfolio performance.

6 Conclusions and Policy Implications

People invest in securities with an anticipation to earn a decent return on their investments. The three-step model proposed by Mintzberg et al. [12] is expected to guide

them take rational decisions. But, individuals ignore the process as they believe that they could earn good returns at the earliest applying shortcuts. But, in real-life situations, the outcome is different; they end up losing their valuable investment. The task of choosing the right stocks, at the right time, and at the right price, does not appear simple. Institutional investors get adequate information and have analytical acumen due to resources at their disposal. They gather and process information to arrive at an objective investment decision. In contrast, the high degree of uncertainty in the market makes individual investors skeptic. Therefore, individual investors adopt shortcuts as an easy alternative to rational analysis of facts and figures. Based on available information, they make impromptu decisions to buy, hold, or sell stock. They apprehend that if they fail to act today, they miss the opportunity and the probable gain. In such circumstances, heuristics offer them an easy solution and the investors move away from rationality.

This study has discussed the investment behavior of the individual investors trading on the NSE of India in the light of representativeness, availability, anchoring, gambler's fallacy, and overconfidence heuristics. The researchers find that the heuristic biases provoke the investors behave irrationally. The results further reveal that all the heuristics have negative influence on the investors' decisions. The investors fall into the traps of various biases, leading to inappropriate investment decisions and poor investment returns. The result of this study coincides with that of the findings of Waweru et al. [19], Park et al. [38], Yaowen et al. [26], Arthur et al. [25], Clark [51], and Loweis et al. [52].

The implications of this study are that the individual investors should be made aware of the heuristic biases underlying their investment decisions. This will prevent them from committing errors leading to erosion of their investment. Besides, it is necessary to provide them a need-based securities-investments training to help them apply rational decision making before any further investments. This paper studied the interrelationship of various heuristic variables with that of the rational decision-making process of individual investors in the stock market. The scope could be further extended to analyze the influence of other variables on investment decision making, investment performance, and on the market efficiency.

References

1. Weintraub ER <http://www.econ.lib.org/library/enc/NeoclassicalEconomics.html>(2008)
2. Neumann JV, Oskar M (1944) Theory of games and economic behaviour. Princeton University Press, Princeton, New Jersey
3. Ackert LF, Richard D (2016) Understanding behavioural finance. Cengage Learning India
4. Kahneman D, Tversky A (1974) Judgment under uncertainty: heuristics and biases. J Sc 85(4157):1124–1131
5. Bernstein PL (1998) Against the Gods: the remarkable story of risk. Wiley, USA
6. Shefrin H (2000) Beyond greed and fear: understanding behavioral finance and the psychology of investing. Harvard Business School Press, Boston, MA
7. Subash R (2012) Role of behavioural finance in portfolio decisions: evidence from India. Diploma thesis, Charles University, Prague

8. Parikh P (2009) Value investing and behavioural finance. Tata McGraw Hill, New Delhi
9. Baker HK, Nofsinger JR (2010) Behavioral finance: an overview, behavioral finance: investors, corporations, and markets, pp 1–21
10. Dutta A, Sinha M, Gahan P (2020) Perspective of the behaviour of Retail Investors: an analysis with Indian Stock Market Data. In: Computational intelligence in data mining
11. Chandra A, Kumar R (2012) Factors influencing indian investors behaviour: survey evidence. SSRN Electron J <http://doi.org/10.2139/ssrn.2029642>
12. Mintzberg H, Raisinghani O, Theoret A (1976) The structure of unstructured decision process. *Adm Sci Q* 21:246–275
13. Kubilay B, Bayrakdaroglu A (2016) An empirical research on investor biases in financial decision-making, financial risk tolerance and financial personality. *Int J Financial Res* 7(2):171
14. Arora M, Kumari S (2015) Risk taking in financial decisions as a function of age, gender: mediating role of loss aversion and regret. *Int J Appl Psychol* 5(4):83–89
15. Oliveira A (2007) A discussion of rational and psychological decision making theories and models: search for a cultural, ethical decision making model. *Electron J Business Ethics Organizat Stud* 12(2):12–17
16. Uzonwanne FC (2016) Rational model of decision making, global encyclopedia of public administration. In: Farazmand A (ed) Public policy and governance. Springer International Publishing AG
17. Bazerman MH (2002) Judgment in managerial decision making, 5th edn. Wiley, New York
18. Ritter JR (2003) Behavioral finance. *J Pacific-Basin Finance* 11(4):429–437
19. Waweru NM, Munyoki E, Uliana E (2008) The effects of behavioral factors in investment decision-making: a survey of institutional investors operating at the Nairobi Stock Exchange. *Int J Business Emerging Markets* 1(1):24–41
20. Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational Cascade. *J Polit Econ* 100(5):992–1026
21. Pompian MM (2006) Behavioral finance and wealth management (how to build optimal portfolio that account for investor biases), 1st edn. Wiley, New Jersey, Canada
22. Shefrin H (2008) Behavioral approach asset pricing, 2nd edn. Elsevier, London
23. Dangol J, Manandhar R. (2020) Impact of heuristics on investment decisions: the mediating role of Locus control. *J Bus Soc Sci Res* V(1):1–14
24. Chen G, Kim KA, Nofsinger JR, Rui OM (2007) Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investor. *J Behav Decis Mak* 20(4):425–451
25. Athur AD (2014) Effect of behavioural biases on investment decisions of individual investors in Kenya, doctoral dissertation. University of Nairobi, Nairobi
26. Yaowen XUE, Suqing SUN, Zhang P, Tian MENG (2015) Impact of cognitive bias on improvised decision-makers' risk behavior: an analysis based on the mediating effect of expected revenue and risk perception. *Manage Sci Eng* 9(2):31–42
27. Toma FM (2015) Behavioral biases of the investment decisions of Romanian investors on the bucharest stock exchange. *Procedica Economics and Finance* 32:200–207
28. Brahmana R, Hooy CW, Ahmad Z (2012) The role of herd behaviour in determining the investor's Monday irrationality. *ASIAN Acad Manage J Accounting Finance* 8(2):1–20
29. Ikram Z (2016) An empirical investigation on behavioral determinants, impact on investment decision making, moderating role of locus of control. *J Poverty Investment Develop* 26
30. Luong LP, Ha DTT (2011) Behavioral factors influencing individual investor's decision making and performance: a survey at the Ho Chi Minh stock exchange, pp 1–103
31. Murithi DK (2014) The effect of anchoring on investment decision making by individual investors in Kenya, doctoral dissertation. University of Nairobi, Nairobi
32. O'Neill B, Puza BD (2004) Dice have no memories but I do: A defense of the reverse gambler's belief. *Math Sci* 30(1):13–16
33. Singh R (2019) Behavioural finance. PHI Learning Private Ltd., Delhi
34. Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psychol Rev* 115(2):502
35. Odean T (2002) Volume, volatility, price and profit when all traders are above average. *J Finance*

36. Das U, Mohapatra SR (2016) Behavioral biases in investment decision making: a research on individual's overconfidence. *Asian J Res Bus Econ Manage* VI:28–35
37. Bakar S, Yi ANC (2016) The impact of psychological factors on investors' decision making in Malaysian stock market: a case of Klang Valley and Pahang. *Procedia Econ Finance* 35:319–328
38. Park J, Konana P, Gu B, Kumar A, Raghunathan R (2010) Confirmation bias, overconfidence, and investment performance: evidence from stock message boards
39. Kengatharan L, Kengatharan N (2014) The influence of behavioral factors in making investment decisions and performance: study on investors of Colombo Stock Exchange Sri Lanka. *Asian J Finance Accounting* 6(1):1
40. DeBondt WFM, Thaler RH (1995) Financial decision making in markets and firms: a behavioral perspective. *Handbooks Oper Res Management Sci* 9(13):385–410
41. Gervais S, Simon H, Odean T (2001) Learning to be overconfident. *Rev Financial Stud* 14(1):1–27
42. Kansal P, Singh S (2018) Determinants of overconfidence bias in Indian stock market. *Q Res Financial Markets* 10(4):381–394
43. Madaan, G Singh S (2019) An analysis of behavioural biases in investment decision making. *Int J Financial Res* 10(4):55–67
44. Simon HA (1992) Economics, bounded rationality, and the cognitive revolution. Elgar, Aldershot Hants, England
45. Anderson A, Henker J, Owen S (2005) Limit order trading behaviour and individual investor performance. *J Behavioural Finance* 6(2):71–89
46. Kim K, Nofsinger J (2008) Behavioural finance in Asia., *Pacific Basin Finance J* 169(1–2), 1–7
47. Abdin SZ, Farooq O, Sultana N, Farooq M (2017) The impact of heuristics on investment decision and performance: Exploring multiple mediation mechanisms. *Res Int Bus Financ* 42:674–688
48. Babajide AA, Adetiloye KA (2012) investor's behavioural biases and the security market: an empirical study of the Nigerian security market. *Account Financial Res* 1(1)
49. Shah AK, Oppenheimer DM (2008) Heuristics made easy: an effort reduction framework. *Psychol Bull* 134(2):207
50. Sharma M, Firoz, Md (2020) Do investors exhibit cognitive biases: evidence from Indian equity market. *Int J Financial Res* 11(2):26–39
51. Clark S (2014) The availability bias: how the news can hurt your investment decisions. In: An excerpt from perspectives, vol. 3(1)
52. Lowies GA, Hall JH, Cloete CE (2016) Heuristic-driven bias in property investment decision-making in South Africa. *J Property Investment Finance* 34(1):51–67
53. Kumar S, Goyal N (2016) Evidence on rationality and behavioural biases in investment decision making. *Q Res in Financial Markets* 8(4):270–287

Exploratory Review of Applications of Machine Learning in Finance Sector



Sandip Rakshit , Nyior Clement, and Narasimha Rao Vajjhala 

Abstract The finance sector is one of the key pillars of any nation's economy. However, with the emergence of big data and rapid advancements in technology, the finance sector is processing significant amounts of heterogeneous data. Institutions in the finance sector are increasingly using machine learning algorithms and techniques to process these heterogeneous data. This exploratory review provides an in-depth look at the machine learning applications in the finance sector. The state-of-the-art machine learning applications in the finance sector were reviewed in this exploratory study. The primary research question addressed in this study was to explore the machine learning algorithms and techniques applied to the applications in the finance sector. Various machine learning algorithms and techniques used in finance sector were broadly discussed in this study. This study also provides some suggestions about how machine learning can maximize productivity in the finance sector.

Keywords Machine learning · Supervised learning · Unsupervised learning · Finance · Security · Algorithmic trading · Artificial intelligence · Data science

1 Introduction

Machine learning (ML) algorithms and techniques in conjunction with other technologies can help process and use large volumes of heterogeneous data [1, 2]. Financial sector, in particular, can benefit significantly from the use of machine learning algorithms and techniques. The financial sector is one of the key pillars

S. Rakshit · N. Clement
American University of Nigeria, Yola, Adamawa, Nigeria
e-mail: sandip.rakshit@aun.edu.ng

N. Clement
e-mail: nyior.clement@aun.edu.ng

N. R. Vajjhala (✉)
University of New York Tirana, Kodra e Diellit, Tirana, Albania
e-mail: narasimharao@unyt.edu.al

of the economy of any nation. Some might even argue that an economy's health relies majorly on its financial sector. Thus, there is a linear relationship between an economy's health and its financial sector's strength. While most people limit their understanding of what the financial sector is to exchanges made at the marketplace, there is much more to the financial sector than that. The financial industry is itself made up of smaller industries/sectors.

The financial sector's sub-sectors are usually financial institutions and firms like banks, real estate firms, insurance, and investment companies, providing financial services to their commercial and retail customers. Some of the services offered by the institutions in this sector are but not limited to providing loans to businesses for expansion, provision of mortgages to homeowners, insuring lives and assets, and building up savings for retirement. The value created by this industry increases as the interest rate drops. That is because a reduction in interest rate attracts more investment.

1.1 An Overview of Machine Learning

The technology industry's focus has been drifting toward building intelligent machines, machines that can have the ability to automate repetitive tasks with little or no human involvement. Machine learning is a methodology relying on a premise that machines should learn from the provided data as well as experiences relying primarily on data mining and learning algorithms [1]. Machine learning, a subset of artificial intelligence, uses general methodology to solve a number of problems. Some of the areas where machine learning algorithms have been successful, include mail filtering, optical character recognition, and computer vision [3].

Even though machine learning has been here for decades, its relevance rose due to the immense availability of data and the emergence of powerful computers at a lesser cost. In machine learning, an algorithm is usually trained with a training data set, and a trained algorithm gives rise to a model. The quality of prediction of a machine learning algorithm depends on the quality of the data used in training the algorithm [4].

Machine language algorithms can be classified into three categories to compare the classifiers, namely individual classifiers, homogeneous, and heterogeneous ensembles. The summary of machine learning techniques is provided in Fig. 1. Individual classifiers rely on a single machine learning algorithm [3]. Some of the techniques including decision trees, support vector machines, and neural networks are examples of individual classifiers. Model development and forecast combination are key steps in the ensemble method. Ensemble methods include both homogeneous and heterogeneous ensemble classifiers. Homogeneous classifiers such as bagging and boosting can be useful in increasing the accuracy of prediction of the classifiers [3].

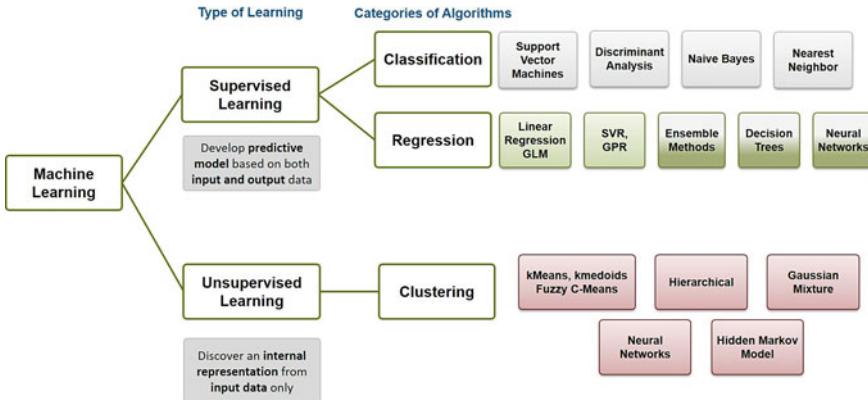


Fig. 1 Summary of machine learning methods

1.2 Classification of Machine Learning Algorithms

The application domain of machine learning algorithms can be divided into several categories, including supervised, unsupervised, semi-supervised, and reinforcement learning techniques [1, 3]. Supervised learning techniques and algorithms should ideally be applied in situations where both the predictors and responses are available. In this learning style, machine learning algorithms are trained with labelled data. Supervised learning algorithms help deduce a function from a labelled training that includes the instances as well as the anticipated outcome for each value of the instances [5]. For instance, an algorithm is provided with an image as input. The expected output is the content in the image. The algorithm learns by comparing its output with the expected output. In this example, the data is the image, and the label is the name of the provided image. There are two major supervised learning problems, namely: classification and regression problems. The way algorithms are trained and are formally called their learning style or machine learning method.

Unsupervised learning techniques and algorithms are suitable for application in situations where only predictors, such as independent or exploratory variables are available [3]. Unsupervised learning techniques help in extracting the most distinct features of the data. Unsupervised learning algorithms help deduce a function to specify the unseen structure of the unlabelled data [5]. In this learning style, an algorithm is trained with unlabelled data. The algorithm explores the unlabelled data and finds some structure within the data that it could use to generalize.

Semi-supervised learning algorithms are trained with both labelled and unlabelled data. It is usually more of the labelled data and less of unlabelled data. Semi-supervised learning techniques help deduce a function grounded on smaller amount of labelled training dataset and a considerably larger amount of unlabelled dataset [5].

The purpose of semi-supervised learning is to make optimal use of the large unlabelled samples [6]. Semi-supervised learning includes semi-supervised clustering and semi-supervised classification [6].

Reinforcement learning techniques are often used for robotics, gaming, and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. This type of learning has three primary components: the agent (the learner or decision-maker), the environment (everything the agent interacts with), and actions (what the agent can do). The objective is for the agent to choose actions that maximize the expected reward over a given amount of time.

2 Applications of Machine Learning in Finance Sector

The use of machine learning algorithms and techniques in dealing with financial data has several advantages over the traditional techniques. Machine learning techniques can automatically identify hidden features in the financial data apart from processing data with nonlinear characteristics [7]. With the rapid growth in finance, it has become necessary to automate some processes and secure existing processes.

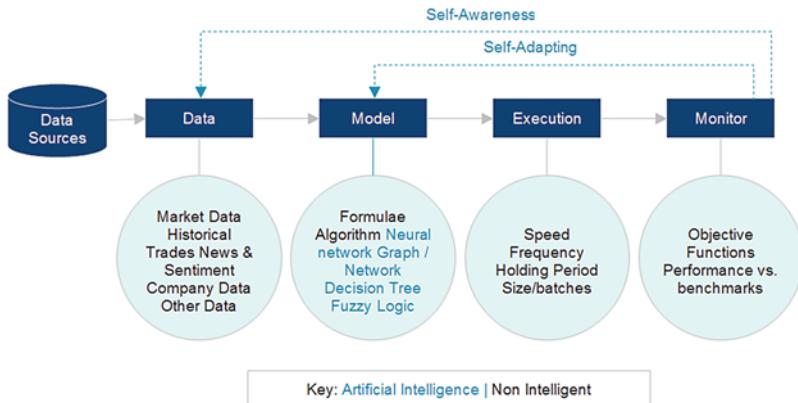
Several applications in the financial sector, including credit risk management, portfolio management, automatic trading, and fraud detection use machine learning algorithms and techniques [3]. Because machine learning primarily concerns itself with parsing and learning from data and big data is one of the important things the finance sector has, this science can be applied in several finance domains. Machine learning has several applications in finance. Some of the machine learning methods frequently used by the finance sector, include k-NN, decision trees, support vector machines, neural networks, and boosting [3].

This paper seeks to answer the following research questions:

- (a) What machine learning algorithms and techniques are applied to the applications in the financial sector?
- (b) What are the practical implications of using machine learning algorithms in the financial sector?

3 Analysis of Extant Literature

Portfolio management is one of the commonest applications of machine learning in finance. They provide an alternative to the traditional and manual consulting individuals do with financial experts concerning their investment options. More formally put, we could say Robo—Advisors are digital platforms or online applications that automate investment processes. They provide financial guidance and services to clients. They achieve this through the use of algorithms and statistics. One of the

Figure 1: Conceptual model of algorithmic trading**Fig. 2** Conceptual model for algorithmic trading

significant advantages of Robo—Advisors lies in the fact that they greatly simplify the entire investment process that could have been otherwise a daunting task.

Algorithmic trading is another key area in the finance sector employing machine learning algorithms and techniques. Before the technological era, trading was purely a paper-based activity. Algorithmic trading has four major components: data component, model component, execution component, and monitor component. Figure 2 depicts these components. Buyers and sellers had to be physically present to purchase or sell goods and services, and certificates for these goods were offered at the end of each transaction. However, after some time, electronic certificates replaced the existing physical certificates, and trading was faster. That notwithstanding, trading in those days was inefficient by far. As a result, there was a need to digitize the trading process and make it more efficient and profitable, bringing about algorithmic trading. Algorithmic trading, as its name implies, refers to the process of using computer programs (algorithms) that instructs the computer to buy and sell stocks on one's behalf when certain market conditions are met. These algorithms take advantage of the computing resources available to them and trade at a speed and frequency that is impossible to match by human traders.

Fraud detection is also another area within the finance sector using machine learning techniques extensively. Fraud is the deliberate misrepresentation of a material fact with the intent of misleading a person or entity into acting upon it, resulting in the harm of the person or entity acting upon it [8]. With the recent emergence of machine learning, this technology has found numerous applications in this domain concerning this use case. Machine learning has been applied in finance to reduce the high occurrences of “false positives” (a situation whereby a financial institution or merchant declines a valid transaction request based on suspecting it to be fraudulent). Several machine learning algorithms including logistic regression, support vector machines, decision trees, and random forests were used in the domain of

credit card fraud detection [9]. Prior research indicates that random forest is one of the simplest and most suitable algorithms in the case of credit card fraud detection. Support vector machines gave the best result when applied to dataset with minimal fraud rates while giving moderate accuracy as the fraud rates increased [9].

Even though chatbots have been in existence for quite some time, machine learning-based chatbots brought about a new chatbot experience in finance. These new breeds of chatbots greatly improved human–computer interaction and, by extension, the end-user experience. This is made possible due to these chatbots' natural language processing capability and their ability to learn from past experiences. As a result of these immense capabilities that these chatbots possess, they can adapt to every customer and his/her behavioral changes. These chatbots can do this because they parse tons of customer finance queries.

Money is also one of the common occurrences in finance. However, with the emergence of machine learning, financial institutions seek ways to use machine learning models to reduce these events. These models will be trained to spot signs of money laundering in these financial transactions. Khac et al. [10] present a solution developed as a tool for identifying and analyzing money laundering using real transaction datasets. The authors used a three-level data mining framework comprising of various components, namely the pre-processing of data, data mining, and knowledge management. Lin-Tao et al. [11] proposed a method for increasing the detection rate of suspicious financial transactions and money laundering attempts with low false positives. Deng et al. [12] proposed active learning through sequential design method for prioritization to improve the process of money laundering detection. Salehi et al. [13] suggest unsupervised machine learning algorithms for effective money laundering detection given the crime's dynamic nature.

4 Conclusion and Future Research Directions

Even though machine learning has a plethora of applications in finance, significant technological advancements are expected in the next decade. Numerous improvements could be made in the application of machine learning in the financial sector. For example, sentiment analysis could be heavily applied in finance to deduce people's emotions at any given point in time. This could, in turn, be used to enhance machine–human interaction. Furthermore, finance is one sector where security is paramount. Everyday existing security protocols are undergoing improvements, and new ones are being implemented to ensure users' financial data are being properly secured. However, passwords and user names remain the primary user authentication mechanism, even in the financial sector for now. Given that user security in this sector is a high-stakes game, there is the need to architect a more stringent mechanism for ensuring user security. Machine learning can be used to enhance user security is through facial recognition, voice recognition, and biometric data. Despite the positive applications of machine learning in finance, it will be wrong to overlook the negative impact this technology could have on this economic sector. However, the

widespread adoption of AI could introduce new systemic and security risks in the financial system. The early big movers offer their AI applications (that includes machine learning) as a “service” to their competitors, attracting users to accelerate their system’s learning and turning cost centers into profit centers. As this trend widens, the financial system may face new risks.

In conclusion, even though machine learning has an increasingly wide range of finance applications, these applications would grow in the forthcoming years. It is as a result of this that finance institutions are now investing in this technology. Their investments are indeed bringing them a lot of benefits. Some of these benefits include a drastic reduction in operational costs. Because of this reduction in operating expenses, there tends to be an equal increase in revenue. Machine learning in finance also leads to an increase in customer loyalty due to a better user experience. In the meantime, machine learning algorithms in finance offer investment advice, reduce fraud, and trade on behalf of financial institutions. While busy with all these tasks, these algorithms are always learning and getting smarter day by day and bringing the world closer to completely automated financial processes.

References

1. Chagas BNR et al (2020) A literature review of the current applications of machine learning and their practical implications. *Web Intell* (2405–6456), 18(1):69–83
2. Agarwal A, Jayant A (2019) Machine learning and natural language processing in supply chain management: a comprehensive review and future research directions. *Int J Business Insights Transf* 13(1):3–19
3. Teng H-W, Lee M (2019) Estimation procedures of using five alternative machine learning methods for predicting credit card default. *Rev Pac Basin Financ Mark Policies* 22(03):1950021
4. Moruff OA, Maruf AO, Toshio A (2020) Performance analysis of selected machine learning algorithms for the classification of phishing URLs. *J Comput Sci Control Syst* 13(2):16–19
5. Ahmad A et al (2020) A systematic literature review on using machine learning algorithms for software requirements identification on stack overflow. *Sec Commun Netw* p 1–19
6. Zhao J et al (2018) Safe semi-supervised classification algorithm combined with active learning sampling strategy. *J Intell Fuzzy Syst* 35(4):4001–4010
7. Chen Y et al (2020) Financial trading strategy system based on machine learning. In: Mathematical problems in engineering, pp 1–13
8. Ogbodo UK, Mieseigha EG (2013) The economic implications of money laundering in Nigeria. *Int J Acad Res Accounting Finance Manage Sci* 3(4):170–184
9. Sravya K et al (2020) Credit card fraud detection using machine learning algorithms—Study of customer behaviour. *Grenze Int J Eng Technol (GIJET)* 6(2):143–150
10. Khac NL, Markos S, Kechadi MT (2010) A data mining-based solution for detecting suspicious money laundering cases in an investment bank. In: Second international conference on advances in databases, knowledge, and data applications. IEEE Menuires, France
11. Lin-Tao L, Na J, Jiu-Long Z (2008) A RBF neural network model for anti-money laundering. In: International conference on wavelet analysis and pattern recognition. IEEE, Hongkong
12. Deng X et al (2009) Active learning through sequential design, with applications to detection of money laundering. *J Am Stat Assoc* 104(487):969–981
13. Salehi A, Ghazanfari M, Fathian M (2017) Data mining techniques for anti money laundering. *Int J Appl Eng Res* 12(20):10084–10094

Prediction of Personality Traits in Facebook Users



Mamta Bhamare and K. Ashokkumar

Abstract Social media is a forum for people to introduce themselves to the world, sharing personal details, and perspectives about their lives. This awareness can be used to improve app performance and application experiences. Personality has been shown to be important to many kinds of interactions; it has been shown to be beneficial for predicting job satisfaction, efficiency, and even preference for specific interfaces. Various information is widely shared through social media, i.e., Facebook and Twitter. User and user data are important research instruments in the fields of behavioral learning and personality via status updates. There is a rapid increase in use of social networks. Similar research has been carried out in this area and continues to grow. This attempts to create a program that can predict a person's character using a dataset. Its research is conducted in the Big Five Model Personality.

Keywords Personality · Personality prediction · Social networks · Big Five Personality · Machine learning

1 Introduction

In recent years, social media has been the most widely used method for interpersonal interaction and communication. Direct contact between people is decreased by using a mobile device. As a consequence, understanding the entity's personality is challenging. Nonetheless, data shared on social media will assist us in collecting the information we need, because people spend a lot of time on social media and express their feelings and opinions through status updates, comments, and tweets. The information available on social media platforms offers a unique amount and wealth of information about human behavior and social interactions [2]. Social media provides

M. Bhamare (✉) · K. Ashokkumar

Department of Computer Science and Engineering, Sathyabama Institute of Science & Technology, Chennai, India

e-mail: mamta.bhamare@mitwpu.edu.in

K. Ashokkumar

e-mail: ashokkumar.cse@sathyabama.ac.in

information that allows people to identify who they are and/or need in order to identify important personality characteristics by examining what is accessible in social media. At about 800 million people spending around 40 min everyday using Facebook, it hits 1.8 trillion users. Users of Facebook usually express their feelings and views by changing their status or posting. While Facebook is more commonly used to exchange photographs and videos at present, this investigation concentrates on the linguistic dimension of users who are updating their state. Psychology studies have shown that the person's temperament and linguistic behavior are associated [2, 3]. It is possible to evaluate and explain this association effectively by using a natural language system. As a result, the aim of this study is to create a prediction system that can automatically predict user personality based on their Facebook activities. Big Five Personality, MBTI, or DISC, a Myers Briggs type indicator, is some of the personality types used in personality research. However, in this article, Big Five Personality is used as the most frequently and correctly informed of personality traits after other parameters and a basis for literature review. Attributes of transparency, perception, extraversion, harmony, and neurotics are part of this model. The making of a model that can anticipate characters precisely by utilizing online networking content is pertinent to an assortment of fields, for example, showcasing, business knowledge, brain research, and human science. The paper consists of the following: Sect. 2 introduces a character conversation and a Big Five Model. Sect. 3 gives a concise review of the principle endeavors in the writing to perform character expectation from internet-based life information. Section 4 introduces a bare essential depiction of the suggested approach. Section 5 explains how the technique can be applied to a dataset that incorporates three separate datasets ranging from writing to Web-based media mining networking, and Sect. 6 concludes the paper with a general discussion of the methodology as well as suggestions for further study.

2 On Personality and Big Five Model

“Personality” originates from the Latin expression *persona*, which means the cover utilized in a play area by the onscreen characters. The thought emerged from the way that character contains the blend of highlights and characteristics. Author [17] contend that character can be characterized as an individual attribute with activities, demeanor, feelings, and brain. In any case, these are generally huge qualities. The assortment of attributes makes it hard to decide the character since it does not give a premise to the arrangement and comparison of individuals. Since human emotions are broad and the necessary emotions for classification are difficult to pick, the same issue occurs when the feelings in a text are told (sentiment analysis). Many researchers, for example, agree to simplify the presentation of feelings by their polarities (positive or negative) in order to automate sentiment analysis [7, 23]. The study of personality shows a similar structure.

To request to permit an assessment of character, different scientists have characterized the key highlights to request to make a model of character. Schultz and

Schultz (2006) assumed that the character could be considered as a definite and unusual collection of highlights that cannot be adapted to different conditions. Each model of forecast of character needs to attempt to furnish these gatherings with names of highlights. As far as the qualities wherein they spoke to a neuropsychic structure, first portrayal of a structure for character was made. Two trademark rates were depicted that figured a hypothesis [21] once in a while alluded to as mental highlights (Hall et al. 2000). Mainly a trademark is known as the main stage, while the second is known as a single trademark. The specific characteristics are not unique to the individual, i.e., they can be shared between different people. Trademark examination empowers us to look at people or individual gatherings when one individual is engaged with the investigation of individual aura (Hall et al. 2000). Highlights can be gotten from the recurrence of a conduct which is not normal for the assortment and force of the conduct of an individual (Hall et al. 2000). For example, if an individual offers wry remarks on an online networking Web site as often as possible, it very well may be induced.

In light of the lexical speculation, the most extreme individual contrast is encrypted in language (Hall et al. 2000). And indication in a mental phrasing lexicon (Garcia 2007) in wording (words) can be systematized. In his investigation of the character structure [22] utilized the phrasing portrayed to begin characterizing a Big Five or Five-Factor Model [21]. The Grand Fifth characterizes a character framework partitioned in five sections, known as OCEAN [22], which are openness to experience. Openness to experience shows a person's ability to accept newness. High levels show wide curiosity and a search for novelty, while low levels indicate a preferentiality for familiarity and comfort. People open to experience frequently use social media. Intelligent means people who are attentive, attentive, punctual, and organized. An awareness is associated with an unattended, reckless, and careless person. Honest individuals like to utilize less informal organizations, claiming that these Web sites are an unnecessary diversion. Extroversion is about outgoing, socially active and talking people, and introversion is about people who are shy and quiet in general. Extroverts tend to make friends and take them on the Internet outside the virtual environment and suggest they view the Web as a way to keep in contact, but not to replace personal links. Conformity is viewed as a measure of people's friendliness, while friendships outside the virtual environment are difficult to start and maintain. Emotional control is determined by neuroticism. Low neuroticism suggests resilience and greater emotional control, while elevated levels propose more noteworthy affectability, tension, and less enthusiastic control. Higher hypochondriacs will in general utilize the Internet as an instrument to lessen the feeling of dejection and to make a feeling of ownership [20].

The Big Five Model is based on a broad spectrum of personality analysis and theory. The results are based on lexicon [21, 22]. In general, this model is seen as a significant advance in researching a general example that characterizes the structure and assessment of the character (Hall et al. 2000). A connection among character and phonetic levels has been set up by a few specialists. Linguistic features related to character traits are explained in Furnham (1990). The extroversions were related to contextualization [19]. In the Big Five Model, many linguistic features associated

with every character feature were identified [15]. This phonetic system licenses AI procedures to make a judgment for character dependent on a lot of writings. Author [15] first researched automated text mining for personality forecasting and a further analysis by a group of computer researchers [18]. In the following year, author [17] demonstrated that computers can predict characteristic traits based on the language used (as illustrated in the Big Five Model).

The characteristics of the Five-Factor Personality (FFM) and the OCEAN model are a taxonomy or characteristics classification. Once factor analysis is applied to personality survey data (the statistical technique), terms sometimes refer to one person to explain personality aspects. The five factors consist of the acronym OCEAN as shown in Fig. 1 together with words of high and low ratings in relation to each characteristic function.

The following segment covers the primary ideas and works identified with character expectations that depend on the Big Five Model using Web-based social networking information.

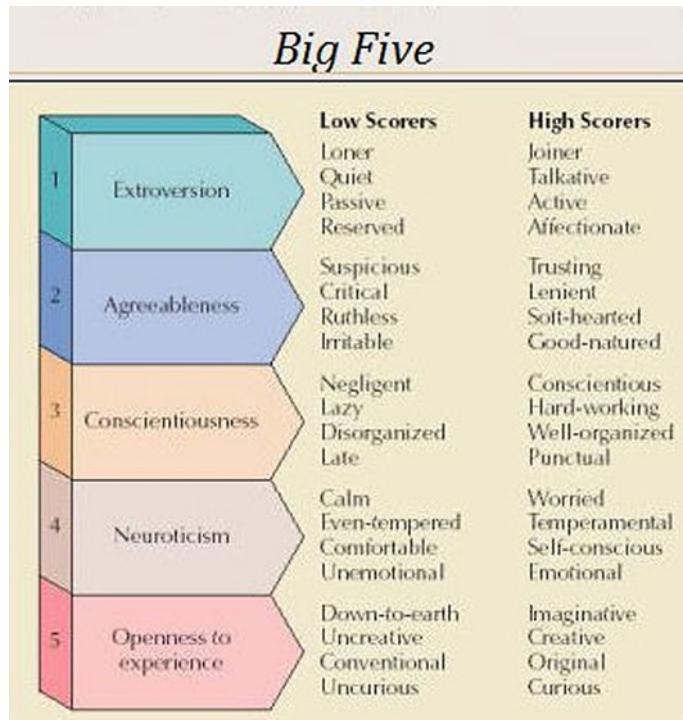


Fig. 1 Big Five Model [24]

3 A Review of Personality Prediction Based on Social Media Data

Linguistic and non-linguistic features are extracted from the text data of social media. LIWC, MRC, speech acts, post tags, H4LvD, sentiment, and others comes under linguistic categories. Structural, behavioral, and temporal are non-linguistic features [25]. Prediction is the activity of removing models that can order obscure information or patterns in predictions. If it is intended to predict categorical values, the classification is called a categorification, and when the target is model qualities or continuous functions, it is stated as a criterion. [5]. The extent of this work incorporates the expectation of unmitigated marks which compare to the character of subject gatherings and subsequently portray them as an undertaking for grouping. Machine learning predictive specialists were utilized in various Internet-based life mining fields, for example, assumption examination [6, 7] and event detection [9, 11]. Machine learning classification can be regulated, uncontrolled, or semi-controlled. The supervised learning consists of an implementation of a collection of preclassified data to train the predictor, called the training set. The classifier will learn how these previously classified data are classified and how new data can be generalized. For unregulated learning, data are unlisted, and therefore, learning is based on the input data's statistical regularity. An unregulated approach to learning produces internal representations that encode input data's characteristics [5]. Semi-monitored learning is intermonitored and unmonitored learning, as labeled and unlabeled data are used for the training of a classification system [12]. Albeit numerous information via Web-based networking media are contemplated, character gauge depends on the model, and at present, the Big Five Model is the most famous one [4]. Different machine-students, for example, KNN calculations, Bayesian classifiers, models of relapse, neural systems, bolster vector machines (SVM), etc., can therefore be used for this task. Authors Golbeck, Robles, and Edmondson refer to the relationship between social media profiles and personality traits as the first one they investigate. The authors analyzed Twitter account by creating a Twitter form with an inventory of 45 large five individuals. The inventory and the most recent tweets have been analyzed for each app. The Linguistic Inquiry and Word Count (LIWC) and MRC Psycholinguistic Databases collected language knowledge from their texts. Authors [15] have established the LIWC to extract 81 various textual features grouped into five different categories: regular ranking, psychological [16]. The MRC is a 150,000-term database of linguistic and psychological characteristics.

Social networking languages have been used to forecast personalities across a range of realms, including Facebook, Twitter, and YouTube [8]. In all domains, the results indicate a clear connection between demographic characteristics such as age and gender and personality outcomes. The corelationship between gender and Agreeableness is positive on Facebook, but not on Twitter or YouTube. The word count has a positive correlation on Facebook and Twitter, but it has a negative correlation on YouTube. Scientists have used machine learning and data extraction methods

and algorithms to facilitate users' personal discovery through social networks and systems, according to a review of related work.

4 Proposed Work

The personality classification issue is considered on the basis of knowledge from the following contents-textual content written by the individual and on request by means of social networking or other methods. The following steps define the standard approach to resolving the problem based on the content: A. Data collection, B. Determination of participants' personality characteristics, and C. Model building. There are ways in which large quantities of personal data are available in this proposed system. The investigation uses learning algorithms and sophisticated data mining to collect data from users and learn from patterns. This research can now be used to predict users based on classifications of the past. Based on observation trends, the program analyzes comprehensive user characteristics and habits and stores user patterns in a database. The program now foresees the new personality of users based on personality information stored in the previous user data classification.

4.1 *Methodology*

After splitting the original dataset into training and test sets, we can use the training set to construct a model that predicts new values or labels. We forecast new values using the previously established model and the test dataset (or labels). Finally, we can test the accuracy of the classifier by comparing predicted and expected labels. Scikit also provides a range of packages for extracting vectorizing functionality from documents. The objective is to construct a classification system that can predict personality traits from continuous and binary text input. To accomplish this, we built a system that uses machine learning techniques. Preprocessing, data analysis, feature extraction, and classification model are the four main stages. In this work, my Personality dataset was explored after preprocessing involving the technique and processes for making the texts appropriate for the classification purposes which was carried out. In the preprocessing step, the stemming is formulated, the stop words are removed, and the written datasets are normalized, and these datasets are given as frequency values. Finally, the results are calculated by using a particular algorithm to classify the characteristics [Big Five] (Fig. 2).

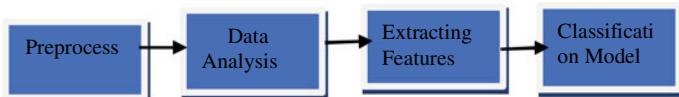


Fig. 2 Methodology for personality prediction

Table 1 Distribution of my Personality dataset

Value	OPN	CON	EXT	AGR	NEU
Yes	176	130	96	134	99
No	74	120	154	116	151

4.2 Dataset

The dataset is split into two parts. The first sample of my Personality is made up of 250 Facebook users with approximately 10,000 statuses based on the Big Five Personality Traits model. The distribution of my Personality data by personality type is given in Table 1.

4.3 Preprocessing

Since the dataset is made up of several CSV files, preprocessing is required to merge them into a single data frame.

4.4 Data Analysis

We may begin working on the data analysis until we have just one data frame. By concentrating on the characteristics, we can see some overview stats, associations, distributions, and so on.

4.5 Model Classification

As described above, classification was based on traditional machine learning. Current algorithms included in the learning process are logistic regression (LR), linear discriminant analysis (LD), K-nearest neighbors (KNN), decision tree, naive Bayes (NB), and support vector machines (SVM).

Fig. 3 Prediction values for extraversion traits

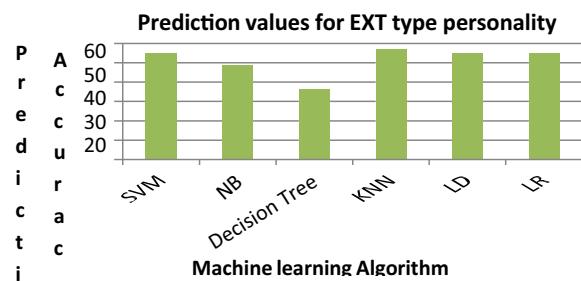
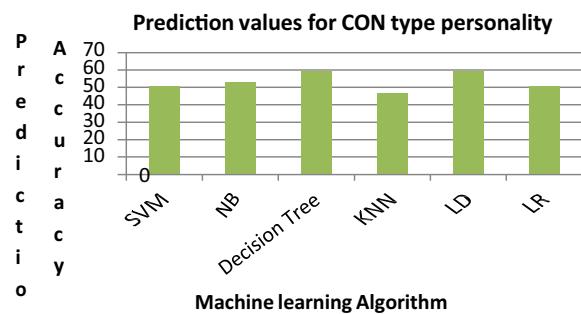


Fig. 4 Prediction values for conscientiousness traits



5 Results

my Personality dataset is used for predicting the personality using Big Five Model in the proposed work. Different machine leaning algorithms namely SVM, logistic regression (LR), linear discriminant (LD), KNN, and decision tree are applied on the dataset for checking accuracy of the personality prediction. Graph has been plotted using the prediction values for all five types of Big Five Personality Model using different machine learning algorithm for comparative study. Figures 3 and 4 show the prediction results for Big Five Personality using SVM, NB, decision tree, KNN, LD, and LR. We can say that SVM gives better accuracy compared to all other machine learning algorithm that are used in this work for neuroticism and openness type of personality traits.

6 Conclusion

Written text, online status updates, Facebook messages, and Twitter texts are all examples of factors that can be used to assess an individual. The research focuses on the output of automatic personalities prediction. Facebook posts are used in this work to predict and identify requirements for personality. Our aim is to use various techniques to isolate the features and encapsulate them to decide which types of

written texts are the features. There are various benefits to personality forecasts including business intelligence, employees, marketing for the understanding of the customer's behavior, advertising departments that like the most sort of brand that knows the personalities of people. Different machine learning algorithm has been applied on Facebook post of users to predict personality. Results of the same are studied and discussed.

References

1. Prantik H, Kuntal KP, Alfredo C, Madhu Kumar SD (2018) Predicting Facebook-user's personality based on status and linguistic features via flexible regression analysis techniques. In: Proceedings of the 33rd annual ACM symposium on applied computing (SAC '18). Association for computing machinery. New York, NY, USA, 339–345. <https://doi.org/10.1145/3167132>
2. Barbier G., Liu H (2011) Data mining in social media. In: Aggarwal C (ed) Social network data analytics. Springer, Boston, MA
3. Golbeck J, Robles C, Edmondson M, Turner K (2011). Predicting personality from Twitter, 149–156. <https://doi.org/10.1109/PASSAT/SocialCom.2011.33>
4. Golbeck J, Robles C, Turner K (2011) Predicting personality with social media. Conf Hum Factors Comput Syst Proc 253–262:0001. <https://doi.org/10.1145/1979742.1979614>
5. Han J, Kamber M (2011) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann, Burlington
6. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of twitter data. In: Proceedings of the workshop on languages in social media, 30–38
7. Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. Intell Syst IEEE 28:15–21. <https://doi.org/10.1109/MIS.2013.30>
8. Golnoosh F, Geetha S, Sharu S, Fabio C, Michal K, David S, Sergio D, Marie-Francine M, Martine De Cock (2016) Computational personality recognition in social media. User Modeling User-Adapted Interaction 26(2–3):109–142.
9. Abel F, Gao Q, Houben GJ, Tao K (2011) Semantic enrichment of twitter posts for user profile construction on the social web. In: Antoniou G et al. (eds) The semantic web: research and applications. ESWC 2011. Lecture notes in computer science, vol 6644. Springer, Berlin, Heidelberg.
10. Kosinski M, Matz S, Gosling S, Popov V, Stillwell D (2015) Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. Am Psychol 70(6):543
11. Earle P, Bowden D, Guy M (2012) Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys Annali di geofisica. 54:0001. <https://doi.org/10.4401/ag-5364>
12. Chapelle O, Scholkopf B, Zien A (eds) (2006) Semi-supervised learning. MIT Press, Cambridge. <https://doi.org/10.7551/mitpress/9780262033589.001.0001>
13. Pennebaker J, Boyd R, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015
14. Moffitt K, Giboney J, Ehrhardt E, Burgoon J, Nunamaker J (2010) Structured programming for linguistic cue extraction [Online]. Available from: <http://splice.cmi.arizona.edu/>
15. Pennebaker JW, King LA (1999) Linguistic styles: Language use as an individual difference. J Pers Soc Psychol 77(6):1296–1312. <https://doi.org/10.1037/0022-3514.77.6.1296>
16. François M, Marilyn AW, Matthias RM, Roger KM (2007) Using linguistic cues for the automatic recognition of personality in conversation and text. J Artif Int Res 30(1):457–500
17. Mairette F, Walker M (2006) Automatic recognition of personality in conversation. Proc HLT-NAAACL

18. Schler J, Koppel M, Argamon S, Pennebaker J (2006) Effects of age and gender on blogging. In: Computational approaches to analyzing weblogs—papers from the AAAI spring symposium, technical report. (AAAI spring symposium—technical report, vol. SS-06–03), pp 191–197
19. Heylighen F, Dewaele J (2002) Variation in the contextuality of language: an empirical measure. Found Sci 7:293–340. <https://doi.org/10.1023/A:1019661126744>
20. Hughes DJ, Rowe M, Batey M, Lee A (2012) A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. Comput Hum Behavior 28(2):561–569. <https://doi.org/10.1016/j.chb.2011.11.001>
21. Allport GW, Odberth HS (1936) Trait-names: a psycho-lexical study. Psychol Monogr 47(1):i–171. <https://doi.org/10.1037/h0093360>
22. Cattell RB (1957) Personality and motivation structure and measurement. World Book Co
23. Gangemi A, Presutti V, Diego RR (2014) Frame-based detection of opinion holders and topics: a model and a tool Comput Intell Magz IEEE 9(20):30. <https://doi.org/10.1109/MCI.2013.2291688>
24. <https://sites.psu.edu/leadership/2016/03/29/big-five-relationships/>
25. Mamta B, Ashok Kumar K (2019) Personality Prediction from social networks text using machine learning. Int J Recent Technol Eng (IJRTE) 8(4). ISSN: 2277–3878
26. Dey N, Borah S, Babo R, Ashour AS (2019) Social network analytics: computational research methods and techniques, Elsevier. ISBN: 9780128156414

Software Quality Prediction Using Machine Learning



Aparna Mohapatra, Saumendra Pattnaik, Binod Kumar Pattanayak, Srikanta Patnaik, and Suprava Ranjan Laha

Abstract Since twenty-first century, software quality is considered as a vital factor in the global competitive position for any software product in order to ensure quality and to ensure the dependency of software product. Software fault proneness has made tremendous progress in predicting the faults. On the other hand, the prediction model helps to create an accurate model and also helps the developers to deliver the software in time to their customers [4]. A software quality prediction model seeks to predict the quality factor that whether the software is prone to fault or not. In the early stage of software development, the users use the fault prediction model to detect the faults. So, our main aim should be to create reliable, portable, and robust software that minimizes the errors that occur when a program runs. Software quality prediction model helps us to know which components are at fault which can be corrected by detailed testing. To improve the quality, reliability, efficiency, and maintenance cost, the software fault should be predicted beforehand. It is difficult to develop fault prediction software. The cost of detecting and correcting the errors becomes extremely higher as we move from requirement analysis to maintenance phase, where defects might even lead to loss of life [3]. Fault detection in the beginning stages help the stakeholders [5] to converge their resources on modules that are likely to cause fault in the beginning phase.

Keywords Software quality prediction · Software fault prediction · Machine learning techniques

A. Mohapatra · S. Pattnaik (✉) · B. K. Pattanayak · S. Patnaik · S. R. Laha
Department of Computer Science and Engineering, ITER, Siksha ‘O’ Anusandhan University,
Bhubaneswar, Odisha, India

B. K. Pattanayak
e-mail: binodpattanayak@soa.ac.in

S. Patnaik
e-mail: srikantapatnaik@soa.ac.in

1 Introduction

The justifiability of software quality is an important part in software project. Since the software project is immensely growing in the fast changing world of software development, it has become essential to give a software quality model which would bypass the software quality mark. Out of all the attributes of the software system, the quality is the most compulsory feature. The team of developers is always concerned or focuses a lot on the challenges they come across when it comes to maintain the quality of the software. Maintaining the quality of the software throughout the software development process requires a lot of resources, which consumes a lot of time, cost, and effort. Therefore, predicting the bugs at the early phases of software development lifecycle helps to reduce cost, time, resources, and effort. This also leads to user satisfaction and improves the software version and leads to resource optimization. The presence of faults affects the software reliability, maintenance cost, and quality. Software bug prediction is a very important activity in developing the software. Dealing with the bugs is a necessary problem in the process of software development. This causes high risks which can lead to project failure. This also causes the decrease of quality of software and increase the maintenance cost. Hence, the quality prediction plays a vital role which contributes to the development of the software. Software development is a laborious process which includes planning, analysis, designing, implementing, testing, and maintenance. A software engineer should develop software within the time and cost which are determined during the planning phase. So, it is amply clear that a software project would produce the highest level of efficiency that will enhance the software development organization to efficiently use the resources. The software quality as well as the reliability should achieve the expectations of the customer so that the product could be delivered on time. Efficient software product is based on the applicability of software product. The quality prediction model helps us to identify whether software is faulty or not. Hence, the above process would help the project manager to use the availability of resources at his disposal to correctly target the error in the software model.

Rest of the paper has been organized as follows. Section 2 includes various prediction approaches. Software quality prediction is comprehensively described in Sect. 3. Various machine learning techniques are detailed in Sect. 4. The advantages of ML are covered in Sect. 5. Section 6 incorporates the related work in the field of software quality prediction. A comprehensive analysis of software quality prediction using ML is elaborated in Sect. 7, and Sect. 8 concludes the paper.

2 Types of Prediction Approaches

In software engineering, there are different types of prediction [2] approaches. They are:

1. Security prediction

2. Quality prediction
3. Effort prediction
4. Fault prediction
5. Test effort prediction
6. Correction cost prediction
7. Reusability prediction.

The above application would benefit the developer in decreasing the expense so that it would be viable in the software market and its expansion.

3 Software Quality Prediction

The software quality prediction is an activity that seeks to predict the quality of the software that is whether the project or model is fault prone or not. Here, the quality software refers to the software which is convincingly defect-free and that is delivered within the time frame and the specified budget, that also meets the user or customer requirements and their expectations and can easily be maintained. In the context of software engineering, quality refers to both functional quality as well as structural quality. A software quality prediction model helps the development team for tracking and detecting the software defects during development cycle that saves a lot of time and effort. A quality prediction model is an essential tool in order to meet the objectives of the applicability of the software models. It ensures the reliability of the product that is delivered. The viability of the software project can be determined by using the software quality prediction model. The applications would benefit the developers in decreasing the expense that would be viable in the software market and its expansion.

The correct prediction of buggy modules enables the accurateness of the software modules that is why the developers of the software take an interest in developing the quality models or projects. The bedrock of the successful computer model rests on the set of codes which are fault-free. So, these activities are used in planning and scheduling testing activities.

Nowadays, the justifiability of software quality has become an essential part in the software development project. Assurance of software quality and reliability has become very important as that of delivering the model within the given time and fixed cost. To get the desired software quality, it is essential to enrich the software quality models. The software test estimator is a management activity that roughly decides how much time a task would take to complete. Estimating the effort for the test is a major and a very important task in software test management.

Software fault is an error, defect, flaw, malfunction, or a mistake that causes it to create unexpected or erroneous outcome. Faults are essential properties of system. These are the programming errors or mistakes that cause different performance compared with what is expected. The major faults are from source code or design, and some are from the incorrect code generated by the compiler. Software faults are

dangerous problems for both the software developers and the clients. These defects not only decrease the software quality, and it also increases the cost and delays the development schedule which leads to late delivery of the software product or module. Therefore, the software fault prediction models are proposed to sort out this problem. SFP can efficiently progress the effectiveness of software testing and direct the allocation of resources. To develop quality software, it is very important to detect and correct the faults in early phase of software development life cycle (SDLC) [1].

The software quality, reliability, and maintenance cost depend upon the flawless software quality. Developing a fault prediction model is a challenge to developers to obtain a flawless model by identifying faulty modules.

Software faults are predicted beforehand to innovate an advance software model which can stand the test of the time [3]. Sediments of error namely miscommunication, errors in programming, time pressure, and poor documentation all together makes a software model faulty.

Detecting the defects in an early phase enables in delivering high-quality and low-cost software to the customer [4]. Therefore, it is very important to evaluate various techniques and determine the best technique for prediction of defects which will be very useful for practical application for software practitioners and researchers. It would enable the researchers to use the best techniques to develop the defect prediction model. To predict the dependent variable, the defect prediction models are used on the basis of independent variables. The metrics are used as independent variables, and software quality attributes are used as dependent variables.

The growing demand of industries for compatible software programs is a challenge for the researchers to upgrade their models to meet the demand of the customers, thereby the quality is examined carefully and undertaken cautiously before the software is released to different industries [5].

The quality of the software is the principal feature of a software model. A large sum of resources is required for maintaining the quality of software to improve the design throughout the software development life cycle. A large sum of resource is required for maintaining the quality of software for the improvement of the design in the development life cycle [5].

4 Machine Learning Techniques Used

Some machine learning techniques are as follows:

- Naïve Bayes (NB): NB is a simple and viable classification technique which depends on Bayes theorem that independently assumes among the features. It is a group of algorithms that is based on common principles [3].
- Artificial neural network (ANN): It is a powerful, strong, and complicated machine learning technique which resembles with human brain and its functions [3].

- Decision tree (DT):- DT represents tree with many decision nodes, that has many branches and leaf nodes, and all the decisions are taken by the branches and the leaves[3].
- Backpropagation: This learning algorithm is applied to multilayer feed-forward network with continuous differential activation function. The basic concept is that the weights are updated by the gradient descent method [4].
- LVQ1: LVQ1 algorithm is a simplest algorithm because the position of only one of the codebook vectors is updated at a time.
- Multipass LVQ: Multipass LVQ is used for Parkinson's disease prediction.
- SOM: It is based on unsupervised learning, that means no human intervention is required during the learning process. It can also detect the features inherent to the problem, that is why also called SOFM.
- Multipass SOM: Multipass SOM belongs to neural network algorithm. It is recommended for better results where two passes are executed on the same underlying model.
- LCS: It is a learning component that performs supervised learning, reinforcement learning, and unsupervised learning.
- ZCS: It was proposed by Wilson to present in order to reduce the original framework of LCS without making any changes to its features.
- XCS: It stands for accuracy-based system and was the first classifier that was proposed by Wilson. It uses the accurateness to finalize the appropriate action and parents.
- Feature selection: It is a process of selecting the relevant features or we can say that filtering the irrelevant or noisy features from the dataset.
- Data balancing: The process of balancing the imbalanced data is called data balancing.
- Support vector machines (SVMs) are supervised learning models which are associated with the learning algorithms that analyze data used for classification and regression.
- NN: Neural networks are nonlinear algorithms for data and image processing. In the general purpose computer, these are implemented.
- Multilayer perceptron: A multilayer perceptron is a class of feed-forward artificial neural network. It uses a supervised learning technique called backpropagation for training.
- Random basis function (RBF): RBF is means to approximate multivariable functions by linear combinations of terms based on a single invariable function.
- Multinomial Naïve Bayes uses the term frequency, which is the number of times a given term appears in the document.
- K-nearest neighbor (KNN): It is a classification algorithm. It calculates K-nearest data points from data point X and then uses these points to determine which class X belongs to.
- Auto-sklearn: It is an automated machine learning toolkit and a drop-in replacement for a scikit-learn estimator (Fig. 1).

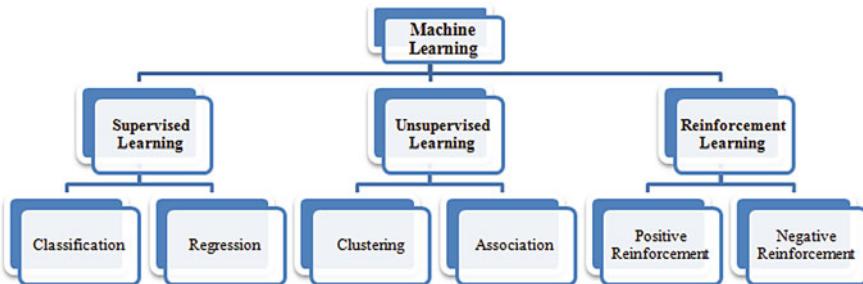


Fig. 1 Types of machine learning algorithm

5 Advantages of Using Software Quality Prediction Model

The advantages of using software quality prediction model are:

1. To create an accurate model.
2. Helps the developers to deliver the software in time to their customers.
3. Predict whether a model is buggy or not.
4. It reduces the maintenance cost.

6 Related Work

We detail here various software quality prediction techniques as reported in the literature.

Awni Hammouri in the year 2018 aimed [3] for exploring the optimum utilization, three ML techniques is used such as Naïve Bayes, decision tree, and ANN. As future work, other ML techniques can be involved and this would give an extended similarity or differences between them. To upgrade the viability of the software prediction model that can be obtained by including a lot of software metrics.

R. Malhotra in the year 2017 choose [4] 14 object-oriented, inheritance, and other metrics on the basis of correlation-based feature selection (CFS) technique, and the relevant data was calculated using the constructed automatic tool. CFS is a function which ranks the attributes according to a heuristic evaluation function that is based on correlation. As the future work, the conclusion drawn from this study can be used for practical applications by software practitioners to determine best technique for defect prediction and consequently carry out effective allocation of resources.

C. Wondaferaw in the year 2017 [5] thoroughly analyzed and appreciated ELA on the basis of its viability of its performance when it is adjoined with techniques such as FS and DB in order to determine the viable methods which would intact better for SFP. It prophesies to invent many ELA including vote and stacking. They have also thought of having a framework which would help them to detect more viable ELA and that would improve their classification performance. Ensemble learning algorithm

is a meta-algorithm that combines many ML techniques into one prediction model which improves the machine learning results by combining many models that would decrease the bias, variance and would increase the prediction capability of the model.

Amod Kumar and Ashwini Bansal in the year 2019 [6] aim to have a relation among the dependent and independent variable using genetic-based machine learning practice. It prophesies in creating a working tool that can be used in comparing the results with other ML algorithms to prove its performance.

C. Lakshmi Prabhas and Dr. N. shivakumar in the year 2020 [7] did a time-consuming research by taking many parameters such as confusion, accuracy, and recall which are measured as well as compared with them. It prophesies that these parameters stand the test of several datasets. If the dataset is increased, then it enhances the findings.

B. Yalciner and M. Ozdes in the year 2019 used [8] seven ML algorithms to identify the fault present in the software much before delivering to the customer. As the future work, an additional experimental study would be conducted by using different dataset.

Lov Kumar and Ashish Surekha in the year 2017 [9] aimed to apply the source code and ML techniques to identify aging bugs. SMOTE method is used to count the impact of imbalanced class in the proposed ML solution. Synthetic minority oversampling technique (SMOTE) is a statistical technique that increases the number of cases in our dataset but in a balanced way. It takes the entire dataset as an input and increases the percentage of only the minority cases, not of majority cases.

S. Delphine Immaculate in the year 2019 [10] used three supervised ML techniques to construct a software model that would identify the happening of the software errors based on historical data. They have thought of training the model in ANN to boost the accuracy of the prediction process.

R. Malhotra and Megha Khanna in the year 2015 [11] aimed for examining the implementation of statistical, ML techniques, and SBT to build the models in order to identify the proneness changes attribute of a class using object-oriented metrics. This study is used by developers and researches to optimize their work. In the future, it has been planned to create an exact model based on ML and SBT that would perform even better.

Thi Minh Phuong in the year 2019 [12] aimed for conducting seven ML techniques to have a fault-free model using PROMISE dataset. The paper foresees to deal with the imbalanced data. Kazuya Tanaka, Akito Monden, and Zeynep Yucel in the year 2019 [13] aimed to efficiently evaluate the auto-sklearn so that it can predict the defects in software modules. In the future, their analysis would be why auto-sklearn did not perform to the desired standard (Table 1).

7 Analysis

Software quality prediction being a vital phase in software development life cycle (SDLC), it necessitates a rigorous and an appropriate software testing strategy to

Table 1 Summary of the software quality prediction techniques

Title of the paper	Methods used	Advantages	Disadvantages
Software Bug Prediction using ML Approach [3]	Naïve Bayes, decision tree, and artificial neural network	It has been concluded that DT is superior among NB and ANN	NB and ANN do not show best results as that of DT
Empirical comparison of ML Algorithm for Bug Prediction in open-source software [4]	Perception, backpropagation, LVQ1, Hierarchical LVQ, Multipass LVQ, SOM, Multipass SOM AIRS1, AIRS2 parallel, CLONAL-G	Results show that nonparametric Friedman test is superior among the single layered perceptron over other technique	Algorithms such as backpropagation, LVQ1, Hierarchical LVQ, Multipass LVQ, SOM do not show better results
Ensembles-based combined learning for improved software fault prediction: A Comparative study [5]	Feature selection and data balancing	Optimum level of operation of single learning algorithm is dependent on ensemble learning algorithm (ELA)	The challenges like in-equal class problem in buggy software dataset and noisy features cannot be solely handled by ELA
Software fault proneness prediction using genetic-based ML techniques [6]	Learning classifier system (LCS), zeroth level classifier system (ZCS) and Accuracy Based System (XCS)	Machine learning-based model is used on neural network approach. ZCS and XCS help to increase the accuracy of the predicted model	Backpropagation and Levenberg Marquardt does not provide strength to the approach as that of ML
Software defect prediction using ML techniques [7]	Principal component analysis, random forests, Naïve Bayes, support vector machine, and NN	The technique linear classification has the greatest accuracy	NB, RF, and NN have maximum precision in only one dataset
Software defect estimation using ML algorithms [8]	Bagging, random forests, multilayer perceptron, radial basis function, Naïve Bayes, multinomial Naïve Bayes, one discriminative classifier SVM	Random forests and bagging algorithms are better than other algorithms	Other methods such as bagging and NB do not perform well in KC2 dataset
Aging-related bug prediction using extreme learning machine [9]	Extreme learning machine and three kernels such as linear, polynomial, and RBF	Synthetic minority oversampling technique (SMOTE) is applied to recon the effect of imbalanced data distribution	SMOTE does not improve the performance for polynomial or RBF kernel

(continued)

Table 1 (continued)

Title of the paper	Methods used	Advantages	Disadvantages
Software bug prediction using supervised ML algorithms [10]	Logistic regression, Naïve Bayes, and decision tree	Random forests algorithm is used because of its ensemble nature	Models that are built using singular classifiers usually have less accuracy
Mining the impact of object-oriented metrics for change prediction using ML and research-based techniques [11]	Logistic regression, random forests, bagging, multilayer perceptron, adaptive boosting	The model predicted using bagging technique outperformed the techniques such as ML and search-based techniques (SBT)	LR, RF, BG, multilayer perceptron, adaptive boosting did not perform well
Experimental study on software fault prediction using ML models [12]	Logistic regression, K-nearest neighbor, DT, random forests, Naïve Bayes, SVM, and multilayer perceptron	SVM achieved a highest performance. Multilayer perceptron produced a highest accuracy	LR, KNN, DT, random forests, and Naïve Bayes did not perform well
Prediction of software defects using automated ML [13]	Auto-sklearn, RF, DT, and linear discrimination analysis	Auto-sklearn displayed same performance as that of RF and is capable of identifying the defects in the module without the knowledge of ML	Auto-sklearn did not show the similar result as that of DT and LDA

make the implementation features of the software optimal. Various strategies are available though, machine learning techniques have been proven to be the most appropriate ones so far as the literature study reveals. However, of all machine learning techniques, ANN, SVM, and decision tree approaches have been the most desirable ones as claimed by various authors that we have detailed in Sect. 6. These methods provide a higher level of accuracy which makes them most preferable. Nevertheless, RBF network approach also could be useful so far as the accuracy is concerned. Moreover, any machine learning technique can be carefully applied to software quality prediction if the various characteristics of the software are optimally chosen.

8 Conclusion

A software quality prediction model seeks to predict the quality factor that whether the software is prone to fault or not. The prediction model helps to create an accurate model and also helps the developers to deliver the software in time to their customers. It helps us to know which components are at fault which can be corrected by detailed testing. To improve the quality, reliability, efficiency, and maintenance

cost, the software fault should be predicted beforehand. The table has been prepared by taking consideration of the survey report conducted by 11 research papers. This also conveys the advantages and disadvantages of the methods used in the papers. As predicting the fault before the actual testing helps the software developers to complete the work within the fixed time and the project is delivered to the customer within the timeframe which in turn also reduces the cost and time of the project, the time left can be used in software maintenance. So, the minimum is the fault, the maximum time for software maintenance.

References

1. Kalaivani N, Beena R (2018) Overview of software defect prediction using machine learning algorithms. *Int J Pure Appl Mathe* 118(20):3863–3873
2. Singh M, Salaria DS (2013) Software defect prediction tool based on neural network. *Int J Comput Appl* 70(22)
3. Hammouri A, Hammad M, Alnabhan M, Alsarayrah F (2018) Software bug prediction using machine learning approach. *Int J Adv Comput Sci Appl* 9(2):78–83
4. Malhotra R, Bahl L, Sehgal S, Priya P (2017) Empirical comparison of machine learning algorithms for bug prediction in open source software. In: International conference on big data analytics and computational intelligence (ICBDAC), pp 40–45
5. Yohannese CW, Li T, Simfukwe M, Khurshid F (2017) Ensembles based combined learning for improved software fault prediction: a comparative study. In: 12th international conference on intelligent systems and knowledge engineering (ISKE), pp 1–6
6. Kumar A, Bansal A (2019) Software fault proneness prediction using genetic based machine learning techniques. In: 4th international conference on internet of things: smart innovation and usages (IoT-SIU), pp 1–5
7. Prabha CL, Shivakumar N (2020) Software defect prediction using machine learning techniques. In: 4th international conference on trends in electronics and informatics (ICOEI) (48184), pp 728–733
8. Yalçiner B, Özdeş M (2019) Software defect estimation using machine learning algorithms. In: 4th International conference on computer science and engineering (UBMK), pp 487–491
9. Kumar L, Sureka A (2017) Aging related bug prediction using extreme learning machines. In: 14th IEEE India council international conference (INDICON), pp 1–6
10. Immaculate SD, Begam MF, Floramary M (2019) Software bug prediction using supervised machine learning algorithms. In: International conference on data science and communication (IconDSC), pp 1–7
11. Malhotra R, Khanna M (2015) Mining the impact of object oriented metrics for change prediction using machine learning and search-based techniques. In: International conference on advances in computing, communications and informatics (ICACCI), pp 228–234
12. Ha TMP, Tran DH, Hanh LTM, Binh NT (2019) Experimental study on software fault prediction using machine learning model. In: 11th international conference on knowledge and systems engineering (KSE), pp 1–5
13. Tanaka K, Monden A, Yücel Z (2019) Prediction of software defects using automated machine learning. In: 20th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD), pp 490–494

Analysis of Stock Market and Its Forecasting



Sunil Wankhade, Adarsh Kaul, Sanjana Mohile, and Ruchira Kadam

Abstract The stock market is very volatile. Therefore, it is crucial to get accurate future trends and price predictions. Forecasting nonlinear data such as stock values requires a great calculation task that can be achieved easily with the help of machine learning algorithms. In this paper, we collect the data from Yahoo Finance. We perform data analysis and then compare different machine learning algorithms. We score each model with R^2 method. We also try different parameters with the help of hyperparameter tuning. The result is based solely on numbers and takes several axioms that could or could not follow as the prediction time in the real world.

Keywords Yahoo Finance · Data analysis · Machine learning · R^2 method · Hyperparameter

1 Introduction

In general, stock is one of the ways to raise money for the quick expansion of the company. In this research, we try to analyze the past data and try to forecast the current price of the stock. To achieve this, we will be using different machine learning algorithms and try to get an accurate result. With the help of this, we would be able to conclude which machine learning algorithms are best for forecasting and which machine learning algorithms we should avoid. Before you invest your money, you are required to do extensive research about companies and try to analyze the return or future trends of the company. With the help of this research, we would try to eliminate the time consumption in studying the trends and put it in terms that even non-commercial people can understand. There are two basic types of stock analysis:

- Fundamental analysis
- Technical analysis.

S. Wankhade (✉) · A. Kaul · S. Mohile · R. Kadam
Rajiv Gandhi Institute of Technology, Versova, Mumbai, India
e-mail: sunil.wankhade@mctr.git.ac.in

Fundamental analysis: In this type of analysis, we use math and statistics to find out the growth and trend of the company. We use tools like profit and loss statements, variance, and income statements to achieve the price.

Technical analysis: In this type of analysis, we rely on historical data and present price estimation. We calculate the risk percentage, the rate of return, and the value of return. The stock market mainly runs on technical analysis because of changes in the volume and price of a stock. For technical analysis, we used tools like Tableau and Python which help us to calculate the values and plot them in a graphical manner to make it easily understandable.

2 Literature Survey

Over the past two decades, many important changes have taken place in the environment of financial markets. Investors have been extended to include the innovations in potent contact and trade services. The projection of stock return has drawn researchers' interest for several years. This is a significant financial issue. It assumes a statistical connection with potential inventory returns that the fundamental knowledge that was open to the public in the past. Data mining techniques are modern techniques that can be used to collect information from these data to derive those relationships from available data. This is why many researchers centered their efforts on technical and advanced mathematics and science. The domain of artificial intelligence and data preprocessing techniques has acquired considerable attention. The two differences between the approach developed by the authors and further researchers are that the judgment prototype was first updated into a partial model to reduce the classification error and secondly, a decision tree with two-layer bias was used to enhance the precision of buying. The study results suggest that the decision model proposed provided better consistency. The findings of this analysis revealed that artificial neural networking gives investors the ability to boost their probability in stock. More specifically, the estimation of returns tends to be higher than the prediction of a multivariate model.

Prediction models for Indian Stock Market from Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016) publication are by Aparna et al. [2]. In this paper, the author tries to find out the trend of the stock by running different sentimental analysis and some machine learning algorithms. However, they do not give the exact value; they only show whether the price will go up or down.

A machine learning model for Stock Market Prediction is published by International Journal of Computer Science and Telecommunications. Written by Osman et al. [3], we learn from this paper how numerous SVR models can be used to project stock prices. The author uses the least square SVM algorithms in this article to estimate the stock price. The writers used a mean square error to validate the model's consistency.

Stock market prediction using machine learning classifiers and social media, news published by Journal of Ambient Intelligence and Humanized Computing (2020) and written by Wasiat et al. [4]

In this paper, they run algorithms on social media and financial news data to discover the impact of this data on stock market prediction accuracy for ten subsequent days. For improving performance and quality of predictions, feature selection and spam tweets reduction are performed on the datasets.

Recent advances in Stock Market Prediction Using Text Mining: A Survey by Faten Subhi Alzazah and Xiaochun Cheng, submitted on March 8, 2020, reviewed on March 24, 2020, and published on June 1, 2020 [5]. This study aims to compare many machine learning (ML) and deep learning (DL) methods used for sentiment analysis to find which method could be more effective in prediction and for which types and amount of data.

Stock Movement Prediction from Tweets and Historical Prices Published by Yumo Xu and Shay B. Cohen School of Informatics, University of Edinburgh 10 Crichton Street, Edinburgh EH8 9AB, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 1970–1979 Melbourne, Australia, July 15–20, 2018. c 2018 Association for Computational Linguistics. [6] In this study, the author forms a cluster of similar kinds of companies and tries to identify a trend for the entire set of companies in a group. The model introduces recurrent, continuous latent variables for a better treatment of stochastic and uses neural variations inference to address the intractable posterior inference.

3 Problem Statement

While working on stock market analysis and prediction, the foremost problem is accuracy. To make sure we assess and get the output for the data on the proposed algorithms, data whose findings have already been disclosed would be used. We will study the algorithms that give us the most relevant outcomes. An important challenge is the processing of data, whether we produce or gather our metrics. We got the data from the API call as it was easier. After that, there is the cleaning of the data, for any unusual/unwanted data that might be replicated or for null values. This will affect the result of the overall prediction. After that, we compare the different models which have different parameters and still get the results so that we can compare and decide the most suited one.

4 Proposed System

The figure below, Fig. 1, shows that the six steps involved in the system of stock market prediction and analysis are as follows:

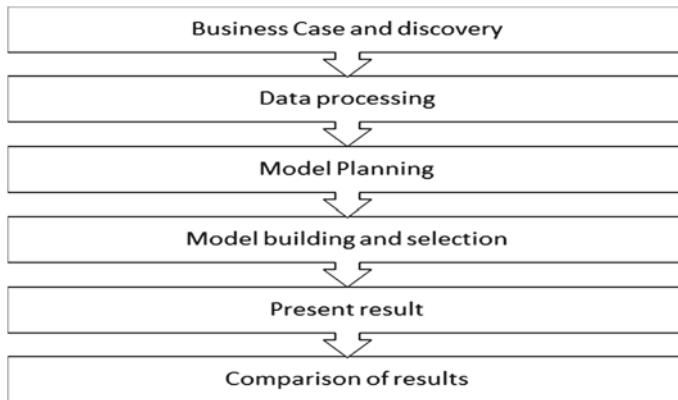


Fig. 1 Architecture for stock market analysis

The first step is the business case and discovery: In this, we first create the map of the project and try to find out how to approach the work. The issues in the problem statement are dealt with in this step. We calculate the time required for each stage and how should one approach it.

The second step is data processing: In this step, we try to eliminate the null values and remove the noisy data. This is an important step because the noisy data can affect the results and prevent us from getting accurate results.

In the third step, model planning, we discuss all the possible models available to us for regression and select the top 5 models for this research.

In the fourth step, model building and selection, we use different models and parameters available to us out of which we select the parameter of each model that gives us the most accurate results.

In the fifth step, present result, we show the results of the predicted price of each model with the original price of the model. In this step, we calculate the predicted price and with help of the line graph, we try to show the results.

In the final step, comparison of results, we compare the results of all and conclude which is the best and most suitable model for our research.

4.1 Scope of the Project

Scanning of stock data using data mining will be useful for new investors to invest in the stock market based on the various factors considered by the algorithms used in the research. Through this research, we can predict the future trend based on past data analysis. Based on the past data analysis, a person can also understand whether the value of the company will rise higher or drop lower. Another use of this analysis will be if any person is wanting to create a trading bot, they can directly use the research and know the best prediction method used to obtain the stock prices. We

would also be able to provide a rough estimation of the return if the person invests in this stock.

5 Data Analysis

It is vital to analyze the data before applying a machine learning algorithm because if the data is noisy, we do not get an accurate result. Data analysis involves data collection, cleaning of data, applying mathematical formulae, and reducing the data redundancy.

5.1 *Reading of Data*

Pandas DataReader

For reading the data, we have used the Python Pandas Datareader API call. Datareader allows reading the data from various sites such as Quandl, Yahoo Finance, World Bank, and many more sites. It allows us to use the data from these sites and directly import to our environment without downloading. This saves a lot of space because sometimes the data is in megabytes or gigabytes and downloading and uploading are time consuming. This is also better when you have frequent entries in the dataset, the latest entries can be obtained easily [7].

Yahoo Finance

It is a site maintained by Yahoo that gives us the financial news and all the latest updates. We have used Pandas DataReader and obtained the data from Yahoo Finance because it is a reputed site and can be trusted with the data. The data present in Yahoo Finance was already cleaned so we did not have to clean the data which is why we decided to go with it [8].

5.2 *Statistics*

After obtaining the data, we perform some statistics to understand the nature of the data. Statistics also allows us to modify and drop the columns which do not have any impact on the prediction. For this research, we have used correlation. Correlation helps us in finding out how different columns are correlated to each other. If any two columns are highly correlated then, either of them can be dropped.

5.3 Train and Test Split

This is a machine learning technique in which the data is divided into two sets, train set and test set. Train set is used to train the machine learning algorithms, and a test is used to test the output after training. The results are noted and sent back to machine learning algorithms, which are later tuned [9].

6 Modeling

There are mainly five algorithms which help in predicting the price of the stock, and those are as follows.

6.1 Linear Regression

Linear regression is a linear model, e.g., a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). This method tries to find a pattern or a relationship between the two variables, x and y , and predicts the future trend on its basis.

In Fig. 2, it portrays the difference of actual price and the predicted value of linear regression algorithm. Blue line is used to represent the original price, and red line is the predicted value of the used algorithm. In linear regression, the value has a fluctuation of 3 Rs (on a mean) in the start but then has a very negligible difference. Due to the difference in starting price, this algorithm is not suited.

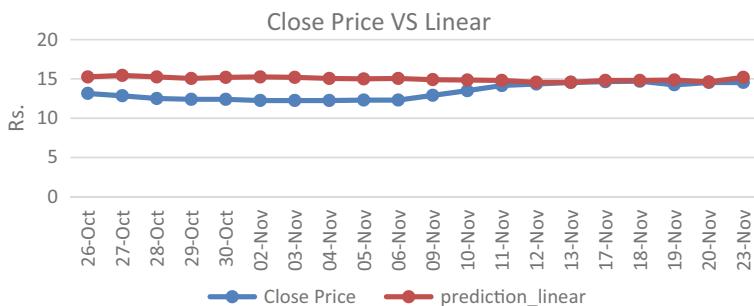


Fig. 2 Prediction of linear regression versus original

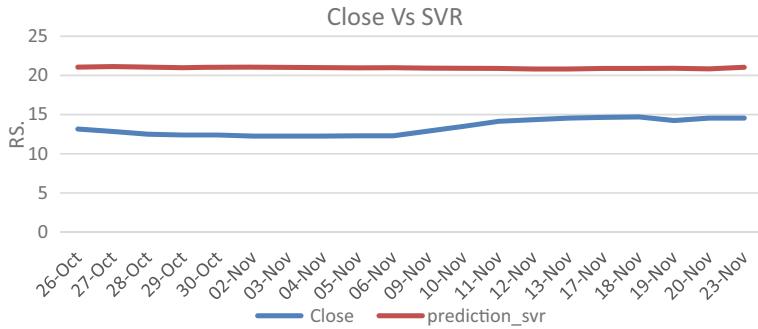


Fig. 3 Prediction of SVR versus original

6.2 SVR

Support vector regression model is a supervised machine learning algorithm. In the SVR algorithm, we plot each data as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then we regress by finding a hyperplane which distinguishes very well between two groups. Support vectors are just independent observation coordinates.

In Fig. 3, it depicts the price difference of original value and the predicted value of SVR model. Blue line is used to represent the original price, and red line is the predicted value of the used algorithm. In SVR, as we can see the original value is never meet with the predicted price, hence is not at all advised to be used.

6.3 XGBoost

XGBoost is an ensemble learning method. Ensemble learning is a structured approach for integrating different learners' predictive intensity. Although these two methods can be used in many mathematical models, decision-making bodies were primarily used.

In Fig. 4, it depicts the price difference of original value and the predicted value of XGBoost algorithm. Blue line is used to represent the original price, and red line is the predicted value of the used algorithm. XGBoost algorithm shows the same variance as linear regression and is in fact more inaccurate than linear regression.

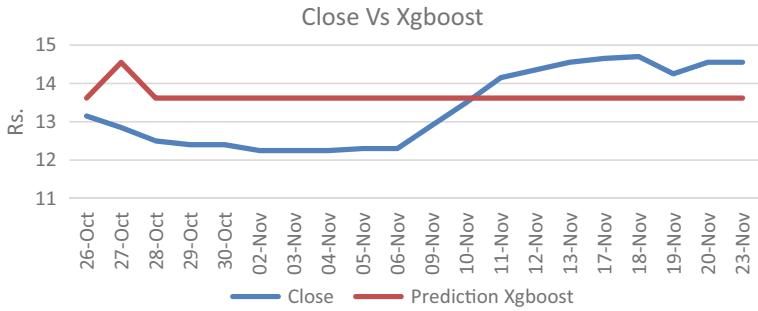


Fig. 4 Prediction of XGBoost versus original

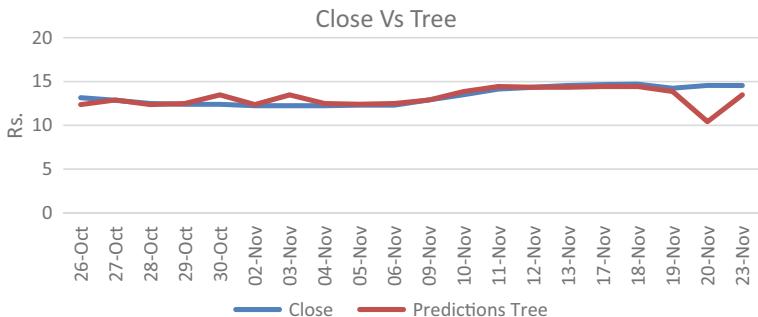


Fig. 5 Prediction of decision tree versus original

6.4 Decision Tree

A decision tree with its base at the top is drawn upside down. By cutting, the efficiency of a tree is additionally enhanced. It requires deleting the branches that use low-level functionality. This decreases the tree's complexity and thus increases its predictive ability by reducing overcrowding.

In Fig. 5, blue is the initial price of the line and the red is the expected algorithm value. Decision tree shows an almost trivial predictable price gap and appears to be more accurate than linear, SVR and XG Raise in both situations.

6.5 Random Forest Regression

Random forest is a type of supervised, ensemble-based learning machine algorithm. The combination is a method of learning in which you apply various algorithms or algorithms to a stronger prediction model multiple time. The random forest algorithm incorporates several algorithms of the same shape, including many determine trees

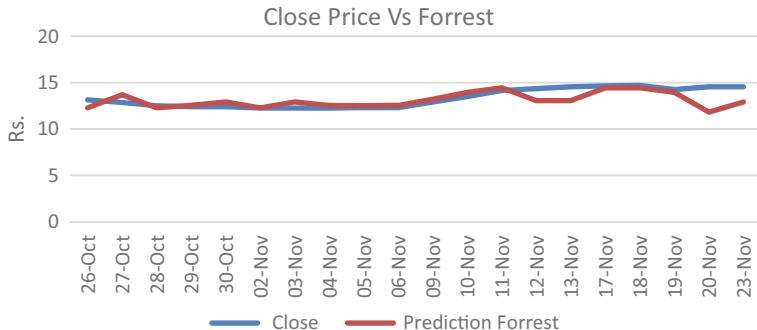


Fig. 6 Prediction of random forest versus original

that lead in a wood of trees, and that is why the random forest was named. The random forest algorithm can be used for regression and classification functions.

Figure 6 indicates the price difference between the original value and the estimated value of the random forest algorithm

7 Hyperparameter Tuning

They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as hyperparameters. These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. Of course, you must select from a specific list of hyperparameters for a given model as it varies from model to model. [10] Often, we are not aware of optimal values for hyperparameters which would generate the best model output. So, what we expect from the model is to explore and select the optimal model architecture on its own. This selection procedure for hyperparameter is known as hyperparameter tuning. Hyperparameter tuning is required because in the earlier part of the research we have used default parameters which might not give accurate results. So, in order to get better and optimum results, we use hyperparameter tuning with the help of the range of parameters.

There are two types of hyperparameter tuning:

- GridSearchCV
- RandomSearchCV.

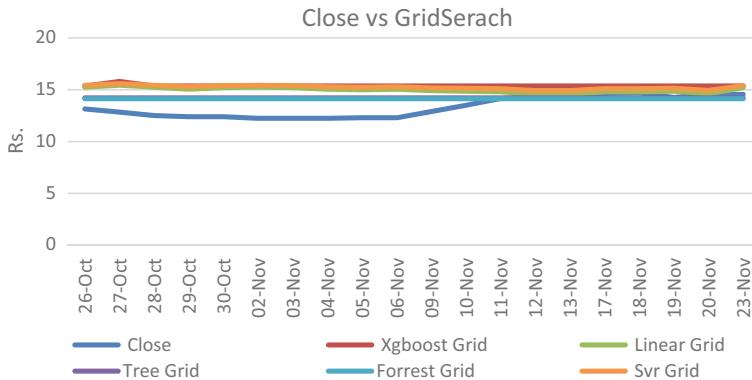


Fig. 7 Prediction of GridSearchCV versus original

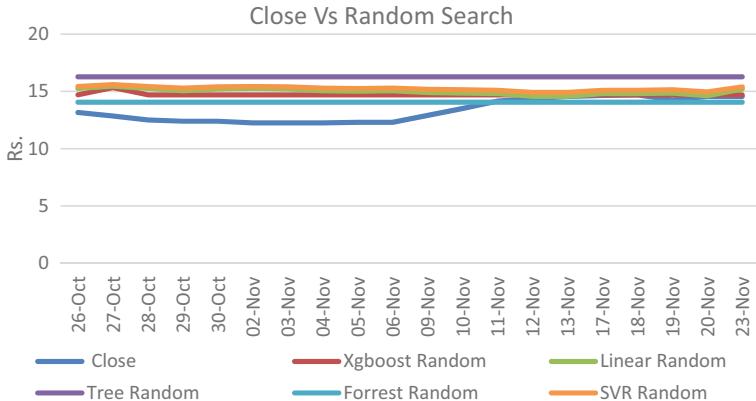
7.1 *GridSearchCV*

The GridSearchCV instance implements the usual estimator API: when “fitting” it on a dataset all the possible combinations of parameter values are evaluated and the best combination is retained. The grid search provided by GridSearchCV exhaustively generates candidates from a grid of parameter values specified with the param_grid parameter. Grid search is used to find the optimal hyperparameters of a model which results in the most “accurate” predictions. Grid search is the most convenient way of tuning hyperparameters. With this approach, we can easily propose the design for each possible combination of all the values of the hyperparameter, test each model, and determine the best parameters (Fig. 7).

7.2 *RandomSearchCV*

While using a grid of parameter settings is currently the most widely used method for parameter optimization, other search methods have more favorable properties. RandomizedSearchCV implements a randomized search over parameters, where each setting is sampled from a distribution over possible parameter values. The main benefit over grid search is adding parameters that do not influence the performance and does not decrease efficiency. Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. It is similar to grid search, and yet it has proven to yield better results comparatively. The random search differs from grid search by having a distinct collection of values for each hyperparameter; alternatively, we provide for each hyperparameter a statistical distribution, from which values can be sampled randomly.

As we see in figure number 8, the random forest, decision tree, and the XGBoost results are very far from the original price at the beginning, however at the end, i.e.,

**Fig. 8** Prediction of random search CV/s original

from 2020-11-13, they move close to meet the original price. On the other hand, SVR is never close to the original price (Fig. 8).

8 Results

For the scoring of the algorithms, we have R^2 scoring method. This method is used for scoring regression problems. What R^2 does is it sums the square of residual (RSS) value divides with total sum of squares (TSS) and the answer is subtracted by 1 [11].

$$R^2 = 1 - (\text{RSS}/\text{TSS})$$

Table 1: It shows the R^2 result of models used with the help of default parameters. The table shows the random forest algorithm gives us the most accurate result followed by decision tree and XGBoost. The least accurate result is given by SVR

Table 2: It shows the R^2 result of models used with the help of GridSearchCV parameters. It gives us more accurate results compared to the default parameters. The results of the table depict that the random forest algorithm gives us the most accurate result followed by the decision tree and XGBoost.

Table 1 Simple model R^2 method

Model	Train	Test
Linear	0.960118	0.964879
SVR	0.958118	0.860560
XGBoost	0.991506	0.957736
Tree	0.999274	0.944105
Forest	0.992728	0.959457

Table 2 GridSearchCV model R^2

Model	Train	Test
Linear	0.960118	0.964879
SVR	0.958118	0.900560
XGBoost	0.991506	0.958736
Tree	0.999274	0.943105
Forest	0.992728	0.974457

Table 3 RandomSearchCV model R^2

Model	Train	Test
Linear	0.960118	0.964879
SVR	0.958118	0.900560
XGBoost	0.991506	0.967736
Tree	0.999274	0.954105
Forest	0.992728	0.969457

Table 3: It shows the result R^2 of models used. As you can see the results of the table show the random forest algorithm gives us the most accurate result followed by the decision tree and XGBoost. The least accurate result is given by SVR.

9 Conclusion

This paper provides a clear picture of how to execute machine learning. After the analysis of some different kinds of regression models on the data of Yes Bank, we have come to a conclusion that the most suitable model is random forest regression which gives the most accurate result. Random forest regression is a model that gives the most accurate result even after hyperparameter tuning.

The R^2 score of random forest regression model was 0.97959457 before hyperparameter tuning and whereas after tuning it increased to 0.974457. This paper is limited to supervised machine learning only, and therefore only attempts to explain the fundamentals of this nonlinear system.

References

1. Kunal P, Neha A (2019) Stock market analysis using supervised machine learning. In: International conference on machine learning, big data, cloud and parallel computing (Com-IT-Con). Amity University, Uttar Pradesh
2. Aparna N, Manohara Pai MM, Radhika MP (2016) Prediction models for Indian stock market. In: Twelfth international multi-conference on information processing. India

3. Osman H, Omar SS, Mustafa AS (2013) A machine learning model for stock market. *Int J Comput Sci Telecommun* 4(12)
4. Wasiat K, Mustansar AG, Muhammad AA, Amin K, Khaled HA, Ahmed SA (2020) Stock market prediction using machine learning classifiers and social media, news. *J Ambient Intell Hum Comput*
5. Faten SA, Cheng X Recent advances in stock market prediction using text mining: a survey. <https://doi.org/10.5772/intechopen.92253>
6. Xu Y, Shay BC (2018) Stock movement prediction from tweets and historical prices. In: Proceedings of the 56th annual meeting of the association for computational Linguistics (Long Papers), pages 1970–1979 Melbourne, Australia, July 15–20
7. Overview of machine learning technique and algorithms, https://www.udemy.com/share/101_Qd4BkQccl1QQXo. Last Accessed 21 Dec 2020
8. For collection of data <https://in.finance.yahoo.com/quote/YESBANK.NS/>. Last Accessed 23 Nov 2020
9. For data reading APIRetrieved, <https://pandas-datareader.readthedocs.io/en/latest/>. Last Accessed 23 Nov 2020
10. Agarwal T (2020) Hyperparameter optimization in machine learning. Apress, India

Evaluation of the Technology Acceptance Model for Lean Six Sigma Approach—The Main Study



Slawomir Switek, Ludoslaw Drelichowski, and Zdzislaw Polkowski

Abstract This work is a continuation of an article: Empirical evaluation of the revised Technology Acceptance Model for Lean Six Sigma approach—a pilot study. The purpose of the paper is to test rated opinions of Lean Six Sigma users in order to conclude about external factors affecting effectiveness of implementation of the management conception and at the same time from theories related to individual perception standpoint to describe internal motivation of its users by applying revised Technology Acceptance Model. In order to achieve the goal, the gotten data during the main study have been analyzed by use covariance based structural equation modeling (CB-SEM) applied to resolve multiple regression equations describing the TAM model. The conducted research and analyses allow us to conclude that the technology acceptance model is an interesting proposal supplementing the set of diagnostic tools in the field of appropriate change management, which is the implementation of the Lean Six Sigma management conception. This is the pragmatic value of this work.

Keywords Lean six sigma · Enterprise resource planning (ERP) system · Technology acceptance model (TAM) · Structural equation modelling (SEM) · Change management

S. Switek
WSB University, Dabrowa Gornicza, Poland
e-mail: sswitek@wsb.edu.pl

L. Drelichowski
WSG University, Bydgoszcz, Poland
e-mail: ludoslaw.drelichowski@byd.pl

Z. Polkowski (✉)
Wrocław University of Economics and Business, Wrocław, Poland
e-mail: zdzislaw.polkowski@ue.wroc.pl

1 Introduction

The assessment of the effectiveness of the use of various methods and resources supporting the management of organizations is an extremely important criterion in assessing the accuracy of the selection of management methods and techniques in a given organization, as well as the correctness of the implemented implementation works. It seems justified to assume that one of the key evaluation criteria is the values of economic and financial indicators of the enterprise in the implementation period. It may be difficult to take into account the impact of other innovative activities carried out in parallel to management and organizational activities because it is difficult to clearly assign the source of the improvement of the indicator values to individual innovative activities. A certain answer in the search for a solution to this problem, which can be obtained indirectly, is to conduct a survey of the opinions of users who apply the implemented solutions at various management levels in the organization.

Studies of the literature concerning the methodology of assessing the degree of acceptance of various management methods as well as organizational and IT techniques indicate the TAM model as a widely used tool in achieving this goal [1].

2 Theoretical Background

The explanatory power of the theories discussed in this chapter varied and grew as modifications were introduced by successive researchers. On the one hand, it is important that the theoretical model explains the behavior of an individual, but on the other hand, it would take into account many aspects resulting from and influencing the environment in which the individual operates. This type of interface, which are external (exogenous) factors, in the case of the TAM model can be easily adapted to the given application [2].

Due to the wide use of this model in many areas and the possibility of its adaptation to external conditions, it will be used for empirical research in this work. Przeclewski also confirms that the TAM model has been the most frequently used theory explaining the use and acceptance of technology in research practice in relation to other single acceptance models. In the meta-analysis of the model, which this researcher conducted on the basis of a quantitative review of the literature, he concludes that the values of the correlation coefficients between the model factors, although they are average high, are characterized by high variability, which is related to the influence of external variables. He describes the assessment of the impact of these factors as valuable from both a theoretical and practical point of view [3].

As noted earlier, the literature research has so far failed to demonstrate the use of perceptual theories to explain behavior in the case of lean six sigma acceptance, which provides an opportunity to recognize this new research area and, consequently, to understand human actions during user interaction with tools and methods used in lean six sigma.

In this work, the first Davis' technology acceptance model, often marked in the literature as TAM I, has not been re-mentioned—interested researchers refer to the first part of our consideration [4].

The relationships of the model are theorized to be linear. The model can be expressed using the following four equations [5]:

$$\text{PE} = \sum i = 1, n \beta_i X_i + \varepsilon \quad (1)$$

$$\text{PU} = \sum i = 1, n \beta_i X_i + \beta_n + 1 \text{ PE} + \varepsilon \quad (2)$$

$$\text{AT} = \beta_1 \text{ PE} + \beta_2 \text{ PU} + \varepsilon \quad (3)$$

$$\text{BI} = \beta_1 \text{ AT} + \varepsilon \quad (4)$$

where:

X_i design feature $i, i = 1, n$

PE perceived ease of use.

PU perceived usefulness.

AT attitude toward behavior.

BI behavioral intention of use.

β_1 standardized partial regression coefficient.

ε random error term.

3 Conceptual Model and Research Hypotheses

Structural equation modeling, replacing the system of many multiple regression equations, allows to test research hypotheses with a high complexity of relationships between variables. It enables testing of these relationships between variables that the theory assumes as cause-effect. Since SEM can deal with latent variables, it is also great for statistical inference in the TAM model. Moreover, from a technical point of view, SEM is good at failing to meet the normality assumption of a multivariate distribution, so that SEM algorithms can be applied to data that comes from a discontinuous scale. This feature allows to analyze the results of surveys, where the number of responses is limited (e.g., Likert's scales).

In general, the TAM model examines the behavior of market participants, i.e., users of a specific e-product, analyzing both psychosocial factors (behaviors, attitudes) and considering the technical, organizational, and process side of the tested product or service. Similarly, according to the authors of this study, it should be stated that the Lean Six Sigma management conception, based on the DMAIC process, can be treated as a virtual technology, in which also technical factors, i.e., its tools, its

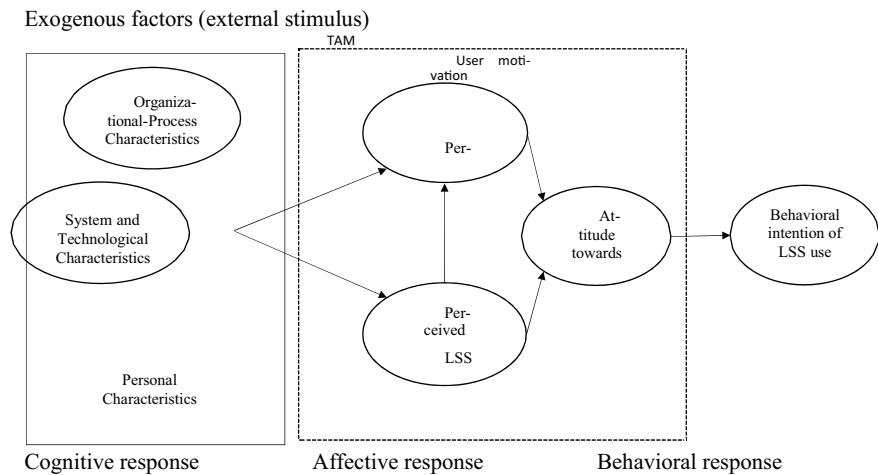


Fig. 1 The TAM model adapted to ERP systems. *Source own work*

structure and specific practices related to this program will influence the attitudes and behavior of users—in this case, trained Lean Six Sigma Belts of all levels.

External factors were classified into three subgroups: organizational and process characteristics, system and technological characteristics, and personal characteristics of the participant. This division is quoted by Bobek and Sternad [6] after Davis [7], who dealt with the assessment of the acceptance of ERP systems, where, as in the case of the lean six sigma conception, standardization is process-based and much less dependent on the entity. On a general level, it is universal for all TAM models. The input (observable) factors change, depending on the scope of the model application (see Fig. 1).

Lean Six Sigma programs are implemented globally in various organizations around the world, so often the implementations involve customers (users) from different cultures. As a result, it is a huge group, diverse in terms of the needs and perception of this type of initiative. Organizations implementing such solutions should examine the opinions of their users with regard to use of this system, not only in terms of the product itself (e.g., quality of training materials, simple and clear methodology) but above all to prepare better the implementation, taking into account local conditions and preferences. Therefore, in order to examine the opinions of LSS users, the following hypotheses are stated:

- H1: Perceived six sigma ease of use (PE) has a positive and direct effect on perceived six sigma usefulness (PU),
- H2: Perceived six sigma ease of use (PE) has a positive and direct effect on attitudes toward the six sigma (AT),
- H3: Perceived six sigma usefulness (PU) has a positive and direct effect on attitudes toward the six sigma (AT),

- H4: Attitude toward the six sigma (AT) has a positive and direct effect on behavioral intention of six sigma use (BI),
- H5: There is a group of external factors which have an influence through the conceptual factor personal characteristics (CPU) on the use of the six sigma,
- H6: There is a group of external factors which have an influence through the conceptual factor system and technological characteristics (CST) on the use of the six sigma,
- H7: There is a group of external factors which have an influence through the conceptual factor organizational-process characteristics (COP) on the use of the six sigma.

4 Methodology

The questions validated by the pilot study were used to create the main study questionnaire form. In the adopted research model, they contain input factors of the exogenous variables of the model (the first two characteristics) resulting from the aspects of implementing the Lean Six Sigma conception and the characteristics of organizations that have taken up the challenge of implementing this improvement methodology.

Questions about psychosocial factors (personal characteristics of the user, perceived ease of use, attitude toward the program, and intend to use) were taken from the empirical research of Park [6], Alharbi and Drew [7].

Accordingly, the identified external factors after the pilot study determining the structure of the lean six sigma solutions acceptance model by users, have been classified into three subgroups: organizational and process characteristics, system and technological characteristics, and personal characteristics of the user.

External factors are:

- organizational and process characteristics (COP):
- strategy (O1),
- process management (O2),
- change management (O3),
- trainings (O4),
- management involvement (O5),
- practices that sustain the LSS program and its development (O6),
- system and technological characteristics of the LSS conception (CST):
- project selection (S1),
- resource availability (S2),
- DMAIC respected and supervised (S3),
- personal characteristics of the participant (user) of the LSS conception (CPU):
- previous knowledge in the field of improvement techniques (P1),
- previous knowledge in the field of statistics (P2),
- personal openness to the application of new ideas and practices (P3).

To explain the process of acceptance (acceptance) of the LSS program using the TAM model, the methodology of structural equation modeling with the ML, maximum likelihood estimator was used, which is offered by the IBM SPSS Amos v. 24 program. CB-SEM with the ML estimator is the most frequently used SEM method, which copes well with failing to meet the normality assumption of a multivariable distribution, although it requires relatively large samples ($N > 250$) [8].

Davis, using the previous research of Ajzen and Fishbein [9], applied to operationalize attitudes in the TAM model, the 7-point Likert scale, for which the Alpha-Cronbach coefficient reached the value of 0.96 [10].

For the purposes of this study, a 7-point Likert scale was also used. Czakon cites (13) two interesting aspects in this regard [12]:

- questionnaire studies conducted for several years at the Department of Entrepreneurship of the University of Economy in Katowice showed that the 5-point scale is too narrow and limits the respondents' ability to assess the phenomenon, while the ten-point scale gives a too wide range of assessments and there is no mathematical mean in the form of one number which, on the one hand, affects the indecisiveness of the respondents, and on the other hand, overestimates or underestimates the feelings of the managerial staff as to the assessment of a given phenomenon,
- although a questionnaire based on such scale examines the feelings of management personnel referring to the level of a given phenomenon, but not its actual level (which is a limitation of such questionnaire), there are studies that show that subjectivism and attitudes of managers largely determine about the actual level of the phenomenon.

The LSS conception creates its own organization consisting in building a training network and certifying its users at various management levels, therefore only those respondents who had experience in implementing this conception (experts) were invited to the study.

Therefore, the answers on the 7-point Likert scale were coded from 1 to 7, where: 1—I strongly disagree, 2—I disagree, 3—I partially disagree, 4—I neither agree nor disagree, 5—I partially agree, 6—I agree and 7—I completely agree.

5 Main Study

The adoption of appropriately grouped model input factors made it possible to present its formal structure (see Fig. 2).

The main study was conducted using the CAWI/CATI method during the period from February, 2017 to the first week of July, 2017. Initially, the form was sent out (by courtesy of consulting companies dealing with Lean Six Sigma—SBTI Breakthrough Technologies in San Marcos, TX—USA and the White Raven Academy from Wrocław—Poland) in the form of an internet link to clients of these institutions.

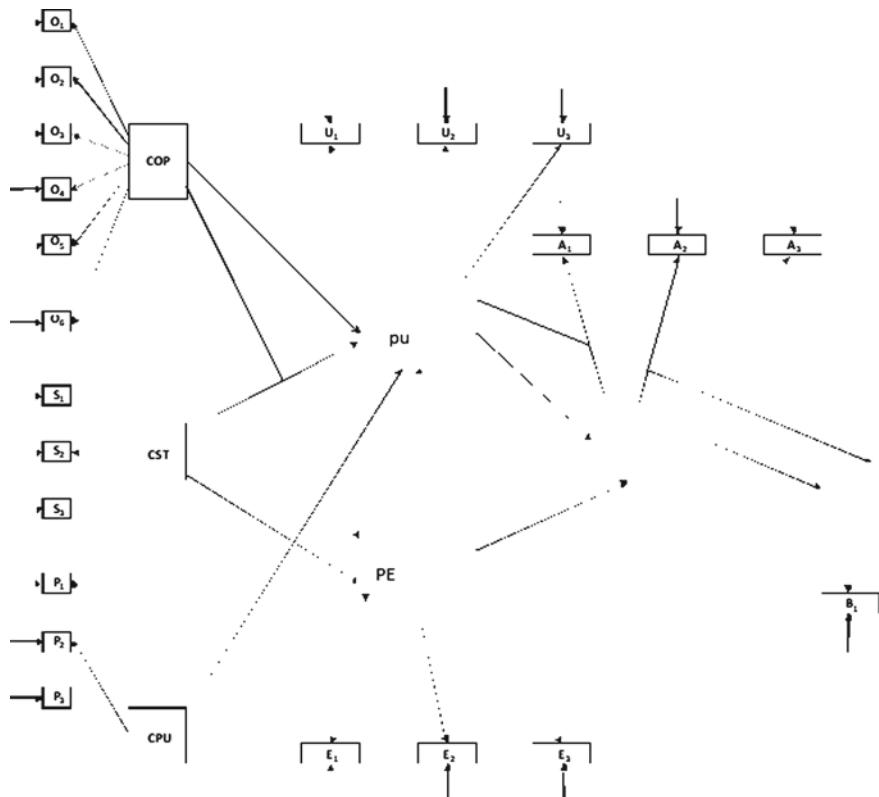
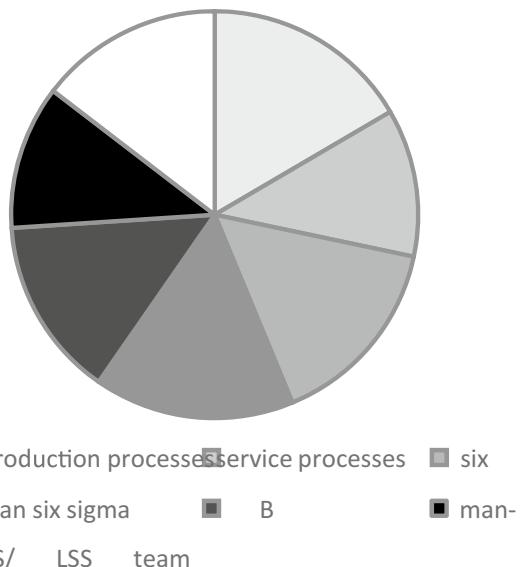


Fig. 2 The formal structure of a theoretically interesting model. *Source* own work

Moreover, the link was placed on professional portals Goldeline.pl and LinkedIn.pl. Despite these actions, the number of responses was quite moderate and, in order to exceed the sample size threshold, to be considered large for the SEM-ML methodology ($N > 250$), a professional interviewing company was engaged, which collected 200 responses. The size of this non-random sample was 317 (249 from Poland, 38 from other European countries, 24 from North and South America together with Australia, 5 from Asia, and one respondent from an undefined region).

External interviewers also encountered difficulties in gathering opinions of respondents, as usually, large companies tend to protect information about improvement programs and not participate in surveys, besides potential respondents often had to wait for their superiors to approve participants in the survey. After the interviews have been conducted, the interviewers concluded that, in general, people who work in an LSS environment on a daily basis are satisfied with it. The respondents from the main study most often came from electro-technical industry, consulting and training, automotive, and transactional services (BPO, SSC, and ITO), the category that included many other industries.

Fig. 3 Respondent professional experience.
Source own work



Virtually half of them were experienced in the pure Six Sigma approach and the other in the integrated Lean Six Sigma approach. As far as the implementation of tools and techniques in processes is concerned, most of them—though not decisive (55%)—came from production processes. The roles performed (Belt, manager, team member) were evenly distributed (see Fig. 3).

6 Data Analysis and Results

The statistical analysis of the data was performed in the IBM SPSS AMOS version 24.0 program and started with checking the fulfillment of assumptions about the normality of the distribution of observable variables, i.e., the multivariate distribution. The verification of normality of the distribution of a given variable consisted in checking whether its skewness and kurtosis differ significantly from expected values. The calcs made for skewness and kurtosis show values decreased by 3 for each variable. In a normal distribution, skewness is 0 and kurtosis is 3. For each variable, the data shows that all values of the test statistic to verify the hypothesis for skewness are outside the range $[-2; 2]$. Similarly, most of these statistic values for kurtosis do not fall within this range, so the hypothesis that the partial distribution is normal should be rejected. For the given multivariate distribution the multivariate coefficient of kurtosis is 268,113, and the test statistic is 70,383 (it, therefore, goes well beyond the range $[-2; 2]$). Due to the above, the null hypothesis that multivariate kurtosis is 3 should be rejected. This is obviously caused by the discontinuity of the adopted

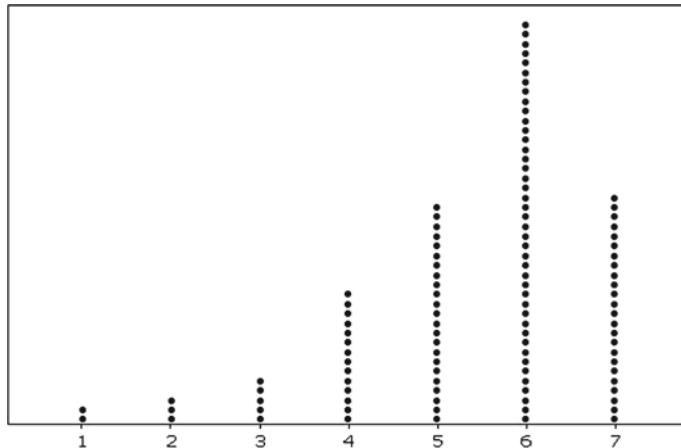


Fig. 4 Descriptive statistics of the obtained multivariate distribution. *Source* own work

Likert scale and the respondents' answers, where most of the observations are in the upper half of the scale, which results in an asymmetry of distributions. (see Fig. 4).

In order to loosen the assumption of normality of the multivariate distribution, the ML parameter estimation method with bootstrap sampling was used [13], which is compliant with the method proposed by Bollen and Stine [14]. As mentioned earlier, Konarski states, citing Nevitt and Hancock [15], that the number of 250 is optimal for estimating parameters of the SEM model. Thus, the value of 250 was adopted for the study. An alternative method would be to use the ML estimator with the data transformed to normal distribution, unfortunately, the transformation according to Johnson's algorithm was not successful.

In terms of scale reliability, the calculated Alpha-Cronbach coefficients for each variable showed values above 0.8 in each case. Therefore, it should be assumed that there is no issue with the accepted questions within a given construct.

In Confirmative Factor Analysis (CFA), as the tool used to verify the measurement model, it is assumed that standardized regression weights should exceed the desired value of 0.7. Only P3 (personal openness to the application of new ideas and practices) does not achieve this value. At the same time, the value of 0.663 can be considered acceptable (18). All values of these loadings are positive, so external factors will strengthen the explained variable and are statistically significant at the significance level $\alpha = 0.05$. Moreover, according to the quoted authors, squared multiple correlations should exceed the desired value of 0.5 for those factors with loads exceeding 0.7. This condition is met for all factors. Only for P3, which did not reach the load value of 0.7, the value of the explained variance is 43.9%. This is not a bad result.

Overall, the measurement model is correct and of course acceptable. For all external factors, the values of loads and the percentage of explained variance reach

the desired values. For one factor only, they are below the thresholds, but not much below.

The next step in the SEM analysis is to assess the quality of the model. In order to determine the degree of fit of the model to empirical data, many statistics would be used together what is recommended in assessing the quality of structural equation models [17].

The value of the CMIN fit test statistic is 927,276, which at 219 degrees of freedom gives a significance level of 0. This means that the null hypothesis stating that the model reproduces the sample covariance matrix well should be rejected. Due to the weakness of these statistics, it is not possible to base the fit assessment solely on this one. CMIN/DF is often used to correct for the complexity of the model. In this case, it is 4234. Some researchers recommend rejecting models in which this value exceeds 2, others take less restrictive limits—5 or 10. It should also be remembered that CMIN favors complex models—for the best model (saturated) it suggests 276 parameters included in the model. However, the obtained CMIN values here are closer to the saturated model (see Table 1).

Table 1 Fit statistics

Fit statistics	Default model	Saturated model	Independence model
CMIN	927,276 at degrees of freedom of 219	0	6,226,367
Hoelter 0.05	87	–	15
Hoelter 0.01	93	–	16
RMSEA	0.101	–	0.273
FMIN	2934	0	19,704
FMIN LO 90	1955	0	18,100
FMIN HI 90	2551	0	19,726
RMR	0.437	0	0.740
GFI	0.821	1	0.173
AGFI	0.774	–	0.098
PGFI	0.651	–	0.158
NFI	0.851	1	0
RFI	0.828	–	0
IFI	0.882	1	0
TLI	0.863	–	0
CFI	0.881	1	0
PGFI	0.866	0	1
PNFI	0.737	0	0
PCFI	0.763	0	0

Source own work

Table 2 Informative criteria

Informative criteria	Default model	Saturated model	Independence model
AIC	1,041,276	552	6,272,367
BIC	1,255,534	1,589,457	6,358,822
ECVI	3295	1747	19,849
ECVI LO90	3009	1747	19,046
ECVI HI90	3605	1747	20,672

Source own work

N Hoelter statistics and RMSEA value didn't exceed the threshold of acceptance while GFI and AGFI basically could be accepted. NFI, IFI, and CFI do as good as AGFI does.

The Akaike AIC, Bayes-Schwartz BIC, and ECVI information criteria inform that the adopted research model is much closer to the saturated model (describing the perfect fit) than to the independence model (see Table 2).

As you can see from the presented above tables, the model used explains a large part of the observed data variability (82.1%), but many criteria for assessing the model fit do not reach the acceptability threshold, although it has to be pointed out, the obtained values of the statistics do not differ significantly from the minimum requirements. This is a prerequisite for further optimization of the model.

In order to improve the degree of model fit, the output data elements from the program SEM algorithm were used that suggested a correlation relationship between COP and CST, CST and CPU, and lastly COP and CPU. Due to the above, the model has been graphically modified. The mentioned exogenous variables were appropriately connected with arcs ending with an arrow on both sides in the path chart. This corresponds to covariance, which can be treated as a non-standardized correlation.

The SEM calculation algorithm based on ML estimator with bootstrap sampling was restarted. For example, the N Hoelter statistic came very close to the required threshold of 200, the RMSEA was a satisfactory value, the RMR value dropped significantly and the model came much closer to the saturated one. In parallel, the information criteria values dropped by half, approaching the saturated model strongly. The model turned out to be adequate and fit, and the percentage of variance explained by the model ($GFI = 87.4\%$) is a high result as only 12.6% of the variance remains unexplained. This result confirms that the identification of external factors is correct.

For the adopted external factors, the obtained values of standardized path coefficients exceed the requested value of 0.7 and they're statistically significant at the α level of 0.05. Moreover, squared multiple correlation values exceeded the value of 0.5 for the factors. Only for P3, these parameter values were not achieved, but at the same time are acceptable for the model [16]. All values of these loadings are positive, so external factors will strengthen the explained variable (acceptance).

Table 3 shows that all the hypotheses, except for H2, were met at the assumed significance level $\alpha = 0.05$.

Table 3 Regression weights—the final model

			Estimate	S.E	C.R	P	Label
PU	←	PE	0.391	0.074	5317	***	par_22
T	←	PU	0.995	0.081	12317	***	par_24
AT	←	PE	-0.070	0.065	-1080	0.280	par_25
BI	←	AT	0.938	0.231	4053	**	par_26
BI	←	PU	200	0.224	0.893	0.372	par_27
O6	←	COP	1000				
O5	←	COP	0.972	0.065	14,993	***	par_1
O4	←	COP	0.854	0.061	13,959	***	par_2
O3	←	COP	1169	0.062	18,993	***	par_3
O2	←	COP	0.811	0.058	14,036	***	par_4
O1	←	COP	1132	0.064	17,681	***	par_5
S3	←	CST	1000				
S2	←	CST	1221	0.084	14,604	***	par_6
S1	←	CST	1036	0.074	13,994	***	par_7
P3	←	CPU	1000				
P2	←	CPU	1342	0.105	12,757	***	par_8
P1	←	CPU	1401	0.107	13,091	***	par_9

Source own work

Table 4 Covariances between exogenous variables

Kowariancja			Estimate	S.E	C.R	P	Label
COP	↔	CST	1218	0.133	9165	***	par_28
CST	↔	CPU	0.398	0.069	5809	***	par_29
COP	↔	CPU	0.397	0.071	5601	***	par_30

Source own work

Table 5 Correlations between exogenous variables

Korelacja			Estimate
COP	↔	CST	0.931
CST	↔	CPU	0.465
COP	↔	PU	0.414

The added elements (paths) of the research model taking into account the correlation between the exogenous variables of the model turned out to be statistically significant, and the relationship between COP and CST is the strongest (correlation coefficient close to 1), while the relationships between CPU, COP, and CST turned out to be twice weaker (see Tables 4 and 5).

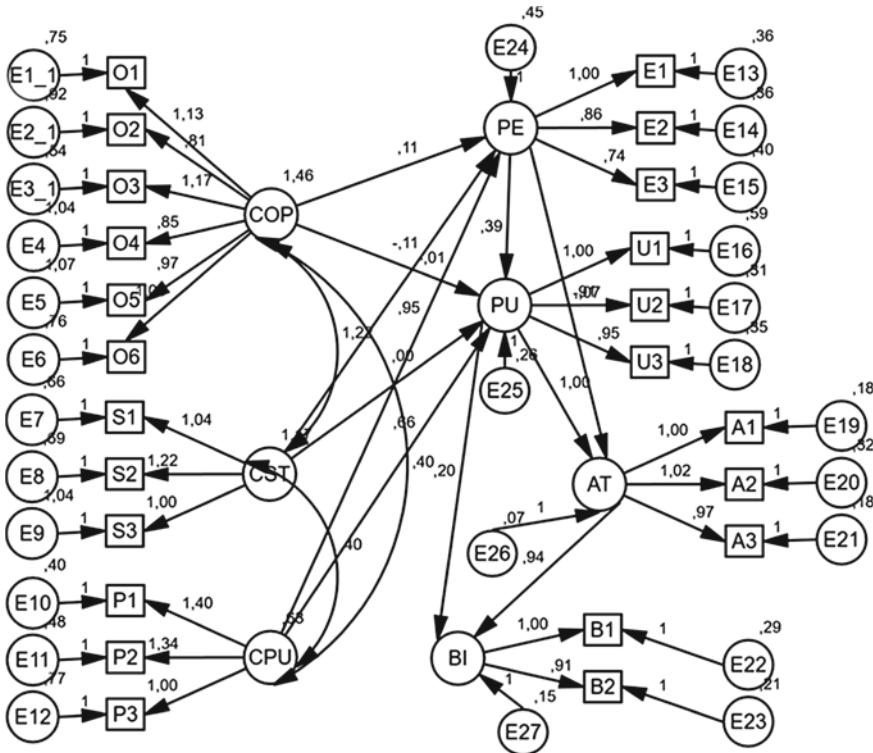


Fig. 5 The final form of the path model—non-standardized values. *Source* own work

In case of SEM method, stability of the obtained results should be also confirmed. To evaluate the stability of the obtained parameter values, significantly different sizes of bootstrap samples and the final form of the path model were used (see Fig. 5). In the analyzed case, the values of the path parameters were generated for the following size of the bootstrap sample—250, 1000, and 10,000. Values of estimated parameters and values of the standardized S.E. error remained the same, which confirms the stability.

The SEM model presented above is graphically represented by the path chart, where the cause-effect relationship is symbolized by an arrow pointing from the independent variable to the dependent variable in a given relationship.

Each arrow corresponds to one path factor. The path coefficients here are the regression coefficients. Random components, or residuals, are represented in circles with the description of E .

7 Discussion and Conclusions

The technology acceptance model TAM is a methodological and purposeful tool to support the analysis of factors (both external and internal) determining the effectiveness of Lean Six Sigma application, which was demonstrated in the theoretical considerations and the results of empirical research contained in the paper. The adopted final model explained 87.4% of the variability of the observed variance-covariance matrix, which was the input to the SEM methodology used. The GFI value of 87.4% is a high figure as only 12.6% of the variation remains unexplained. This result confirms that the identification of external factors is correct. The R^2 multiple correlation coefficients for endogenous variables of the model - AT, BI amounted to 92.8% and 89.2% respectively, which proves a high explanation of the variance of the variables included in the model.

The analysis of the revised TAM model made it possible to establish in particular:

1. No impact of organizational and process characteristics and system and technology characteristics on the perceived ease of use and perceived usefulness; both latent variables are strongly correlated with each other, which may lower their actual significance, and the correlation suggests that they are not separately distinguished by system users.
2. The strong influence of the user's personal characteristics on the general perceives of the LSS conception was confirmed. This should provide a strong impetus to those responsible for developing the LSS methodology in organizations to properly take care of selection criteria for potential candidates to participate in this strategic action for each company.
3. No impact of perceived ease of use on attitude toward the LSS management conception. Ease of use does not inspire an attitude, while usefulness does.
4. Strong influence of perceived usefulness on attitude with negligible influence of perceived ease of use on the attitude. This proves that the respondents believe that the use of the LSS conception is useful for their organizations and at their workstations. Training plays a similar, strong role in enhancing perceived ease of use, and therefore the attitude strongly influences the behavioral intention to use.
5. The influence of all exogenous variables of the model stimulates the acceptance of Six Sigma.

User acceptance and the consolidation of employees' attitudes are explained by changes in cultural, internal factors, including interactions in terms of external factors influencing the organization. The main study was carried out on a group of 317 people and the interpretation of the results brought results similar to those presented in the publications on the application of the TAM model in the implementation of ERP systems, which means that it is advisable to continue the research work started.

References

1. Szmigierska B, Wolski K, Jaszczak A (2012) Modele wyjaśniające zachowania użytkowników internetu. *E-mentor* 3(45)
2. Bobek S, Sternad S (2012) End user's knowledge issues in ERP solutions use. In: Studies & proceedings of polish association for knowledge management, no. 58
3. Przeclewski T (2011) Akceptacja oprogramowania open source. Metody i modele informatyki ekonomicznej, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk, pp 88–102
4. Switek S, Drelichowski L, Polkowski Z (2020) Empirical evaluation of the revised technology acceptance model for lean six sigma approach—a pilot study. In: This paper is currently in process of publishing at innovations in information and communication technologies—IICT-2020
5. Venkatesh V, Davis FD (1996) A model of the antecedents of perceived ease of use: development and test., *Decis Sci* 27(3):451–481
6. Park SY (2009) An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. *Educ Technol Soc* 12(3):150–162
7. Alharbi S, Drew S (2014) Using the technology acceptance model in understanding academics' behavioural intention to use learning management systems. *Int J Adv Comput Sci Appl* 5(1):143–155
8. Nevitt J, Hancock GR (2001) Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Struct Equ Model* 8(3):353–377
9. Ajzen I, Fishbein M (1980) Understanding attitudes and predicting social behavior. Prentice Hall, Englewood Cliffs, NJ
10. Davis FD (1993) User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *Int J Man Mach Stud* 38:479
11. Czakon W (2016) Podstawy metodologii badań w naukach o zarządzaniu. Wydawnictwo Nieoczywiste, Piaseczno, p 315–316.
12. Lyon DW, Lumpkin GT, Dess GG (2000) Enhancing entrepreneurial orientation research: operationalizing and measuring a key strategic decision making process. *J Manag* 26(5):1055–1085
13. Engel K, Moosbrugger H (2003) Evaluating the fit of structural equation models: test of significance and descriptive goodness of fit measures. *Methods Psychol Res Online* 8(2):23–74
14. Bollen KA, Stine RA (1993) Bootstrapping goodness of fit measures in structural equation models. In: Bollen KA, Long LS (eds) Testing structural equation models. Sage, Newbury Park, pp 111–135
15. Nevitt J, Hancock GR Performance of bootstrapping approaches, pp 353–377
16. Bedyńska S, Książek M (2012) Statystyczny drogowskaz 3—praktyczny przewodnik wykorzystania modeli regresji oraz równań strukturalnych. Wydawnictwo Naukowe PWN, Warszawa, p 225
17. Hopper D, Coughlan J, Mullen R (2008) Structural equation modelling: guidelines for determining model fit. *Electron J Bus Res Methods* 6(1):53–60

Recognition of Compound Characters from Degraded Kannada Documents



T. N. Sridevi and Lalitha Rangarajan

Abstract Mapping to editable fonts in machine interpretable codes such as ASCII/ISO/Unicode of the recognized or labeled images by classifier is one of the challenging tasks and comes under post-processing stage of optical character recognition protocol. Proposed here are sequences of steps that takes any Kannada compound character retrieved from a degraded document and produces editable text of the character images. The key steps of the proposed method are simple preprocessing of degraded documents, segmentation of characters, extraction and ordering of connected components, and mapping to Unicode based on correlation analysis of the character components, and finally, editable texts of characters are obtained. Datasets employed for experimentation include 1866 characters and components in the training data, and 1651 simple, 848 multi-, and 105 complex compound characters. The accuracy of the ordering based on correlation analysis is around 82.55% for simple compound characters, 61.43% for multi-compound characters, and 57.14% for complex compound characters.

Keywords Recognition of degraded Kannada characters · Old Kannada documents · Optical character recognition (OCR) · Segmentation · Compound characters · Unicode · Template matching

1 Introduction

Optical character recognition is automatic or electronic adaptation of images of handwritten or printed text (scanned document/photo of a document) into machine-encoded text. Offline handwriting recognition includes the automatic translation of text in the form of image to letter codes which are functioning in computer and text-processing presentations. Recognition of compound Kannada characters is one of the most challenging tasks in the process of scan converting a character image to editable Unicode format.

T. N. Sridevi (✉) · L. Rangarajan

Department of Studies in Computer Science, University of Mysore, Mysuru, Karnataka, India

Compound characters are broadly categorized into simple, multi- and complex compound characters. A consonant is always a base character and that can be combined with vowel modifier or consonant modifier or a combination of both. A base character combined in conjunction with a vowel forms a simple compound character. A base character combined with one or more vowels or consonant modifiers is multi-compound. Complex compound characters are formed as a result of combining two or more consonants or vowels with base characters.

The proposed method works on degraded documents that have undergone some preprocessing as discussed in [1].

Figure 1 shows some samples of simple, multi-, and complex compound characters extracted from degraded documents.

Processing of compound characters involves interpretation of character and its composition in details. The analysis of character components is possible only when components are detected. Various steps in compound character processing include preprocessing, connected component analysis (CCA), detection of components and recognition, ordering of character components, and mapping of Unicode.

From the existing literature, it may be inferred that there is scope for processing of complex characters in Kannada script, even when the characters are not degraded. There are several reasons for the complexity involved in this process, such as presence of multiple consonants and vowel conjuncts along with base character. In the proposed work, characters assumed as inputs are extracted from degraded documents. Some examples of degraded documents are aged documents and documents

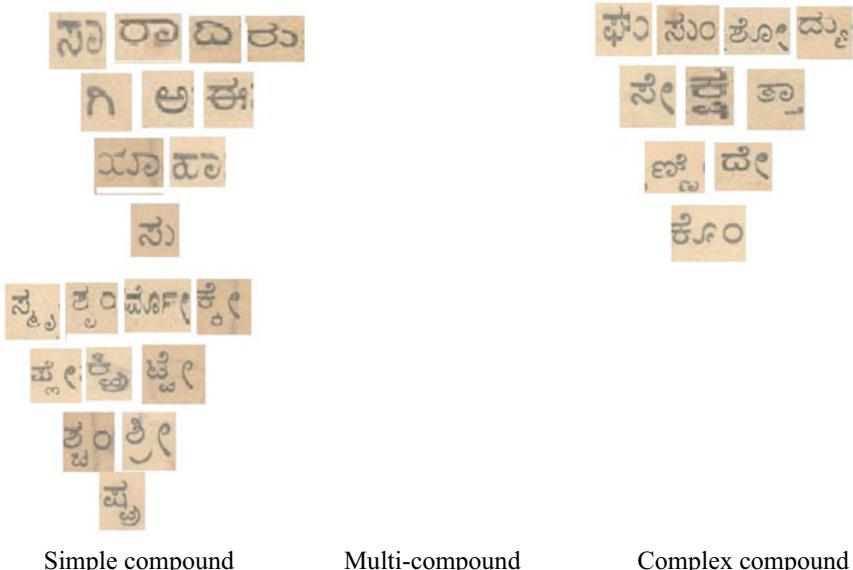


Fig. 1 Compound Kannada characters from degraded document

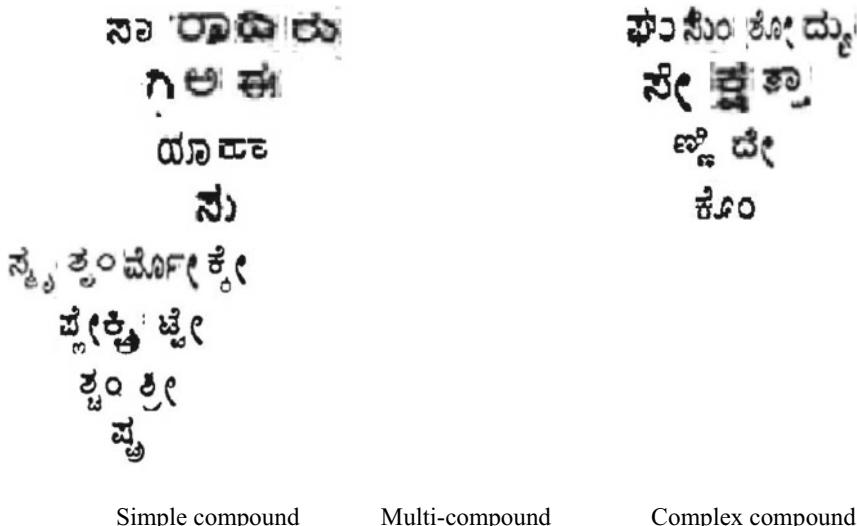


Fig. 2 Preprocessed compound characters in Kannada script

with marks and annotations. Types of noises present in the degraded documents are foxing effect, stain marks, aging marks, annotations, pale yellow shades and show through effect, etc. which is as shown in Fig. 2. The preprocessing procedure of the degraded documents is addressed in our work [1]. The characters are extracted from these documents, and a database is created to perform the recognition of Kannada characters. Character set is extracted from the degraded documents using projection profiles and built as a knowledge base. Some characters are split characters and are segmented manually. The resultant character set built from the degraded documents is consisting of characteristics such as breakages and merges even after preprocessing using binary image analysis [1]. The reasons for breakages and merges include excessive degradations in the document; therefore, in this work, a technique that possibly perform post-processing of simple, multi-, and complex compound characters with variety of degradations is addressed.

Plenty of research has been reported in the literature in the area of printed character recognition for Kannada, and the results are satisfactory for printed fonts or text particularly if they are not degraded. Adapting the behavior of these algorithms to fit the requirements of degraded printed fonts particularly for aged Kannada documents requires some remodeling combined with heuristic rules. Some of the existing works reported for printed character recognition of non-degraded documents are discussed here.

OCR system based on skeleton matching for portions of text is proposed by Li [2], wherein conventional and some novel methods are combined. Here, they have

employed thinning algorithm followed by template matching strategy for classification of characters. Implementation strategy involves the use of two stages of classification, first character dimensions-based broad classification followed by fine classification using similarity values. Subsequently, the method employs post-processing step based on spell check. Template matching based on compressed textual template images for pattern comparison is proposed by Inglis and Witten [3]. Work by Hogervorst et al. [4] employs a neural network. The method proposed works on electronic images of documents. Storage of characters happens subsequent to interpretation of characters. Many data entry systems can be replaced by this automatic generation of data. A six-step algorithm for Bangla character recognition is discussed in Pal et al. [5]. The sequence of steps in the suggested algorithm are: flatbed scanning of document, skew correction, text graphics separation, line segmentation, zone detection, word and character segmentation using available and proposed novel techniques. Adaptive OCR that works by extracting models for specific documents is proposed by Xu and Nagy [6]. New algorithms devised can approximate character widths and locations of characters within words. Training samples of text images are generated automatically. This is tolerant to transcription errors, and hence, an automatic transcript produced by imperfect omni font OCR system can be used.

An OCR template matching to recognize characters that works on non-degraded printed documents proposed by Muda et al. [7] has been claimed to produce good accuracy and comparable to some contemporary literature. Research by Acharya et al. [8] addresses recognition of printed characters after the preprocessing steps of noise removal, segmentation of lines, words, character components. A single Euler number feature seem is used by Dhanda et al. [9] for Kannada numeral classification. The proposed algorithm works for multi-font multi-size characters and is free from thinning, size normalization. Classification method is nearest neighbor with Euclidean distance. The work in [9] is followed by another contribution Dhanda et al. [10] which is designed to take restricted set of handwritten Kannada and English characters as input for recognition algorithm. The recognition system developed addresses the recognition of Kannada numerals and vowels, English uppercase letters. Features extracted are directional spatial features such as stroke density, stroke length, and the number of strokes. Results of experimentation using KNN classifier and four-fold cross validation demonstrate the success of the proposed method. However, the work does not address the complex and compound characters of Kannada language. Yet, another research by Dhanda et al. [11] is about recognition of Kannada/Arabic numerals using zone features. Pradeep et al. [12] have focused on neural network algorithm for handwritten character recognition. The algorithm proposed does not require the step of feature extraction. Rasheed et al. The [13] is a research by which is specifically designed to recognize characters on number plates of standardized number plates of Islamabad. The method devised uses Hough lines and template matching. An extensive survey of character recognition algorithms using a variety of feature extraction and classification methods is available in Dedgaonkar et al.'s [14] work. Choudhary et al. [15] have discussed offline English handwritten character

recognition method that employs features extracted from binarized images. A multi-layered feedforward ANN does the recognition. Some preprocessing and normalization steps performed are background, foreground noise removal, cropping, and size normalization. The authors have claimed the method works well particularly on offline cursive handwritten characters.

Performance of template matching for typewritten and handwritten character recognition is analyzed by Kumar and Sharma [16]. Two parameters, namely accuracy and execution time, are taken to study performance. It is to be noted that the authors have not considered compound characters in this work. Usage of DWT and template matching is explored in Sarungbam et al. [17]. DWT is chosen as the transformation is invariant to rotation, scaling, and translation. The authors have restricted the experiment to frequently occurring characters. Neural networks are designed to recognize broken Kannada characters by Sandhya and Krishnan [18]. However, they have reported lot of challenges due to breakage and complicated shapes of the characters, despite using an algorithm devised to fill breakages using end points. Choudhary et al. [19] have explored handwritten English character recognition through a multi-layer feedforward neural network. The training data are of size 1300 with 50 instances of each character.

It is noticed from literature survey that most of the works are reported are on printed Kannada characters extracted from non-degraded text rather than degraded text. Also, the recognition tasks cover mostly printed numerals and non-compound characters. Thus, it is important to carry out an investigation in the area of printed compound character recognition from degraded documents and also the mapping of Unicodes for the characters which are subjected to excessive degradations.

The rest of the paper is organized in three sections. Sections 2 and 3 cover the details of proposed method and experimentation. Section 4 concludes the paper.

2 Methodology

Recognition of Kannada characters is carried out in three phases in this proposal. They are (i) preprocessing, (ii) segmentation and ordering, and (iii) Unicode mappings. Figure 3 shows the phasewise block diagram of the proposed methodology (Fig. 4).

2.1 Preprocessing

Initially, the character image is assumed as input to the system, where inputs can be acquired in the forms of simple compound or multiple compound or complex compound characters. The characters acquired for processing are initially subject to Otsu's thresholding for conversion to binary form. Followed by which a morphological erosion with line structuring element as line with length and degree are assumed to be 1 and 0, respectively, to uncover the degradations such as breakages within the

Fig. 3 Sample of Kannada degraded document

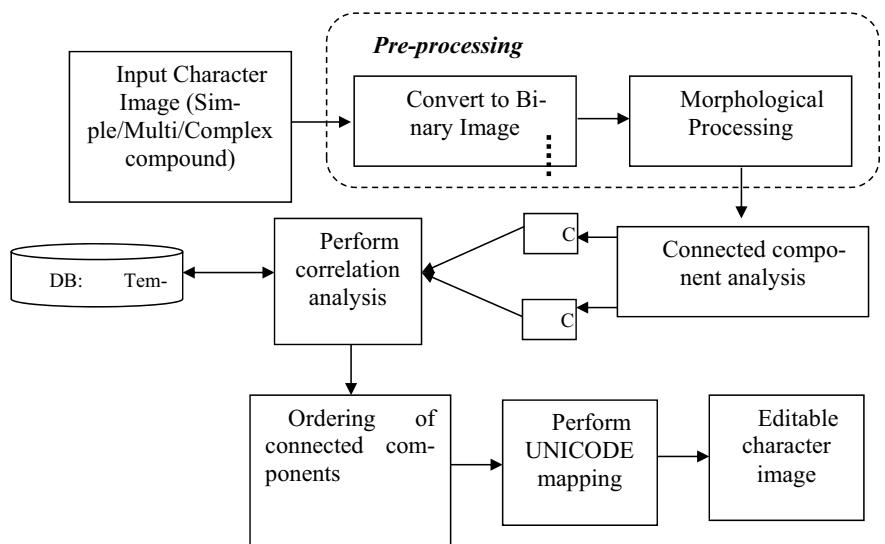
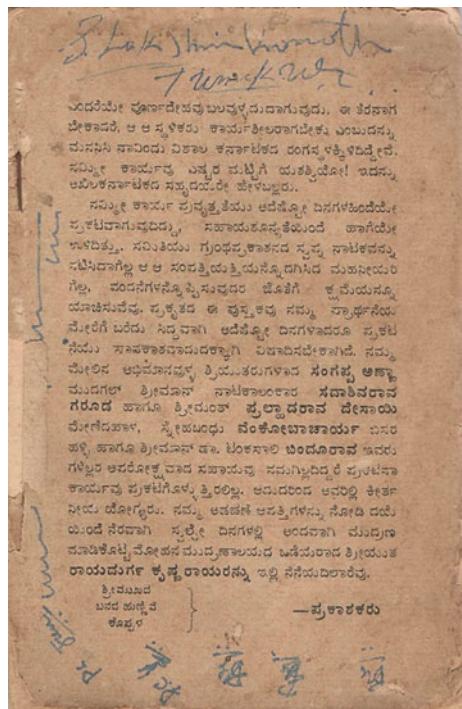


Fig. 4 Block diagram of processing compound characters

character. Further, the preprocessed character is forwarded to character component analysis.

2.2 Character Component Analysis and Unicode Assignment

Consider a character C with n components $C_1, C_2, C_3 \dots C_n$ which is obtained as an outcome of connected component analysis of C . For a C_i , an arbitrary component of C correlation with t_l (a template in the database) is determined.

If t_m represents the template that returned the maximum correlation with character component C_i , then the Unicode model returned for template t_m is returned.

Similarly, each component C_i of C will be processed, and the highest correlation template and the corresponding Unicodes are identified. Unicodes are assigned to components in the order of labeling of components.

2.3 Challenges Encountered in Processing Compound Characters

CCA performs the scanning of characters row-wise from top to bottom.

It is to be noted that in the proposed model, we come across few challenges in processing of compound characters when subjected to CCA, particularly with degraded characters. It is noticed that the compound characters possess merges and splits within a character due to high degradations. Some incorrect segmentations of CCA are shown in Fig. 5. The characters are broken into more components than what is present (Fig. 5a). The order of components in the output is incorrect (Fig. 5b). The components are not split (Fig. 5b). Further, CCA does not correctly perform the

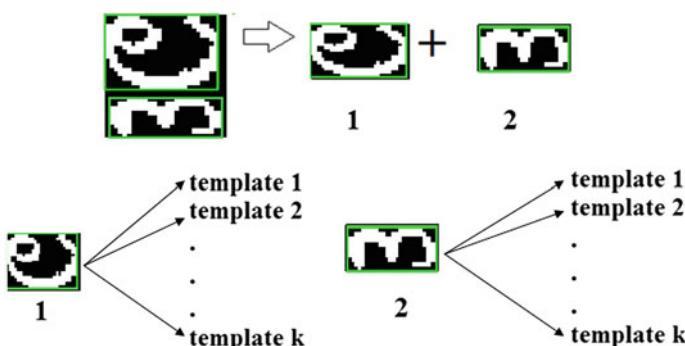
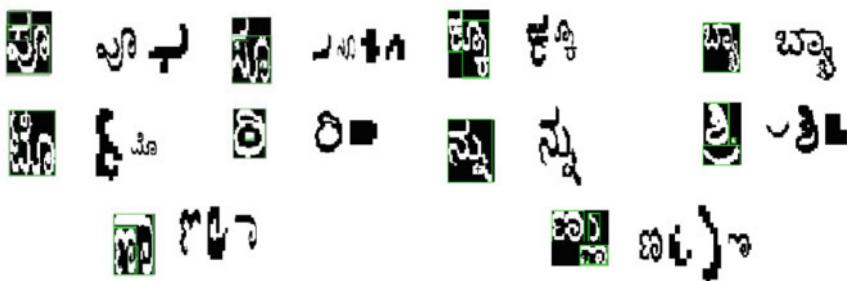


Fig. 5 Correlation analysis of components 1 and 2 with template set



a) Samples of Simple compound characters b) samples of multi compound characters

Fig. 6 Results of connected component analysis. **a** Samples of simple compound characters
b Samples of multi-compound characters

component identification if the separation between components is not clear in the preprocessed image (Fig. 6).

The proposed work does not address the complex compound segmentation problems. The reason being number of such characters in the degraded documents we have collected is very few.

In the proposed approach, order provided by CCA is retained as the actual ordering to be adapted to carry out the Unicode mappings.

2.4 Unicode Mappings

In this work, focus is mainly on the recognition and Unicode mappings of simple/multi- and few available complex compound characters. Once the template with maximum correlation measure is being identified, Unicode of corresponding components will be interpreted by our algorithm based on the ordering provided by CCA. Figure 7 shows the instances of compound characters and mapping of their corresponding Unicode (Fig. 8).

Fig. 7 Merges within complex compound character components

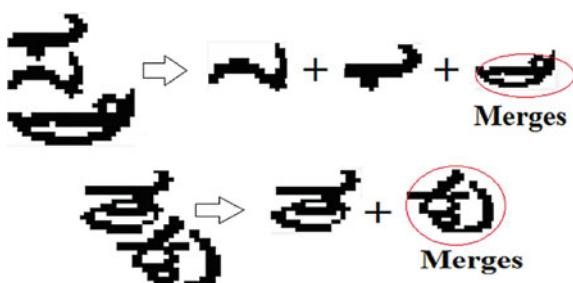


Fig. 8 Procedure of Unicode mapping

$$\begin{aligned}
 \text{ଦୟ} &= \text{ଦ} + \text{ୟ} + \text{ଦ} \\
 \text{ବ୍ୟ} &= \text{ବ} + \text{ୟ} + \text{ବ} + \text{ୟ} \\
 \text{ଶ୍ୟ} &= \text{ଶ} + \text{ୟ} + \text{ଶ} + \text{ୟ} \\
 \text{ଷ୍ୟ} &= \text{ଷ} + \text{ୟ} + \text{ଷ} \\
 \text{ତ୍ୟ} &= \text{ତ} + \text{ୟ} + \text{କ} + \text{ୟ} \\
 \text{ହ୍ୟ} &= \text{ହ} + \text{ୟ} + \text{କ} + \text{ୟ} + \text{ର} + \text{ୟ} \\
 \text{ନ୍ୟ} &= \text{ନ} + \text{ୟ} + \text{ତ} + \text{ୟ} + \text{ର} + \text{ୟ}
 \end{aligned}$$

3 Experimental Analysis

Datasets for experimentation are characters extracted from preprocessed documents similar to the ones shown in Fig. 1b of this paper. A total of 4470 characters are collected for experimentation, out of which 2604 characters are considered for testing with 1651 simple compound characters, 848 multi-compound characters, and 105 complex compound characters. Simple compound characters have one component, and other compound characters have two or more components. Recognition of segmented character component is done through computation of correlation coefficients of the input component with all templates in the database, and the component with highest correlation coefficient is returned. Number of templates in database being used is 1866. Accuracy of template matching technique toward recognition of segmented character components is defined as number of characters recognized correctly to total number of characters for which correct Unicode mappings generated. Table 1 shows the details of datasets for experimentation process. Table 2 shows the correct order of character components. Table 3 shows the performance of template matching technique.

Table 4 shows the accuracy of online OCR.

From Table 4, it is observed that the efficiency of the proposed sequence of steps outperforms the efficiency of currently available online OCR (<https://www.newocr.com/>) when tested across 10 characters each class from simple, multi-, and complex compound characters.

Table 1 Training and testing data

Size of testing data		Size of training data
Single component characters	Characters with more than one component	
No. of simple compound characters	No. of multi-compound characters	1866
1651	848	
		105

Table 2 Correct order of character components

Type of character	Character instance	No. of character components	Ordering as per CCA	Output of correct/desired processed order
Simple compound		1		
Multi-compound		2		
Complex compound		3		

Table 3 Performance of template matching

Type of compound characters	Total no. of characters	No. of characters recognized correctly	Accuracy in %
Simple compound character	1651	1363	82.55
Multi-compound character	848	521	61.43
Complex compound character	105	60	57.14

Table 4 Accuracy of online OCR

Sl. no.	Type of character	Accuracy of online OCR with 10 character samples (%)
1	Simple compound	20
2	Multi-compound	20
3	Complex compound	0

4 Conclusion

In this work, primary focus is to develop a model for segmentation and Unicode mappings for degraded complex Kannada characters. Degradations existing within a character increase the number of components. Connected component analysis is carried out to segment the characters, and template matching is employed to perform approximate shape matching, and the template with maximum correlation is being used to perform componentwise Unicode mappings.

A limitation of the proposed work is that complex compound characters with their components being merged due to dilations in the document and splits within the connectivity of components of a particular character have not been addressed. A generic solution for corrections of such characters requires a lot of additional work in the segmentation stage, and this task is planned to be our next research work.

References

1. Sridevi TN, Rangarajan L (2019) Binary image analysis technique for preprocessing of excessively dilated characters in aged Kannada. *Int J Recent Technol Eng* 8(4)
2. Li N (1991) An implementation of OCR system based on skeleton matching
3. Inglis S, Witten IH (1994) Compression-based template matching. In: Proceedings of data compression conference DCC'94, IEEE, pp 106–115
4. Hogervorst ACR, Dijk MK, Verbakel PCM, Krijgsman C (1995) Handwritten character recognition using neural networks. In: Neural networks: artificial intelligence and industrial applications, Springer, London, pp 337–343
5. Pal U, Chaudhuri BB, Belaïd A (2006) A complete system for Bangla handwritten numeral recognition. *IETE J Res* 52(1):27–34
6. Xu Y, Nagy G (1999) Prototype extraction and adaptive OCR. *IEEE Trans Pattern Anal Mach Intell* 21(12):1280–1296
7. Muda N, Ismail NKN, Abu Bakar SA, Zain JM (2007) Optical character recognition by using template matching (alphabet). In: National conference on software engineering and computer systems 2007 (NACES 2007)
8. Acharya DU, Subbareddy NV, Makkithaya K (2008) Hierarchical recognition system for machine printed Kannada characters. *Int J Comput Sci Network Sec* 8(11):44–53
9. Dhanda BV, Benne RG, Hangarge M (2011) A single euler number feature for multi-font multi-size Kannada numeral recognition. ArXiv preprint [arXiv:1111.4290](https://arxiv.org/abs/1111.4290)
10. Dhanda BV, Hangarge M, Mukarambi G (2010) Spatial features for handwritten Kannada and English character recognition. *IJCA, Special Issue RTIPPR* 3:146–150
11. Dhanda BV, Mukarambi G, Hangarge M (2011) Kannada and English numeral recognition system. *Int J Comput Appl* 975:8887
12. Pradeep J, Srinivasan E, Himavathi S (2011) Neural network based handwritten character recognition system without feature extraction. In: 2011 international conference on computer, communication and electrical technology (ICCCET), IEEE, pp 40–44
13. Rasheed S, Naeem A, Ishaq O (2012) Automated number plate recognition using hough lines and template matching. In: Proceedings of the world congress on engineering and computer science, vol 1. pp 24–26
14. Dedgaonkar SG, Chandavale AA, Sapkal AM (2012) Survey of methods for character recognition. *Int J Eng Innovative Technol (IJEIT)* 1(5):180–189

15. Choudhary A, Rishi R, Ahlawat S (2013) Off-line handwritten character recognition using features extracted from binarization technique. *Aasri Proc* 4:306–312
16. Kumar S, Sharma P (2013) Offline handwritten and typewritten character recognition using template matching. *Int J Comput Sci Eng Technol* 4(06):818–825
17. Sarungbam JK, Kumar B, Choudhary A (2014) Script identification and language detection of 12 Indian languages using DWT and template matching of frequently occurring characters. In: 2014 5th international conference confluence the next generation information technology summit (Confluence), IEEE, pp 669–674
18. Sandhya N, Krishnan R (2016) Broken Kannada character recognition—a neural network based approach. In: 2016 international conference on electrical, electronics, and optimization techniques (ICEEOT), IEEE, pp 2047–2050
19. Choudhary A, Ahlawat S, Rishi R (2015) A neural approach to cursive handwritten character recognition using features extracted from binarization technique. In: Complex system modelling and control through intelligent soft computations, Springer, Cham, pp 745–771

Demand Forecasting and Design Thinking for a New Product Using Neural Networks and Generative Adversarial Networks



Shweta Upadhyaya, Prankul Kumar, and S. Ushasukhanya

Abstract The top trouble spot for business heads is demand instability. Such a large number of elements from climate vacillations to posts by web-based media influencers—sway purchasers—making them oftentimes adjust their perspectives. More terrible still, things reshape clients aim very suddenly. There is no magic wand to foresee situations that arise around us, for example, if any famous actor or actress scorns the use of make-up items from a particular brand then that brand might see a major fall in their sales just because a particular person was offended. However, there are advances to improve the exactness of demand forecasting. Truly, it will never be 100% exact, yet it very well may be exact enough to assist you with accomplishing your business objectives. Various companies come in with their products to the market every other day thinking that the digital market space will be cheaper for them; however, the digital market is clogged with problems of space, feature similarity, cost variation, quality, and uncertainty of demand. New products have little to no brand value which plays another major role in determining the demand for the product. This project will try to make a tool that will analyze the existing market data to identify the viability of the market segment before launch and then suggest required/additional features to improve the success of the product launch using design thinking creating massive ideation through GAN-generated alternatives.

Keywords Deep learning · Neural network · CNN · Generative adversarial network (GAN) · Demand forecasting · Market segment analysis

S. Upadhyaya · P. Kumar (✉) · S. Ushasukhanya

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

e-mail: pp7805@srmist.edu.in

S. Upadhyaya

e-mail: sa7406@srmist.edu.in

S. Ushasukhanya

e-mail: ushasukhanya.s@ktr.srmuniv.ac.in

1 Introduction

The market is full of similar products with little or no quality or brand value, in that case how do we launch a product which is both different and effective even if similar products exists in the market world. On top of that a wide array of tools are available to forecast the demand for new products in the market but almost have proven to be very uncertain when coping with huge amounts of data, the error percentage is usually very high when making such forecasts.

Demand forecasting is the evaluation of likely future interest for any item (product) or service. The term is generally used with demand planning, yet the latter is broader cycle that starts with deciding/determining, however, is not limited to it. Likewise, generative adversarial network (GAN) is an AI (ML) model in which two neural networks contend with one another to turn out to be more exact in their forecasts. GANs typically run unsupervised and use a cooperative zero-sum game framework to learn. In our project, we will be making use of deep learning instead of regression to take care of the availability of a large amount of data.

To solve all these issues, we will use neural networks and backpropagation to get more accurate and stable results. An algorithm of gradient descent which is Adagrad will be used which will give faster convergence. Design thinking is used in the ideation phase as well as GAN. Our use of GAN is to suggest design options, by making a computer to create several similar but not exactly the same designs.

Using GAN, we advise on product shape, size, and color. For this purpose, we click multiple pictures of the product from various perspectives, and then, GAN will make minor distortions and merge the pictures to create a new product. We can create thousands of such designs in a short span of time which allows us a variety of options to choose from, and it is possible that a few of the new designs will be better than the existing design which fulfills our purpose of creating a different and effective product.

2 Literature Review

The digital market is very unpredictable, and many players come with their new products to the market, thinking that the digital market space will be cheaper for them; however, the digital market space is having two problems—the market is overcrowded with several products in every segment with a wide variety of cost, quality, and features, and any new product from an entrepreneur is launched with little existing reputation and brand value so, for such a product, it is very important to analyze the market data in digital space before the actual launch. A lot of research has been conducted on-demand forecasting and GAN in various industries. These researches are mostly focused on finding segmentation in the market, and demand forecasting is based on *K*-means clustering which performs poorly with large amounts of data. Moreover, the impact of the bullwhip effect on the supply chain has not

been taken into consideration by the existing systems. The existing systems for design ideation produce less accurate results and are purely based on the learning of GANs. Generally, K -means clustering is used, and it is not good in differentiating advanced technological features. Lau et al. [1] demonstrated how stochastic factors in demands become the reason for the alarming bullwhip effect which affects the entire supply chain. They developed a mathematical approach by using an algorithm called the minimum description length (MDL) which helps in selecting the most optimal neural network, and they complimented this technique with the surrogate data method to figure out the characteristics in the demand and their nature. They were able to use these techniques for demand forecasting with a wide variety of real-world data. However, they still could not establish how the bullwhip effect impacts the entire supply chain and future research has been suggested on the same. Liuqing et al. [2] studied various artificial intelligence and data mining techniques to enhance design ideation. They came up with two models—the semantic ideation network and visual concept combination model. These models were executed separately, and the results were used to create deductions about whether or not design ideation had been enhanced or not. They used to step forward and path track algorithms to prepare these experiments and finally concluded that the semantic ideation network did not impact the design ideation but the visual concept combination model enhanced the design ideation by 80%. So it was concluded that GAN can create a major shift in the design ideation process.

Several empirical studies have focused on deals determining in industries, for example, materials and dress design, books, and gadgets. Notwithstanding, not many investigations have been made on-demand forecasting in the healthcare industrial sector which manufactures products for disabled people like wheelchairs, crutches, etc. Our study focuses on improving the aesthetics of a product-centered around helping differently abled people and forecasting the market segment for it.

Chenxi Yuan et al. proposed design automation to decrease the human efforts that are put into designing garments in the fashion industry using GANs. The fact that designers need a distinct mind-set and innovating thinking every time when they are designing something sometimes hinders the very process they are dedicated to. Their paper explores the potential of GANs. This paper demonstrated the use of attribute GAN to manipulate some specific attributes to alter the designs of the garments. The performance of this model was tested with a large fashion dataset. The conclusion drawn from these tests was that GANs work differently in different domains, and therefore, the algorithm should be customized based on a specific domain.

The use of neural networks in demand forecasting has proven to be one of the best combinations since neural networks have also been mathematically demonstrated to be universal approximates of functions, and they provide accurate mathematical results. Kochak and Sharma [3] demonstrated the importance of selecting proper forecasting techniques and concluded that “the learning algorithm of the prediction imposes to be a better prediction of time series in the future.” When the prediction performance of recurrent neural networks in simulated time series data and practical sales data was used, it was found that the influence of several factors on-demand function in the retail trading system resulted in the forecasting period

becoming smaller which shows that the ANN approach provides more accuracy in forecast [3]. Similarly, various other scholars like Zhu Ying et al. and Davide Mezzogori et al. proposed systems that used deep neural networks, entity embedding, and backpropagation algorithms to forecast the market segment.

Hilal Kilimci et al. [4] proposed a novel model to improve the interest determining measure which is one of the principal issues of supply chains. For this reason, they utilized nine diverse time arrangement strategies, uphold vector relapse calculation (SVR), and DL approach-based interest estimating model were developed. To get an official choice of these models for the proposed framework, nine distinctive time arrangement techniques, SVR calculation, and DL model are mixed by another joining system which is suggestive of boosting group procedure. “In this way, the final decision of the proposed system is based on the best algorithms of the week by gaining more weight which made their forecasts more reliable with respect to the trend changes and seasonality behaviors” [4]. At the same time, researchers have also concluded that the time series method is not a reliable way to demand forecast since it is a traditional statistical method, the dataset is very sparse, and a lot of contradictory evidence has been found in both regards.

Few studies have also used big data platforms and various visualization techniques to perform market segment analysis on the available dataset [5]. One of the significant disadvantages of predictive analytics is that it is not the easiest method as it includes complex AI calculations. Moreover, it is designed to create estimates for at least a month out and is inappropriate and not intended to envision the closer future. Nonetheless, our work will be a step in the direction of using demand forecasting and design ideation for various sectors using deep neural networks and generative adversarial networks.

3 Proposed Methodology

The proposed system for demand forecasting is based on neural networks which can perform better with large amounts of data which is an important factor as due to data explosion, we need to process more data to get better inference these days. We will be using multilayer perceptron for creating the classifying model for market segment analysis and convolution neural network for GAN. The proposed system for design ideation will be utilizing the design thinking process along with the learning of the general adversarial networks to produce better and accurate results. We shall be dividing the project into these modules:-

1. Collecting the data and pictures for market segment analysis and GAN.
2. Developing a neural network—A multilayer perceptron and generative adversarial network.
3. Training, validating, and testing with test data.
4. Design thinking for analyzing and choosing the best output images (manually).

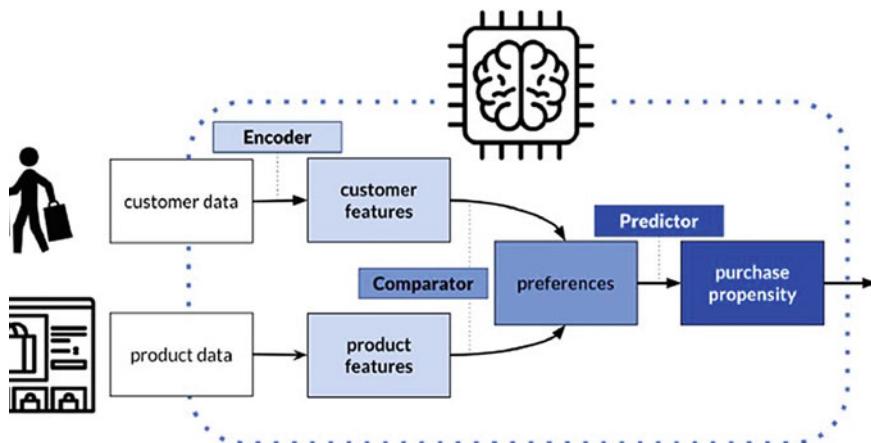
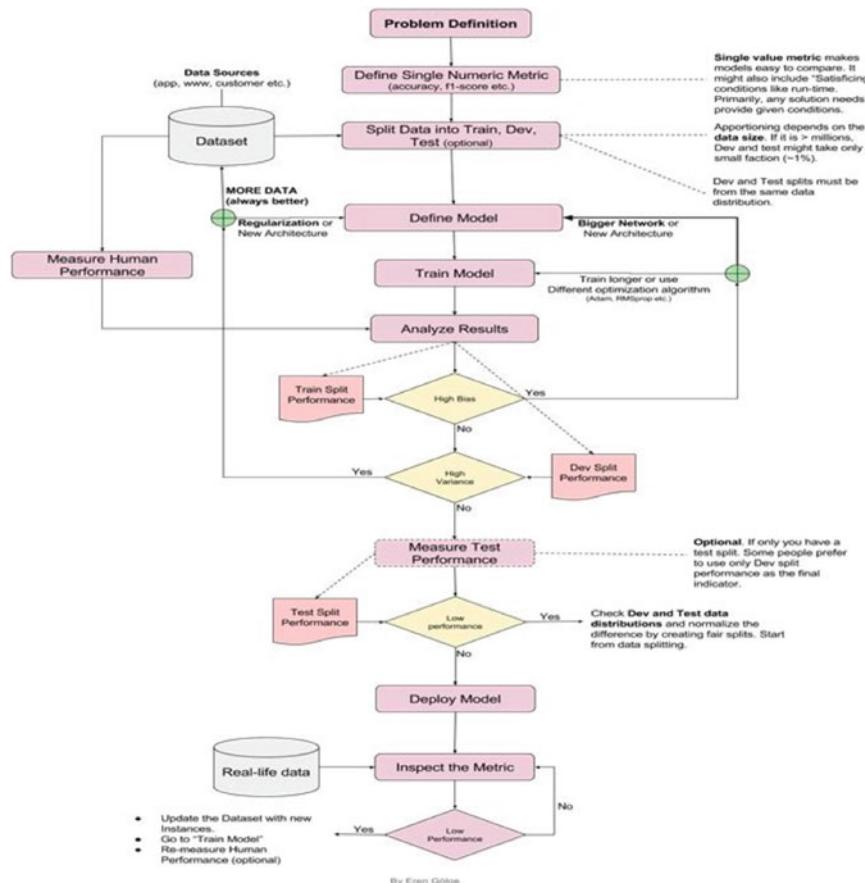


Fig. 1 Architecture diagram for demand forecasting

We create a database of pictures of the product by clicking the photos of the product from various perspectives, which is followed by the creation of a neural network for GAN to reduce the error. After GAN is created and functional, we utilize the software to develop various images of the product with minor distortions (the distortions depend upon probability) and read multiple new images of the same product. We then utilize design thinking and manually choose the image of the product which gives a new and better design for the product (in context with aesthetics). We use web scraping to collect the data and utilize that to create a neural network that can handle a huge amount of data and then predict the demand for the product (market segment analysis) (Fig. 1).



4 Comparative Analysis

Subsequent to perusing all these research papers we gathered that the information accessible is extremely high as are the boundaries for making the investigation which makes regression a complicated instrument, it was discovered that on multiple occasions because of regression underfitting of the ends were occurring while with neural networks a typical issue discovered was overfitting. Since neural network needs a huge database, it was discovered valuable for demand forecasting. A neural network can likewise be utilized for cross-arranging, for example, with the assistance of administered learning, it can perform examination for different diverse informational indexes, after gaining from one though relapse could not matter investigation gained starting with one informational collection then onto the next.

We additionally comprehended that regression by and large requires a visual model for determining yet since there are an excessive number of elements/boundaries into play, and it makes a hyperplane which thus entangles the entire cycle and again causes the issue of underfitting. Though this is a reward for neural networks, the odds of overfitting with neural networks are low which cements our target for utilizing neural networks in this venture. Other factors that went over are despite the fact that time series forecasting is a generally utilized strategy; it is not extremely valuable since the informational index is exceptionally scanty for it. We sorted out that computational creativity theory can be utilized for the design thinking part about our undertaking, and the incitement for the design thinking segment must be visual {image based}. The literature surveys us to comprehend the informational index and how we can sort out our intended interest group.

5 Conclusion

Most of the studies on-demand forecasting demonstrated the use of time series methods which do not work well with sparse datasets. Other systems for finding segmentation in the market and demand forecasting used K -means clustering which performs poorly with large amounts of data. The existing system for design ideation produces less accurate results and is purely based on the learning of GANs. This paper proposes the use of deep neural networks for demand forecasting which can perform better with large chunks of data and produce accurate predictions along with the use of the design thinking principles for design ideation which will be accomplished by the optimized learning of the GANs. Nonetheless, our work will be a step in the direction of using demand forecasting and design ideation for the healthcare sector using deep neural networks and GANs.

References

1. Lau HCW, Ho GTS, Zhao Y (2013) A demand forecast model using a combination of surrogate data analysis and optimal neural network approach. *Decis Support Syst* 54(3):1404–1416. <https://doi.org/10.1016/j.dss.2012.12.008>
2. Chen L, Wang P, Dong H, Shi F, Han J, Guo Y, Childs PRN, Xiao J, Wu C (2019) An artificial intelligence based data-driven approach for design ideation. *J Vis Commun Image Rep* 61:10–22. ISSN 1047-3203. <https://doi.org/10.1016/j.jvcir.2019.02.009>
3. Kochak A, Sharma S (2015) Demand forecasting using neural network for supply chain management. *Int J Mech Eng Robot Res* 4(1):96–104

4. Hilal Kilimci Z, Okay Akyuz A, Uysal M, Akyokus S, Ozan Uysal M, Atak Bulbul B, Ali Ekmis M (2019) An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. Complexity 2019. Article ID 9067367, 15 p. <https://doi.org/10.1155/2019/9067367>
5. BeschiRaja J, Pamina J, Madhavan P, Sampath Kumar A (2018) Market behavior analysis using descriptive approach. Int J Pure Appl Math 118(7):171–175. ISSN: 1311-8080 (printed version). ISSN: 1314-3395 (on-line version). URL: <http://www.ijpam.eu>. Special Issue

An Efficient Method of Predicting the Average Fuel Consumption in Automobiles Using Ensemble Stacking in Python



S. Anandamurugan, R. Deenadhayalan, B. Venkatesan, S. Sakthivel, and S. Rajesh

Abstract In this article, we estimated the mpg of the car by stacking the whole. The ensemble technique is used to approximate the mpg and increases the predictive efficiency since it requires multiple learning machines. The stacked regressors used are XGBoost, LightGBM, the Root Mid-Mean Squared Logarithmic Error (RMSLE) function for the purposes of compute the relation between the values predictable for the auto-learning model given by this model and the actual target value, that is, the mpg of the automatic learning model.

Keywords mpg prediction · Regression techniques · Ensemble stacking · Stacked regressors · LightGBM · XGBoost · Automobile · Predictive performance · Fleet management · Fuel consumption

1 Introduction

Today, fuel use in the automotive market and the fleet organisation has been one of the greatest obstacles. Think of an interurban vehicle operating in a fleet between a big city and a rural town. During this phase, the vehicle will travel through heavy traffic and various environments such as hill stations, highways, muddy areas, etc. During this time, in addition, depending on the day of the week, traffic strength, weather conditions and car loads can differ. Every driver will have their way of driving the vehicle. Fuel efficiency of vehicles can be affected by careless maintenance and improper driving habits such as time duration of cruise mode, number of breaks applied, acceleration speed by the driver, etc. There is no existing reliable machine learning technique for predicting fuel consumption in automobile vehicles effectively.

S. Anandamurugan
Kongu Engineering College, Erode, Tamil Nadu, India

R. Deenadhayalan (✉) · B. Venkatesan · S. Sakthivel · S. Rajesh
Paavai Engineering College, Namakkal, Tamil Nadu, India

For controlling and maintaining the fuel consumption rate on average, a machine learning technique is introduced in this paper. In this paper, ensemble stacking, a machine learning theory has been developed which merges the machine learning models into one predictive model and gives more accurate results higher than the expectation of the machine learning models alone. Our proposed approach to evaluate the fuel consumption in automobile vehicles using the following predictors listed below:

- This paper is developed with the vehicle data that is acquired from the UCI repository.
- A range of cylinder, horsepower, displacement, acceleration and model year are used in the dataset. Fuel intake in mpg is the performance vector received.
- **Displacement:** It is the volume of the engine of the vehicle, usually in litres or cubic centimetres.
- **Origin:** It is a discrete value from 1 to 3. This dataset is not clear about this so we can assume these numbers as countries.
- **Model year:** The four-digit year is indicated as a decimal number containing the last two digits. (For example, 1970 is the year of model = 70).
- **Weight:** The weight of the car is mentioned as total vehicle mass, including the passengers and luggage.
- **Acceleration:** Acceleration is the rate of change of velocity as a function of time. Here it is modified into a decimal value.
- **Methods:** Kernel Ridge Regression, Gradient Boosting Regressions, XGBoost, LightGBM, LASSO regression, Elastic net regression.
- Our model dataset is from a wide variety of vehicles which can provide us with a reasonable approximation of our undisclosed vehicles mpg. The study is evaluating machine learning approaches from a set of refined predictors to fuel consumption in gallons per litres.

2 Literature Review

A. Cappiello, in his paper, developed a model, which is instantaneous statistical emission of gases such as CO, HC, NO_x and the fuel consumption in the automobiles based on physical weight of the car. They collected data in three different ways as regional-based, facility-based transportation and microscopic traffic which in term is called as CMEM model. The CMEM model has three levels where the driving cycle data are collected as macro-level, meso-level and micro-level, whereas in macro-level they will collect data for every hour and in meso-level the data is collected for every minute and in micro the data is collected for every second. They performed the model on aggressive data cycle and tested on another driving cycle to estimate its capabilities. Their model helps in the Intelligent Transportation System where traffic conditions, vehicle emissions can be controlled [1].

Geoffrey I. Webb and ZijianZheng, in their paper proposed that ensemble learning strategies have improved the accuracy of base learning algorithms. They followed

the basic principles of boosting and bagging to improve the efficiency. They have stated that the diversity of algorithms using in the ensemble model may increase the test error of the individual members in the ensemble, and it is hard to measure those. However, the success of this ensemble depends on the how the ensemble techniques are equally distributed. So they have succeeded in reducing the error in the internal models in ensemble by working with Wag algorithm, Sasc algorithm, MB algorithm and adaboost algorithms. They have demonstrated the hierarchy between internal errors, bias and variance of the different algorithms chosen [2].

Sandareka Wickramanayake, H. M. N. Dilum Bandara, in their paper, developed a comparative study, on the fuel consumption of fleet vehicles using three different machine learning algorithms, namely Random Forest, Neural Networks and Gradient Boosting. They state the analysis of fuel consumption in fleet vehicles will reduce the fraudulences. In this study, they have chosen the public bus data in time series. The challenge they face is that the dataset available is so less and yet they need to provide a comprehensive result on their paper. They say that the fuel consumption may be affected by both the internal and external factors. So, they studied the pattern in the data and performed an analysis and proved that the random forest algorithm predicted more accurately rather than the two others [3].

S. Sailaja, in her paper, suggested an ensemble stacking model by combining the unsupervised model and supervised model. They have used the diabetes dataset. Before developing the model, they pre-processed their dataset by clearing the null values and outliers. They have studied the inter-relationship between the data. K -means as the unsupervised model performed the elbow method to find the number of clusters in their data. KNN as the supervised model stacked along with logistic regression acted as the meta-classifiers and K -means as the base classifier performed the ensemble stacking and found the result that the stacked prediction's accuracy is higher than the individual models [4].

We are experiencing a fascinating age in which human intelligence is augmented by computer intelligence by data analytics artificial intelligence (AI) and machine learning (ML). However, latest research indicates that personally trained DNN models are susceptible to serious inputs [5].

Remotely sensed techniques Modified methods for calculating subterranean biomass (live organs) have been explored, but limited experiments have used remotely sensed tools to estimate the dry weight of subterranean dead organs. For this reason, a method of assimilation of the leaf area index (LAI) derived from the radiative model to the WOrldFOodSTudies (WOFOST) model was introduced [6].

Rahul Goel, Dinesh Mohan, Sarath K. Guttikunda, Geetam Tiwari, in their paper proposed a new reliable method to collect the data to estimate the emission and fuel consumption in the fleet vehicles. In their method, they followed two ways, namely primary surveys where they collect data directly from the customers in the petrol bunk and secondary data resources available from already developed data. Their method provides fleet size, annual mpg and fuel efficiency of cars and bike in their three chosen city Delhi, Vishakhapatnam and Rajkot. They have distinguished and the studied the factors affecting the fuel consumption in the new vehicles and vehicles in fleet. They have explained that the number of vehicles owned in India is numerously

higher than the individual fleets in the country. Permanent motor synchronous magnet is a robust electric vehicle system that can produce maximum torque at low power starting and is difficult to control [7]. They included that more number of fleets should be developed and proper fuel efficiency methods are introduced, so that the fuel consumption rate of the vehicles can be minimized [8].

3 Proposed Methodology

Fleet management faces a lot of problem in dealing with driver settlement, resource sharing, vehicle maintenance, and lifespan costing. Most of these factors are out of their control, which results in accounting and predicting complications. In vehicle maintenance, the loss will be very high since the fleet management depends mainly on that. Therefore, the vehicle should give efficient mileage and it is needed to check that periodically if there is any variation in the mileage given by the vehicle. There are some mechanism or machine-learning model to predict the mpg of the vehicle but they have certain per cent error in their accuracy.

This proposed system focuses on predicting the mpg of the vehicle more precisely by the concept of ensemble stacking. In this model, the dataset is obtained from the UCI Repository is analysed and pre-processed for the better prediction of the result. Ensemble learning is a technique that combines some classification or regression for better accuracy. The base-level models are trained are provided as input to the meta-model. In this model, regressors are used as the base model and XGBoost, LightGBM as the meta-model. This model works on the weighted ensemble. A weighted ensemble is a model-averaging group where the output of each member's choice of value between 0 and 1 is determined by the participation of any member in the final forecast. The RMSLE function will provide a predicted value to the actual value by adding weight to those values in the algorithms chose, the final output value could be derived.

4 Data Analysis

4.1 Data Collection

This dataset is a changed version of the dataset imposed within the StatLib library. This dataset is found in the UCI Repository with eight attributes. Fuel consumption in mpg is projected in terms of three dissimilar multivalued and five continuous attributes.

4.2 Attribute Information

Attributes	Types
MPG	Continuous
Cylinders	Multi-valued discrete
Weight	Continuous
Horsepower	Continuous
Displacement	Continuous
Model year	Multi-valued discrete
Acceleration	Continuous
Origin	Multi-valued discrete
Car name	String

4.3 Data Preprocessing

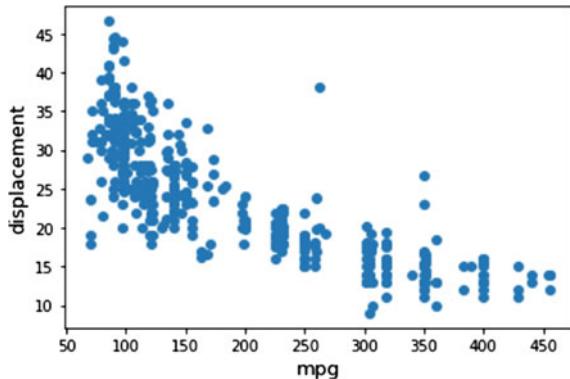
The goal of preprocessing data is to reduce possible mistakes in this model. In general, the model is just as strong as the knowledge it produces and we aim to ensure that the model results on the data set as correctly as possible during data preprocessing. While we cannot clean the data set completely, we should at least take some fundamental measures to ensure that our data set has the greatest potential to build a successful model.

4.4 Data Cleaning

The purpose of data cleaning is that some dataset may contain a null value or unknown value such as symbols and without clearing those, the model does not predict well. Because the machine cannot understand those. In our dataset, the horsepower variable has ‘?’ in some of their rows and those valued rows are removed before modelling.

4.5 Missing Data

The dataset should not possess any null value, and if any found, it should be deleted. In our data, we didn't find any such.

Fig. 1 Outliers

4.6 Outliers

Outliers, as the name says some data points in the dataset may be found far away from the observations and such points may give erroneous results in prediction so that needs to be removed. Figure 1 which is the auto-mpg dataset outlier's evaluation, we did not find any outliers in this so we need not want to remove any data points. But Outliers removal is not always safe. There square measure in all probability others outliers within the coaching knowledge. The elimination of them could, however, have a bad impact on our models if the test results still contain outliers. So we should only make any of our templates stable on them, rather than deleting them everywhere.

4.7 Target Variable Analysis

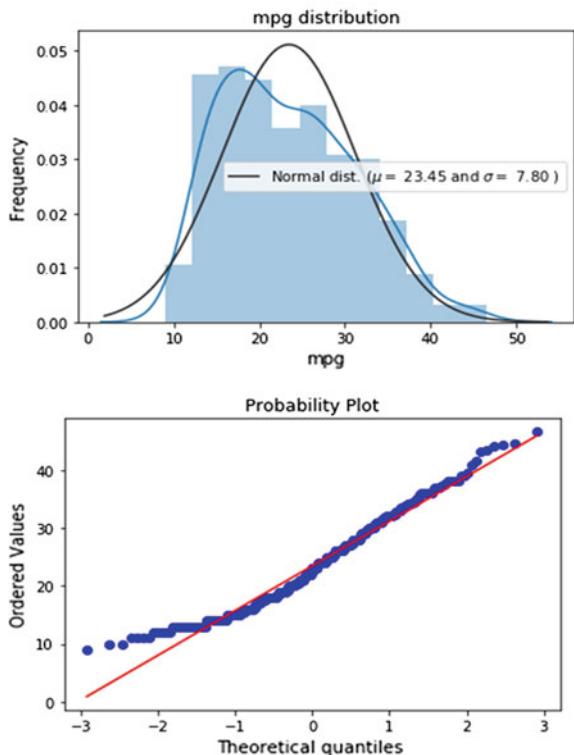
In this dataset, mpg is the target variable that we need to predict. So before modelling, we've performed an analysis on that variable. Skewedness in a symmetrical bell curve corresponds to a distortion in a collection of results. It is assumed to be bent whether the curve is rotated to the right or to the left. Figure 2 reveals mpg is right skewed and could mislead the outcome of the intended variable. Since (linear) models love the common knowledge, we must reformulate and distribute this variable more normally. The data is usually distributed and the skew has been corrected (Fig. 3).

4.8 Exploratory Data Analysis

EDA is a secondary approach, examining datasets to review their main characteristics, with visual strategies. It refers to the essential method of playing initial investigations on the information. It is used to discover patterns, to classify anomalies, to

Fig. 2 Right skewed

mu = 23.45 and sigma = 7.80



check postulates, to picture assumptions with the assistance of outline statistics and graphical representations.

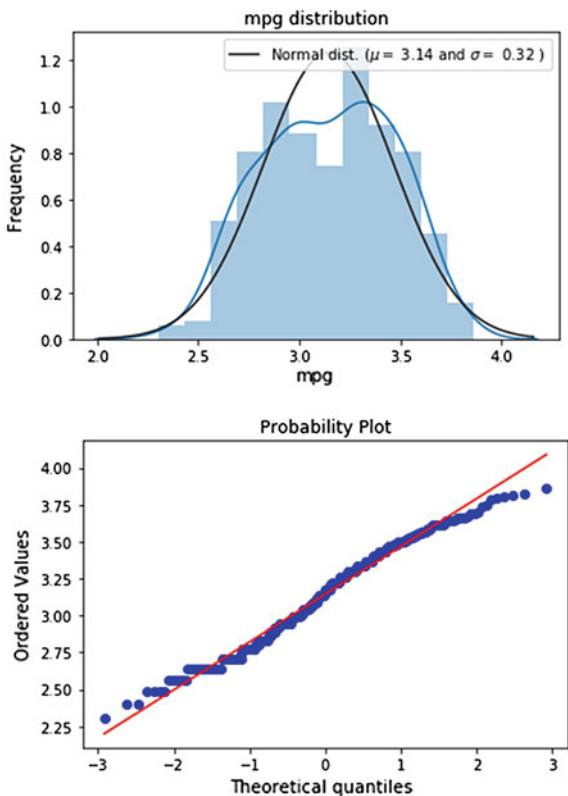
A matrix of correlation is a visual instrument that describes the interaction between the various columns. The feature selected for predictions is: mpg, cylinders, horse-power, weight, acceleration which should all be self-explanatory. Figure 4 represents the correlation between the variables in the dataset.

4.9 Feature Engineering

Feature architecture ensures that new alternatives from current data are typically generated through many associated tables. Feature engineering needs to retrieve the knowledge from the data in a single table from which you can practice your machine learning algorithm. In our dataset, we have removed the car name and selected remaining for the train and testing of models since the car name is not numeric and also not needed for the modelling.

Fig. 3 Skew corrected

mu = 3.14 and sigma = 0.32



5 Evaluating Method

In K-Fold, the selected data set is divided into K number of folds, where each fold is used as the testing set at certain point and remaining points as the training set, which is done iteratively K times. Here, the data set is split into five folds.

Root Mean Squared Logarithmic Error (RMSLE) (Fig. 5) is a function for forecasting the relationship between machine learning values expected by the algorithm and the real objective value.

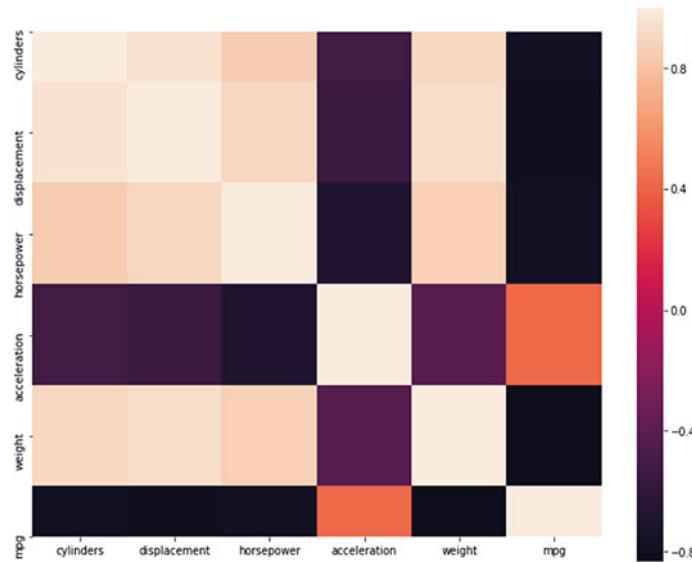


Fig. 4 Heat map of the auto-mpg dataset

Fig. 5 RMSLE function

ROOT MEAN SQUARED ERROR(RMSE)

$$\sqrt{1/n \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

ROOT MEAN SQUARED LOG ERROR(RMSLE)

$$\sqrt{1/n \sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2}$$

6 Modelling

6.1 Base Model

Regression analysis is a powerful statistical tool which enables the relationship to be determined between two or more interest variables. Although a variety of regression analyses exist, all of these study an effect on a vector quantity of one or more independent variables. And here in our model we have taken four separate regression models-LASSO Regression, Elastic Net Regression, Core Recovery, Gradient Boosting Regression and LightGBM-Before we stack the models we have worked to search for RMSLE error in our dataset. The scores represent the mean and standard deviation (Table 1).

Table 1 RMSLE values of the models

S. No.	Model	Mean	Std.
1	Lasso	0.0008	(0.0002)
2	Elastic Net	0.0008	(0.0002)
3	Kernel Ridge	0.0082	(0.0044)
4	Gradient Boosting	0.0538	(0.0225)
5	XGBoost	0.0767	(0.0177)
6	LGBM	0.0697	(0.0123)

6.2 Simple Stacking

In this approach, we prefer to use the out-of-fold projections of these simple models to train our meta-model to incorporate a meta-model on an ordinary base model.

The process for the training of the model is stated below:

1. Divide the whole course into two different sets—training and testing
2. First, many models should be trained—**training**
3. Second, the base models should be checked—**testing**
4. Use the results from the step 3 because they are the correct values, and Feed the learners to prepare a more experienced student called a meta-model.

The first three measures are quadratically calculated. The training details were separated into five folds. Then, we can do five iterations. Every model is trained in 4 folds and we predict for each iteration the remaining fold (test fold).

Thus, after five iterations, we would be satisfied that all data will be used to encourage the consumer to use out-of-fold predictions to train our meta-models in step 4.

We average all predictions of the simple models for the prediction portion. Take a look at data and use it as meta-functions for the final meta-model forecast (Fig. 6).

Here, the Elastic Net Regression, the Kernel Ridge Regression and the LASSO Regression are known as the basic model and the Meta-model Gradient Boosting Control. And with a meta-learner we observed the score of this stack.

Stacking Averaged models score: 0.0340 (0.0223).

And we can see that the score is higher than the previous base model stacked.

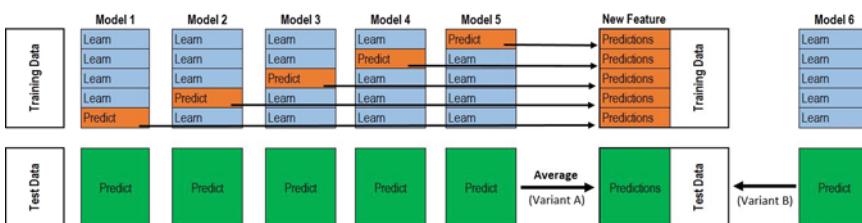


Fig. 6 Ensemble technique

6.3 Ensembling

Regression Now, we ensemble the Stacked Regressors, XGBoost and LightGBM. Previously, we checked how the prediction of these models gives the value close to the actual mpg variable by evaluating the output from the RMSLE function (Table 2).

An associative degree extension of one model averaging ensemble is given to a weighted group wherever the output of the model weights each member's contribution to the final projection. The weight values are chosen between 0 and 1 for each ensemble member. By the value obtained from Table 2, we can choose the weight of each model by trial and error method and thereby predicting the ensemble output value.

$$\text{Ensemble} = \text{stackreg_p} * 0.70 + \text{xg_p} * 0.15 + \text{lg_p} * 0.15 \quad (1)$$

7 Result and Discussion

We have taken 400 tuple data sets and imported the heat map of Seaborn to create a correlation map for this dataset. Between each indicator there are some strong associations. For cylinder, displacement, strength and weight it makes reason that mpg is negatively connected. The model year and source also make sense because, owing to the various fuel costs, the non-American/European countries in these regions could have greater fuel efficiency. The number of cars can be estimated depending on their source (US = 1, Asia = 2, Europe = 3).

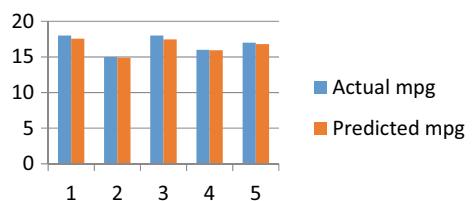
And the target variable has been evaluated for the skew and it was corrected. Feature selection was done to remove the attributes which do not help predict the result. Then, by keeping the RMSLE evaluation function as the key function, we performed the performance of all the models that have been chosen for their greatness. The stacking was performed in the base model and assessed the results. The ensemble was performed on the base model and meta-model, thereby predicting the mpg value (Fig. 7).

Table 2 RMSLE Values of the ensemble model

S. No.	Model	RMSLE value
1	Stacked Regressors	0.000789015706381
2	XGBoost	0.05162735435577
3	LGBM	0.03018659859537

Fig. 7 Predicted mpg value

	Predicted_mpg
0	17.568693
1	14.868792
2	17.470673
3	15.936917
4	16.792574

Fig. 8 Predictive results**Table 3** Predictive results

S. No.	Actual mpg	Predicted mpg
1	18	17.56
2	15	14.86
3	18	17.47
4	16	15.94
5	17	16.79

7.1 Comparison Chart

In this paper, we have predicted the mpg value by ensemble technique and compared the actual mpg with predicted mpg value. Predictions results are shown ([Fig. 8](#)). Actual mpg and predicted mpg comparison are shown in Table 3.

8 Conclusion

This paper presented a machine learning model developing a stack of regressors and XGBoost and LightBGM as the meta-models to predict the automobile vehicles in a fleet. The model relies on eight predictors: mpg, cylinders, horsepower, weight, acceleration, car name, model year, origin. An experiment was completed using ensemble stacking of the regressors and thereby proving that the combination of models will give high accuracy in the prediction.

References

1. Cappiello A, Chabini I, Nam EK, Lue A, Abou Zeid M (2002) A statistical model of vehicle emissions and fuel consumption. In: Proceedings. The IEEE 5th international conference on intelligent transportation systems, Singapore, pp 801–809
2. Webb GI, Zheng Z (2004) Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. *IEEE Trans Knowl Data Eng* 16(8):980–991
3. Wickramanayake S, Dilum Bandara HMN (2016) Fuel consumption prediction of fleet vehicles using machine learning: a comparative study. In: 2016 Moratuwa engineering research conference (MERCon), Moratuwa, pp 90–95
4. Sailaja S, Sowjanya AM (2019) *Int J Innov Res Sci Eng Technol* 8(10):23–30
5. Liu L (2019) Deception, robustness and trust in Big Data fueled deep learning systems. In: 2019 IEEE international conference on Big Data (Big Data), pp 3–3. IEEE
6. Zhang Y, Shu Q, Wang L, Quan X, Liu X, Lu B (2019) Estimation of fuel biomass for grasslands using data assimilation technique. In: IGARSS 2019–2019 IEEE international geoscience and remote sensing symposium, pp 9988–9991. IEEE
7. Ramesh GP, Pandiaraj R (2017) MPC based hybrid battery and fuel cell powered PMSM drive for electric vehicle applications. *Int J MC Square Sci Res* 9(3):8–19
8. Goel R, Mohan D, Guttikunda SK, Tiwari G (2015) Assessment of motor vehicle use characteristics in three Indian cities (2015)

An Empirical Study of Privacy Concern and Trust in the Decision to Revisit Personalized Social Networking Websites



Darshana Desai

Abstract Social networking websites adopt personalization to cater to users' needs to serve better are widely researched, but little is known about privacy concerns of the user and trust affecting user's decision making to revisit with personalized social networking websites. This research develops a parsimonious model to predict users' intention to revisit social networking websites using personalization as a result of the trade-off between information personalization, control, trust, and privacy concern. In addition to this trade-off, the research identified that a users' intent to use personalized services is positively influenced by trust toward SN websites. Our findings suggest that firstly, SN websites can improve their abilities to acquire and use users' information through trust-building with highly relevant personalized information; secondly, it is of critical importance to understand the value of privacy of users and provide more control to the users for personalized service.

Keywords Content personalization · Trust · Control · Privacy concerns · Revisit intentions

1 Introduction

Over the last decade, Technology has influenced socializing in society in amazing and innovative ways with social networking sites (SNS) like Facebook, Instagram, LinkedIn, etc. More than 3.96 billion people use various social networking sites to communicate worldwide in 2020 in real time to share photos and videos [1]. As Internet access in India has become easily available and more users are active on social media covering all age groups, it is expected to reach it from 330 to 448 million by 2023 [2]. The use of social networking sites has grown to include all strata of individuals with different mindsets, backgrounds, and demographics. Users from the various age groups are offered personalized content based on their search history, contacts, interactions with their friends, and content that is liked while interacting

D. Desai (✉)

Department of MCA, Indira College of Engineering and Management, Pune, India

on these SNS. Social media websites are also considered an effective medium for marketing and influence users with personalized information. Users' interactions with such websites are tracked and the identification of users' preferences which have also raised privacy concerns among the users. Users can control the personalized information and offerings with privacy settings provided on the SNS to some extent. The hacking of social networking websites and users' information breaching has raised more privacy concerns and motivated users to have control over information sharing [3]. However, personalization is researched in recent years but has a deficit in the study of the effect of personalization on users' information processing and behavioral intentions affecting control, privacy concerns, and trust on social networking websites. This research attempts to fill this gap in exploring various factors affecting users' decision making to revisit social networking websites having personalization offerings. The main objective is to explore the attitude of the users toward personalization in social networking sites with privacy concerns, desire to control, and trust.

2 Related Work

Personalization is the process of providing tailor-made content to the users to cater to their explicit or implicit needs identified through customization with choices or analyzing users browsing behavior and interaction with the websites. Information personalization is the extent to which information can be catered according to the user's implicit or explicit requirement [3–5]. Information personalization refers to the degree to which customers are provided with uniquely tailored information based on their individual needs as gathered from the consumer's interaction with the provider [6–8]. Personalized content with high relevance to users' needs reduces cognitive efforts needed for information access and services leading to higher satisfaction [5]. The high relevance of personalized information builds more trust in the user which inspires users to revisit the social networking websites [4]. Social networking sites have at their disposal an increasing amount of personal information about using their behavioral trail on this website interaction.

Personalizing online interactions improves customer relationships and increases desirable behaviors, such as positive word-of-mouth and increased purchase intent [3]. Users develop more cognitive and emotional trust with a significant increase in personalization [6, 9, 10]. However, other research suggests that the use of personal information stimulates privacy concerns in the user and affects the decision to continue using websites. The research examines constructs affecting users' behavioral intentions in the form of continue to revisit SNS and also studies inter-relations of users' control, privacy concerns, and trust toward websites during the interaction. Results show that increasing perceived information control reduces the negative effect of privacy concerns on behavioral intentions.

3 Conceptual Framework and Hypotheses

3.1 Content Personalization and Privacy Concern

Perceived privacy of users is the subjective probability in which users believe that collection and subsequent use, access, and disclosure of their personal information are consistent with their expectations [11]. Highly relevant information personalization reduces the cognitive efforts of the user and increases satisfaction in the user [5]. However, users experience higher privacy concerns when the personalization process uses users' information without consent and have negative feelings about personalization.

To provide personalized content, applications require user's personal information [1, 9, 12] and understand users' implicit need by observing interaction with the website [4] like social media behavior; likings have the potential for an invasion of privacy. Individuals might have higher privacy concerns about online personalization if they are not aware of the intentions behind content personalization which reduces trust in social networking websites.

3.2 H1: Users' Privacy Concerns Are Positively Associated with Content Personalization

Content Personalization and Control

Personalization provided with the choices to users generates a high level of perceived control, and users who experience a sense of involvement with the external personalization process [5, 10, 13] are more likely to have comfort levels and enjoy the interaction with social networking websites. Users experience the intrinsic feeling of control when given a choice of information sharing, viewing of commercial advertisement, setting of privacy, accessibility of information, consent asked for cross-app communication. So our work hypothesizes.

3.3 H2: Users' Experience Higher Control Over Customization Choices in Personalization

Privacy Concern and Trust

Personalization of websites reduces cognitive efforts by the user to search for information, and users feel higher satisfaction with more relevant personalized information [5, 7] eventually developing trust toward websites. Users with higher privacy concerns need more control over personalization to develop trust over social

networking websites. Privacy concerns of the user lead to trust-building up on fulfillment [9, 13]. Personalization has cognitive benefits and includes the cost of information disclosure. In the process to achieve personalization, websites need to collect more information about users' personal and behavioral characteristics like interests and implicit needs. This increases privacy concerns when privacy risks are associated particularly in wary of personalized content [10].

3.4 H3: Users' Trust is Highly Correlated with Privacy Concerns

Control and Trust

Dwyer et al. [7] defined trust as the “willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action which is important to the users trust irrespective of the ability to monitor or control that other party. Users having control over information flow, profile visibility, and protection of their personal information are more likely to develop trust in social networking websites.” Shin [14] defined privacy as control over the flow of one's personal information, including the transfer and exchange of that information.

3.5 H4: Users Experience More Trust with Higher Control Over Personalization

Highly relevant personalized content offering on websites like a targeted advertisement, recommendation, location tracking, and offering suggestions based on geographic-location increase user's satisfaction and induces intrinsic feeling trust amongst users.

Trust and Revisit Intention

Earlier research shows a positive association between users' trust in digital interaction and willingness to use [13, 15]. We can infer the significant role of trust in behavioral intention to revisit social networking sites offering online personalization. Trust in the website instigates users to share more personal information for personalization to feel more confident and motivated to revisit social networking websites. Users with less trust are less likely to perceive and value benefits from the personalization of websites, on the contrary, likely to visit frequently with the trusted websites having personalization features [10]. So postulates:

3.6 H5: Users' Intention to Revisit a Website is Highly Associated with Trust in Personalized Social Networking Websites

Control and Revisit Intention

Users experience higher control provided with explicit choices and preference settings of personalization and play an important role in users' decision-making process to revisit a website through cognitive and hedonic belief [2, 7]. Users experience less control when websites collect and use users' personal information for their benefit without when information control is low (i.e., the firm has collected and used personal information about the consumer without their consent which leads to higher privacy concern and lowers user's intention to revisit the website). So research proposes:

3.7 H6: User Intention to Revisit Social Networking Websites is Positively Associated with Control

Privacy Concern and Revisit Intention

Users are made aware of the privacy policy of social networking sites at the time of registration to continue the use of features of the websites and are known to users' information sharing. Users having higher privacy concerns, limit their interaction with the sites in the form of posting contents and disclosing their behavioral attributes like preferences, thought sharing on expression of thoughts on social media, on the other hand users with less privacy concerns are more likely to share personal information, their living styles, habits, location and show openness in thought sharing. This behavioural attribute exhibits their inbuilt trust in the social networking websites. References [14, 16] identified that users' control over information, is a prime factor in establishing trust in social media online environments. On the contrary users' loss of control over personalization of information over social networking sites like control of people who can view post or profile, feel more vulnerable and develop less trust toward websites consequently it affects in behavioral intention to revisit the website.

H7: Users' intention to revisit is highly associated with privacy concerns.

4 Research Method

4.1 Data Collection and Sampling

Research focuses on the behavioral intention of users with personalization, so we adopted a survey-based method for data collection and test proposed research hypotheses. The random sampling method is used to target a population of users who are using the social networking sites like Facebook, Google+, Yahoo.com, MSN.com for more than two years and experienced personalized service by the website in different forms like a recommendation of friends, post, and content based on users liking and interactions. All the constructs identified from the prior studies are content personalization, trust, privacy concerns, control, and revisit intentions, measured with the five-point Likert scale. All the responses were collected through online questionnaires from the respondents who are using Facebook's social networking website having personalization as a prime feature. Table 1 depicts the demographics of the responses.

Table 1 Sample demographics

Measure	Item	Frequency	Percentage (%)
Gender	Male	204	66.6
	Female	198	33.4
Age (years)	18–25	368	78
	26–35	70	14.4
	36–50	34	7.2
	>50	1	0.4
Education	Undergraduate	177	37.4
	Graduate degree	109	23.2
	Postgraduate	184	62.8
	Doctorate	3	0.6
Occupation	Student	339	71.7
	Service	103	21.8
	Self-employment	14	3.0
	Homemakers	17	3.6
Experienced	Directly	184	45.8
Personalization	Indirectly	197	49

4.2 Measurement Model

The reliability of the constructs was checked through a pilot study of 50 responses to ensure the validity of variables and survey questions. The population of data collection is the respondents from India having Facebook accounts and has experienced personalization directly or indirectly. A total of 473 valid responses were collected from a total of 550 responses after data preprocessing by removing noise and cleaning incomplete and inconsistent data. Responses with a standard deviation below 0.30 were removed to fetch final valid responses for further data analysis. Factors were identified and confirmed using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) using SPSS 20.0. The structural equation modeling (SEM) technique is used to check the model fit of the proposed model.

The Cronbach's Alpha coefficient value of all the constructs are in the range of 0.70–0.90 which is above the 0.70 showing higher internal consistency of constructs and internal consistency of the scales of questionnaire items used within this survey. Factor loading scores of the construct items are also satisfactory which are more than 0.60 in confirmatory factor analysis shows higher constructs validity and reliability.

The validity of the model is checked using confirmatory factor analysis (CFA) and two-step validity measurement identifying convergent validity and discriminant validity of the measurement model. Subsequently, testing of the hypothesis is performed, and the model is identified using the structural equation modeling technique.

4.3 Confirmatory Factor Analysis and Validity Test

Confirmatory factor analysis is used to test the measurement model using SPSS AMOS 21.0. Model fit is proved if the model-fit indices reach accepted standards and recommended values; the result shown in the table exhibits adequate fit as the model-fit index exceeds the recommended value.

Composite reliability measure shows internal consistency of construct item; CR value is above 0.7 indicating good reliability and measurement items stability of each construct. Every CR scored above 0.8, which is above the recommended value by Fornell and Larcker [17], indicating good reliability and stability for the measurement items of each construct.

Convergent validity of measurement model standards recommended by Bagozzi and Yi [18] is as follows: (1) Factor loadings of the construct should exceed 0.5 [19]; (2) CR should be above 0.7; and (3) average variance extracted (AVE) score of each construct should exceed 0.5 [17]. The result shown in Table 2 depicts the factor loading score of each construct item exceeded 0.7, CR of the construct is above 0.8, and AVE score ranges from 0.46 to 0.95 which satisfies all the conditions for convergent validity showing measurement items that correlate strongly with its theoretical constructs.

Table 2 Statistics of construct items

Construct	Items	Factor loadings	Composite reliability (CR)	The average variance extracted (AVE)	Cronbach's Alpha
Content personalization	SNIP1	0.781	0.864	0.516	0.863
	SNIP2	0.717			
	SNIP4	0.732			
	SNIP3	0.776			
	SNIP5	0.623			
Control	SNCON1 SNCON2	0.980 0.927	0.973	0.943	0.973
Privacy concern	SNPRIVACY1	0.742	0.85	0.462	0.871
	SNPRIVACY2	0.756			
	SNPRIVACY3	0.652			
	SNPRIVACY4	0.695			
	SNPRIVACY5	0.787			
	SNPRIVACY6	0.720			
Trust	SNTRUST1	0.748	0.828	0.547	0.828
	SNTRUST2	0.718			
	SNTRUST3	0.780			
	SNTRUST4	0.659			
Intention to revisit	SNINT1	0.969	0.967	0.907	0.967
	SNINT2	0.946			
	SNINT3	0.944			

Fornell and Larcker [17] suggested that for the discriminant validity of the model, the AVE of the construct should exceed the correlation coefficients of the constructs. Table 3 shows the correlation coefficient matrix of all the constructs in the research model; also diagonal element values are square roots of the AVE score for the constructs. The correlation coefficient values for any two constructs are lesser than the square root of the AVE score for the constructs. Research shows good discriminant validity of the research constructs in model measurement as constructs are different

Table 3 Discriminant validity

Construct	Trust	Privacy concerns	Content personalization	Intention to revisit	Control
Trust	0.725				
Privacy concerns	0.670	0.709			
Content Personalization	0.447 0.431	0.451 0.459	0.718 0.412	0.942	
revisit intention					
Control	0.394	0.469	0.376	0.497	0.974

from each other. Research result shows that the measurement model is having good construct reliability, discriminant validity, and convergent validity.

4.4 Test Model Fit with Structural Equation Modeling

Research model is tested with structural equation model using AMOS 21.0. The model-fit indices $\chi^2/df = 1.72$, GFI = 0.94, AGFI = 0.92, NFI = 0.95, CFI = 0.97, RMSEA = 0.046 indicate good model as shown in Table 4. All the indices for model fit in structural equation modeling which indicates that is proposed model have good fit.

Figure 1 displays the standardized path coefficients, variance explained (R^2), and path significances values significance for the path as per the proposed hypothesis, and all hypotheses are supported except correlation of trust and control. As with

Table 4 Hypotheses testing result

Hypotheses			Estimate	S.E	C.R	P-value
H1: Privacy_Concern	←	Content_Personalization	0.492	0.040	12.188	***
H2: Control	←	Content_Personalization	0.395	0.045	8.702	***
H3: Trust	←	Privacy_Concern	0.692	0.031	22.263	***
H4: Trust	←	Control	0.052	0.029	1.752	0.080
H5: Revisit_Intension	←	Trust	0.211	0.059	3.593	***
H6: Revisit_Intension	←	Control	0.313	0.038	8.320	***
H7: Revisit_Intension	←	Privacy_Concern	0.208	0.057	3.665	***

*** indicates p -value < 0.001

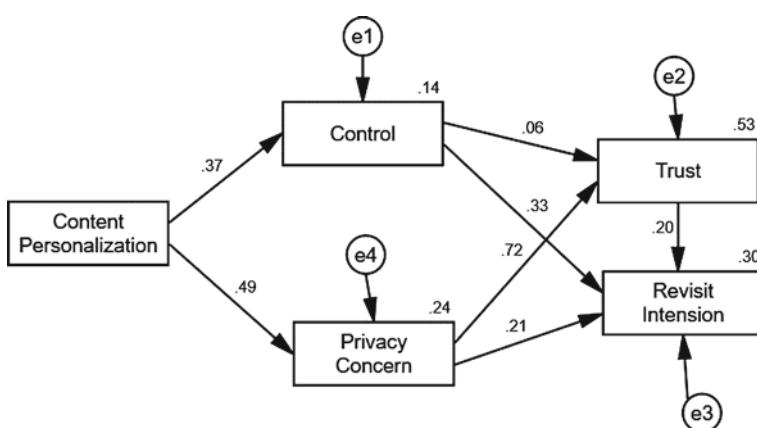


Fig. 1 Structural equation modeling result

variance explained (R^2), R^2 values of revisit intention and trust are 0.34 and 0.54 which is above 0.3 shows a good research model.

5 Result and Discussion

SEM result in Table 4 shows the high correlation between constructs content personalization, privacy concerns, control, trust, and revisit intention. All the hypotheses proposed in the model satisfy users' trust dependency on control over the personalized website. Users' privacy concern and desire to control over the personalization of the website are highly dependent on content personalization. Users given the choices for customization and settings for personalization and information sharing develop more privacy concerns and more desire to control the personalization of social networking website. Research shows that users develop more trust toward the personalized website and privacy concerns. Users with high privacy concerns are less likely to develop trust in social networking websites having personalization. Users are more likely to revisit social networking websites with more trust and control over the personalization.

6 Conclusions and Future Scope of Research

This research is a qualitative study on users' behavior toward personalization in social networking websites like Facebook. Users' privacy concerns and control over the personalization process in the form of explicit customization play a significant role in developing trust toward social networking websites. Users with higher desire to control show high privacy concerns and are less likely to share personal information; also such users avoid more interaction with the website in the form of nondisclosure of information and restricting users to view their post on social media sites. Personalized content with high relevance is the key factor in developing trust mediated by users' privacy concerns and desire to have control over websites subsequently develop trust and motivate users to repeat visits to the website. Research is more useful in designing the personalization of the website and understanding users' attitudes and key factors affecting decision making to revisit the website.

The survey-based research methodology adopted for the study can be further researched with users' behavioral intention in a controlled laboratory environment with live interaction with the personalization of social networking websites.

References

1. Acquisti A (2004) Privacy in electronic commerce and the economics of immediate gratification. Proc ACM Conf Electron Commer 5:21–29. <https://doi.org/10.1145/988772.988777>
2. Al Qudah DA, Al-Shboul B, Al-Zoubi A, Al-Sayyed R, Cristea AI (2020) Investigating users' experience on social media ads: perceptions of young users. Heliyon 6(7). <https://doi.org/10.1016/j.heliyon.2020.e04378>
3. Taylor DG, Davis D, Jillapalli R (2009) Privacy concern and online personalization: the moderating effects of information control and compensation. Electron Commer Res 9:203–223. <https://doi.org/10.1007/s10660-009-9036-2>
4. Desai D (2019) Personalization aspects affecting users' intention to revisit social networking site. Int J Trend Sci Res Dev (IJTSRD) 4(1):612–621. ISSN: 2456-6470. <https://doi.org/10.31142/ijtsrd29631>
5. Desai D (2019) An empirical study of website personalization effect on users intention to revisit E-commerce website through cognitive and hedonic experience. In: Balas V, Sharma N, Chakrabarti A (eds) Data management, analytics, and innovation. advances in intelligent systems and computing, vol 839. Springer. https://doi.org/10.1007/978-981-13-1274-8_1
6. Appel G, Grewal L, Hadi R, Stephen AT (2020) The future of social media in marketing. J Acad Mark Sci 48(1):79–95. <https://doi.org/10.1007/s11747-019-00695-1>
7. Dwyer C, Hiltz SR, Passerini K (2007) Trust and privacy concerns within social networking sites: a comparison of Facebook and MySpace'. In: Americas conference on information systems, proceedings of the thirteenth Americas conference on information systems, Keystone, 9–12 Aug, Colorado, USA, p 339. <https://aisel.aisnet.org/amcis2007/339>
8. Dey N, Borah S, Babo R, Ashour AS (2019) Social network analytics: computational research methods and techniques. Elsevier. ISBN: 9780128156414
9. Aldhafferi N, Watson C, Sajeew ASM (2013) Personal information privacy settings of online social networks and their suitability for mobile internet devices. Int J Secur Privacy Trust Manag 2(2):1–17. <https://doi.org/10.5121/ijspmtm.2013.2201>
10. Stevenson D, Pasek (2015) Privacy concern, trust, and desire for content personalization. In: The 43rd research conference on communication, information and internet policy. <https://doi.org/10.2139/ssrn.2587541>
11. Chellappa RK, Sin RG (2005) Personalization versus privacy: an empirical examination of the online consumer's dilemma. Inf Technol Manage 6:181–202. <https://doi.org/10.1007/s10799-005-5879-y>
12. Komiak S, Benbasat I (2006) The effects of personalization and familiarity on trust and adoption of recommendation agents. MIS Q 30(4):941–960. <https://doi.org/10.2307/25148760>
13. Senthil Kumar N, Saravanakumar K, Deepa K (2016) On privacy and security in social media—a comprehensive study. Phys Procedia 78:114–119. <https://doi.org/10.1016/j.procs.2016.02.019>
14. Shin D (2010) The effects of trust, security, and privacy in social networking: a security-based approach to understand the pattern of adoption. Interact Comput 22(5):428–438. <https://doi.org/10.1016/j.intcom.2010.05.001>
15. Tucker CE (2014) Social networks, personalized advertising, and privacy controls. J Mark Res 51(5):546–562. <https://doi.org/10.1509/jmr.10.0355>
16. Mohamed N, Ahmad IH (2012) Information privacy concerns, antecedents and privacy measure use in social networking sites: evidence from Malaysia. Comput Hum Behav 28(6):2366–2375. <https://doi.org/10.1016/j.chb.2012.07.008>
17. Fornell C, Larcker D (1981) Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 18(1):39–50. <https://doi.org/10.2307/3151312>
18. Bagozzi RP, Yi Y (1988) On the evaluation of structural equation models. JAMS 16:74–94
19. Hair Jr. JF et al (1998) Multivariate data analysis with readings. Englewood Cliffs, NJ, Prentice-Hall

20. Chellappa RK, Shivendu S (2006) A model of advertiser—portal contracts: personalization strategies under privacy concerns. *Inf Technol Manage* 7(1):7–19. <https://doi.org/10.1007/s10799-006-5726-9>
21. Gupta A, Dhami A (2015) Measuring the impact of security, trust, and privacy in information sharing: A study on social networking sites. *J Direct Data Digit Mark Pract* 17(1):43–53. <https://doi.org/10.1057/dddmp.2015.32>
22. Naresh M, Malhotra KK, Kim SS, Agarwal J (2004) Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. *Inf Syst Res* 15(4):336–355. <https://doi.org/10.1287/isre.I040.0032>
23. Sutanto J, Palme E, Tan C, Phang C (2013) Addressing the personalization-privacy paradox: an empirical assessment from a field experiment on smartphone users. *MIS Q* 37(4):1141–1164. Retrieved 5 Mar 2021, from <http://www.jstor.org/stable/43825785>
24. Social Media Statistics and Facts (2020, November 11) <https://market.us/statistics/social-media/>
25. Digital and Social Media Landscape in India (2020, December 12) <https://sannams4.com/digital-and-social-media-landscape-in-india/>

Distribution System Voltage Stability Index Determination with Nature-Inspired Meta-heuristic Cuckoo Search Algorithm



M. Sridhar Bhatlu, Satyajit Panigrahy, and Ashwani Kumar Chandel

Abstract The integration of distributed power generators are in the distribution system is being done to control the power losses and enhance the voltage stability of the existing system. Consequently, the sizing and placement of DGs are very crucial. The task of finding the locations for DGs and the proper sizes of DGs can be accomplished with the help of meta-heuristic techniques. Subsequently, in the present paper optimal DG sizing, placement is addressed using a cuckoo search optimization algorithm. Radial 33- and 69-bus distribution systems are considered to perform the robustness of the proposed technique. The results are then compared with other conventional techniques. Test results, thus, demonstrated that the Cuckoo search optimization algorithm (CSO) outclasses the former algorithms in aspects of the index terms of the voltage stability and loss minimization.

Keywords Distributed generator (DG) · Meta-heuristic techniques · Loss minimization · Optimal placement · Binary particles swarm optimization (BPSO) · Ant-lion optimization (ALO) · Cuckoo search optimization (CSO)

1 Introduction

Integration of distributed power generation such as hydro, ocean energy, wind, biomass, solar, geothermal, and micro-turbines into the distribution system has one of the important studies. The penetration of DGs into the existing system affects performance. So, the study of DGs designing, sizing and proper placement, etc., is very popular among the researchers. Hence, investigators are searching for sustainable resolutions for DG integrated systems that are capable of optimal placement of DGs, loss minimization of the distribution system, loading ability, voltage profile improvement, cost minimization, and finally, it will improve the reliability of the system. These are considered to be DG penetration level factors. Different DGs are categorized on the rating, namely large, medium, small, and micro-DGs [1].

M. S. Bhatlu (✉) · S. Panigrahy · A. K. Chandel
National Institute of Technology, Hamirpur, HP 177005, India

A lot of approaches are developed for optimal location and sizing in [2, 3]. However, the investigators have failed to a larger system when determining the optimal solution. A lot of mathematical approaches have been developed in [4–6] like Kalman filter algorithm, improved analytical method, and exhaustive load flow for optimizing the distribution system disabilities. These methods are computationally faster but having drawbacks which are complexity in the load flow, high-power losses, and less accuracy and took a high amount of time for convergence. The meta-heuristic optimization algorithms are most efficient to solve the multi-objective problems. The other popular methods for multiple DG placement are the Artificial Bee Colony Algorithm (ABC), Ant Colony Search (ACS), and Harmony Search Algorithm (HSA) [7–9]. Further, the optimal sizing and placement of DGs CSO [10, 11] are considered in the paper. CSO is one of the latest methods which is developed based on the brood parasitism of a cuckoo bird. The main advantage of this algorithm is that the number of parameters considered is very less, and the rate of convergence of the algorithm is faster. CSO a population-based algorithm like GA and PSO, and it follows elitism as same as HSA. One of the major constraints is that the CSO follows levy flights. To check the efficiency of the method, it has been tested on a 69-bus radial distribution system, and the voltage stability index is compared with PSO and ALO, respectively.

This paper has the following sections. Section 1 outlines the literature and Sect. 2 deliberate upon the problem formulation. Section 3 explains about ALO for loss minimization. Section 4 describes a Cuckoo search for DG allocation and sizing. Section 5 deliberate upon the test cases on which the performance of the above-mentioned algorithm has been evaluated. Sections 6 and 7 proved the simulation results and conclusion, respectively.

2 Problem Formulation

Voltage stability improvement (VSI) and power loss reduction in the distribution network are achieved by optimal sizing and locations of the DGs which is obtained in the present paper with the CSO algorithm. Suitability and efficiency parameters are considered for the implementation point of view. To implement this, the weight method that is proposed in [12] is considered. The multi-objective optimization problem is converted into a traditional single-objective problem by weight method. The mathematical expression of the objective function is:

$$\min f = w_1 f_1 + w_2 f_2 \quad (1)$$

where f_1 =Overall active power loss of the system and measured from 2, f_2 =voltage stability index (VSI) of the system which is radial and measured from 3, and w_i =weighing constants which are altered in the middle of 0 & 1 and $\sum_{i=1}^k w_i = 1$.

The objective function decides the weighing factors which is best part of objective function. Here two weights are considered for sending of quality power to consumer end, VSI is appropriated to 0.3, and 0.7 is appropriated for active power loss. Load flow equations as formulated in [13] are used.

2.1 Formulation of the Objective Function

Power loss coefficient f_1 can be formulated as

$$f_1 = \frac{P_{L(DG)}}{P_L} \quad (2)$$

where P_L = overall power loss (real), and $P_{L(DG)}$ = power loss in distribution network. Power loss in network with DG and without DG is expressed as $P_{L(DG)} = \sum_{ni=1}^N I_{ni}^2 R_{ni}$. Here, I_{ni} = current, and R_{ni} = resistance, $I_{ni} = \frac{V_{mi}-V_{ni}}{R_{ni}-jX_{ni}}$, $P_{ni} - jQ_{ni} = V_{ni}^* I_{ni}$. Further, the VSI is:

$$VSI_{ni} = |V_{mi}|^4 - 4[P_{ni}R_{ni} + Q_{ni}X_{ni}] |V_{mi}|^2 - 4[P_{ni}R_{ni} + Q_{ni}X_{ni}]^2$$

In which, V_{mi} and V_{ni} are sending and receiving side voltages, P_{ni} and Q_{ni} are active and reactive power, R_{ni} and X_{ni} are resistance and reactance, respectively. By considering the minimum (VSI_{ni}) in denominator, VSI can improve objective function, and this is expressed below:

$$f_2 = \frac{1}{\min(VSI_{ni})} \quad (3)$$

where VSI_{ni} should have a value greater than zero, for $ni= 2, 3..., nn$ that can provide a viable solution. To avoid instability in voltage levels, it is necessary to find out the weak buses that have a minimum value of VSI.

2.2 Constraints

Because of penetration of DGs power reversal is takes place in the network frequently so that voltage levels will change extremely. Therefore, threshold values of the voltage are taken in the ranges of 0.95 per unit to 1.05 per unit and mathematical relation for voltage ranges shown as $V_i^{\min} \leq V_i \leq V_i^{\max}$. Like wise for real and reactive power, thresholds are considered from 0 to 5MW ($P_{dg}^{\min} \leq P_{dg} \leq P_{dg}^{\max}$) and 0 to 1MVAR ($Q_{dg}^{\min} \leq Q_{dg} \leq Q_{dg}^{\max}$).

2.3 Standard Deviation

Bessel's standard deviation is considered to investigate the missing data in test cases, why because the cause of missing terms may deviate the solution quality, i.e., $S = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}}$ where, x_i =Values from the iterations, \bar{x} =Value (mean), N =Samples count.

3 Ant-lion Optimizer for Power Loss Minimization

Seyedali Mirjalili [14] has worked on a population-based algorithm to tune the control parameters of an ant-lion optimizer like the number of variables are six, the population of ants is 40 and the random number in between 0 to 1. In this, the total 500 iterations are considered for the best comparison. Ant-lion optimizer is consisting of five basic steps as explained below.

1. *Random Walk of Ants*: The randomized walks have already set to an initial value, and the ants would keep on updating their locations in the process of optimization.
2. *Building of the Trap*: In cuckoo search algorithm, each cuckoo lay one egg only like in ant-lion optimizer only one ant can be hunted. In the trap building process, the fitness values of the roulette wheel selection method are used for the ant-lions.
3. *Sliding Off the Ants Near to the Ant-lion*: - The physical behavior of ants in the ant-lion technique is, actually pit is made with sand when ants enter inside the pit, ants automatically slide down to the middle of the pit.
4. *Catching of the Prey and rebuilding the Pit*: When an ant falls inside the pit, the last phase of this process will start, and it gets caught by ant-lion. In order to depict this process, we presumed that catching of the ants will be done when an ant becomes a supplementary fitter than the corresponding ant-lion. In that case, ant-lion necessitates updating its old position to a new position to increase the probability of holding the prey.
5. *Elitism*: In EA (Evolutionary Algorithms), Elitism is the most important characteristic, and it allows us to uphold the best challenging solutions all through the optimization procedure.

4 CSO for DG Placement and Sizing

Cuckoo search is useful for optimal placement of DG. Xin-she Yang and Deb [15] in 2009 recommended CS optimization algorithm. CSO has been found to be suitable for solving d welded beam design and spring design applications. In this paper, it has been considered that only one egg is present in each nest. For the purpose of explanation, CSO is described by considering three flawless rules [16].

- In an instant, a cuckoo bird laid only single egg and which is randomly placed in pre-defined nest.
- Each egg is a fresh solution and good qualities of eggs in the nest are used for the upcoming generation.
- The possibility of noticing the cuckoo bird's egg which does not belongs to him is $P_a = [0, 1]$, while the nests are already defined by programmer. Here the good thing is if the host bird recognizes the egg is not belongs to him, then immediately it will choose two options either it will leave the nest or it will throw the egg from nest.

4.1 CSO for DG Placement

For implementing the CSO algorithm, some parameters are considered beforehand. The possibility of noticing the cuckoo bird's egg which does not belongs to him is (P_a), nest count (n), and step size, i.e., levy flight action (α). These restrictions have a set values of $n=20$, $(\alpha)=1$ and $(P_a)=0.6$ in existing work. For estimating the objective function and easiness of the method, to consume less time for performing load flow, and it should be depends on the type of network, so that forward backward sweep algorithm is picked. It is identified in the study [17]. Cuckoo search optimizer consists of five basic steps as discussed below.

1. Firstly, initialize the population (n) of cuckoo which is already defined by programmer so that new user does not specify any values of population of cuckoo.
2. Levy flight generation is the best part of the algorithm with this levy flight programmer can take as much as possible step length. After performing load flow, the obtained results must be compared with objective function values, so that best solutions can be obtained.
3. Fresh solution replacement has to be done when the selected nest having good solution then only solution replacement takes place otherwise old solution only considered for the next step.
4. The possibility of noticing the cuckoo bird the egg in the nest does not belongs to him is P_a so that waste nest will not considered, and it will go to next nest or throw the egg. This is called as generating new nest.
5. Finally, the stopping criteria, i.e., termination. In this step, programmer has to predefine the iterations. If CSO reached to termination point with the closeness of $1 \times e^{-6}$, the process then will be stopped, and the results are displayed.

5 Test Cases

Three cases of testing are done in this study. These are as follows.

Case A: Without placing a DG.

Case B: Reconfigured network with DG placement.

Case C: Here, a maximum of two numbers of DGs are allocated.

Case C further divided into two sub-cases given below.

Case (i): Real power supplying to the load by means of micro-turbines and photo-voltaic (DG).

Case (ii): Real and reactive power supplying to the load with 0.98 (lagging) power factor by means of gas turbines and biomass.

6 Simulation Results and Discussion

The total active and reactive load powers for 33- and 69-bus systems are taken as 3.80MW, 3.75 MW, 2.69 MVAR, and 2.3 MVAR, respectively. The test data are taken from [17]. The values obtained by running a program for IEEE 69-bus system without DG allocation are considered as reference. Further, it has been analyzed with DG allocation. In the 33-bus test system, VSI has improved efficiency, and the 17th node voltage is 0.9180 p. u before DG placement, it increased to 0.9899 when the DG placed in the network, and it is shown in Fig. 1 by considering the test case C. Test case C of 69-bus VSI is plotted in Fig. 2. It shows that the VSI is improved after DGs integrated into distribution network. Node voltage of 65 bus has increased from 0.9092 to 0.9834 p.u. after optimal DG sizing.

Standard deviation and mean of the objective function are displayed in Table 1 for test case C obtained from PSO and CSO in the 69-bus system. After running

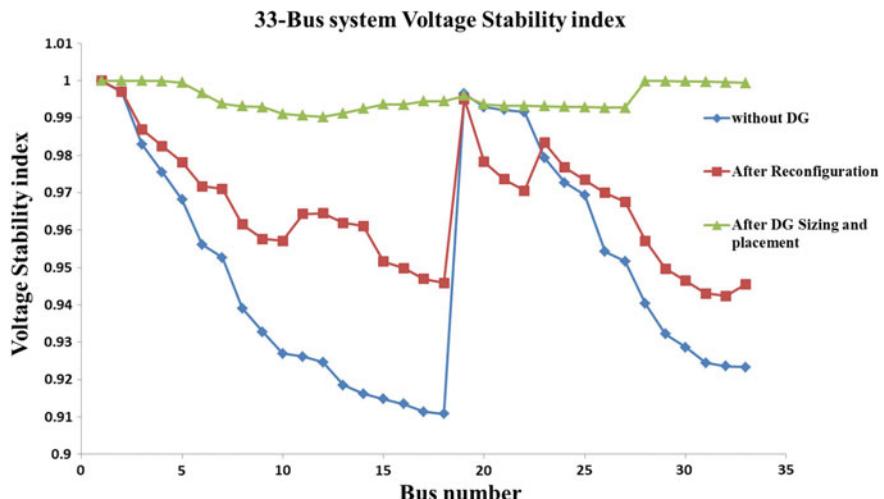


Fig. 1 Voltage stability index of the 33-bus system

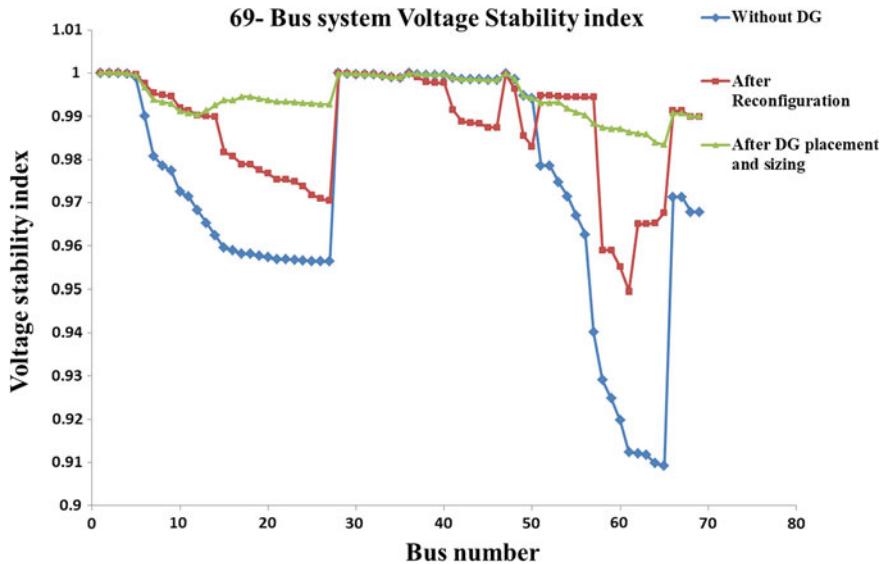


Fig. 2 Voltage stability index of the 69-bus system

Table 1 Objective function values for 69-bus system

Methods	Objective functions	Values			Standard deviation
		Best	Worst	Mean	
CSO	f	0.4279	0.4745	0.4369	0.0088
PSO	f	0.4308	0.9956	0.5845	0.2103

the program 30 times continuously, the obtained iteration results have been demonstrating that CSO is more efficient, and the obtained objective function value from CSO is 0.0088; but in PSO, it has 0.2103 which is somewhat higher. ALO, PSO, and CSOs convergence characteristics are plotted in Fig. 3. From the above discussions and observations, CSO is performing efficiently than ALO and PSO.

The excellence and efficiency of CSO results are related to the PSO algorithm and DG integrated with the distribution network which is already reconfigured by the BPSO algorithm. Tables 2 and 3 display for test case C for 33 & 69-Bus systems, respectively. In the 33-bus system, only multiple DG test cases are considered for effective results. Table 2 shows the result of the second sub-case only while Table 3 shows the result of both the sub-cases, respectively. In every platform, CSO shows the effectiveness and efficiency by considering the parameters like the reduction of power loss and maintaining the voltage stability index within the threshold range. In multi-DG and single allocations, CSO and PSO have alike characteristics in terms of convergence, though CSO still has a better solution quality than ALO and PSO.

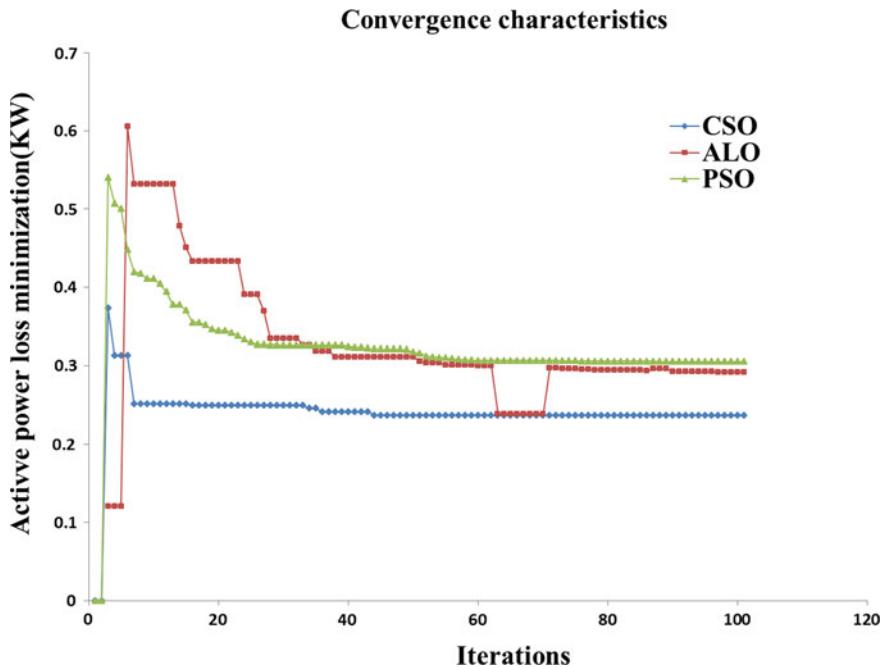


Fig. 3 Active power loss minimization with PSO, ALO, CSO algorithms for 69-bus system

Table 2 The results of the subtest case (ii) of case C for 33-bus system

Test case C	Bus number	Size of DG (MW)	Active P_{loss} (MW)	Change in loss reduction (%)	Change in load (%)
Case (ii)	6,31	1.1, 0.57	0.0288	58	80
	6,31	1.4, 1.1	0.036	82.3	100

Table 3 The results of the subtest case (i) and (ii) of case C for 69-bus system

Test case C	Bus number	Size of DG (MW)	Active P_{loss} (MW)	Change in loss reduction (%)	Change in load (%)
Case (i)	61	2	0.0839	62.79	0.884
	22, 61	0.6, 2.1	0.0763	65.92	0.9612
Case (ii)	61	2.3	0.0526	76.7	0.8947
	18,61	0.8,2.0	0.0399	82.3	0.9771

Table 4 Per unit voltage change at the node of DG

Bus system	Bus number	Normal case (Neglecting DG)	Reconfiguration (Neglecting DG)	Reconfiguration (DG placement and sizing with CSO)
33	6	0.9561	0.9782	0.9795
	31	0.9241	0.9432	0.9512
69	18	0.9581	0.9788	0.9945
	22	0.9568	0.9753	0.9933
	61	0.9123	0.9650	0.9860

Table 5 Percentage of loss minimization related to former 2 methods

Methods	$P_{\text{loss}}(\%)$	
	33-bus	69-bus
Reconfiguration with PSO	32.95	56.18
Reconfigured Network considering DG	34.79	56.06
DG placement and sizing with CSO	Case (i)	–
	Case (ii)	58
		82.3

Table 4 illustrates that the p.u. node voltages of 33 and 69 test systems at which DGs are placed in a distribution network. Both with DG and without DG nodes have shown improvement in p.u. voltage. It visually demonstrates that CSO is a better approach for improving voltage profile than other approaches.

Table 5 is telling only about a % reduction in a power loss of the 33 and 69 radial distribution bus system. Various significant observations of the 69-bus system are stated next. Comparison results of DG integrated (reconfigured network) with the network (without DG) are expressing that reconfigured network with DG is having poor outcomes why because with the changing of tie switches the topology of the network is changed immediately. But in the case of the 33-bus system because of DG placement % of loss reduction slightly improved because of the reduction of the robustness of the bus system reduced from 69 to 33. This is considered to be the downside of BPSO. Hence, to improve all aspects with one algorithm in less time with minimum losses, CSO is considered to be best.

7 Conclusion

Distribution system voltage profile, loading ability, and loss minimization using meta-heuristic CSO have been developed for 33 and 69 radial distribution bus systems. To demonstrate the highness of the developed algorithm for optimal sizing

and placement of the DG's 33 and 69-bus systems are selected as test systems. The percentage of power loss is improved from 56.18 to 82.3% and 32.95% to 58% in 69 and 33-bus systems, respectively. A convergence characteristic of CSO is enhanced than ALO and PSO.

References

- Padhy NP, Jena P (2018) NPTEL: Introduction to smart grid-1. <https://nptel.ac.in/courses/108107113/7>
- Atwa YM, El-Saadany EF, Salama MMA, Seethapathy R (2010) Optimal renewable resources mix for distribution system energy loss minimization. *IEEE Trans Power Syst* 25(I):360–370
- Abu-Mouti FS, El-Hawary ME (2011) Heuristic curve-fitted technique for distributed generation optimisation in radial distribution feeder systems. *IET Gener Transm Distrib* 5(2):172–180
- Lee SH, Park JW (2009) Selection of optimal location and size of multiple distributed generations by using Kalman Filter Algorithm. *IEEE Trans Power Syst* 24(3):1393–1400
- Hung DQ, Mithulanthan N, Bansal RC (2010) Analytical expressions for DG allocation in primary distribution networks. *IEEE Trans Energy Convers* 25(3):814–820
- Hung DQ, Mithulanthan N (2013) Multiple distributed generators placement in primary distribution networks for loss reduction. *IEEE Trans Ind Electron* 60(4):1700–1708
- Rao RS, Ravindra K, Satish K, Narasimham SVL (2013) Power loss minimization in distribution system using network reconfiguration in the presence of distributed generation. *IEEE Trans Power Syst* 28(1):317–325
- Abu-Mouti FS, El-Hawary ME (2011) Optimal distributed generation allocation and sizing in distribution systems via Artificial Bee Colony Algorithm. *IEEE Trans Power Del* 26(4):2090–2101
- Wang L, Singh C (2008) Reliability-constrained optimum placement of reclosers and distributed generators in distribution networks using ant Colony System Algorithm. *IEEE Trans Syst Man Cybern C Appl Rev* 38(6):757–764
- Yang XS (2010) Nature-inspired meta-heuristic algorithms, 2nd edn. Luniver Press
- Yang XS, Deb S (2009) Cuckoo Search via Levy flights: Proceedings: World Congress on Nature and Biologically Inspired Computing. IEEE Publications, USA, pp 210–214
- Chakravorty M, Das D (2001) Voltage stability analysis of radial distribution networks. *Int J Electr Power Energy Syst* 23:129–135
- Vovos PN, Bialek JW (2005) Direct incorporation of fault level constraints in optimal power flow as a tool for network capacity analysis. *IEEE Trans Power Syst* 20:2125–2134
- Mirjalili S (2015) The Ant Lion Optimizer. *Adv Eng Softw* 83:80–98
- Yang XS, Deb S (2009) Cuckoo search via Levy flights. In: Nature & biologically inspired computing. NaBIC 2009. World Congress on 2009, pp 210–214
- Yang XS, Deb S (2010) Engineering optimisation by Cuckoo Search. *Int J Math Modelling Numer Optim* 1:330–343
- Jasthi K, Das D (2018) Simultaneous distribution system reconfiguration and DG sizing algorithm without load flow solution. *IET Gener Trans Distrib* 12:1303–1313

Effect of E-training on Employee Performance in IT Industry



Bidush Kumar Sahoo, Smruti Rekha Sahoo, Jyoti Prakash Mishra, and Binod Kumar Pattanayak

Abstract In this contemporary world, the IT industry is growing in faster rate, with proper learning and performance of employee. The main aim of the current study is estimating the performance of the employee in particularly in Indian IT sector. This study was based on the executive employee of IT industry, and response was collected by questionnaire. This study showed that e-training has no significant role on the performance of employee by using structural equation model.

Keywords Confirmatory factor analysis · Structural equation model · Cloud computing · RMSE and e-training

1 Introduction

In modern competitive business environment, the change globally takes place; so every company needs to concentrate on implementing the changes in the required places of their organizations. If the company adopting changes then that company can sustain in this competitive market. One of the major aspects that require to be developed that is the manpower of the organization as without their proficiency and commitment towards organizations cannot touch success. Thus, it is necessary that the employees get trained on the innovative skills and filter the existing skills and knowledge. Training is a very important aspect to measure the performance of employee as well as the management system of an organization. To satisfy all the

B. K. Sahoo (✉)

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

S. R. Sahoo

College of Engineering and Technology, Bhubaneswar, India

B. K. Pattanayak

Siksha 'O' Anusandhan University, Bhubaneswar, India

J. P. Mishra

Gandhi Institute for Education and Technology, Baniatangi, India

e-mail: jpmishra@gietsbsr.com

needs of learning and training, nowadays, there are several methods and instruments have been introduced. Presently, many organizations use e-learning as a main instrument to cope up with the current competence levels. Through that employee's skill levels also match with the changing business environment. E-learning is comprised much more than online learning, and it is extra about how we can interact virtually through employees. Training is taken as an important measure that can be identified the employee performance gaps. According to Armstrong et al. [1], training is the most excellent way to give learning opportunities as well as career development to employees. The rebellion of e-training provides individual elasticity and easy technique to be taught. Thus, it has become an simple way to teach employees individually or groups that provide different learning opportunities, career advancement and it also reduces the cost of learning which is very in limited financial plan. E-training can be grouped into a number of categories, such as online, self-study, self-study with expert, web-based, computer-based (CDROM), and video/audiotapes. E-training can be delivered using streaming video and audio, satellite transmission, audiotape, e-exam and video conferencing and teleconferencing. This objective of this study was to investigate the effect of E-training on employee performance in IT industry.

2 Literature Review

The pedestal of virtual training was based on the viewpoint of flexible accessibility [2]. From side to side, this is a great benefit to the employees who are engaged at the workplace and they can study at any time which is more relaxed to them. According to Collin [3], Servage [4] they have described the e-learning is suitable conceptual inter-disciplinary study. But e-learning remains as the most challenging area of research in this competitive world. When keeping this in mind, it has been exposed that workplace learning is much challenged with the difficult dimensions such as motivation, individual responsibilities, interest and several activities [5].

Illeris [6] explains the some necessary components of workplace learning for better understanding the learning atmosphere such as learning infrastructure, methods of learning, efficiencies of learner and expertise required to execute, social environment along with the team and group in that environment. Cassidy and Eachus [7] experiment that there is a significant and positive relationship between self-efficacy and computing experience.

According to Meelissen and Drent [8], that learner of higher socioeconomic status has positive and significant relationship of computer attitudes than lower socioeconomic status.

According to Bonk [9], in a very short duration, the use of online learning, digital technology, digital notes, and wireless learning trends may convert to e-learning. Schlag [10] confirmed that employers are switching from manuals booklets to e-training and they created in an innovative way and implemented as e-learning in an efficient and cost-effective way. And this was done very successfully by converting the manuals to e-training and which was very easily accessible by employees. Schlag

[10] e-learning and e-training are identical. The success of e-learning has the benefit to both customers as well as employees. To recapitulate best results in e-training, it has to be utilized with limited budget.

Newton and Doonga [11] stated the employers' opinions and justifications regarding e-training for corporate involvement. And the development increases in skills and knowledge, efficiency and efficiency of employees, easiness of execution, consumption of time, and cost effective. On the other hand, there are so many benefits of e-training like it can deliver at any time, any place or to any one. It is just like just-in-time training through e-training the organization can retain effective learners and higher collaboration and interactivity; it will better supervising system on the employee's performance and their continuous progress.

Ozturan and Kutlu [12] investigated that e-learning is statistically significant with employee satisfaction. Many more authors like Yap et al. [13] examined the relationship between e-training and the effectiveness of organizational commitment and employee satisfaction. They confirmed that e-training is very effective and more committed towards organizations. Author Sarmento [14] stated that e-learning has the major role in field of productivity in hotel industry. But some of the researchers like Moller et al. [15] stated that e-learning implementation may productive or unproductive. Though the researchers' illustrate that employees are finding, it is complicated to make balance between quality and progress of e-learning. Here, e-learning looks like a sword which has double-edged. Now, e-learning continues as a dynamic force in many organizational sectors as learning and training instrument.

There are many more studies on e-training which are based on the contact in hostile environments. Korossy [16] shows his theory of knowledge on the level of employee performance competency. Later on Korossys et al. (2003) studied on the same competencies of learning process. In accordance to e-training, a lot of study, analysis, concepts and processes have been approved (Paquette 2006; Xu et al. 2005). The essential elements for e-learning are both ontology and semantic web technologies. Through the help of these elements, they examine a better study and structured the activities for the better learning environments ([17, 18], Vassileva 1995).

Based on the past study surveys, following hypothesis was formulated (Fig. 1).

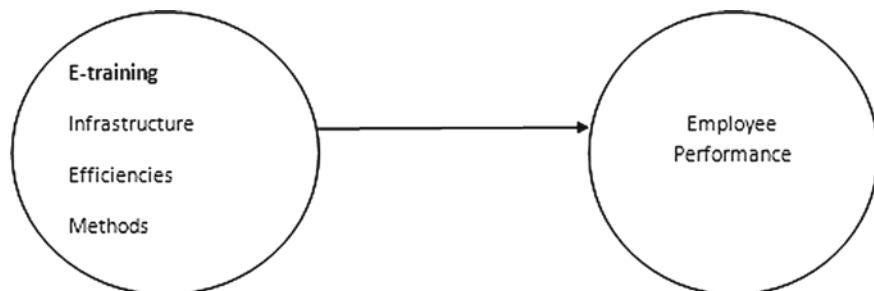


Fig. 1 Conceptual model

- H1: E-training has the impact on employee performance.
 H2: Dimension infrastructure of e-training has significant impact on employee performance.
 H3: Dimension of efficiencies of trainers has positive relation with employee performance.
 H4: Dimensions of training method have positive impact on employee performance.

3 Research Methods and Design

The current research was mainly conclusive in nature and also based on field survey. The sources of data were mainly primary data, and data were customized and structured questionnaire. The study was conducted by collecting responses from executive employees of IT sector in the India country. The study was experimented by collecting responses from executives from different regions of India country like Odisha, Karnataka and Mumbai. This study has selected a total of 153 sample respondents comprising of executive employees from IT industry. Primarily, 250 samples were planned covering only executive employees. Because of less convenience to employees and incomplete responses, and budgetary constraints limited the response size to 153. The questionnaire was planned considering two constructs of study, and for which the items were collected from the scale of Baldwin and Ford [19] and employee performance items were taken from the model developed by Thomson (2008). The responses were collected in 5-point Likert scale.

Pilot study was made to test the reliability and validity of survey before final analysis (Tables 1 and 2).

Table 1 shows that the reliability of construct was found above 0.80, which is normally considered as reliable and enough for final analysis.

Table 1 Scale reliability
Cronbach's Alpha score

Scale reliability Cronbach's Alpha score		
Variables	E-training	Employee performance
Score	0.851	0.864

Table 2 CFA output results of the dimensions of E training and employee performance

Factors	CMIN/df	GFI	CFI	RMSE	AVE	CR
Infrastructure	2.486	0.913	0.923	0.073	0.7227	0.737
Efficiencies	1.395	0.907	0.897	0.0713	0.6804	0.832
Training methods	1.131	0.926	0.939	0.053	0.6845	0.7951
Employee performance	2.813	0.958	0.906	0.062	0.7538	0.795

Table 2 shows the CFA to observe the validity of data. It was found that all the test of fit indices has come between the satisfactory levels. Lastly, data analysis and explanation were made by using regression by using SPSS 20 software and SEM using Amos 20 software package.

4 Results and Discussions

The key purpose of this study was to investigate the effect of e-training on employee performance in Indian IT industry. Though, as per the previous studies, it is hypothesized, the E-training may have the major impact on employee performance. Some of the dimensions of e-training have the major influence of employee performance. Hence, this study used multiple regression analysis and SEM analysis to find the relationship between e-training and employee performance.

Table 3 shows that the dimension infrastructure and efficiencies of E-training have the positive and significant relationship with employee performance at level of 5% while training methods of e-training have no significant relationship with employee performance.

Figure 2 shows the path diagram of SEM, showing the direct relationship between e-training and employee performances. E-training was having two observed variables, and employee performance was having five observed variables.

Table 4 shows the SEM results, and it was found that there is no statistical significant effect of e-training on employee performance.

Overall, it was concluded that, in Indian IT sector, e-training has no significant role to develop employee performance.

Table 3 Effect of e-training dimensions on employee performance by using multiple regression analysis

E-training dimensions	Indices of multiple regression analysis					
	R2	F	Unstandardized coefficient B	Std. error	t	Remarks
Infrastructure	0.166	9.873*	0.119	0.048	2.469*	H2 hypothesis is accepted
			0.155	0.06	2.578*	H3 hypothesis is accepted
			-0.023	0.058	-0.394	H4 hypothesis is rejected

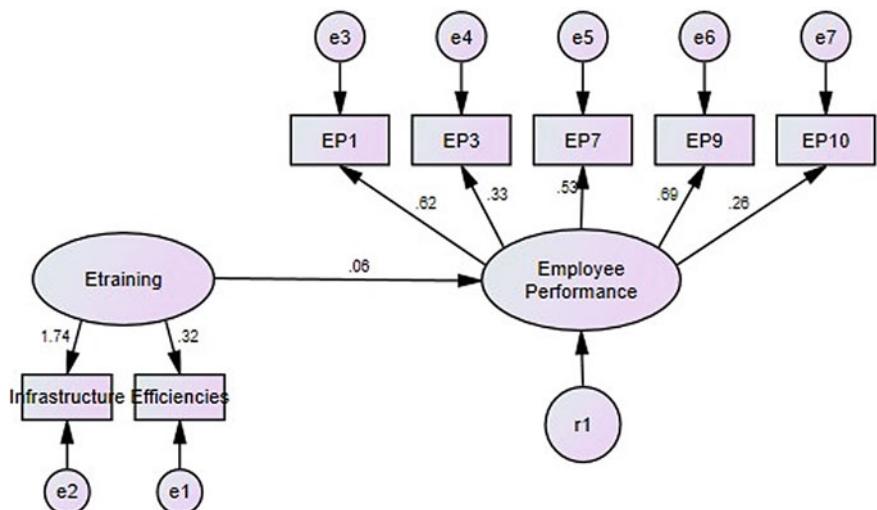


Fig. 2 Structural equation model showing the direct relationship between e-training and employee performance

Table 4 SEM results showing the direct relationship between e-training and employee performance

Direct relationship	Estimate	S.E	C.R	P	CMIN/df	GFI	CFI	RMSEA
Employee \leftarrow E-training performance	0.111	0.105	1.054	0.292	2.11	0.952	0.918	0.073

5 Conclusion

As said by the current study, it was accomplished that the IT industry in India should be focused on employee performance as well as e-training. The organization must give an attention on what the employee demands from the workplace and what should be the e-training methods, training infrastructure and the efficiencies of the trainer, it may have the influence on the performance of an employee. If the industry focuses on these above factors, then the employee performance improves day by day and which will be helpful for developed the organizational productivity.

References

1. Armstrong MB, Landers RN, Collmus AB (2016) Gamifying recruitment, selection, training, and performance management: game-thinking in human resource management. In: Emerging research and trends in gamification. IGI Global, pp 140–165
2. Willems J (2005) Flexible learning: implications of “whenever”, “wherever” and “whatever”

3. Collin K (2006) Connecting work and learning: design engineers' learning at work. *J Workplace Learn* 18(7–8):403–413
4. Servage L (2005) Strategizing for workplace e-Learning: some critical
5. Sadler-Smith E, Evans C, Boström L, Lassen LM (2006) Unraveling learning, learning styles, learning strategies and meta-cognition. *Education + Training*
6. Illeris K (2003) Workplace learning and learning theory. *J Workplace Learn*
7. Cassidy S, Eachus P (2002) Developing the computer user self-efficacy (CUSE) scale: investigating the relationship between computer self-efficacy, gender and experience with computers. *J Educ Comput Res* 26(2):133–153
8. Meelissen MR, Drent M (2008) Gender differences in computer attitudes: does the school matter? *Comput Hum Behav* 24(3):969–985
9. Bonk CJ (2009) The world is open: how web technology is revolutionizing education. Association for the Advancement of Computing in Education (AACE), pp 3371–3380
10. Schlag PA (2001) E-training: an integrative model for the John T
11. Newton R, Doonga N (2007) Corporate e-learning: Justification for implementation and evaluation of benefits. a study examining the views of training managers and training providers. *Educ Inf* 25(2):111–130
12. Ozturan M, Kutlu B (2010) Employee satisfaction of corporate e-training pro- grams. *Procedia Soc Behav Sci* 2(2):5561–5565
13. McGuire, D., Bagher, M., Yap, M., Holmes, M. R., Hannan, C. A., & Cukier, W. (2010). The relationship between diversity training, organizational commitment, and career satisfaction. *J Eur Ind Train*
14. Sarmento M (2010) E-Learning as a tool to improve quality and productivity in hotels. *Worldwide Hospitality and Tourism Themes*
15. Moller L, Foshay WR, Huett J (2008) Implications for instructional design on the potential of the web. *TechTrends* 52(4):67
16. Korossy K (1997) Extending the theory of knowledge spaces: a competence performance approach. *Zeitschrift fur Psychologie* 205(1):53–82
17. Knight P, Tait J, Yorke M (2016) The professional learning of teachers in higher education. *Stud High Educ* 31(03):319–339
18. Stojanovic L, Staab S, Studer R (2001) E-learning based on the semantic web. In: WebNet2001-world conference on the WWW and Internet, Oct 2001, pp 23–27
19. Baldwin TT, Ford JK (1988) Transfer of training: a review and directions for future research. *Pers Psychol* 41(1):63–105

Analysing Odisha Turmeric Price and Other Major Turmeric Producing States of India: A Longitudinal Approach



Bidush Kumar Sahoo, Rojalin Pani, Bikash Chandra Pattanaik, and Saumendra Pattnaik

Abstract This study proposes the price behaviour of turmeric in major producing states of India using a longitudinal research design. The results of the co-integration and vector error correction model (VECM) test are showing the presence of a short-run relationship among some states. Then, Granger causality test illustrates the directional movement between the dependent variables (Odisha) and independent variables (Andhra Pradesh, Karnataka, Maharashtra, Tamilnadu and Telangana). Auto-regressive integrated moving average (ARIMA) model is used to predict weekly turmeric prices for the state of Odisha, India, during the year 2020. From the AGMARKNET Website, price data for all states has been taken, which is from January 2004 to December 2019. The model is validated using MAPE and RMSE. The study outcome is beneficial for the farming community and other related stakeholders.

Keywords Co-integration · VECM · Granger causality · ARIMA

1 Introduction

Turmeric (*Curcuma longa* L.) otherwise famous as “Indian Saffron” is a vital commercial crop grown all over India. Turmeric is used as curry powder in Indian culinary due to its varied forms as a condiment, flavouring and colouring agent. The drug industry and the cosmetic industry are widely using it due to its anti-cancer and anti-viral activities. On religious and ceremonial occasion’s turmeric and its by-product, Kum-Kum is used. Turmeric is becoming an ideal product as a food

B. K. Sahoo (✉)

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

B. K. Sahoo · B. C. Pattanaik

Gandhi Institute for Education and Technology, Bhubaneswar affiliated to Biju Patnaik University of Technology, Rourkela, India

R. Pani · S. Pattnaik

Siksha ‘O’ Anusandhan University, Bhubaneswar, India

colourant due to the growing demand for natural products as food additives. “It is native of South Asia, particularly India, and is the dried rhizome of *Curcuma longa* L., an herbaceous perennial belonging to the family Zingiberaceae”.

India is the major producer of turmeric that contributes 78% of the world’s turmeric production. China, Myanmar, Nigeria, Bangla Desh and others contribute 8%, 4%, 3%, 3% and 4%, respectively. India exports turmeric to countries like the U.S.A, IRAN, U.A.E, Malaysia, etc.

Turmeric producing states of India are West Bengal, Odisha, Meghalaya, Maharashtra, Kerala, Andhra Pradesh, Karnataka, Tamilnadu, Assam, Bihar, Tripura, Uttar Pradesh and Arunachal Pradesh. (<http://www.indianspices.com>). Among them, Andhra Pradesh, Karnataka, Maharashtra, Telangana and Tamilnadu are market competitors of Odisha in the context of turmeric.

According to Gooijer and Hyndman, 2006 forecasting using the time series model is used as a basis for manual and automatic planning in many application domains and is an important statistical analysis technique. “Forecasts are calculated using mathematical models that capture a parameterized relationship between past and future values to express the behaviour and characteristics of a historic time series” [1].

2 Literature Review

Every forecasting process is followed by two phases like diagnostic and testing [2–4]. Before forecasting, appropriate model is established [1]. Price forecasting for agricultural commodities is very important. It is due to its perishable and semi-perishable nature [5, 6]. While reviewing kinds of literature, it is noticed that Boxjenkins ARIMA technique is widely used to forecast price for agricultural commodities [6–9]. Validation for the forecasting model is done by using MSE, MAPE and Theils U coefficient criteria [10–12].

In the above context of research gap analysis, the current study identified the following research problem statement. The research problem of the current work is to study the appropriate techniques for forecasting agricultural commodity price data with accuracy study and comparing different price series to get and suggest proper market integration to members in the agricultural supply chain, which leads to increased efficiency of the agricultural industry.

Research Objective

- To forecast the agricultural commodity price for Odisha with an appropriate model.
- To determine the seasonal index of crop price in the Odisha market.
- To study the market Integration of major respective crop-producing states in India.
- To find out the causal relationship between the prices series of selected crop-producing states in India.

3 Research Methodology

Indian markets are not sufficiently integrated; this is the main argument against the agricultural trade liberalization of India. The current study makes an organized attempt to measure the degree of integration amongst selected agricultural commodity markets in India. Now, when they talk of price forecasting, market integration and causality of agricultural commodity price forecasting arise weekly price for each commodity seems suitable. Hence, it has been decided in the current study to collect commodity price for each state weekly.

3.1 Data

The time series that we consider in our analysis of our study are price data for the crop turmeric for selected states of India. The used data series are taken from January 2004 to December 2019. Data is taken as per the availability. A month is considered as 4 weeks, which implies that we have 768 data points for each state. The source of data is the official Website of AGMARKNET. We have selected major turmeric states; among them, these six state's data is available, namely Odisha, Andhra Pradesh, Karnataka, Maharashtra, Telengana and Tamilnadu. In our study, the Odisha state is taken as a dependent variable. As the data points are secondary are valid as well as reliable.

In this study, results are taken for the log form data, i.e. the first differenced data. The results and discussion section present descriptive statistics, unit root test using ADF test and PP test, market integration using correlation coefficient test, co-integration test, Wald test and Granger causality test. Forecasting is done using ARIMA model for the state of Odisha.

4 Results and Discussion

4.1 Descriptive Statistics and Correlation Coefficient Results

Descriptive statistics like mean, median, maximum, minimum, standard deviation, skewness, kurtosis and Jarque–Bera are presented under Table 1. The calculated value of Jarque–Bera (JB) has been depicted in Table 1.

Table 1 Descriptive statistics of price data for major turmeric producing states of India

States	Mean	Median	Std. Dev	Skewness	Kurtosis	Jarque–Bera	P-value
Odisha	4976.744	4839.965	1851.797	0.925078	4.280725	151.9	0.00*
Andhra Pradesh	5110.339	4806.575	2874.364	1.201377	4.59773	249.779	0.00*
Karnataka	5737.307	5482.37	3026.492	1.180891	4.81791	266.4844	0.00*
Maharastra	6505.953	6066.83	3331.243	0.791095	3.22185	76.57631	0.00*
Telengana	5208.046	5037.83	2921.365	1.137101	4.419805	215.6351	0.00*
Tamilnadu	5957.728	5992.195	3168.087	1.015735	3.99914	153.7546	0.00*

Data for turmeric analysis

4.2 Unit Root Test

It is found that the data series are not normally distributed, i.e. they follow a trend. Now, test for stationarity is to be done. By applying the ADF test and PP test, it is found that data points are non-stationary at their level form. The probability values are given in the respective table which is followed by ADF and PP test results at first difference. The null hypothesis is data series are non-stationary.

The price data for turmeric (Table 2) is also found stationary after first-order differencing. All series that are found non-stationary after first-order differencing are eligible for co-integration test.

Table 2 Unit root result for price data for major turmeric producing states of India

States	ADF test P-values at level data	PP test P-values at level data	Computed value (ADF test)	Critical value @ 1% level (ADF test)	P-value (ADF test)	Computed value (PP test)	Critical value @ 1% level (PP test)	P-value (PP test)
Odisha	0.0309	0.0000	-3.44	-11.98	0.00*	-3.44	-60.66	0.00*
Andhra Pradesh	0.4381	0.3270	-3.44	-35.74	0.00*	-3.44	-35.13	0.00*
Karnataka	0.1719	0.0207	-3.44	-28.22	0.00*	-3.44	-50.20	0.00*
Maharastra	0.4759	0.2247	-3.44	-22.49	0.00*	-3.44	-32.58	0.00*
Telengana	0.3969	0.2842	-3.44	-25.45	0.00*	-3.44	-40.73	0.00*
Tamilnadu	0.4724	0.3214	-3.44	-23.83	0.00*	-3.44	0.00*	0.00*

Note Null hypothesis: There is unit root. Alternative hypothesis: There is no unit root.

*Null

4.3 Market Integration

Tests like the co-integration test, Granger causality test and ARIMA tests can be done as it is found that all the data points for turmeric are found not normally distributed and stationary.

Before going for other tests, correlation test is to be conducted to get the correlation coefficient as to get the strength of a linear fit amongst two variables together geometrically and statistically.

4.4 Johansen Co-integration Test

The test for co-integration can be done with two different techniques like Engle and Granger test of co-integration and Johansen's test of co-integration. It is found that the data series are not normally distributed and series are made stationary after the same order differencing. In this study, Johansen's test of co-integration has been used, and the test result for turmeric crop has been given under Table 3.

Table 3 Johansen co-integration results for major turmeric producing states of India

Johansen co-integration test

Trace test

Hypothesized number of co-integration equations	Eigenvalue	Trace statistic	0.05 critical value	Probability	Significance at 5% level
Odisha	0.168195	418.7372	95.75366	0.0001	Yes
Andhra Pradesh	0.124356	286.6961	69.81889	0.0001	Yes
Karnataka	0.114147	191.4819	47.85613	0	Yes
Maharastra	0.077757	104.5788	28.79707	0	Yes
Telengana	0.058288	46.54012	15.49471	0	Yes
Tamilnadu	0.004842	3.480183	3.841466	0.0621	No

Max-Eigenvalue test

Odisha	0.168195	132.0411	40.07757	0	Yes
Andhra Pradesh	0.124356	95.21419	33.87687	0	Yes
Karnataka	0.114147	86.90308	27.58434	0	Yes
Maharastra	0.077757	58.03873	21.13162	0	Yes
Telengana	0.058288	43.05993	14.2646	0	Yes
Tamilnadu	0.004842	3.480183	3.841466	0.0621	No

From Table 3, it is found that the trace statistics and max eigenvalues are more than the critical value at 0.05 so it can be confirmed as the data series are co-integrated. In this study, the null hypothesis is rejected as the probability value is less than 0.05. The analysis confirmed the existence of co-integration among turmeric markets. The study rejects the null hypothesis as the probability “value is less than 0.05”. As the analysis confirmed the existence of co-integration, the study can move further to identify the relationship that exists between the variables for the long-run and short-run dependency. This can be measured through the VECM model as the data series are co-integrated and implement the Wald test to measure the short-run relationship.

4.5 Wald Test

The test shows that there is a presence of short-run relationship among some states and no long-run relationship is there. However, the analysis is unable to predict the directional movement from a dependent variable to independent variables, so the cause and effect relationship can be measured through Granger causality to predict the future price movement.

4.6 Granger Causality Test

To determine whether the price of one market is useful for forecasting another, the Granger causality test is used. Data series are made stationary before running the Granger causality test. To test the null hypothesis, F -statistics is appointed. The null hypothesis, in this case, is taken as one time series does not granger causes another. If the p -value is less than 0.05, then we reject the null hypothesis, i.e. accept the alternative hypothesis, and if the p -value is more than 0.05, then accept the null hypothesis, i.e. reject the alternative hypothesis. Here, our objective is to forecast Odisha market price, and hence, causality of the Odisha market with other markets is given under Table 4.

4.7 Seasonal Index

For monthly data, a twelve-month moving average is expected to eliminate the seasonal movements if they are of constant pattern and intensity. Hence, for finding out the seasonal indices, the percentage of 12-month moving average is found out. As per the norms of the multiplicative model, each observation in a time series is the product of T , C , S and I (trend, cyclical component, seasonal component and irregular component).

Table 4 Granger causality test for turmeric

Granger causality test		
Null hypothesis	F-statistic	Prob.
TURAP does not Granger Cause TUROD	19.7295	0.00000
TURAP does not Granger Cause TURAP	1.21174	0.29830
TURAP does not Granger Cause TUROD	22.2877	0.00000
TURAP does not Granger Cause TURKAR	12.983	0.00000
TURAP does not Granger Cause TUROD	20.7795	0.00000
TURAP does not Granger Cause TURMAH	1.76908	0.17120
TURAP does not Granger Cause TUROD	24.1226	0.00000
TURAP does not Granger Cause TURTEL	1.37251	0.25410
TURAP does not Granger Cause TUROD	22.2029	0.00000
TURAP does not Granger Cause TURTN	1.39534	0.24840

$$\frac{T \times C \times S \times I}{T \times C} = S \times I$$

$$\text{Seasonal Index} = \frac{\text{The time series at Time } 't' }{\text{Centered Moving Average at Time } 't'}$$

The seasonal index for selected commodities is presented under Table 5.

Table 5 Seasonal index for major turmeric producing states of India

	Andhra Pradesh	Karnataka	Maharashtra	Odisha	Tamilnadu	Telengana
January	103.20	100.66	99.35	97.35	102.06	92.42
February	95.04	92.89	110.42	91.90	97.70	87.86
March	94.06	97.90	111.90	92.11	99.09	87.84
April	96.33	99.86	106.73	92.93	98.8	90.15
May	98.89	100.02	100.83	102.48	98.45	94.03
June	96.93	99.18	92.49	104.61	95.76	94.54
July	100.48	101.92	94.69	100.30	100.63	97.99
August	102.98	101.87	93.9	100.83	101.29	100.97
September	97.92	99.25	91.52	107.03	97.78	100.25
October	100.61	100.11	94.86	103.03	99.56	100.59
November	105.68	102.82	99.6	102.51	104.79	105.29
December	107.87	103.53	103.71	104.93	104.09	106.56

Table 6 ARIMA (p, d, q) model for turmeric price in Odisha

	AR	MA	AIC	SC
Turmeric	1	1	16.3479	16.3606

Table 7 ARIMA (p, d, q) (P, D, Q) model for turmeric price in Odisha

Crops	(p, d, q)(P, D, Q)	AIC	AICc	BIC
Turmeric	(1, 1, 1) (1, 1, 1)	14.5172	14.52013	13.54268
	(1, 1, 1) (2, 1, 1)	14.509	14.51191	13.54077
	(1, 1, 1) (2, 1, 2)	14.4908	14.4938	13.529
	(1, 1, 1) (1, 1, 2)	14.5046	14.50758	13.53644

4.8 Forecasting

ARIMA model non-seasonal (p, d, q) part is identified in Table 6.

Out of the 24 ARIMA models for each series tested for non-seasonal component, some are found suitable as they are having the lowest AIC and SC which are shown in Table 7. From the unit root test for these price series, all are found stationary after the first difference. From these two results, ARIMA (p, d, q) for turmeric is ARIMA (1, 1, 1). Afterwards seasonality test is performed, and results are presented in Table 7.

From Table 7, it is found that turmeric has suitable ARIMA (p, d, q) (P, D, Q) model is ARIMA (1, 1, 1) (2, 1, 2). These models are selected by comparing consecutive models and considering the lowest AIC, AICc and BIC values. AIC, AICc and BIC values for turmeric are 14.4908, 14.4938 and 13.529, respectively.

From Table 8, it is found that turmeric price will persist stably during the year. All the pictorial representations are depicted in Fig. 1.

It is clarified from Fig. 1 that these are the ACF of residuals, standardized residuals, p -values for Ljung–Box statistic of ARIMA model and normal Q - Q plot of standardized residuals. The majority of the correlation function coefficients are within the limit which signifies them as good models. Quantile plots of standardized residuals are in good fit.

4.9 Validation of Models

MAPE and RMSE for turmeric price in the Odisha market are observed as 6.24 and 438.62, respectively (Fig. 2).

Table 8 Price forecasting for turmeric during the year 2020 in Odisha

Months	Weeks	Turmeric	Months	Weeks	Turmeric price (Rs./Qti)
		Price Rs./Qti			
January	First week	5865.85	July	First week	5949.72
	Second week	5907.62		Second week	5988.91
	Third week	5972.55		Third week	6053.04
	Fourth week	5946.98		Fourth week	6027.77
February	First week	5880.58	August	First week	5961.74
	Second week	5937.37		Second week	6000.59
	Third week	6010.88		Third week	6064.55
	Fourth week	5981.91		Fourth week	6039.34
March	First week	5911.22	September	First week	5973.40
	Second week	5938.77		Second week	6012.40
	Third week	5996.80		Third week	6076.43
	Fourth week	5973.93		Fourth week	6051.20
April	First week	5910.93	October	First week	5985.22
	Second week	5954.13		Second week	6024.18
	Third week	6020.34		Third week	6088.19
	Fourth week	5994.25		Fourth week	6062.96
May	First week	5927.19	November	First week	5997.00
	Second week	5965.33		Second week	6035.96
	Third week	6028.94		Third week	6099.97
	Fourth week	6003.87		Fourth week	6074.74
June	First week	5938.11	December	First week	6008.77
	Second week	5976.91		Second week	6047.74
	Third week	6040.84		Third week	6111.76
	Fourth week	6015.64		Fourth week	6086.53

5 Conclusion and Policy Implications

The study says that turmeric in Odisha will remain stable throughout the year. Hence, farmers are advised to increase their cropping area to fetch maximum benefit. Price forecast for the crop will help farmers to take decisions regarding increasing cropping area and storing, selling harvested produce, respectively, at a stipulated period to meet loss due to overproduction. This model of ARIMA can be used for forecasting for the upcoming years. Future research work can be done based on per day activity also. Using the model of the study, forecasting for the next years can be done. A comparative study can be conducted by collecting the same year crop price.

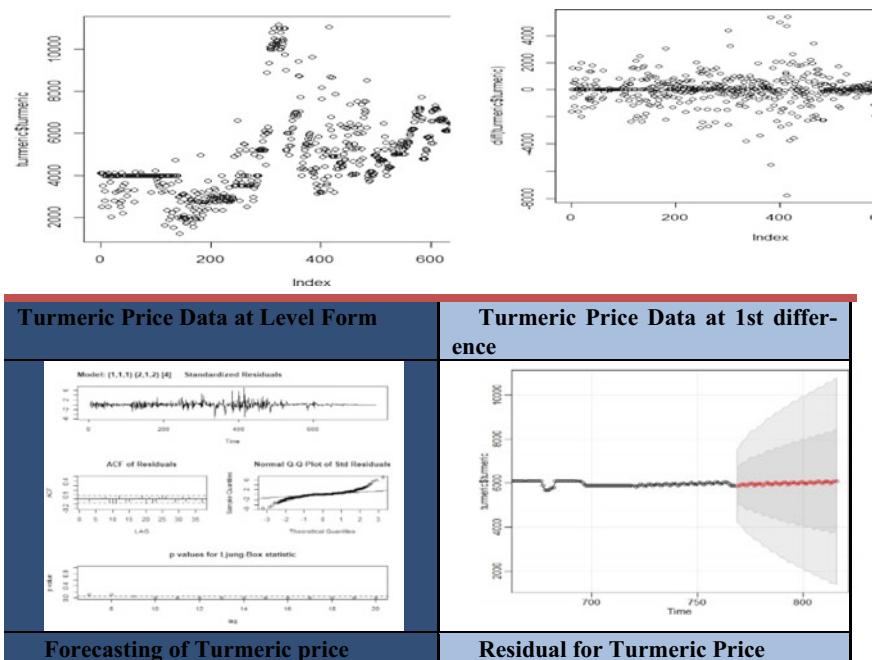


Fig. 1 Turmeric price forecasting for Odisha for the year 2020

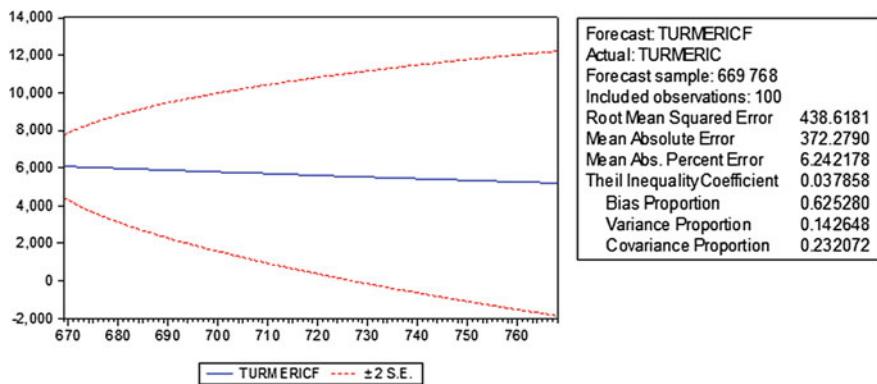


Fig. 2 Validation of model

References

- Mitra D (2017) Hierarchical time-series models for forecasting oilseeds and pulses production in India. *Econ Affairs* 62(1):103–111. <https://doi.org/10.5958/2230-7311.2017.00045.9>
- Pani R et al (2019) Green gram weekly price forecasting using time series model. *Revista ESPACIOS* 40(06)

3. Pani R, Biswal SK, Mishra US (2019) J Adv Res Dyn Control Syst 11:236–243
4. Mohapatra S et al (2018) Price forecasting of groundnut in Odisha. Pharma Innov J 7(3):111–114. Available from: <http://www.thepharmajournal.com/archives/2018/vol7issue3/PartB/7-2-33-193.pdf>
5. Dash A et al (2017) Forecasting of food grain production in Odisha by fitting ARIMA model. J Pharmacognosy Phytochem 6(6):1126–1132. <http://www.phytojournal.com/archives/2017/vol6issue6/PartP/6-5-541-759.pdf>
6. Darekar A, Reddy AA (2017) Price forecasting of pulses: case of pigeon pea. J Food Legumes 30(3):42–46
7. Darekar A, Reddy AA (2017) Price forecasting of maize in major states. Maize J
8. Darekar A et al (2017) Forecasting of common paddy prices in India. J Rice Res 10(1). <http://www.icar-iirr.org/journal%202017-10-13.pdf>
9. Venkatesh Panasa R, VijayaKumari GR, Kaviraju S (2017) Maize price forecasting using auto regressive integrated moving average (ARIMA) model. Int J Curr Microbiol App Sci 6(8):2887–2895
10. Darekar A, Reddy AA (2018) Forecasting wheat prices in India. Forecasting wheat prices in India. Wheat Barley Res 10(1)
11. Ramanujam V, Viswanathan T (2018) An empirical analysis of forecasting volatility of pepper price in the spot market in India. Research Gate
12. JadHAV V et al (2018) Application of ARIMA model for forecasting agricultural prices. J Agr Sci Tech 19:981992

Document Classification Using Genetic Algorithm



Samarjeet Borah, Needhi Kumari Singh, Passang Uden Yolmo,
Rahul Kumar, and Ranjit Panigrahi

Abstract Document classification involves categorization through labelling of documents based on the contents. In this paper, a genetic algorithm-based document classification approach is proposed. It is assumed that there is a predefined group of documents and a set of keywords corresponding to each group. Whenever a new document is taken for classification, a new gene is created consisting of one-third of the unique words in file, followed by crossover and mutation. Finally, a function is used to find the appropriate group for a document based on gene matching. The experiment is conducted using NetBeans IDE 7.3. It has been found that the proposed approach could classify all the files under consideration, with utmost accuracy.

Keywords Document · Classification · Genetic algorithm · Supervised classification · Parameters

1 Introduction

A document can be considered as a record or a medium for capturing of some event or thing to prevent information loss. There are several forms of documents, like traditional handwritten manuscripts, printed document, etc. A document can also be maintained into an electronic format and stored in a computer as one or more file(s). An entire document or individual parts may be treated as single data item. Like files or data, a document can be a part of the database. In this digital edge, millions of documents are generated every day. Therefore, there is always a need for organizing the available documents to facilitate the search process. Document classification is

S. Borah · N. K. Singh · P. U. Yolmo · R. Kumar · R. Panigrahi (✉)
Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India
e-mail: ranjit.p@smit.smu.edu.in

S. Borah
e-mail: samarjeet.b@smit.smu.edu.in

N. K. Singh
e-mail: nidhi.sh@smit.smu.edu.in

an automatic document organization process. It also serves for topic extraction and fast information retrieval or filtering.

1.1 Classification of Documents

Several techniques are available in literature for classification of documents. These classification techniques can be broadly categorized into two categories: *supervised* and *unsupervised*. Some of the supervised classification techniques include statistical methods, neural networks, decision tree including classification and regression trees, genetic algorithms, etc. Unsupervised classification technique primarily includes clustering. There is a third category of techniques known as semi-supervised classification techniques.

Again, based on the working principle, the document classification techniques can be categorized into two types, namely online and offline. Clustering methods can be used to group the retrieved documents into a group of meaningful categories, as is achieved by enterprise search engines such as Northern Light [1] and Vivisimo [2], consumer search engines such as Polymeta [3] and HelioID [4], or open-source software such as Carrot2 [5].

The document classification approach narrated in this paper is a semi-supervised one of supervised and unsupervised techniques. Some of such techniques available in literature are similarity-adapting methods and search-based methods.

1.2 Parameters Used in Document Classification

The documents can be classified based on their subjects or similar other attributes. Some of the attributes include document type, author, printing, text rich, etc. Other various parameters for document clustering may be appropriate numbers of clusters, optimal path to the destination, heuristic cost of the clusters, etc. There are two main philosophies behind subject classification of documents—content-based approach and request-based approach.

1.3 Genetic Algorithm (GA)

Genetic algorithms are exploratory algorithms based on the mechanism of natural selection and natural genetics. They combine survival of the fittest among string structures with structured yet randomized information exchange to form search algorithms. It is basically search heuristics that imitates the process of natural evolution. There are different phases and components which take part in implementation of

GA in any document classification technique. These are also known as GA operators. Components of genetic algorithm include initial population, fitness function, selection, crossover and mutation.

In this work, emphasis is mainly given on document classification using genetic algorithm. It is known that genetic algorithms are best suitable for finding optimal solution. Therefore, on applying this technique in combination with other classification may yield more accurate and improved results.

2 Related Works

Few important and influential works from literature have been included in this section to highlight current knowledge and findings in the domain.

Support vector machine (SVM) is found as a useful tool in document classification [6]. In several studies, it is found that SVM outperforms the contemporary methods. An experiment on one class classification of documents using SVM can be found in [7] with Reuters data set. The experiment was conducted for binary vectors' representation of data, TF-IDF representation and Hadamard which is a variation of TF-IDF. A comparative analysis was made among one-class SVM, naïve Bayes, nearest neighbour and one-class neural network classification. The one-class SVM is found outperforming the others except neural network. Also the performance is found comparable to neural network. Manevitza et al. present a method for using feed-forward neural network for effective classification and salvage of interest in Internet by filtering documents [8]. They explained that this method is superior to one-class SVM, nearest neighbour, Rocchio, naïve Bayes, etc. The usefulness of naïve Bayes classifier in document classification has been presented by Ting et al. in [9]. They used pre-processing and feature extraction for improving the document mining quality. They tried to find out the features which are useful and critical. The hierarchical attention network is found useful for classification of documents [10]. It represents the hierarchical structure of documents, and attention mechanisms are applied both at the word level and the sentence level. It is a successful approach which outperforms already existed methods by a substantial margin.

Some of the major issues involved in document classification are curse of dimensionality, presence of irrelevant and redundant features, etc. During the classification process, the main problem encountered is the selection of parameters out of many complex parameters. Appropriate keyword selection based on content is also a major issue faced by the approaches for a particular group of documents.

3 Proposed Methodology

The proposed method works with a predefined groups of documents. For each group, a set of keywords is maintained separately. The first step was selection of the chromosomes from the individual population and giving a particular keyword to a group of documents. It follows the crossover of documents. Since documents contain a large set of texts, choosing an optimal crossover-point is found difficult. The next task is the mutation, making some random changes in chromosomes. Other problem is taking a maximum gene matching percentage. A general algorithm of the proposed method is given below:

Algorithm

```

Step 1 Begin
Step 2 Predefine groups (e.g. G1, G2, G3, and G4).
Step 3 Define list of keywords for each group
Step 4 List of file (e.g. F1, F2, F3....Fn)
Step 5 For each file do the following:
      a. Create a gene consisting of 1/3rd of the unique word
          in file (say gene1)
      b. Find the match of this gene with the file and not the
          %match.
Step 6 Similarly, create new genes using crossover (take any 2
      genes and break them at 3:2 level and do the crossover;
      fill the remaining length in new genes with random word
      in file).
Step 7 Create new genes using mutation (i.e. replace some ran-
      dom word in the initial gene with remaining word in the
      file) [generate 6 genes using mutation)
      a. If the match percent is greater than the threshold
          (say 60% for file) then: take that file and put it in
          to the category defined by the gene.
      b. Else: Take 2 genes with max percent match and use it
          as parent for second generation.
Step 8 Repeat steps 2 to 6
Step 9 End

```

Based on the algorithm stated above, several pseudocodes have been developed to execute the classification process. All of these pseudocodes are stated below. The outermost function is the documentClassification() which represents the general algorithm. It is accompanied by several other functions such as createNewGene() for creation of new genes, crossover() to perform crossover, mutation() to generate mutated genes, etc. Finally, computeGeneMatchPercentAndGetBestGroupForFile() is used to find the appropriate group for a document based on gene matching.

Function: documentClassification()

1. Start
2. Read all groups and the corresponding keywords-
 - 2.1 For each group g in groups ($G_1, G_2, \dots G_n$)
 - 2.2 groupAndKey(G_1, K_1)
 - 2.3 End loop
3. Read the stop-word list in stopwords
4. For each file f in the file list ($F_1, F_2, F_3, \dots F_n$).
 - 4.1 Read the file f and the corresponding words ($w_1, w_2, \dots w_n$).
 - 4.2 Remove the words from file f which are stopwords and common words(s).
 - 4.3 Remove the word that is empty:
If ($w=null$) $f.remove(w);$
5. Set generationCount = 0;
6. Loop: while(generationCount<noOfGenerations)
 - 6.1 If generationCount=0 (/it's the 1st generation)
 - Create parent genes1: createNewGene(f)
 - Create parent genes2: createNewGene(f)
 - Else
 - Parent genes1=bestParentGeneFromPreviousGeneration
 - Parent genes2=secondBestParentGeneFromPreviousGeneration
7. Add the genes to allGeneList
8. Create 2 new gene using crossover: crossover(allGeneList)
9. Create 4 new gene using mutation:
 - 9.1 Randomly select 4 genes
 - 9.2 For each gene do the mutation: mutation(geneSelected, f)
 - 9.3 Add the new genes to allGene
10. twoBestGenesWithHighestMatch =computeGeneMatchPercentAndGetBestGroupForFile(allGene)
11. if matchPercentOfBestGene>thresholdForDefiningCategory
break;
12. generationCount = generationCount + 1;
13. groupAndFiles.add(groupFoundForBestGene , filename)
14. Print the categorized value to file.
15. Stop

Function: createNewGene(fileWords)

1. Start
2. geneLength := fileWords.length/3
3. Generate random numbers between 1 and fileWords.length
4. Get words fromfileWords at position specified by random number and put in the gene
5. return the new gene.
6. Stop

Function: crossover(allGeneList)

1. Start

2. noOfGenes = allGeneList.size()
3. sizeOfSwap = gene.length/3
4. Generate random start point between 0 and (2*gene.length/3) (startPointOfSwap)
5. endPointOfSwap = startPointOfSwap + sizeOfSwap -1
6. Select 2 random genes(gene_i, gene_j) from allGeneList by generating two different random number in 0 to noOfGenesLength
7. Swap the gene1 and gene2 between startPointOfSwap and endPointOfSwap
8. Add the new genes to allGeneList and return
9. Stop

Function: mutation(geneSelected, fileWords)

1. Start
2. noOfChromosomesToBeMutated = gene.length/3
3. Generate random numbers between 0 and gene.Length
4. For i=0 to noOfChromosomesToBeMutated
 - a. Replace the word at positon i(random number) with the word from file(that has not be used in the genes) specified by the same random position
5. Return the new mutated genes
6. Stop

**Function: computeGeneMatchPercentAndGetBestGroupForFile
(allGenes, groupAndKeyWords)**

1. For each gene in allGenes:
 - 1.1 For each group in groupAndKeywords:
 - 1.1.1 noOfCommonWords := common words between group's key and the gene
 - 1.1.2 tempPercent = noOfCommonWords*100/ gene.length
 - 1.1.3 if(tempPercent>percentMatch)
 - 1.1.3.1 percentMatch = tempPercent
 - 1.1.3.2 tempGroupName = group
 - 1.1.4 end loop
 - 1.2 if(highestMatchPercent<percentMatch)
 - 1.2.1 get the position of two genes with highest
 - 1.2.2 secondhighestmatchPercent = highestMatchPercent
 - 1.2.3 highestMatchPercent = percentMatch
 - 1.3 else if(secondhighestmatchPercent<percentMatch)
 - 1.3.1 secondhighestmatchPercent = percentMatch
 - 1.3.2 get position of genes with 2nd highest match percent
- 2 PercentList = 2 max % computed for gene
- 3 GroupList = 2 corresponding groups for the gene with max % match
- 4 GeneList = 2 corresponding genes
- 5 Return the percentList, groupList and GeneList
- 6 Stop



The screenshot shows a Windows Notepad window titled "bus1.txt - Notepad". The content of the document is as follows:

India's forex reserves soar \$5 billion to \$291 billion.
MUMBAI: India's forex reserves surged by a whopping \$5.04 billion to \$291.3 billion in the week ended November 29 on account of a robust jump in
This is the fourth consecutive week when the country's reserves have jumped and logged one of the sharpest spikes in recent times.
In the week ago period, the reserves had risen by \$2.691 billion to \$286.26 billion.

FCAs, which form a major part of the overall reserves, rose by \$5.071 billion to \$263.736 billion for the week under review, the RBI said.
FCAs, expressed in dollar terms, include the effect of appreciation or depreciation of the non-US currencies such as the euro, pound and yen, held in
During the week under review, the gold reserves remained unchanged at \$21.227 billion, while the special drawing rights rose by \$12.2 million to \$4.
India's reserve position with the IMF fell by \$46.2 million to \$1.905 billion during the period, the apex bank data showed.

Fig. 1 A sample input document

4 Results and Discussion

The proposed document classification technique enables read documents from the disc, write separate scripts for different processes involved such as file reading, document matching, crossover, mutation and other computations involved during the classification process in a convenient way. The reusability feature allows the algorithm to be enhanced later, and more modules can be added with additional features that can be released. The system has been experimented with various documents to examine its accuracy and efficiency. Documents were taken from various domains such as business, sports and technical to check versatility of the algorithm and found desired results. A total of 28 documents are tested, and all documents were classified appropriately. A sample test document is presented in Fig. 1, which is followed by the set of input documents in Figs. 2 and 3.

5 Conclusion

In this paper, a simple document classification task based on genetic algorithms is discussed. This is a simple experiment to understand the usability of genetic algorithms in document classification. However, difficulties are encountered during the crossover of documents belonging to two different sets. During mutation, one-third of the gene length is taken, and the gene has been changed randomly. Again, the work has certain limitations, such as it is experimented on a larger-sized group, domain dependency, etc. The classification which is performed is based on grouping of the documents; it can be modified to document classification without grouping any set of documents. With this enhancement, the processing time of computations can also

<input checked="" type="checkbox"/>	bus1.txt	07/12/2013 03:09	Text Document	2 KB
	bus2.txt	07/12/2013 03:10	Text Document	2 KB
	bus3.txt	07/12/2013 03:11	Text Document	2 KB
	bus4.txt	07/12/2013 03:11	Text Document	4 KB
	bus5.txt	07/12/2013 03:12	Text Document	2 KB
	bus6.txt	07/12/2013 03:12	Text Document	2 KB
	bus7.txt	07/12/2013 03:13	Text Document	1 KB
	bus8.txt	07/12/2013 03:14	Text Document	3 KB
	bus9.txt	07/12/2013 03:14	Text Document	3 KB
	bus10.txt	07/12/2013 03:15	Text Document	2 KB
	bus11.txt	07/12/2013 03:16	Text Document	2 KB
	bus12.txt	07/12/2013 03:17	Text Document	3 KB
	bus13.txt	07/12/2013 03:17	Text Document	2 KB
	bus14.txt	07/12/2013 03:18	Text Document	5 KB
	bus15.txt	07/12/2013 03:19	Text Document	3 KB
	bus16.txt	07/12/2013 03:23	Text Document	3 KB
	bus17.txt	07/12/2013 03:24	Text Document	3 KB
	bus18.txt	07/12/2013 03:25	Text Document	4 KB
	sport1.txt	07/12/2013 02:15	Text Document	3 KB
	sport2.txt	07/12/2013 02:24	Text Document	1 KB
	sport3.txt	07/12/2013 02:31	Text Document	3 KB
	sport4.txt	07/12/2013 02:32	Text Document	4 KB
	sport5.txt	07/12/2013 02:33	Text Document	4 KB
	sport6.txt	07/12/2013 02:33	Text Document	4 KB
	sport7.txt	07/12/2013 02:34	Text Document	2 KB
	sport8.txt	07/12/2013 02:35	Text Document	4 KB
	sport9.txt	07/12/2013 02:35	Text Document	1 KB
	sport10.txt	07/12/2013 02:36	Text Document	4 KB

Fig. 2 Set of input documents

```
{business.txt=[bus1.txt, bus10.txt, bus11.txt, bus12.txt, bus13.txt, bus14.txt, bus15.txt, bus16.txt, bus17.txt, bus18.txt, bus2.txt, bus3.txt, bus4.txt, sports.txt=[sport1.txt, sport10.txt, sport11.txt, sport12.txt, sport13.txt, sport14.txt, sport15.txt, sport16.txt, sport17.txt, sport18.txt, sport2.txt, sport3.txt, sport4.txt, sport5.txt, sport6.txt, sport7.txt, sport8.txt, sport9.txt, sport10.txt]}
```

Fig. 3 Classified documents

be improved further. Incorporating parallel processing technique may enhance the classification process with minimized time complexity.

References

1. Casillas A (2001) Document clustering into an unknown number of clusters using a Genetic Algorithm, pp 1–6
2. Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3(1):1–2

3. Doval D (2001) Automatic clustering of software systems using a genetic algorithm, pp 4–7
4. Naldi MC (2007) Clustering using a genetic algorithm combining validation criteria, pp 1–3
5. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20:53–65
6. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Nédellec C, Rouveirol C (eds) Machine learning: ECML-98. ECML 1998. Lecture notes in computer science (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin. <https://doi.org/10.1007/BFb0026683>
7. Manevitz LM, Yousef M (2001) One-class SVMs for document classification. *J Mach Learn Res* 2:139–154
8. Manevitz L, Yousef M (2007) One-class document classification via neural networks. *Neurocomputing* 70:1466–1481
9. Ting SL, IP WH, Tsang AHC (2011) Is Naïve Bayes a good classifier for document classification. *Int J Softw Eng Its Appl* 5(3):37–46
10. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) 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, association for computational linguistics. <https://doi.org/10.18653/v1/N16-1174>

Analysis of Biomedical and Biological Data

Prediction of Chronic Kidney Disease by Best Accuracy Using Supervised Classification Machine Learning Approach



Diddi Priyanka, Diddi Anusha, T. Anandhi, P. Indria, E. Brumancia, and R. M. Gomathi

Abstract Chronic “kidney disease” mean lasting harm to the kidneys that can decay later few periods of time. On the off chance that damage is exceptionally awful, your kidneys may resign operative. This is called kidney dashing hopes or end-stage kidney disease (ESRD). Kidney disease patients can possibly get into the constant stage, and chronic kidney disease stands a diminishing in kidney work bit by bit. In this way, specialist can to diagnosing of the kidney disease patients. Thus, our is anticipating whether enduring with kidney illness have move in a period of chronic kidney disease or not by indicating best exactness aftereffect of looking at directed arrangement AI calculation continuously applications. The point is to explore AI-based procedures for CKD determining by expectation brings about best precision. The examination of dataset by directed supervised machine learning technique (SMLT) to catch a few data resembles, variable recognizable proof, uni-variate investigation, bi-variate and multi-variate investigation, missing worth medicines and break down the information approval, information cleaning/getting ready, and information perception will be done on the whole given dataset. Moreover, to think about and examine the exhibition of different AI calculations from the given emergency clinic dataset with assessment grouping report, recognize the disarray framework and to classifying information from need, and the outcome shows that the adequacy of the proposed AI calculation method can be contrasted and best exactness with accuracy, callback.

Keywords Chronic kidney disease · Machine learning · Classification algorithms

1 Introduction

Chronic kidney sickness is also called by means of kidney dashing hopes. One out of ten one-on-one overall are experiencing kidney sickness. 10% of the populace overall experiences interminable kidney sickness; one of every five men and one extinct of four ladies from age bunch from sixty to seventy have CKD according

D. Priyanka (✉) · D. Anusha · T. Anandhi · P. Indria · E. Brumancia · R. M. Gomathi
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

to National Kidney Foundation [1, 2]. The present work centers around foreseeing climate an individual is experiencing CKD or not utilizing data mining utilizing machine learning. Information mining [3, 4] is the way toward looking through enormous informational indexes and finding examples and inclines and changing it into justifiable information utilizing data pre-handling, visualization. Machine learning is the field of PC which utilizes factual systems to enable to figure out how to PC [5, 6]. Machine learning can be both supervised and unsupervised learning. Managed learning can be characterized as when we map a contribution to an ideal yield [7, 8]. Machine learning calculations are given to help future forecasts [9, 10]. There are different managed AI calculations similar provision relapse, multiclass characterization, bolster vector machine, K-closest neighbor, Naïve Bayes, random woodland, and some more [11–14]. In unsupervised learning calculation, we train the information utilizing data which is not named [15]. In this calculation, we separate the information into two gatherings dependent on similitude and decreasing the spatiality [16, 17]. Almost basic solo eruditeness movement are bunching calculations. There are different grouping calculations like hierarchical bunching, K-implies grouping, and a lot more [7, 18]. Highlight selection additionally called as factor determination, trait choice, or highlight extraction [2, 19]. It utilizes applicable informational indexes and maintains a strategic distance from repetitive and immaterial information [20–22]. It decreases the dimensionality by utilizing little subsets from the first dataset it helps in simple count of results and achieving shorter preparing times [4].

The random forest grouping strategy gives 100% exactness in expectation contrasted with different classifiers. Zeng et al. (108) have determined that huge information for CKD has imperative chances and needs develop innovation and approach structure to help and give better human services through cross-disciplinary occasions. A robotized deep learning calculation is developed subsequent to looking at different grouping model and its exhibition on the proposed dataset [23, 24].

The proposed work utilized with machine learning algorithm to recognize CKD and non-CKD by utilizing 10 ascribed which contribute to kidney disease.

1.1 Related Work

In 2017, Yildirim et al. [11] made a characterization model for interminable kidney disease forecast on labile collection by utilizing multilayer perceptron [12]. This work centers around eventual outcome of class imbalance in preparing information for anticipating CKD or non-CKD. Multilayer perceptron calculation is used to compute exactness. Resample, SMOTE calculations have been utilized. The work was performed utilizing 0.8 WEKA 3.7.3 Software, and information for look into was taken from UCI machine learning repository of 400 patients with 25 properties. According to the outcome, resample method with multilayer perceptron was progressively precise; however, for performance case, extended grinder instance algorithm was quick with instance of 0.0509.

Mohammad et al. [13] performed mining of activity rules for chronic kidney disease prediction utilizing Naïve Bayes. Gullible Bayes with OneR trait selector was utilized for expectation CKD status of a patient [14]. The thought was to choose a subset from input information by end inactive information which conveyed practically no prescient information. Informational collections were taken from UCI ML repository. The outcome and examination proposed Naïve Bayes with OneR with most noteworthy improved exactness and furthermore diminished number of credited to 80% which is 05 from aggregate of 25 ascribes contrasted with other property evaluators.

Radha et al. [14] in 2016 played out a finding of chronic kidney illness utilizing machine learning [23] utilizing R instrument and calculations like back propagation neural system, random backwoods, radial basis work, ANN. The information for this exploration was medicinal reports of patients taken from various labs in South India. They have utilized 1000 occurrences with 15 CKD related traits. Their model is assessed on various estimates like sensitivity, accuracy, and specificity. The trial results demonstrated that radial basis function performed superior to different calculations and got an exactness of 85.3%.

Wickramasinghe et al. [23] in 2017 proposed dietary expectation of patients with CKD by considering blood potassium level [25]. Their work proposes diet designs by taking patients potassium level in thought. The trial is performed utilizing multiclass jungle, forests, and neural systems in Microsoft Azure machine learning studio. In their outcomes, multiclass decision forest performed with a precision of 99.17%.

Shaikhina et al. [24] in 2017 created characterization model for result forecast in neutralizer contrary kidney transplantation [26]. The base target is to autonomously recognize chance connected with kidney transplant inside initial 30 days of transplant that how much the kidney is acknowledged by the patient's body. This work would help specialists to anticipate results of kidney transplant at early stage. Decision tree, random forest characterization calculations were utilized for this expectation. Their work for anticipating kidney transplant disappointment is performed with a precision of 85%.

1.2 Existing System

Constant kidney disease (CKD) is a genuine general well-being condition worldwide that attached to horrendous well-being results, especially in low-to-center pay nations where millions pass on because of absence of moderate treatment. CKD is a long-haul condition incited by harm to the two kidneys. Kidney harm alludes to any sort of kidney pathology that gives the likelihood to diminish the limit of kidney capacities, especially the decrease in glomerular filtration rate (GFR). Kidneys have a huge number of small veins fill in as channels to expel squander items from the blood. Prescient examination empowers us to present the ideal subset of parameters to nourish machine learning to manufacture a lot of prescient models. This examination begins with 24 parameters notwithstanding the class property and winds up by 30%

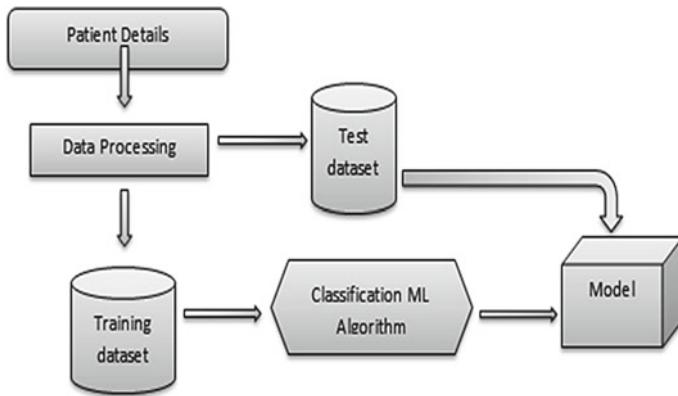


Fig. 1 System architecture

of them as perfect sub set to anticipate chronic kidney disease. An aggregate of four AI-based classifiers has been assessed inside a managed getting the hang of setting, accomplishing best results of AUC 0.995, affectability 0.9897, and explicitness 1.

1.3 Proposed System

The point is to explore AI-based methods for CKD anticipating by forecast brings about best exactness. The investigation of dataset by machine learning (SMLT) to catch a few data resembles, variable recognizable proof, uni-variate examination, bi-variate and multi-variate examination, missing worth medications and break down the information approval, information cleaning/getting ready and information representation will be done on the whole given dataset. The outcome shows that the viability of the proposed machine learning algorithms method can be contrasted and best exactness with accuracy, recall, and *F1 Score* (Fig. 1).

1.4 Module Description

1.4.1 Variable Identification Process/Data Validation Process

Approval methods in AI are used to return the botch pace of the machine learning (ML) model, which can be thoughtful as neighboring the authentic fault pace of the dataset. In the event that the content mass is sufficiently huge to be demonstrative of the public, you may not need the support methods. Notwithstanding, in genuine situations, to occupation with tests of accumulation that may not be a documented agent of the number of inhabitants in given dataset. To finding the missing worth,

copy worth and portrayal of information type whether it is glide variable or whole number. The example of information is used to give a fair-minded assessment of a model fit on the preparation dataset while tuning model hyper parameters. The accompanying outline is given dataset.

1.4.2 Data Validation/Cleaning/Preparing Process

Transfer in the collection bundles with good deal given dataset. To work the variable identifying amount by message shape, message type and assessing the missing qualities, copy regard. An approval dataset is an example of info broken out from fix your framework that is utilized to give a standard of kind expertise, while standardization model's and systems that you can use to apply approving and test datasets while assessing your models.

1.4.3 Exploration Data Analysis of Visualization

Information representation is a significant expertise in applied insights and machine learning. Insights does without a doubt center around quantitative portrayals and estimations of information. Information representation gives a significant suite of apparatuses for increasing a subjective comprehension. This can be useful when investigating and finding a good pace dataset and can help with distinguishing designs, degenerate information, anomalies, and considerably more. With a little space information, information perceptions can be utilized to express and show key connections in plots and graphs that are more instinctive and partners than proportions of affiliation or hugeness. Information perception and exploratory information investigation are entire fields themselves, and it will suggest a more profound jump into some the books referenced toward the end.

1.4.4 Outlier Detection Process

Many machine learning procedure is light-handed to the scope and count of trait regard in the input message. Anomalies in input message can stand and deceive the activity process of machine learning process transfer around person fix business, more skillful models and eventually more fatal outcomes. Indeed, equal earlier discerning models are set up on fix information, and person can take about leading portrait and, thusly, unreal version of collected information. Anomalies container stand the summary spreading of typical choice in explain measuring like mean and standard deviation and in plots, for example, bar chart and scatterplots, packed the body of the message. At extended past, example can communicate to position of message example that are applicable to the issuance, for example, abnormalities on history of misrepresentation identification and PC certificate.

1.4.5 Comparing Algorithm with Prediction in the Form of Best Accuracy Result

It is essential to look at the presentation of various diverse machine learning calculations reliably, and it will find to make a test saddle to analyze numerous distinctive machine learning calculations in Python with scikit-learn. It can utilize this test outfit as a layout all alone machine learning issues and add more and various calculations to look at. Each model will have distinctive execution attributes. Utilizing resampling strategies like cross approval, you can get a gauge for how exact each model might be on inconspicuous information. It should have the option to utilize these appraisals to pick a couple of best models from the suite of models that you have made. When have another dataset, it is a smart thought to picture the information utilizing various systems so as to take a gander at the information from alternate points of view.

2 Conclusion

The explanatory procedure began from information cleaning and handling, missing worth, exploratory examination lastly model structure, and assessment. The best exactness on open test set is higher precision score is will be discovering. This brings a portion of the accompanying bits of knowledge about analyze the CKD infection. To give an expectation model, the guide of man-made consciousness to improve over human exactness and furnish with the extent of early location. With our proposed forecast model, we intend to make it simpler for specialists to do exact finding and expectation of CKD, the two of which have human restrictions because of the technique for discovery of CKD that is utilized at this point. It tends to be gathered from this model, and zone examination and utilization of AI strategy are valuable in creating forecast models that can enable a specialist to diminish the long procedure of determination and kill any human mistake. To isolate crafted by location and expectation strategies to recognize and quantify the zone of cerebrum that is influenced due to CKD and utilize that information in AI to make the forecast model with exactness is higher contrasting different models.

References

1. Serpen AA (2016) Diagnosis rule extraction from patient data for chronic kidney disease using machine learning. *Int J Biomed Clin Eng (IJBCE)* 5(2):64–72
2. Madni HA, Anwar Z, Shah MA (2017) Data mining techniques and applications—a decade review. In: 2017 23rd international conference on automation and computing (ICAC), Sept 2017. IEEE, pp 1–7
3. Ghotra B, McIntosh S, Hassan AE (2017) A large-scale study of the impact of feature selection techniques on defect classification models. In: 2017 IEEE/ACM 14th international conference on mining software repositories (MSR). IEEE, pp 146–157

4. Ma L, Li M, Gao Y, Chen T, Ma X, Qu L (2017) A novel wrapper approach for feature selection in object-based image classification using polygon-based cross validation. *IEEE Geosci Remote Sens Lett* 14(3):409–413
5. Selvan MP, Chandra Sekar A, Lokeshwaran S, Kalai Selvan P (2016) Query optimization technique for videos in relational database. *ARPN J Eng Appl Sci* 11(13):8447–8449
6. Mary AVA, Samuel SJ (2016) Laccase: production and purification by *Pseudomonas aeruginosa*. *J Pure Appl Microbiol* 10(2):1613–1618
7. Dutt A, Ismail MA, Herawan T (2017) A systematic review on educational data mining. *IEEE Access* 5:15991–16005
8. Dean J, Patterson D, Young C (2018) A new golden age in computer architecture: empowering the machine-learning revolution. *IEEE Micro* 38(2):21–29
9. Nagarajan G, Minu RI, Devi AJ (2020) Optimal nonparametric Bayesian model-based multimodal BoVW creation using multilayer pLSA. *Circuits Syst Signal Process* 39(2):1123–1132
10. Jacob, Prem T, Pravin A, Nagarajan G (2019) Efficient spectrum sensing framework for cognitive networks. *Concurr Comput Pract Exp* e5187
11. Yildirim P (2017) Chronic kidney disease prediction on imbalanced data by multilayer perceptron: chronic kidney disease prediction. In: 2017 IEEE 41st annual computer software and applications conference (COMPSAC), vol 2. IEEE, pp 193–198
12. Gunaratne WHSD, Perera KDM, Kahandawaarachchi KADCP (2017) Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In: 2017 IEEE 17th international conference on bioinformatics and bioengineering (BIBE). IEEE, pp 291–296
13. Dulhare UN, Ayesha M (2016) Extraction of action rules for chronic kidney disease using Naïve Bayes classifier. In: 2016 IEEE international conference on computational intelligence and computing research (ICCIC). IEEE, pp 1–5
14. Ramya S, Radha N (2016) Diagnosis of chronic kidney disease using machine learning algorithms. *Int J Innov Res Comput Commun IJIRCCE* 4(1):812–820
15. Pravin A, Jacob TP, Nagarajan G (2019) Robust technique for data security in multicloud storage using dynamic slicing with hybrid cryptographic technique. *J Ambient Intell Humaniz Comput* 1–8
16. Selvan MP, Viji Amutha Mary A, Jancy S (2019) Automatic user domain classification based on support vector machine (SVM). *J Comput Theor Nanosci* 16(8):3327–3331
17. Asha P, Jebarajan T (2015) Association rule mining and refinement using shared memory multiprocessor environment. In: Artificial intelligence and evolutionary algorithms in engineering systems. Springer, New Delhi, pp 105–117
18. Thirumalai C, Duba A, Reddy R (2017) Decision making system using machine learning and Pearson for heart attack. In: 2017 international conference on electronics, communication and aerospace technology (ICECA), vol 2. IEEE, pp 206–210
19. Prince Mary S, Lakshmi SV, Anuhya S (2019) Color detection and sorting using internet of things machine. *J Comput Theor Nanosci* 16(8):3276–3280
20. Jayashree R, Christy A (2018) Re-ranking in user-driven reputation systems with splay tree. *Int J Mater Prod Technol* 56(1–2):3–22
21. RajaPavan V, Suhas BSS, Kanni A (2017) Tracking the calorie consumption and location using existing sensors and mobile applications. In: 2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC), Feb 2017. IEEE, pp 562–566
22. Indra MR, Govindan N, Satya RKDN, Thanasingh SJS (2020) Fuzzy rule based ontology reasoning. *J Ambient Intell Humaniz Comput* 1–7
23. Wickramasinghe MPNM, Perera DM, Kahandawaarachchi KADCP (2017) Dietary prediction for patients with chronic kidney disease (CKD) by considering blood potassium level using machine learning algorithms. In: 2017 IEEE life sciences conference (LSC). IEEE, pp 300–303
24. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N (2017) Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed Signal Process Control*

25. Sheng G, Hou H, Jiang X, Chen Y (2018) A novel association rule mining method of big data for power transformers state parameters based on probabilistic graph model. *IEEE Trans Smart Grid* 9(2)
26. Choi T-M, Chan HK, Yue X (2017) Recent development in big data analytics for business operations and risk management. *IEEE Trans Cybern* 47(1):81–92

System for Full Immunization Coverage



Akanksha Pradhan, Rutuja Patil, Vrushali Alugade, and Varsha Patil

Abstract Universal Immunization Program in India reveals a coverage below 50% in most heavily populated states of the country (Bihar, Uttar Pradesh, West Bengal), and most of the health indicators are low in states where immunization coverage is low. The current mechanism for keeping track of immunization details is manual and tedious. To overcome this issue and maintain integrity, a system is proposed to monitor the schedule from pregnancy of the mother till 5 years of age of child. The schedule may include the due dates for check-up during pregnancy as well as the dates of vaccination till 5 years of age of child. The system will provide notification to the beneficiaries through SMS/Emails. DEO/Health workers can directly capture the beneficiary data into the system using the web portal. A unique barcode will be generated and assigned to each beneficiary to track the immunization schedule. A separate login access is provided to the beneficiary to check their profile details and get information about the scheduled check-ups and immunization campaigns. Dashboards displaying visualization of immunization coverage of various regions can help the government to plan their campaigns. System is extended to collect blood test samples from different labs, and CNN models can be used for detection of Malaria in particular areas, thus, reducing delays in taking action in infected regions. The proposed system connects different healthcare sectors and helps in digitization of the paperwork and ease in maintaining records, thus, providing better healthcare services.

Keywords Immunization · Barcode scanning · Convolutional neural network (CNN) · Database · Image classification

1 Introduction

India has the largest number of births in the world at more than 26 million a year and also accounts for more than 20% of child mortality worldwide. Nine million

A. Pradhan · R. Patil · V. Alugade (✉) · V. Patil

Department of Computer Engineering, SIES Graduate School of Technology, Navi Mumbai, India

immunization sessions are organized each year to target these infants and 30 million pregnant women for routine immunization (RI). Though some improvement has taken place in the past few years, the country still accounts for the largest number of children who are not immunized (approximately 7.4 million). It has been observed that every year in India, 5 Lakh children die due to vaccine-preventable diseases. Another 89 Lakh children remain at risk, because they are either unimmunized or partially immunized against vaccine-preventable disease [1, 2]. The immunization coverage is low because of various reasons such as (i) Lack of awareness, (ii) Parents miss the vaccination schedule of their child due to busy and hectic lifestyle, (iii) The vaccination record handed over to the parents are paper based which are not formally maintained and are being lost by parents [3, 4]. The present system of maintaining immunization detail is manual. The updating of the system, post immunization is done manually on the Reproductive and Child Health (RCH) register by the health workers. The private sectors report the immunization coverage through manual data collection which is then fed into the RCH system by the Data Entry Operators (DEOs). The Government of India has installed Arogya centres in order to keep the track of immunization coverage of the area assigned to that Arogya centre. Apart from this, various immunization campaigns are organized by the Arogya centres. Thus, the Arogya centre has to keep track of the immunization details generated during the immunization campaigns organized by them as well as the details generated at each private and government hospitals under it. The details generated at the hospitals are forwarded to the Arogya centre in pdf form. The Nurses of the Arogya centre have to feed in the data forwarded by the hospitals as well as data generated during immunization campaigns into the RCH portal manually. In order to carry out an immunization campaign, the nurse has to initially carry out a survey in order to find out the total count of beneficiaries in that area. After the survey, a call list is generated which lists out the name of the beneficiaries. The nurse then has to notify the beneficiaries regarding the date of vaccination. On the day of vaccination, the nurse maintains a manual record of the upcoming beneficiaries and their vaccination details, inventory, schedule immunization campaigns, etc. Due to various factors such as manual maintenance, lack of coordination, absence of digitization and absence of interaction with the public leads to low immunization coverage [2, 3]. To overcome all these problems, a centralized full immunization coverage system is designed. Proposed system keeps the track of immunization schedule from pregnancy till five years age of a child. It would monitor the due dates of check-up during pregnancy as well as due dates of immunization of a child. The check-up and immunization dates will be conveyed to the user through SMS or email. By using software barcode technique, vaccination information can be easily updated and also inventory can be easily maintained. The system provides visualization of immunization coverage of various areas which would help government officials to plan campaigns. Each Private Hospital, Government Hospital as well as Arogya centres can feed in the data into the system at their end. Users can view his/her profile which displays the overall details of the check-ups and vaccinations [5, 6]. To do smooth and efficient functioning of the nurses as well as the health workers, barcode technique is incorporated by the proposed system for maintaining and updating data. Proposed system uses

CNN classification model to predict malaria based on blood sample images to take precaution against malaria.

1.1 Review of Literature

Unicef Immunization in India [1]: It talks about the immunization coverage in India over the past few years. Patra [3] proposed Universal Immunization Program in India: The determinants of childhood immunization. According to this research, the immunization coverage differed on the basis of gender, age and education of the population. Vaccination should not be considered the sole responsibility of the health sector. Other sectors should also be informed and influenced by public health considerations. Sharma [2] proposed immunization coverage in India which focuses on examining the immunization coverage in India from 1980 to 2004 in the states of Uttar Pradesh and Uttarakhand. It has been concluded that Uttarakhand has not reached the goal of Universal Immunization coverage. Despite receiving the complete schedule for immunization, only 30% underwent immunization coverage. Ventola [5] proposed immunization in the USA. Recommendations, barriers, and measures are defined to improve compliance. This paper consists of two phases. First phase covered childhood and adult immunization in the USA, whereas the second part covered adult vaccination. The rate of diseases has significantly decreased in Western countries due to immunization strategies aimed at infants and children. Wilson et al. [6] proposed immunization coverage assessment in Canada. The primary objective of the paper was to examine the techniques used to determine immunization coverage in Canada. It was observed that immunization programs are complex and involve many parties. Datta et al. [7] proposed a study to find out full immunization coverage among children aged 12–23 months in rural areas of Tripura. This paper analyses the factors which act as obstacles in achieving 100% full immunization coverage. They have used Lot Quality Assurance Sampling (LQAS) for the same. Panda [8] proposed a study to find out the temporal trend and inequality in India. The study using the NFHS studied the rich-poor inequality, rural–urban inequality and gender-related inequality to understand the equity gap in immunization among regions of India. Srivastava and Shankar [4], a study of immunization coverage and its determinants among under five children residing in urban field practice areas of Karnataka, India. The percentage of children fully immunized was found to be 83%. The main reason for failure of immunization was found to be lack of awareness of the immunization schedule and negligence of parents. Song and Miled [9], Digital Immunization Surveillance: Monitoring Flu Vaccination Rates Using Online Social Networks. In the paper, they demonstrate the feasibility of a real-time, low-cost and labour-efficient alternative monitoring method to the flu vaccination rate surveillance system. This alternative is based on social network data and specifically Twitter data. They demonstrate the viability of this approach by measuring the level of correlation between the frequency of “tweets” mentioning flu vaccinations and HHS data on flu vaccination rates. Siringi and Sharma [10] child immunization using data analysis.

The paper states that data mining helps to understand large complex data sheets in less time, with few risks and remove redundant data to make use of the relevant information. Tropea and Fedele [11] present a comparison between five different classifiers (multi-class logistic regression (MLR), support vector machine (SVM), k-nearest neighbour (kNN), random forest (RF), and Gaussian Naive Bayes (GNB)) to be used in a convolutional neural network (CNN) in order to perform images classification. The dataset composed of images of objects belonging to 256 widely varied categories called Caltech 256. Kakde et al. [12] propose that the main reason behind the performance of convnets is that they are inspired from the mammal's visual cortex. Convolutional Neural Network and an evolutionary approach on highway convolutional neural networks on the basis of train loss, test loss, train accuracy and test accuracy. The models were tested on two datasets that are WANG dataset and Simpson's dataset. Based on experiments, they concluded that in WANG dataset, Alexnet model achieved the highest test and train accuracy. Evolutionary Highway CNN has the least train loss, CNN has the least test loss and in Simpsons dataset, Evolutionary Highway CNN has the Highest test and train accuracy and Evolutionary Highway CNN has the least train and test loss. Sharma et al. [13] conducted an empirical analysis of the performance of popular convolutional neural networks (CNNs) for identifying objects in real-time video feeds. The most popular convolution neural networks used for object detection and object category classification from images are Alex Nets, Google Net and ResNet50. Their analysis demonstrated that Google Net and ResNet50 were able to recognize objects with better precision compared to Alex Net. Moreover, the performance of trained CNN's varied substantially across different categories of objects.

2 Proposed Immunization Method

Proposed Centralized Full Immunization Model is divided into three parts:

(a) Monitor and schedule the immunization coverage

Whenever a new beneficiary visits a hospital or an Arogya centre, the beneficiary registers and unique login id and password is provided. Once the beneficiary gets registered to the hospital, the corresponding Arogya centre can track the record of that beneficiary. Arogya centre will send notification through SMS to the user about the immunization on the corresponding due date. The barcode on the vaccine will be scanned, and the recipient's detail will get updated post vaccination. Also beneficiaries data will be accessed by the nurse visiting the area where the beneficiary resides for the purpose of call list generation. Various functionalities provided by different components are given as below: beneficiary login, profile and vaccination info.

To ease out the functioning of the nurses as well as the health workers, we are generating a system which will incorporate barcode technique. For every beneficiary, a unique barcode is generated in code 128 format which is a high density linear

barcode. Whenever any beneficiary visits the Arogya Center, the ANM only has to scan the barcode, and their details will be fetched and updated automatically into the database. As data will be updated on a regular basis notification regarding due dates can be sent to the users. Barcode technique will be also helpful in maintaining the inventory of vaccines.

ANMs have the following functions:

ANM (Auxiliary Nurse Midwifery)/Hospital:

Update Vaccination Record, Call list, Monthly Report

ANM Head:

Monthly and Yearly Report: The ANM Head can generate monthly and yearly reports.

(b) **Visualization of immunization coverage**

In this phase, the visualization for different users of the system based on their privileges is represented. This visualization will be generated using Tableau Software. Visualization of immunization coverage of various areas can help the government to take strategic decisions to plan their immunization coverage campaigns. It will also give an overview to manage the inventory.

Government Official has the following accesses:

District Officer: Access to visualization of immunization coverage in a particular district.

State Officer: Access to visualization of immunization coverage in a particular state.

Central Officer: Access to visualization of immunization coverage in the entire country.

(c) **Prediction of malaria using blood samples**

Malaria, a life threatening disease is caused by Plasmodium parasites. According to the World Malaria Report 2017, India accounted for 6% of all malaria cases in the world, 6% of the deaths. Study published in The Lancet on 20 November 2010 has reported that malaria causes 205,000 malaria deaths per year in India before age 70 years. The report says that 90% of the deaths were recorded in rural areas, of which 86% occurred at home without any medical attention. There exist many techniques to detect malaria parasites in a patient such as clinical diagnosis, microscopic diagnosis, rapid diagnostic test (RDT) and polymerase chain reaction (PCR). All these diagnostic methods require laboratory settings, and their efficiency largely depends on the human expertise. Such expertise is inadequately present in unreachd remote areas where malaria is predominant. In such circumstances, CNN model can prove to be helpful in early diagnosis of malaria parasites which will enable early treatment.

Methodology

Dataset: Malaria dataset contains 27,558 sample images classified into two groups as parasitized and uninfected. Data was taken from kaggle website.

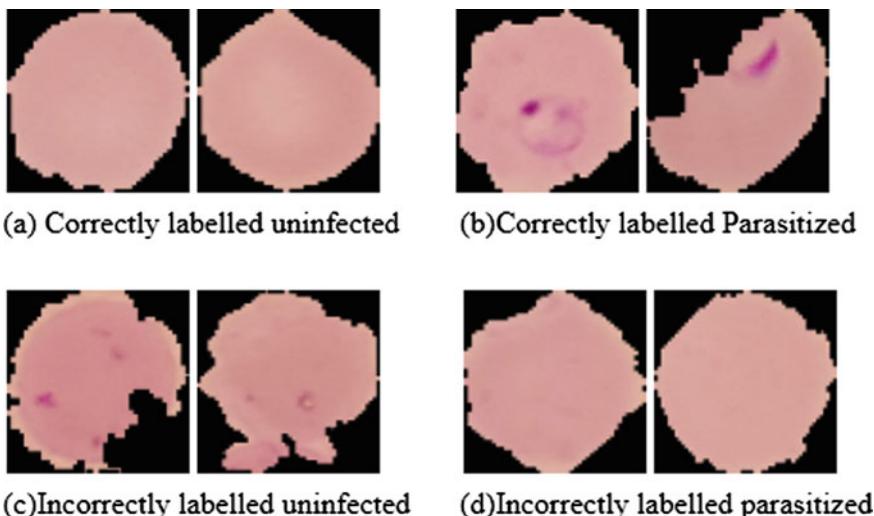


Fig. 1 Correctly and incorrectly labelled blood samples. *Source* From Malaria Cell Images Dataset, by Arunava, 2018, <https://www.kaggle.com/iarunava/cell-imagesfor-detecting-malaria>

Data Pre-processing: The efficiency of the model depends on the data fed for training. Therefore, data pre-processing is vital before performing experiments. In the dataset, we found that there were correctly labelled uninfected images, correctly labelled parasitized images, incorrectly labelled uninfected images and incorrectly labelled parasitized images. The mislabelled data was manually corrected as per the presence and absence of parasites (Fig. 1).

CNN Model: The CNN model is constructed using Keras library for image classification using a sequential model. There are four Conv2D layers each containing 32 nodes. The size of the filter matrix is 3×3 . The activation function used is rectified linear activation. A flatten layer is used that acts as a connection between Conv2D layer and dense layer. Dense layer is the output layer. The model is trained by using infected and uninfected images of malaria.

3 Experimental Results

Proposed centralized full immunization system is implemented with the help of the following software and hardware. Operating System Windows 7, Ubuntu 14 is used. For front end Nodejs, express framework is utilized. Backend is managed with MongoDB, Nexmo API (SMS notifications), and Tableau (Visualization) are also used. For scanning the barcode, QuaggaJs is used. It is entirely written in JavaScript and supports real-time localization. It can be used for decoding different types of barcode such as EAN, Code 128, Code 39 and many more. Using Quagga, barcode scanning

can be performed in 3–4 s. Visualization is generated using business intelligence tool Tableau. The data was taken from MyGov website. Then the data was subjected to the ETL process. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area and then loads it into the data warehouse system. The data was then loaded to the Tableau software where visualization was performed based on key performance indicators (KPI) (Fig. 2).

Similarly, visualization can be created for district officers which represent the count of various vaccinations in a particular district. To check if the sample collected contains malaria or not, CNN model is incorporated to make quick decisions. Sample test is an image given to the system, which will predict if the sample is positive or negative for malaria. The CNN model is constructed using Keras library for image classification using a sequential model. There are four Conv2D layers each containing 32 nodes. The size of the filter matrix is 3×3 . The activation function used is rectified linear activation. A flatten layer is used that acts as a connection between Conv2D layer and dense layer. Dense layer is the output layer. The model is trained by using infected and uninfected images of malaria. The total number of images used as sample input was 27,558. The results obtained were as in Table 1 (Figs. 3 and 4).



Fig. 2 State officer dashboard which gives the immunization coverage of BCG and DPT1 vaccination for the year 2013–16 of all the districts within the state of Maharashtra

Table 1 Validation loss and accuracy of the CNN model which is used to predict if a blood sample image is infected by malaria or not

Validation loss	0.2894
Validation accuracy	0.8821
Loss	0.3602
Accuracy	0.829

Fig. 3 Uninfected blood sample. From C100P61ThinF_IMG_20150918_144104_cell_128.png, by Arunava, 2018, <https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria> [14]



Fig. 4 Infected blood sample. From C100P61ThinF_IMG_20150918_144104_cell_162.png, by Arunava, 2018, <https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria> [15]



4 Conclusion

Unlike existing manual immunization system, proposed centralized full immunization vaccination system maintains the vaccination data on real-time basis, reduces the manual work of the nurses, provides visualization for clear view of immunization coverage in various areas and, thus, promoting the government officials to organize campaigns in order to increase the immunization coverage. Barcode scanning eases maintaining of record. In the proposed system, malaria detection by CNN is used to get immediate results. Future work is to include models for various tests. Use the concept of data mining and machine learning algorithms to find patterns from data and take necessary actions to achieve full immunization.

5 Future Scope

Arogya centres and the clinics are the first point of contact when a person feels unwell mostly in rural areas where health facilities are not up to the mark. The Arogya centres and the general clinics are not well equipped with the medical instruments as well as the required staff. Due to which most of the time either the disease goes undetected or is mistaken for some other which further leads to delay in the proper treatment. To overcome such situations, the system can be further extended for predicting the

diseases based on the symptoms which are recorded in the real time. Once the system predicts the disease based on the given symptoms, the required tests can be carried out, and the treatment can be started at the earliest. This feature will prove effective in the phase of epidemics as during epidemics tests are required to be carried out on a large scale and quick results are expected in order to control the spread. It can be further used to send region specific messages to prevent the spread of contagious diseases. Dedicated vaccination campaigns for specific diseases can also be integrated into the system. Also, the CNN model can further be extended to detect pneumonia, eye, heart and skin related disease.

References

1. Unicef India. Available at <http://unicef.in/whatwedo/3/immunization>
2. Sharma S (2007) Immunization coverage in India. Institution of Economic Growth, University of Enclave, Delhi. Available at URL: www.iegindia.org. Working paper series no. E/283/2007
3. Patra N (2006) Universal immunization programme in India: the determinants of childhood immunization. Available at SSRN: <https://ssrn.com/abstract=881224> or <https://doi.org/10.2139/ssrn.881224>
4. Srivastava AK, Shankar G (2017) Study of immunization coverage and its determinants among under five children residing in urban field practice areas of Karnataka, India. Indian J Forensic Community Med. <https://doi.org/10.18231/2394-6776>
5. Ventola CL (2016) Immunization in the United States: recommendations, barriers and measures to improve compliance: part 1: childhood vaccinations. P T 41(7):426–436
6. Wilson SE, Quach S, Naus M, McDonald SE, Kwong J, Tu K, Desai S, Tran D (2017) Methods used for immunization coverage assessment in Canada, a Canadian Immunization Research Network (CIRN) study. Hum Vaccin Immunother 13(8):1928–1936
7. Datta A, Baidya S, Datta S, Mog C, Das S (2017) A study to find out the full immunization coverage of 12 to 23-month old children and areas of under-performance using LQAS technique in a rural area of Tripura. J Clin Diagn Res JCDR 11:LC01–LC04
8. Panda BK (2019) A study to find out the temporal trend and inequality in India. Available at URL: <https://www.intechopen.com>. <https://doi.org/10.5772/intechopen.88298>
9. Song S, Miled ZB (2017) Digital immunization surveillance: monitoring flu vaccination rates using online social networks. In: 2017 IEEE 14th international conference on mobile ad hoc and sensor systems (MASS), Orlando, FL, pp 560–564. <https://doi.org/10.1109/MASS.2017.8196>
10. Siringi N, Sharma S (2017) Child immunization using data analysis. In: 2017 international conference on inventive computing and informatics (ICICI), Coimbatore, pp 937–943. <https://doi.org/10.1109/ICICI.2017.8365275>
11. Tropea M, Fedele G (2019) Classifiers comparison for convolutional neural networks (CNNs) in image classification. In: 2019 IEEE/ACM 23rd international symposium on distributed simulation and real time applications (DS-RT), Cosenza, pp 1–4. <https://doi.org/10.1109/DS-RT47707.2019.8958662>
12. Kakde A, Arora N, Sharma D (2019) A comparative study of different types of CNN and highway CNN techniques. Glob J Eng Sci Res Manag 6:18–31. <https://doi.org/10.5281/zenodo.2639265>
13. Sharma N, Jain V, Mishra A (2018) An analysis of convolutional neural networks for image classification. Procedia Comput Sci 377–384. ISSN 1877-0509. <https://doi.org/10.1016/j.procs.2018.05.198>
14. C100P61ThinF_IMG_20150918_144104_cell_128.png. [Photograph]. Retrieved from <https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria>

15. C100P61ThinF_IMG_20150918_144104_cell_162.png. [Photograph]. Retrieved from <https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria>

Soft Computing Techniques to Identify the Symptoms for COVID-19



Sujogya Mishra, Aezeden Mohmaed, Pradyumna Kumar Pattnaik,
Kamalakanta Muduli, and Tunku Salha Tunku Ahmad

Abstract COVID-19 since its discovery and spread have caused major disruption in the regular operation of the industries, schools, universities, business, and hospitals and eventually to the national economy of many countries. It is a highly contagious disease and spreading among communities at a very fast pace. Early detection of COVID-19 infected patients and quarantining them are few possible measures to restrict its spread. Persons infected by COVID-19 demonstrate many symptoms. However, these symptoms are somewhat similar to some of the common diseases like a common cold, viral fever, and body ache which we generally face in our day-to-day life. Hence, most of the medical experts confuse COVID-19 with some very common diseases, and when the actual cause is known by that time the situation becomes worse. In this light, this study proposed a new concept based on soft computing techniques to determine the exact symptoms responsible for COVID-19.

Keywords COVID-19 · Pandemic · Rough set · Chi-square · Soft computing

1 Introduction

The occurrence of COVID-19 since its discovery and spread have caused major disruption in the national as well as global economy of the world going by the severe nature of this health-related problem. At this point, there is no known cure

S. Mishra · P. K. Pattnaik

Department of Mathematics, College of Engineering and Technology, Bhubaneswar, India

A. Mohmaed · K. Muduli (✉)

Department of Mechanical Engineering, Papua New Guinea University of Technology, Lae, Papua New Guinea

A. Mohmaed

e-mail: aezeden.mohamed@pnuot.ac.pg

T. S. T. Ahmad

Faculty of Applied and Human Sciences, Universiti Malaysia Perlis, Kangar, Malaysia

e-mail: salha@unimap.edu.my

for COVID-19, and even the proposed vaccines touted by countries such as Russia, China, and the USA have yet to be subjected to the complete testing phase before any form of relieve can be felt. Furthermore, so many countries have gone into recession due to the direct impact of the pandemic on the national economy. In retrospect, it is therefore necessary to understand all possible scenarios to contain the spread of the disease quickly. In doing so, possible initiatives can be developed such advanced techniques for a quick identification of infected people from the symptoms they exhibit and quarantine them.

The symptoms demonstrated by the COVID-19 patients are not much different than the people infected with normal cold, flu, etc. In many occasions, people are found to have ignored these symptoms in the early period of infection, and consult physicians when they observe severe abnormalities in their health condition. There are also many incidents where even, doctors confuse the symptoms of COVID-19 with common cold and flu and prescribed medication accordingly. This not only resulted in severe medical complications for the patients even death in many cases but also rapid community spread of the disease. Hence, in this work, an attempt has been made to develop a model based on rough set theory and soft computing technique to help the medical practitioners in identifying the symptoms of coronavirus quickly. This work proposed an algorithm to find a precise number of attributes as proper symptoms of COVID-19 from vague and imprecise data sets.

2 Review of Literature

The outbreak of COVID-19 from Wuhan, China, has led to several countries imposing lockdown in the hopes of containing the spread of the virus. In most countries, schools, businesses, several industries, hotels, and restaurants have been severely affected, and this has created an atmosphere of economic instability. Many studies also raised concerns whether the health benefits derived from the COVID-19, anti-contagion policies adopted by several countries are worth their economic costs. Meanwhile, several studies also stressed on early detection of COVID-19 based on symptoms and quarantining them could be a game-changer policy to reduce the pace of spread of the disease. Rothan and Byrareddy [1] reviewed the symptoms of COVID-19 patients in the USA. The study revealed that initially, the patients demonstrate mild symptoms which gradually progress to pneumonia on ninth day of illness. In another study, Luo et al. [2] evaluated 1141 COVID-19 cases in China. The study reported that in some patients only gastrointestinal disorders were observed. These patients neither show symptoms of fever nor respiratory abnormalities. Study by Borges do Nascimento et al. [3] revealed that few patients though infected by the virus may not show any kind of symptoms while others might show mild symptoms. Their study also revealed that few patients show symptoms of pneumonia and severe respiratory illness. Allen et al. [4] conducted a survey in the USA to have an understanding of COVID-19 symptoms. This study is based on a collection of information from the general public about the abnormalities they experience relate to their health

and behavior on daily basis. The study gathered responses from 500,000 respondents through mobile application and reported loss of appetite, fever, and loss of taste or smell as some of the symptoms of COVID-19.

3 Methodology

The study employed a survey to collect information from patients regarding what symptoms they exhibit during the period of infection. Information was collected from 20,982 patients from Odisha and analyzed using soft computing techniques. The profile of the respondents is presented in Table 1.

In this work, six conditional attributes given in Table 2 have been considered along with two decision attributes with its values as 1 (higher value) and 2 (average value) and values for decision attribute as P (Positive) and N (Negative). This study has considered several cases of potential coronavirus patients, which are collected from different medical sources of Odisha. The study also included the patient's in quarantine and the confirmed positive case having symptoms.

The objective of this study is to find the exact symptoms from the six conditional attributes given in Table 2 by using the Rough set concept, one of the soft computing techniques.

Table 1 Respondent profile

Respondents based on information taken during six months	Male	Female	Total	Age
Senior	5263	3179	8442	65 years and above
Middle age	2364	1763	4127	45–65 years
Adult	2170	1830	4000	35–45 years
Young	1032	791	1823	15–35 years

Table 2 Conditional attributes

Symbol used	Conditional attributes
α_1	Common cold
α_2	Body ache
α_3	High fever
α_4	Loss of taste or smell
α_5	A rash on skin or discoloration of fingers or toes
α_6	Conjunctivitis

Table 3 Decision table

<i>U</i>	<i>C</i>	<i>D</i>
R_1	1	T
R_2	1	T
R_3	2	F

$\{R_1, R_2, R_3\}$, records {1, 2}, values of the conditional attributes, and $\{T, F\}$, values of the decision attribute

3.1 Basics of Rough Set

Rough set theory was developed by Pawlak in the early' the 80s [5]. It was a soft computing technique like Fuzzy Set the only difference is that it depends more on upper and lower approximation rather than member function.

3.2 Decision Table

The decision table is in the form $\langle U, C, D \rangle$, where U , the universal set, C , conditional attributes, and D , decision attribute as given in Table 3.

3.3 Upper Approximation and Lower Approximation

Upper and lower approximations concerning a set B , respectively, denoted as $\overline{U} = \{x | [x]_B B \cap X \neq \emptyset\}$ and $\underline{U} = \{x | [x] B \subset X\}$.

4 Result Analysis

4.1 Analytical Phase

A decision table is generated for six records as given in Table 4 and mentioned in the Methodology section.

R , records, (α_i) , $i = 1, \dots, 6$, conditional attributes, and D , decision attributes [6]. From Table 4, we have the following group of attribute considered as reduct [7]. 1. $\langle \alpha_1, \alpha_3, \alpha_4 \rangle$ 2. $\langle \alpha_1, \alpha_5, \alpha_6 \rangle$ 3. $\langle \alpha_2, \alpha_3, \alpha_4 \rangle$ and 4. $\langle \alpha_2, \alpha_5, \alpha_6 \rangle$ are four reducts and cannot be separated from the table to give a concrete result core for these reducts that are given as $\cap \text{Reduct} = \cap \langle \alpha_1, \alpha_3, \alpha_4 \rangle \langle \alpha_1, \alpha_5, \alpha_6 \rangle \langle \alpha_2, \alpha_3, \alpha_4 \rangle \langle \alpha_2, \alpha_5, \alpha_6 \rangle = \emptyset$, i.e., we cannot predict which attributes are essential in causing COVID-19 as previously these are known to be organic, so we can predict the nature as on a

Table 4 Initial decision table

<i>R</i>	α_1	α_2	α_3	α_4	α_5	α_6	<i>D</i>
R_1	1	2	1	1	1	1	<i>P</i>
R_2	1	1	2	1	1	2	<i>P</i>
R_3	2	1	1	1	1	1	<i>N</i>
R_4	2	2	2	2	2	2	<i>N</i>
R_5	1	2	1	2	2	1	<i>P</i>
R_6	2	1	2	1	2	2	<i>P</i>

particular atmospheric condition COVID virus symptom is recognized but this time it was not easy to recognize the symptoms. So for this, we calculate the strength from Table 4 using the strength of Rough Set. We have modified strength [8] of Rough Set in form of an algorithm is discussed below.

4.2 Algorithm to Find the Strength

In the following algorithm, we have used, R_s , as reduct space, and n , as number available data sets.

```

1. Initially  $R_s = \varphi$ .
iteration  $i = 1$ 
2. do
{
for  $i = 2 : n$ 
    test
    count  $\geq 45\%$ 
    count =  $\frac{\text{(conditional count with high decision values)}}{\text{(count the positive case)}}$ 

```

Similarly, calculate count with average conditional count with average decision values

```

 $R_s = R_s^{++}$ 
goto
step-2
continue
end
while
    No more classification possible
};
end

```

4.3 Application of the Algorithm

Strength α_1 can be found as,

$$\begin{aligned} R(\alpha_1)_1(D)(P) &= 100\%, \quad R(\alpha_1)_1(D)(N) = 0, \\ R(\alpha_1)_2(D)(N) &= 66\%, \quad R(\alpha_1)_2(d)(P) = 33\%. \end{aligned}$$

Similarly calculating the strength for other attributes, we get,

$$\begin{aligned} R(\alpha_2)_1(D)(P) &= 66\%, \quad R(\alpha_2)_1(D)(N) = 33\%, \\ R(\alpha_2)_2(D)(N) &= 33\%, \quad R(\alpha_2)_2(D)(P) = 66\%. \end{aligned}$$

$$\begin{aligned} R(\alpha_3)_1(D)(P) &= 66\%, \quad R(\alpha_3)_1(D)(N) = 33\%, \\ R(\alpha_3)_2(D)(N) &= 33\%, \quad R(\alpha_3)_2(D)(P) = 66\%. \end{aligned}$$

$$\begin{aligned} R(\alpha_4)_1(D)(P) &= 75\%, \quad R(\alpha_4)_1(D)(N) = 25\%, \\ R(\alpha_4)_2(D)(P) &= 50, \quad R(\alpha_4)_2(D)(N) = 50\%. \end{aligned}$$

$$\begin{aligned} R(\alpha_5)_1(D)(P) &= 66\%, \quad R(\alpha_5)_1(D)(N) = 33\%, \\ R(\alpha_5)_2(D)(P) &= 66\%, \quad R(\alpha_5)_2(D)(N) = 33\%. \end{aligned}$$

$$\begin{aligned} R(\alpha_6)_1(D)(P) &= 66\%, \quad R(\alpha_6)_1(D)(N) = 33\%, \\ R(\alpha_6)_2(D)(P) &= 66\%, \quad R(\alpha_6)_2(D)(N) = 33\%. \end{aligned}$$

From the analysis, it can be seen that attribute α_2 and α_3 has an ambiguous result, but another attribute $\langle \alpha_1, \alpha_4, \alpha_5, \alpha_6 \rangle$ provides some significant result, so we have constructed the next table for reduction [9] with attribute $\langle \alpha_1, \alpha_4, \alpha_5, \alpha_6 \rangle$.

From Table 5, it is clear that record-1 and record-4 are almost similar one, i.e., R_1 record shows all high leads to a positive result and record R_4 shows all average leads to negative results. To do a better analysis, we have taken the mixed values of conditional attributes by dropping R_1 and R_4 which can be observed from Table 6.

From Table 6, we can find the reduct similar to given above for six-attributes space as:

Table 5 Reduct table

R	α_1	α_4	α_5	α_6	D
R_1	1	1	1	1	P
R_2	1	1	1	2	P
R_3	2	1	1	1	N
R_4	2	2	2	2	N
R_5	1	2	2	1	P
R_6	2	1	2	2	P

Table 6 Modified reduct table

R	α_1	α_4	α_5	α_6	D
R_2	1	1	1	2	P
R_3	2	1	1	1	N
R_5	1	2	2	1	P
R_6	2	1	2	2	P

1. $\langle \alpha_1, \alpha_4, \alpha_5 \rangle$
2. $\langle \alpha_4, \alpha_5, \alpha_6 \rangle$
3. $\langle \alpha_1, \alpha_5, \alpha_6 \rangle$
4. $\langle \alpha_1, \alpha_4, \alpha_6 \rangle$

So from the above combination of attributes, we can find the core [9]:

$$\text{Core} = \cap \text{Reduct i.e. } \cap \langle 1, 2, 3, 4, 5 \rangle = \varphi$$

$$\text{Core} = \cap \text{Reduct i.e. } \cap \langle 1, 2, 3, 4, 5 \rangle = \varphi$$

Finding strength by combining three attributes, we have the following result.

$$\begin{aligned}
 R_{\langle \alpha_1, \alpha_4, \alpha_5 \rangle}(D)(1)(P) &= 100\%, \quad R_{\langle \alpha_1, \alpha_4, \alpha_5 \rangle}(D)(2)(N) = \text{Not found}, \\
 R_{\langle \alpha_1, \alpha_4, \alpha_5 \rangle}(D)(2)(P) &= \text{Not found}, \quad R_{\langle \alpha_1, \alpha_4, \alpha_5 \rangle}(D)(1)(N) = \text{Not found}. \\
 R_{\langle \alpha_4, \alpha_5, \alpha_6 \rangle}(D)(1)(P) &= \text{Not found}, \quad R_{\langle \alpha_4, \alpha_5, \alpha_6 \rangle}(D)(2)(N) = \text{Not found}, \\
 R_{\langle \alpha_4, \alpha_5, \alpha_6 \rangle}(D)(1)(N) &= 100\%, \quad R_{\langle \alpha_4, \alpha_5, \alpha_6 \rangle}(D)(2)(P) = \text{Not found}. \\
 R_{\langle \alpha_1, \alpha_5, \alpha_6 \rangle}(D)(1)(P) &= \text{Not found}, \quad R_{\langle \alpha_1, \alpha_5, \alpha_6 \rangle}(D)(1)(N) = \text{Not found}, \\
 R_{\langle \alpha_1, \alpha_5, \alpha_6 \rangle}(D)(2)(P) &= 100\%, \quad R_{\langle \alpha_1, \alpha_5, \alpha_6 \rangle}(D)(2)(N) = \text{Not found}. \\
 R_{\langle \alpha_1, \alpha_4, \alpha_6 \rangle}(D)(1)(P) &= \text{Not found}, \quad R_{\langle \alpha_1, \alpha_4, \alpha_6 \rangle}(D)(1)(N) = \text{Not found}, \\
 R_{\langle \alpha_1, \alpha_4, \alpha_6 \rangle}(D)(2)(P) &= \text{Not found}, \quad R_{\langle \alpha_1, \alpha_4, \alpha_6 \rangle}(D)(2)(N) = \text{Not found}.
 \end{aligned}$$

From the above analysis, the attribute set $\langle \alpha_1, \alpha_4, \alpha_5 \rangle$ has produced some results but still not sufficient to conclude.

From Table 6, we get a set of observation resulting in the following rules:

If

{values of $\langle \alpha_1, \alpha_4, \alpha_5 \rangle$ is 1 and α_6 is 2 → decision is Positive}

or

values of α_1 is 2 and $\langle \alpha_4, \alpha_5, \alpha_6 \rangle$ is 1 → decision is Negative

or

$\langle \alpha_1, \alpha_6 \rangle$ is 1 and $\langle \alpha_4, \alpha_5 \rangle$ is 2 → decision is Positive

or

$\langle \alpha_1, \alpha_5, \alpha_6 \rangle$ is 2 and α_4 is 1 \rightarrow decision is Positive

}

In the future, there may be several cases of COVID with variable symptoms, this leads to two types of possibilities one is normal [10] and another one is trapezoidal forms [11], which can be seen in Fig. 1.

COVID-19 take the shape like the curve from the above analysis in the contrast, and there may be a certain chance that some abnormal symptoms for COVID-19 which was not known may produce a curve shown in Fig. 2.

If the COVID-19 leads to a normal curve, then it will be expected that we recover from the crisis sooner or later. If we have considered Fig. 2 curve, then it will be a million-dollar question when the crisis is going to an end. The second curve is in

Fig. 1 Normal curve

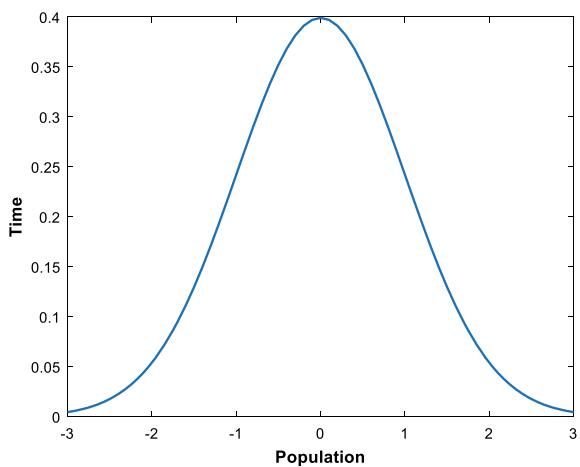
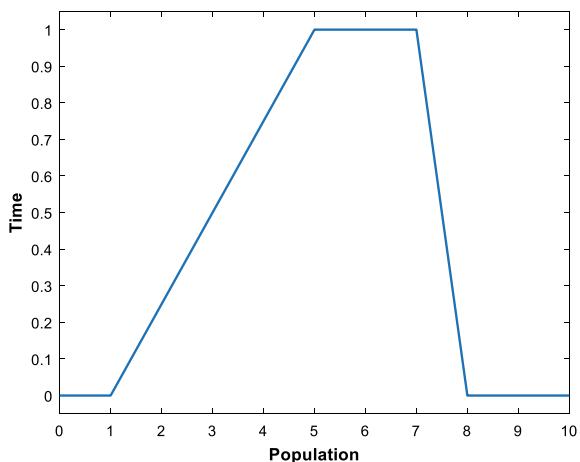


Fig. 2 Trapezoidal curve



the form of a trapezoidal fuzzy number that the height of the curve remains for an arbitrary time scale also and we never know when it starts when it will go to end.

4.4 Verification of Findings Using Statistical Analysis

We verify our claim using the Chi-square test

Null Hypothesis [12]

H_0 : These rules defined above are not sufficient symptoms for COVID-19.

Alternate Hypothesis [12]

H_a : The above rules defined above sufficient symptoms for COVID-19.

Using the formula given in Eq. (1) we calculate the χ^2 value

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

For the above rules, we have the following observed values of 85 cases with expected 75 cases from one set of a sample (scaling with 1000) another set produced observed 15 cases and expected 25 cases.

Now

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(85 - 75)^2}{75} + \frac{(25 - 10)^2}{25} = 1.33 + 4 = 5.33$$

rounding up 2 significant place we have 5.33, as degrees of freedom is $m + n - 1, 1 + 1 - 1 = 1$ (number of rows = 1, number of column = 1), so in this case degrees of freedom will be = 1 and tolerance will be 0.05, 5%. So the result can be verified using the table $\chi^2(0.051, 1)$, i.e., 0.005, will be tolerance and 1 is the degrees of freedom, tolerance is the maximum chance of failure will be 0.05 in our paper.

$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 5.33$ but $\chi^2(0.051, 1) = 3.414 < \chi^2_{\text{cal}}$, so the null hypothesis was rejected [12]. This implies the rules defined above represent the symptoms for COVID-19.

5 Implications

Owing to the continuous increase in the coronavirus infection cases across the world, there is a desperate need to develop solutions toward containing the spread of the disease without restricting the normal operations of business entities, schools, and hospitals [13]. In this regard, rapid and accurate identification of the infected people and isolating them would be a better alternative. Hence, the findings of the research

could be utilized for identification of the COVID-19 infected people and differentiate them easily from other patients suffering from cold and common flu.

6 Concluding Remarks

The world is slowly recovering and still coming to terms with the impact of coronavirus (COVID-19) that claimed hundreds of thousands of life in a short span of time compared to any natural or manmade catastrophe in the last century. COVID-19 has also influenced adversely all aspects of societal growth and development [14]. Hence, numerous studies have conducted and many ongoing to understand COVID-19, its impact, its spread, etc. This research aims to provide a means to confirm COVID-19 infection by recognizing a set of six symptoms for it. This study is essential as we are not sure when the pandemic is going to be over. It could be seen from Fig. 1 that if the COVID leads to a normal curve, then it could be expected that there will relief from the crisis sooner or later. However, if a trapezoidal curve is observed, then it will be difficult to predict its end [15].

References

1. Rothan HA, Byrareddy SN (2020) The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J Autoimmunity* 102433
2. Luo S, Zhang X, Xu H (2020) Don't overlook digestive symptoms in patients with 2019 novel coronavirus disease (COVID-19). *Clin Gastroenterol Hepatol* 18(7):1636
3. Borges do Nascimento IJ, Cacic N, Abdulazeem HM, von Groote TC, Jayarajah U, Weerasekara I et al (2020) Novel coronavirus infection (COVID-19) in humans: a scoping review and meta-analysis. *J Clin Med* 9(4):941
4. Allen WE, Altae-Tran H, Briggs J, Jin X, McGee G, Shi A et al (2020) Population-scale longitudinal mapping of COVID-19 symptoms, behaviour and testing. *Nat Hum Behav* 4(9):972–982
5. Pawlak Z (1991) Rough sets—theoretical aspect of reasoning about data. Kluwer Academic Publishers
6. Mishra S, Mohanty SP, Pradhan SK, Hota R (2015) Rough set approach in finding the cause of decline and down fall of jute industries and the remedy. *Int J Comput Appl* 121(19):35–41
7. Vashist R, Garg ML (2011) Rule generation based on reduct and core: a rough set approach. *Int J Comput Appl* 29:1–4
8. Mishra S, Mohanty SP, Pradhan SK (2015) Generation for student feedback system by the use of rough set theory. *IJCA* 131(18)
9. Vashist R (2015) International conference on computational intelligence and communication networks
10. Gupta SC, Kapoor VK (2020) Fundamentals of mathematical statistics, 10/e
11. Karay FO, de Silva C (2009) Soft computing and intelligent systems design
12. Walpole RE, Myers H, Myers SL (2006) Probability and statistics for engineers and scientists, 8th edn. Prentice Hall
13. Dash M, Shadangi PY, Muduli K, Luhach AK, Mohamed A (2021) Predicting the motivators of telemedicine acceptance in COVID-19 pandemic using multiple regression and ANN approach. *J Stat Manag Syst*. <https://doi.org/10.1080/09720510.2021.1875570>

14. Sahoo KK, Muduli K, Luhach AK, Poonia RC (2021) Pandemic COVID-19: an empirical analysis of impact on Indian higher education system. *J Stat Manag Syst.* <https://doi.org/10.1080/09720510.2021.1875571>
15. Devore JL (2011) Probability and statistics for engineering and the sciences, 8th edn. Cengage Learning

Prediction of COVID-19 Cases in India Using Parametric Curve



Gopal Behera and Ashutosh Bhoi

Abstract Due to novel coronavirus (COVID-19), the world is facing a pandemic situation. Human lifestyle changed drastically during this pandemic period, and everyone is badly affected and do not know when the situation is going to be normal. Though the virus is under control, still there is a uncertainty and unpredictable situation exist not only in India but also all over the world. So it is very important to predict the COVID cases as early as possible so that the best precautionary measure can be taken. In this study, we have designed a parametric estimation curve using linear, exponential, and logistic model for forecasting new cases on 30 days ahead. From experimentation, it is found that the logistic model performs better than the linear and exponential model.

Keywords COVID-19 · Parametric curve · Linear model · Exponential model · Logistic model

1 Introduction

The lifestyle of human is drastically changed during the pandemic period, everybody wear a mask and use hand sanitizer to avoid infection from COVID-19. On January 30, World Health Organization (WHO) [1] declared corona a public health emergency and a pandemic on March 2020. Within a span of few days, this virus quickly [2] spreads all over the world from China and has outbreaks different region of the globe [3, 4]. COVID-19 first case in India was reported on January 30, 2020 in Kerala [5], on a returnee from China. Followed by two more were reported on Feb 2nd and 3rd in the same state. Then this disease gradually spread all over India. The total cases reported of India in August 2020 are 3,080,483 with death and recoveries of 57,263 and 2,313,510, respectively [6].

G. Behera (✉) · A. Bhoi
Government College of Engineering Kalahandi, Bhawanipatna, India
e-mail: gbehera@gcekbpatna.ac.in

Dense populated country like India has taken the precautionary measure by declaring lockdown in last week of March followed by shutdown, as population is the major concern for transmitting or spreading the epidemic. After three months of shutdown, the virus is somehow under the control. At the same moment, the common men are facing problems, as many people have lost their jobs and income. Keeping this scenario in mind, the Government of India has started unlock phases from June to till date, as a result COVID cases have been increasing from July, August, and September with new cases of around 90,000+ daily. The Government of India is more focused on testing and tracing. Till now 35,292,220 samples have been tested with daily 801,147 samples tested [7]. To fight against COVID-19 emergency, India has taken few possible options like vaccination, herd immunity, and plasma therapy.

The primary focus is to recognize the symptoms as soon as possible, otherwise this may lead to spreading to more people [8]. Hence, the administration is trying to minimize the infection of virus by monitoring as well as tracking and tracing of health infected person and in contacted person [9]. The susceptible–infectious–recovered (SIR) model [10] has been applied in different levels of complexity to know the flow between the infected rate, recovery rate, and confirmed cases. This article presents different parametric curve fitting model to forecast the new cases in India prior to 30 days ahead.

The remaining of the paper is organized as follows. Related works are described in Sect. 2. In Sect. 3, the mathematical models are discussed to know the effects of each model for forecasting the new cases. Section 4 presents the simulation results, while conclusion is presented in Sect. 5.

2 Related Work

Machine learning (ML) and statistical models have found in various application of public health domain including prediction of disease and development of relevant drug [11]. ML and deep learning (DL) are successfully applied in classification of cancer tumor, medical imaging applications, prediction of tuberculosis (TB), and analysis of TB [12, 13]. Panda [14] has discussed 20 days ahead forecasting of COVID cases by using ARIMA and Holt Winters model. Behera and Nain [15, 16] explored a comparative study of sales prediction using different machine learning approaches and later used optimization technique to enhance the prediction accuracy. Behera et al. [17] discussed the Holt's winter forecasting approach for univariate time series sales forecast. A robust weighting iterative model is used to predict statistically the severity of corona disease spreading efficiently as compared to the baseline model and could lead to worsening the health situation [18]. Ghosal et al. [19] used multiple regression (MR), autoregression, and linear regression (LR) model for analysis of trends of COVID-19 death cases at the 5th and 6th week in India and found that the performance of autoregression is better prediction than LR and MR. Kumar et al. [20] discussed a reformist inquiry with the thought of the most recent current strategies that may be to battle the coronavirus pandemic. The pattern examination of time

arrangement at a beginning phase of coronavirus was investigated [21] for different nations with the countermeasure approaches to manage the scourge. Kucharski et al. [22] investigated the early spread example of coronavirus in all aspects of China by utilizing diverse datasets and furthermore, they utilize logical strategies to investigate the conceivable spread in different pieces of the world.

3 Model Designing

In this article, we focused on three models namely linear, exponential, and logistic model, where a model is a function that finds the relationship between the independent variable and one or more parameters or coefficients. Mathematically the model is defined in Eq. 1.

$$y = f(\text{time}) + \text{Err} \quad (1)$$

where the Err represents an error on random variations in the data followed by a specific distribution like Gaussian. The main objective of the model is to find optimal parameter combination that can minimize the error. As we are dealing with time series, hence here the time is the independent variable.

3.1 Linear Model

A linear model [19] is a function of the independent variable and one coefficient. This model also built a relationship between the dependent and the independent variables. The curve of linear model is a straight line passing through an intercept, and this line is sometimes known as best fit line. Mathematically linear model is defined in Eq. 2.

$$f(x) = a + b \times x \quad (2)$$

where a is the intercept and b is the coefficients of independent variable.

3.2 Exponential Model

An exponential function is a mathematical model [23] and is defined in Eq. 3, where a is a coefficient, b is the base of the function and x is the variable. The value of a function is increased by a factor e , when the variable is increased by 1.

$$f(x) = a + b^x \quad (3)$$

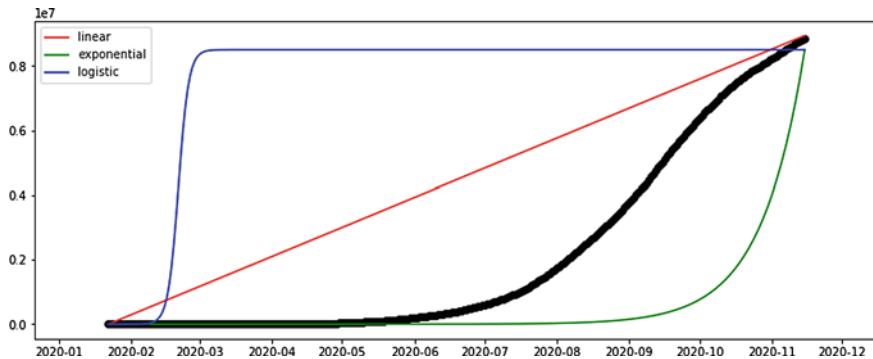


Fig. 1 Parametric curve fitting models of linear, exponential, and logistic model

3.3 Logistic Model

The logistic model [23] is the inverse of the natural *logit* function and can be used to convert the logarithm into a probability. A logistic curve is commonly represented by S-shaped curve. The logistic model is defined in Eq. 4.

$$f(x) = a/(1 + e^{(-b \times (x - c))}) \quad (4)$$

where a , b , and c are the coefficients and x is the dependent variable.

4 Implementation

The experiment is setup with Python and Jupyter Notebook for implementing above models. Further for visualization purpose, we have tried randomly different possible random coefficients to visualize the curve fitting of linear, exponential, and logistic function. Figure 1 represents the visualization of three parametric curve fitting models. Also, Fig. 1 shows that exponential curve fits the data properly. But it is fact that the phenomena has upper limit, because the total infection cannot go beyond the total population of any country. That is, sooner or later the growth of virus infection is going to stop and the curve will flat, which signifies that the logistic curve better than the other models.

4.1 Dataset

In our work, we use the publicly available CSSE COVID-19 dataset.¹ The dataset presents a time series of the number of confirmed cases of contagion reported by

¹ <https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6>.

Table 1 Daily total and new cases as on November 15, 2020

Date	Total cases	New cases
2020-11-11	8,683,916	47,905
2020-11-12	8,728,795	44,879
2020-11-13	8,773,479	44,684
2020-11-14	8,814,579	41,100
2020-11-15	8,845,127	30,548

each country every day since the pandemic started. Further, each row of the dataset contains a time series of confirmed cases in a specific region of the world. Here, we have aggregated all country level and select only one country, that is India.

4.2 Result and Discussion

To find the new confirmed cases at time t of a particular region (India) is defined in Eq. 5.

$$\text{new cases}(t) = \text{total}(t) - \text{total}(t-1) \quad (5)$$

Further, Table 1 shows the daily total and confirmed cases of India. The trend indicates that the cases are decreasing with time and days, and it is possible because of precautionary measures have been taken by Government of India. Figure 2 shows that the total cases were initially increasing slowly and from June to October-2020, the total cases were rapidly increasing with almost around 90,000+, whereas the new cases then started decreasing and now almost the cases are around below 50,000. Further, Fig. 3 shows the daily increase of total cases for next 30 days along with last 7 observation and the trend signifies that the curve is flatten as the total cases is reduced. Also, Fig. 4 shows the weekly prediction of 30 days ahead from today.

5 Conclusion

At present, COVID-19 cases are under control in India. As the virus is more active in cold condition, the Government of India thinking that there may be second wave of this disease. Some of the European countries have already faced the second wave of this diseases. Keeping view of second wave of COVID-19, it is therefore to know in prior the total cases. So that any precautionary measure will be taken to guard against or to fight with COVID-19. In this article, we have discussed three parametric curve fitting models to forecast the COVID-19 cases, which predict COVID-19, 30 days ahead from today in daily basis as well as weekly basis. From Fig. 1, it is concluded

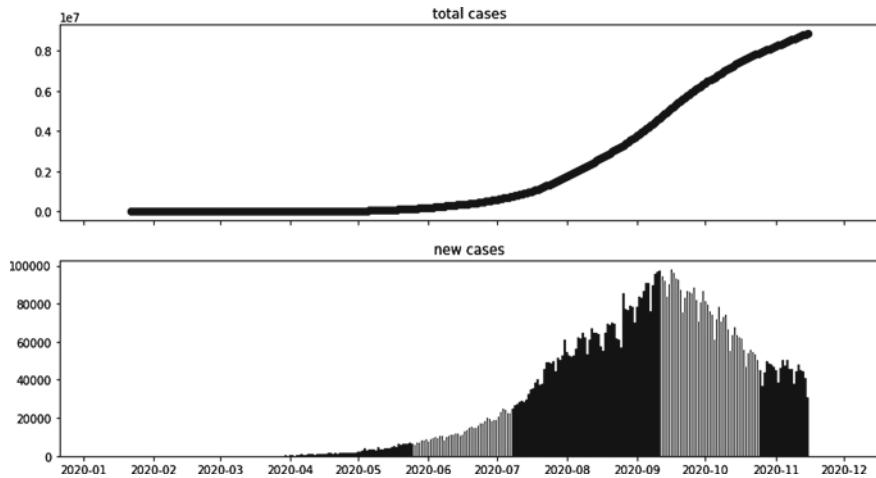


Fig. 2 Total and new cases in India up to November 15, 2020

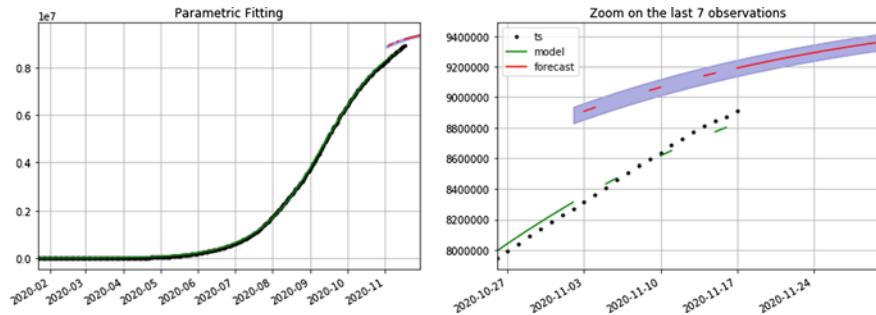


Fig. 3 Daily increase of new cases for next 30 days

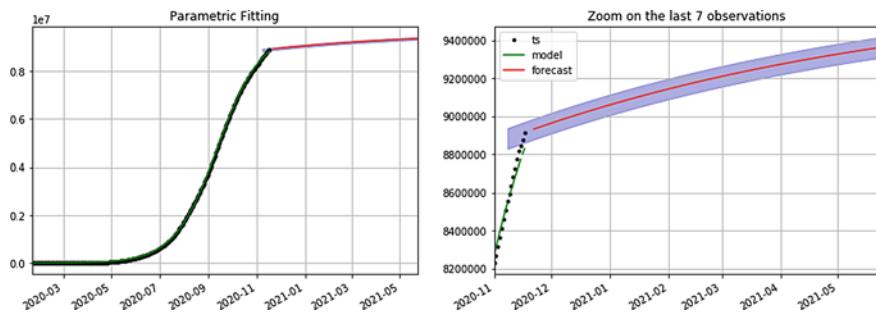


Fig. 4 Weekly prediction of new cases for next 30 days

that the exponential model fits data properly, whereas the phenomena has upper limits that signifies sooner or later the virus growth rate will stop, which signifies that the logistic model more appropriate or better than other two parametric model, as this model's prediction curve flattened and properly fitted than linear and exponential model.

References

1. World Health Organization et al (2020) Naming the coronavirus disease (COVID-19) and the virus that causes it. URL [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet* 395(10223):497–506
3. Hussain AA, Bouachir O, Al-Turjman F, Aloqaily M (2020) Ai techniques for covid-19. *IEEE Access* 8:128776–128795
4. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR (2020) Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access* 8:91916–91923
5. India confirms its first coronavirus case (Jan 30, 2020). Available online: <https://www.cnbc.com/2020/01/30/india-confirms-first-case-of-the-coronavirus.html/>
6. Coronavirus update (live): COVID-19 virus outbreak worldometers. <https://www.worldometers.info/coronavirus/>
7. COVID-19 Indian Council of Medical Research, Government of India. Available at: <https://icmr.nic.in/node/39071>
8. Wang LS, Wang YR, Ye DW, Liu QQ (2020) A review of the 2019 novel coronavirus (Covid-19) based on current evidence. *Int J Antimicrob Agents* 105948
9. Blackwood JC, Childs LM (2018) An introduction to compartmental modeling for the budding infectious disease modeler. *Letters in Biomathematics* 5(1):195–221
10. Zhong L, Mu L, Li J, Wang J, Yin Z, Liu D (2020) Early prediction of the 2019 novel coronavirus outbreak in the mainland china based on simple mathematical model. *Ieee Access* 8:51761–51769
11. Mallapaty S (2020) What the cruise-ship outbreaks reveal about covid-19. *Nature* 580(7801):18–18
12. Jin S, Wang B, Xu H, Luo C, Wei L, Zhao W, Hou X, Ma W, Xu Z, Zheng Z et al (2020) AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system in four weeks. *medRxiv*
13. Tuli S, Basumatary N, Gill SS, Kahani M, Arya RC, Wander GS, Buyya R (2020) Healthfog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated iot and fog computing environments. *Future Generation Computer Systems* 104:187–200
14. Panda M (2020) Application of ARIMA and holt-winters forecasting model to predict the spreading of COVID-19 for India and its states. *medRxiv*
15. Behera G, Nain N (2019) A comparative study of big mart sales prediction. In: International conference on computer vision and image processing. Springer, pp 421–432
16. Behera G, Nain N (2019) Grid search optimization (GSO) based future sales prediction for big mart. In: 2019 15th international conference on signal-image technology & internet-based systems (SITIS). IEEE, pp 172–178

17. Behera G, Bhoi AK, Bhoi A (2020) UHWSF: univariate holt winters based store sales forecasting. In: Intelligent systems. Springer, Singapore, pp 283–292
18. Tuli S, Tuli S, Tuli R, Gill SS (2020) Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet Things* 100222
19. Ghosal S, Sengupta S, Majumder M, Sinha B (2020) Prediction of the number of deaths in India due to SARS-COV-2 at 5–6 weeks. *Diabetes Metab Syndr Clin Res Rev*
20. Kumar A, Gupta PK, Srivastava A (2020) A review of modern technologies for tackling COVID-19 pandemic. *Diabetes Metab Syndr Clin Res Rev*
21. Deb S, Majumdar M (2020) A time series method to analyze incidence pattern and estimate reproduction number of COVID-19. arXiv preprint [arXiv:2003.10655](https://arxiv.org/abs/2003.10655)
22. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Sun F, Jit M, Munday JD et al (2020) Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect Dis*
23. Rzadkowski G, Sobczak L (2020) A generalized logistic function and its applications. *Foundations of Management* 12(1):85–92

Nutritional Ingredients Analyzer for Food



A. Thilagavathy, Tadavarthi Rishi, Veeram Deepak Reddy,
and Sudesh Nimmagadda

Abstract Adequate nutritional regimes have been generally accepted as important prevention and management steps for non-communicable diseases (NCDs). In all cases, little research is now underway into safe food fixtures which support the rehabilitation of NCDs. Right now, significantly investigated the connection between healthful fixings and maladies by utilizing information mining techniques. Initially, in excess of 7000 sicknesses were acquired, and we gathered the suggested nourishment and unthinkable nourishment for every illness. The analyses on genuine information show that our technique dependent on data mining improves the exhibition contrast and the conventional measurable methodology, with the accuracy of 1.682. Moreover, for some basic ailments, for example, diabetes, hypertension, and coronary illness, our work can recognize effectively the initial a few nourishing fixings in nourishment that can profit the restoration of those ailments. These exploratory outcomes exhibit the adequacy of applying information mining in choosing of wholesome fixings in nourishment for ailment examination.

Keywords NCD · Noise intensity · Data mining

1 Introduction

NCD's are ceaseless infections, which are for the most part brought about by word related and ecological elements, ways of life and practices, including corpulence,

A. Thilagavathy (✉) · T. Rishi · V. D. Reddy · S. Nimmagadda
R.M.K. Engineering College, Gummidi poondi, India
e-mail: atv.cse@rmkec.ac.in

T. Rishi
e-mail: rish16312.cs@rmkec.ac.in

V. D. Reddy
e-mail: vde16331.cs@rmkec.ac.in

S. Nimmagadda
e-mail: sud16324.cs@rmkec.ac.in

diabetes, hypertension, tumors, and different maladies. As indicated by the worldwide on non-transferable WHO ailments given, the yearly loss of life from NCDs keeps including, which has made genuine financial weight of the world. Around 40 million individuals passed on from NCDs every year, which is identical to 70% [1–3] of the worldwide loss of life. Measurements of Chinese Occupant's Interminable Ailment and Sustenance shows that, the quantity of some other nations on the planet and the present commonness rate have smothered. What is more, the populace matured 60 or over in China has arrived at 230 million, and around 66% of them are experiencing NCDs as per the official insights [4, 5]. Along these lines, pertinent divisions in every nation, particularly in China, for example, clinical schools, emergency clinics, and infection investigation, focus that all are worried about NCDs.

Appropriate nourishing eating regimens assume a significant job in keeping up well-being and forestalling the event of NCDs [2, 6]. Continuous acknowledgment idea by China is additional effect of nourishment on well-being. Be that as it may, look into dietary fixings in nourishment through information mining, which are helpful for the restoration of maladies that is as yet uncommon quite recently started the IT (Data Innovation) development of keen medicinal services. Most examinations on the connection in nourishment and illnesses are still through costly accuracy preliminaries. What is more, there are additionally numerous anticipation reports; however, they concentrated on just one or a few maladies [7]. In China, considering the connection between nourishing fixings and ailments utilizing data mining is juvenile. Most specialists just prescribe the particular nourishment to patients experiencing NCDs, without giving any important sustenance data, particularly [8].

In the period of large information, information mining has become a basic method for finding new information in different fields, particularly in malady expectation and exact medicinal services (AHC) [9]. It has become a center for preventive medication, essential medication, and clinical medication examination. Regarding the infection investigation through the mining of nourishing fixings in nourishment, we mostly make the accompanying commitments: (I) We separated information identified with Chinese sicknesses, comparing suggested nourishment and forbidden nourishment for every illness; however, many could be expected under the circumstances from clinical and official sites to make an important information base that are accessible on the Web; (ii) Applying clamor power and data entropy to discover which is healthful fixings in nourishment can apply constructive outcomes to maladies; (iii) Right now, information is ceaseless and has no choice traits. Red dependent on an unpleasant set hypothesis, which can more readily choose comparing center fixings from the positive dietary fixings in nourishment. Area II surveys the related work in the field of infection investigation and information mining; portrays the particular information mining calculations utilized right now, why we select the calculations, just as two assessment records; expounds the information, test results, and investigation in detail; presents conversations between techniques. A few ends and potential future research headings are additionally examined.

2 Literature Survey

The most critical advance in the programming method is the literature review. The time factor, economics, and quality of friends must be determined before constructing the unit. If these things are completed, ten next steps must decide the structure and language the device will be used to construct. Whenever the computer programmers start constructing the system, the developers require a kit of outside assistance. Senior software developers, books, and Web sites can obtain this support. The above insights are called to construct the proposed structure before constructing the structure.

- Retrospective study of diabetes testing in a mass meeting in India: consequences for the regulation of non-communicable diseases.

The major case of non-communicable diseases (NCD) mortality [10] in India is heart disease. The national strategy pernicious Development, Diabetes, Cardiovascular Disease, and Strokes Prediction and Management Program aims to extend the NCD management, screenings, and references across India and include network-based initiative and machine tracking engineers. The Indian government regularly finds strict public get-togethers as essential.

In any event, the economic growth paved way for its administrators to provide a hyper testing provider at the Nashik and Trimbakeshwar Kumbh Mela in 2015. In this article, we look at the importance and implications of such a groundbreaking social testing. At the Kumbh, 5760 people deliberately decided on hypertension screening and got a solitary circulatory strain estimation. All in all, the favorable findings were reviewed by 1783 (33.6%), of which 1580 had not been told. Of the 303 newly experienced hypertonic medications, 240 (79%) have earned regulatory clearance and 160 (52.8% under treatment) have been satisfying. Typical circulatory pressure assessment (BP levelled out) was in 55% (18%). The knowledge was also more common (39%) among nicotine consumers with epilepsy, compared with non-users (28%) ($P < 0.001$). Bad telephone chronology (0.01%) has blocked the production of a cell phone. The low levels of tuberculosis understanding, care, and management highlights both Indian and Indian managers' progressive challenge.

- DIETOS: A recommender system for adaptive diet monitoring and personalized food suggestion.

These days there is a far-reaching dissemination of portable weight and diet on the board. Despite the fact that, the most famous applications are not typically tested in clinical settings, just as applications are not bolstered by clinical proof. Most important, the survey highlights the adequacy of applications for weight and diet in public health. Besides, there are not many instances of nourishment that gives the clients healthful realities about reasonable nourishment decisions and consider for the versatile conveyance of nourishment substance to improve the personal satisfaction of both sound individuals and people influenced by constant diet-related infections. The proposed system can assemble a user's well-being profile and gives individualized nourishing suggestions as indicated by the well-being profile. The profile is made

using dynamic constant surveys arranged by clinical specialists and accumulated by the clients. The well-being profile incorporates data about well-being status and possible incessant infections. The main model incorporates a list of run of the mill Calabrian nourishments assembled by sustenance masters. DIETOS can propose not just the utilization of explicit nourishments good with the well-being status, yet in addition it might give dietary signs identified with some particular or well-being conditions.

- Lumping versus division: the need for precise biological data mining.

Biological data mining is assuming an undeniably significant job all through the range of biological and biomedical research with expansive ramifications for the comprehension of life science addresses, for example, the tree of life and functional uses of such information to improve human well-being. Maybe no place is data mining required more than the developing order of exactness medication. The capacity to anticipate singular danger of giving an infection or reaction to treatment is at the center of the idea of exactness medication, which is picking up ever-expanding levels of footing in the period of technology-driven estimation of biological frameworks. This has gotten particularly significant with the new presidential activity on accuracy medication in the USA. It is clear to the pursuers of Bio Data Mining that this will require cautious investigations of huge and regularly complex datasets to best make an interpretation of data into progressively individualized chance. Here, we inquire as to why improved and fitting data mining is not just positive; however, an immense enhancement is for most current examinations of genomic data. The appropriate response deceives some degree in clarifying the current act of—omic examinations and how we should extend it.

- Aquatic ecoscope health evaluation based on the main entropical weight portion assessment: A rising dam test case (Hainan Island, China).

In order to discover if the new approach can undo the calculation coverage which existed for environmental welfare in traditional entropy-driven straight lines, a further evaluation strategy based on the main component analysis (PCA) and randomness weight for pharmacological system good was introduced in the Dwindling Dam, Hainan Area, China. The outcomes demonstrated that the biological system well-being status of Wanling Reservoir indicated an improvement pattern generally from 2010 to 2012; the methods for environment well-being far-reaching file and the biological system well-being status were III (medium), II (great), and II (great), separately. Furthermore, the environment well-being status of the reservoir showed a frail regular variety. The variety of EHCl decreased as of late, demonstrating that Wanling Reservoir would in general be moderately steady. Examining the length of the modern and traditional files indicates that for the traditional, a more embedded relation of 0.382 was feasible than with the new approach, the cumulative load for the four files. The use of PCA with oxidation was advised to retain proper relative weight away again from cover. The interaction study between the photosynthetic status file and EHCl also revealed a big negative correlation ($P < 0.05$), suggesting that the modern entropic-weight PCA-dependent methodology could both further

boost weighting and results precision. The new technique here is reasonable for assessing the environment strength of the reservoir.

3 Methodology

3.1 Data Mining

Data extraction is the way to locate prototypes in large datasets.

Machine Learning, Measuring and Database Convergence Methods. Frameworks. In order to acquire data from a dataset and to turn the information into an accessible structural for additional use, data mining is an emerging subfield of information technology and initiatives. Data mining is the analysis venture of the “knowledge discovery in databases” procedure or KDD.

3.2 Statistical Algorithm (SA)

On the off chance that a specific illness is brought about by the absence of certain healthful fixings, at that point their qualities in suggested nourishment ought to be generally higher in principle. Along these lines, we can make sense of which wholesome fixings esteem are high. The dietary fixings with higher qualities ought to be the fixings, which advantage the restoration of that particular malady, i.e., PNIs. This strategy is alluded to as statistical algorithm (SA) in this paper. The basic documentations are characterized as following, let n be the quantity of the suggested nourishment for a specific illness, and m be the quantity of wholesome fixings. The above thought can be communicated as beneath:

$$\begin{aligned} & \text{sort}\{xi1, xi2, \dots, xim | \text{descend}\} \\ & i = 1 \quad i = 1 \quad i = 1 \end{aligned} \tag{1}$$

$xi1$ shows the primary nutrient value for a certain infection of the suggested i th food; $xi2$ measures the amount of the second food nutrient of the suggested i th food and so forth. Both foods approved for this illness are added to dietary values. Equations are then sorted downward. Therefore, PNIs ought to be the highest of the classified functional foods.

We also obtained different methods for research in order to show whether or not data mining tools can be used for the disease study. Since this paper is trying to take care of another issue from a true application, there is no pertinent work for examination. In any case, we utilized an exploring way to tackle this issue, at the end of the day, we led an impact comparison between various strategies to choose the best one. In order to demonstrate nutrient characteristics in food, we derive

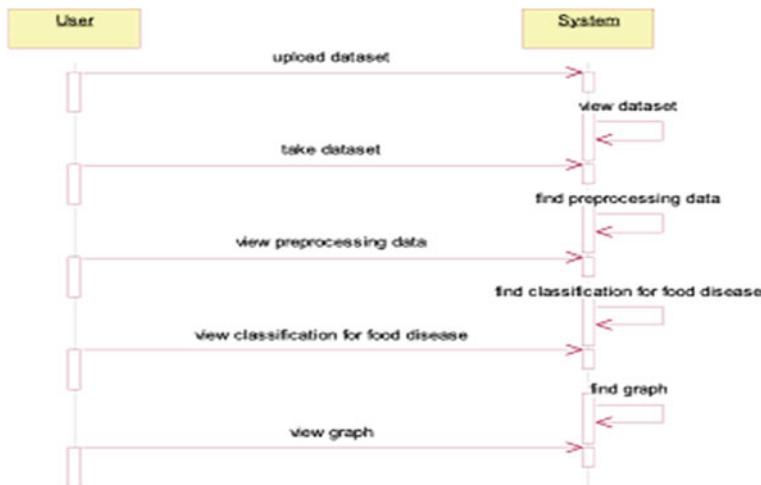


Fig. 1 Overall work flow of proposed diagram

nutrient estimates for four serious illnesses from recommended food and tabú foods. In Fig. 1, nutrient appreciation for four diseases (such as hypertension, coronary sickness, kidneyng, and apoplexy) is seen in the recommended diet and tabu food. The *X*-axis and *Y*-axis are compared and corresponded to the dietary ingredients. Figure 1 indicates that some nutrient estimates are very high, while the others are very poor independent of diets or tabu diets. Broadly speaking, for the most infections in China, there is scarce tabu rice, so we cannot collect more. Because of the less true awareness for limited data mining interventions, we are currently testing foods to decide which nutrient additives may have beneficial effects on a specific disease. In this article, positive nutrient ingredients (PNI) are called nutritious ingredients which help the restoration of illnesses.

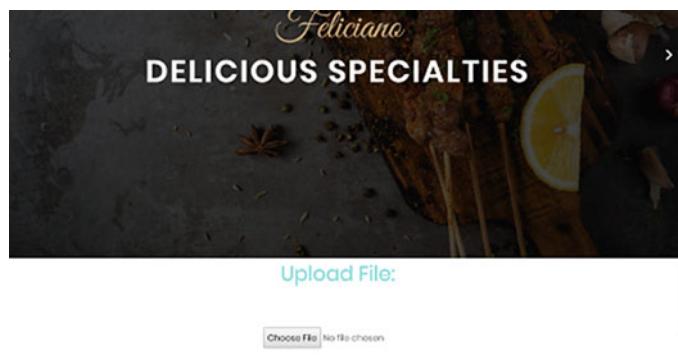
4 Results and Discussions

User module will upload the dataset below.

After entering the appropriate dataset in the upload file option shown in Fig. 2, we have to click the upload dataset (Fig. 3).

Then, user will search or insert the food name which he or she wants positive nutrients to get cure from the disease suffering from. After entering the disease, user will be shown a list of positive nutrients and its cure for the disease as shown in Fig. 4.

The graph (Fig. 5) depicts that eating dairy foods is the major source of causing diseased for the most cases and meat and liquid drinks are the next main source of casing diseases for us when we consumed them.

**Fig. 2** User section home page

Mining of Nutritional Ingredients in Food for Disease Analysis				
Preprocessing				
LAMES	FOOD TYPE	FOOD DISEASE	MINERALS	GRAMS
A	vegetables	Angina	Vitamin A	400g
B	meat	Acne	C	0g
C	fruits	cardiovascular	E	400g
D	Dairy Foods	ovarian	Vitamin B12	1g
E	Grains	Stroke	magnesium	3g
F	Beans and Nuts	tooth decay	potassium	mg
G	Fish and Seafood	Asthma	iron	g
H	liquid drinks	liver disease	copper	g
I	tobacco food	oral cancers	Vitamin A	2g

Fig. 3 User viewing the dataset

The screenshot shows a search interface for the dataset. At the top, there's a search bar with the placeholder "Search...". Below the search bar, there's a text input field labeled "Enter the Food Name classification:" containing the word "Nuts". At the bottom of the form is a green "Submit" button.

Fig. 4 User entering the food item

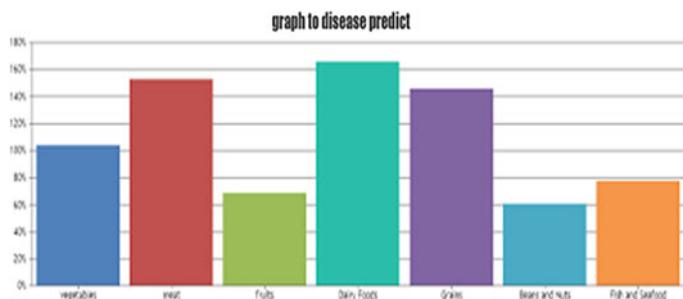


Fig. 5 Prediction graph

Eating fruits, beans, and nuts were causing very least possibility of diseases, so eating more fruits is suggested than liquid drinks and meat in order to stay healthy.

In our paper, we determine the particular vitamin for food type which is the cause of the diseases. The dataset which is presented does contain all the aspects that are related to the food types. In the classification part, we search for a specific food type that helps to easily identify the vitamins that causes diseases (Fig. 6).

The final and satisfactory results are shown in Fig. 7. The first column shows that the results obtained in the first time that means the PNIs for each disease. The results we obtained for the second time are shown in column two, but the results in the DN6 disease are not expected. Generally, this method gives the most accurate results among all the other methods, because we can accurately find out the first two or three nutritional ingredients out of 26 ingredients for each disease. It is the best method that meets our actual needs. For example, if the patient is suffering from the

Fig. 6 Classification

Search....

Enter the Food Name classification:

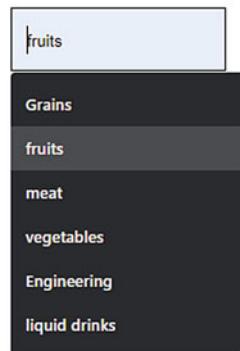


Fig. 7 PNIs for each disease

DN1	H	G	L	P	C	F	E	Y	Q				
DN2	H	F	O	Q	S	V	J	Z	M	Y	B	W	
DN3	F	H	X	K	V	T	U	E	Q	R	G	O	Y
DN4	S	D	X	H	G								
DN5	H	F	K	Q	A	C							
DN6	N	Q	U	K									
DN7	F	J	C	U	N	G	O	Q	B	E			
DN8	H	I	B	D	S	Q	U	Z	R	C	K	F	N
DN9	S	N	W	B									
DN10	H	V	B	L	S	A	D	U	N				

above ten diseases, the recommended food must be rich with the first two or three nutritional ingredients.

5 Conclusion

The key analysis of the article can be separated into two sections: initially, the relationships involving nutrient component and disease that primarily seek to classify which additives play a constructive role in restoring Chinese diseases have been collected and sorted from the biomedical and formal database. To our best understanding, it is the first research in China that uses data mining technologies to disrupt the connection between nutrition and infection nutrient materials. Innovative findings revealed that while the positive dietary components could not be entirely identified by data mining techniques for illnesses, the first two or three have been precisely chosen. Furthermore, if we can merge our vision with tabu food, the outcomes will possibly be smoother and in line with the fact that we will be working on in the future.

The two major aspects of this study include: (1) you can use our Web site to receive illness, prescribed food, tabuic foods, and associated nutrition data; and (2) doctors and illness inspectors may be helped to identify positive nutrient additives that are as effective as possible to rehabilitate illnesses. Actually, some details are not accessible while surgical exams are still in progress. In addition, if investigators discover anything wrong in our work, we expect to notify us and enhance our study. In reality, our knowledge and experience is increasingly improving.

References

1. CNS (2016) Global nutrition report. Chinese Nutrition Society

2. WHO (2014) Global status report on noncommunicable diseases. World Health Organization
3. Balsari S, Vemulapalli P, Gofine M et al (2017) A retrospective analysis of hypertension screening at a mass gathering in India: implications for non-communicable disease control strategies. *J Hum Hypertens* 31(11):750–753
4. DNHFPC of PRC (2015) Chinese resident's chronic disease and nutrition. National Health and Family Planning Commission of the People's Republic of China
5. Tellier S, KiabyLars A, Nissen P et al (2017) Basic concepts and current challenges of public health in humanitarian action. *Int Humanit Action* 229–317
6. Ara F, Saleh F, Mumu SJ, Afnan F, Ali L (2011) Awareness among Bangladeshi type 2 diabetic subjects regarding diabetes and risk factors of non-communicable diseases. *Diabetologia* S379
7. QIANZHAN (2017) Report of market prospective and investment strategy planning on China intelligent medical construction industry. Qianzhan Intelligence Co. Ltd.
8. Ling WH (2017) Progress of nutritional prevention and control on non-communicable chronic diseases in China. *China J Dis Control Prev* 21(3):215–218
9. Margaret MB, Barbara BK, Colette D (2013) Developing health promotion workforce capacity for addressing non-communicable diseases globally. In: Global handbook on noncommunicable diseases and health promotion, pp 417–439
10. Williams M, Moore H (2015) Lumping versus splitting: the need for biological data mining in precision medicine. *BioData Min* 8(16):1–3

Deep Learning Analysis for COVID-19 Using Neural Network Algorithms



V. Vijaya Baskar, V. G. Sivakumar, S. P. Vimal, and M. Vadivel

Abstract The COVID-19 pandemic threatens to devastatingly impact the global population's safety. A successful surveillance of contaminated patients is a crucial move in the battle against COVID-19, and radiological photographs via chest X-ray are one of the main screening strategies. Recent research showed that patients have abnormalities in photographs of chest X-ray that are characteristic of COVID-19 infects. This has inspired a set of deep learning artificial intelligence (AI) programs, and it has been seen that the precision of the identification of COVID-19 contaminated patients utilizing chest X-rays has been quite positive. However, these built AI schemes, to the extent of their author's awareness, have become closed sources and not accessible for further learning and expansion by the scientific community, so they are not open to the general public. This thesis therefore implements COVID-Net to identify COVID-19 cases of chest X-rays images, an open source, accessible to the general public, a deep neural network architecture adapted to the detection. The COVID-Net data collection, which is referred to as COVIDx which includes 13,800 chest X-ray photographs of 13,725 patients from 3 open-access data sources, one of which we launched, are also addressed.

Keywords AI · Chest X-ray · COVID19 · Neural network · Screening · COVID-Net

V. Vijaya Baskar (✉)

Department of Electronics and Communication Engineering, School of EEE, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India

V. G. Sivakumar · M. Vadivel

Vidya Jyothi Institute of Technology, Hyderabad, Telangana, India

S. P. Vimal

Sri Ramakrishna Engineering College, Coimbatore, India

e-mail: vimal.sp@srec.ac.in

1 Introduction

Data science incorporates domain experience, programming capabilities, and quantitative and computational understanding in order to achieve meaningful insights into the information process. In order to produce artificial intelligence systems, data scientists apply machine-learning algorithms to numbers, texts, images, videos, and audio for usual human intelligence tasks. In essence, these programs produce concepts that can be converted into real market value by analysts and business consumers. Particularly for organizations with nearly limitless funding, it is difficult to intensify data science activities. In addition to making data scientists more efficient, the DataRobot autonomous machines' learning network democratizes data analysis and IP, which allows researchers, industrial customers, and other technological experts to become people and data science engineers. It automates repeated modeling activities which once took up much of the time and brain power of data scientists. DataRobot closes the divide between data scientists and the rest of the business and allows the learning of industry more available than ever. The outbreak continues to get a devastating effect on the welfare of the global population and well-being as a consequence of people's contamination with the frequency of coronavirus ARSCoV2 (SARS-CoV-2). Active surveillance of infectious patients is essential for mitigating COVID-19, allowing all affected to seek prompt diagnosis and care and to be separated to reduce the disease transmission. The primary screening technique to detect COVID-19 is to detect SARS-CoV-2 RNA in the reverse transcriptase-polymerase chain reaction (RT-PCR) [1] from respiratory specimens. While the simplest gold standard is RT-PCR checking, the manual process is time-consuming, laborious, and complex.

The alternate mode of monitoring used by COVID-19 scans was the radiogram study, which conducts and analyzes visual markers of SARS-CoV-2 virus exposure via X-ray (e.g., chest X-ray) or computed tomography (CT) imaging by radiologists [2]. Early studies suggest that patients have chest X-ray defects typical of those afflicted with COVID-19 [3, 4] and that radiography evaluation can be used as a crucial tool for COVID-19 screening in epidemic areas [5]. The use of CXR images for COVID-19 screening in accordance with the global COVID-19 pandemic is particularly advantageous.

CXR allows immediate triage of suspected COVID-19 patients and can be conducted in combination with viral testing (which takes time) in order to relieve vast numbers of individual patients that have been more affected by their potential (e.g., New York, Spain, and Italy) or by viral testing (low supply) as an alternative means of relief [3, 6]. In regional areas where patients have to remain home before advanced signs arise, CXR may also be very useful for research, as anomalies are frequently found when COVID-19 prone implementation forms at hospital sites at the time of diagnosis [7, 8]. The CXR is readily affordable and used in various hospital facilities and imagery centers, which several healthcare organizations deem basic devices. The presence of compact CXR-systems means that images can be carried in a single isolation space, thus significantly minimizing the possibility that the COVID-19 can be transmitted during transport to particular structures like a CT

scanner and rooms comprising specific frames [7]. Thus, radiology can be tested more easily and are more applicable in current healthcare facilities despite the prevalence of chest X-ray imaging devices, which renders them an ideal substitute for PCR research (in some instances much more sensitive [9]). However, one of the main bottlenecks confronting radiologists, as visual markers may be ambiguous, is the need to classify the radio graphical pictures. Of this purpose, it is strongly desirable to have computerized diagnoses that will enable radiologists view radiographic images more quickly and reliably to diagnose COVID-19 events.

Motivated by the need to view radiographic images more easily, a variety of deep learning (DI) technologies [10] were introduced and the findings were very strong with respect to the precision in identifying COVID-19 patients through radiography [11–13]. However, in order to gain greater understanding and extension of these systems, the best knowledge of the author was that these built AI technologies were secret sources and unavailable to the scientific community. In fact, such programs cannot be viewed and utilized by the public. As a consequence, recent attempts were made to press for open source and free software AI approaches for radio graphically driven COVID-19 case identification [6, 14]. The paper [13] has made an excellent attempt to build a COVID-19 case dataset with annotated CXR to enable the researcher to locate a dataset comprising COVID-19 cases and SARS and MERS cases.

2 Implementation of COVID-19 Using Deep Learning Algorithms

There we address the implement for programmer design planned COVID-Net, the subsequent configuration of the system, the COVIDx database construction process, and the specifics of deployment of COVID-Net.

Architectural design

In this research, COVID-Net is generated using a human–machine collaborative design approach, where human instructions merge prototyping of the configuration of a network with machine-driven system experimentation in order to produce a network architecture designed to detect CXR images in the case of COVID-19 as shown in Fig. 1. The following are listed in each of the two design phases.

3 Network Design Prototyping

A primary network design prototyping stage is used in the first step of a joint design approach for the human–machine project, which constructs an initial network design system focused upon concepts and best practices in human design. In this analysis,

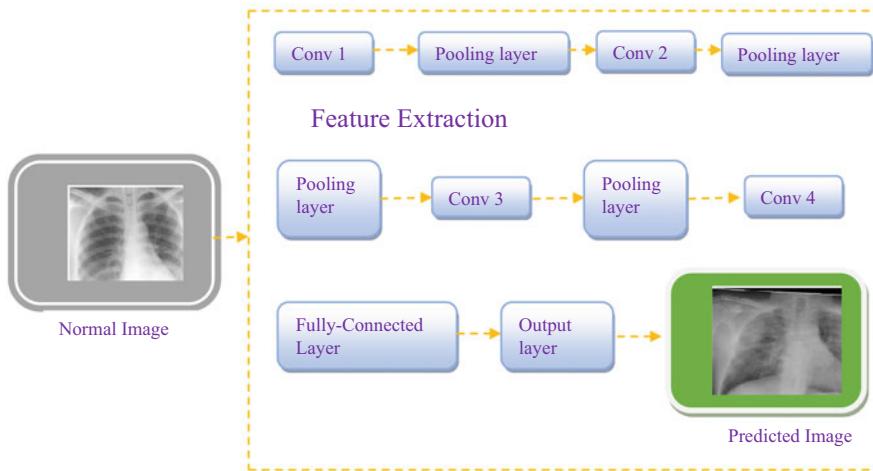


Fig. 1 Structure diagram for COVID-19 prediction using deep CNN algorithm

we have used the concepts of designing residual architecture, as demonstrated time and again in order to allow stable, high-performance trainable neural network architectures to be successfully built. The initial network design concept is built for (a) no infection (normal), (b) non-COVID-19, and (c) COVID-19. Original designs are developed to generate one of the three predictions. The rationale behind these three possible predictions is to help clinicians decide better, as the treatment approach depends on the cause of infection, who has to be identified for PCR testing for COVID-19 validation as COVID-19 and non-COVID-19 requires different plans to treat.

4 Model Creation

A machine-led research process is the second stage in the collective application approach for the human-machine interface used by the new COVID-Net. At this stage, more specifically, the initial network architecture prototype and details, along with the design criteria, serve as a guide to a design discovery strategy to learn and define the optimum micro-architecture designs with which a deep neural network-built architecture will be developed. Such an engineering mechanism powered by the computer allows much greater granularity and greater versatility than achievable through manual software design with a human controlled handling, while ensuring that the resultant deep neural network software satisfies operating criteria unique to domains. This is particularly important in the design of the COVID-Net, where the sensitivity of COVID-19 is needed for the number of cases lost to COVID-19 as much as possible.

5 Dataset Exploration

A minimum of 13,800 CXR photographs for 13,725 patients form the datasets for training and assessment of COVID-Net we recommend. In order to generate the COVIDx dataset, three separate publicly accessible datasets have been combined and revised. The most notable pattern being the small number of cases of COVID-19 infection and related CXR photographs, which illustrates the lack of publicly available COVID-19 case data, but also demonstrates the desire to collect further COVID-19 data as more case data becomes available. More precisely, from 121 COVID-19 medical cases, the COVIDx archive contains images. Much more hospital records and CXR images of no pneumonia and non-COVID-19 pneumonia are available. There are 8066 patients with no common pneumonia at all and 5538 non-COVID-19 patients with pneumonia.

6 Preprocessing

The suggested COVID-Net were pretrained on ImageNet [15] and instead practiced using a learning rate technique on the COVIDx dataset, which reduces the learning rate when for a certain amount of time the learning stagnates. The learning rate = $2e-5$, epoch number = 22, batch volume = 8, factor = 0.7, and patience = 5 for testing. The following hyper parameters were used for training. The data increased was also leveraged by encoding, flipping, horizontal flip, and pressure changing the following forms of increments. Finally, we have implemented a technique to increase the distribution of increasing form of infection by lots. The first COVID-Net implementation was built and tested with a Tensorflow backend using the Keras deep learning library.

7 Results and Discussions

They conduct quantitative and qualitative research in order to determine the efficacy of COVID-Net recommended, in order to achieve an increase in their identification efficiency and decision making. We assessed the precision and vulnerability for each type of infection in the above COVIDx data set in the quantitative analysis of the proposed COVID-Net [16]. The measurement reliable and statistical sophistication (in comparison with the number of parameters) are worked properly.

Those can be noted that by obtaining 92.6% test accuracy, COVID-Net achieves reasonable precision, thereby demonstrating the efficacy of using a joint human-machine modeling approach to speed up, tailor-made mission, data, and organizational necessity development of deep neural network architectures. Some of the sample images are shown in Fig. 2. It is especially relevant for situations like the

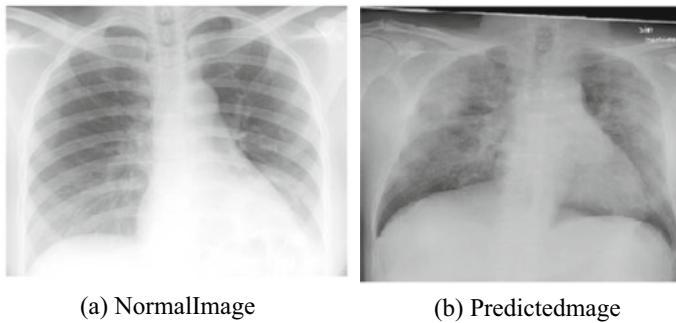


Fig. 2 Sample images in dataset directory

diagnosis of illness, which constantly accumulate new cases and new data and greatly value the opportunity to rapidly produce profound neural network architectures, which are adapted to the ever-changing knowledge base over time. Firstly, For COVID-19 cases (87.1%) that are published, COVID-Net will achieve strong sensitivity, so we want to reduce the number of missing COVID-19 cases as much as possible. While promising, the number of COVID-19 cases available in COVIDx is limited in comparison with other types of infections, which improves efficacy with the additional COVID-19 patient cases.

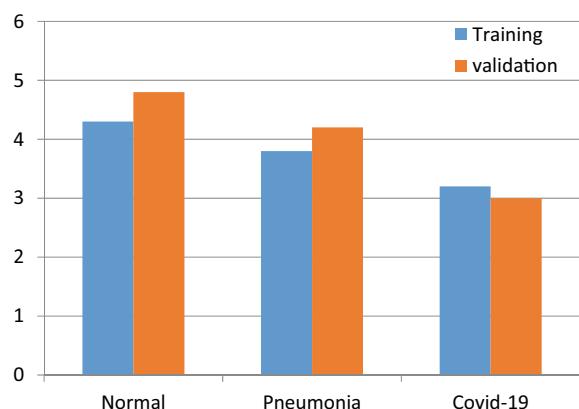
Furthermore, COVID-Net achieves high PPV in COVID-19 instances, suggesting relatively little incorrect COVID-19 positive identification. The high PPV is significant because the pressure on the health sector is compounded by many too many false positives, increasing the requirement for more PCR research and further care.

Third, the frequency of common infections is significantly higher than that of COVID-19 events [17]. For regular and non-COVID-19 instances, this finding can primarily be due to the significantly greater number of pictures.

Thus, based on these findings, while COVID-Net is well supported as a whole in the identification of CXR images in COVID-19 instances, some areas in improvements will gain from the additional knowledge gathering and enhancement of the underlying teaching methodology to generalize them more broadly. The prediction results are showed in Fig. 3. In no way, a production-ready approach is hoped that COVID-Net's promising findings on COVIDx test datasets, together with its open source model and definition on the construction of an open source dataset would enable both researchers and people to speed up the development of higher-quality data systems.

Future recommendations include constantly increasing susceptibility and PPV for COVID-19 infections, gathering new data, and applying the COVID-Net proposal for clinical diagnosis for safety study, estimation of threat status of a patient and estimation of hospitalization period to help in triage study, patient experience management and personalized care planning.

Fig. 3 Analysis report for COVID-19



8 Conclusion

During this research, we presented COVID-Net, a profoundly convolutional, open source as such; we were able to check that COVID-Net did not use inappropriate decision-making information for the detection of COVID-19 cases (e.g., wrong visual markers, embedded body signals, objects of imagery, etc.) which may give rise, for the wrong reasons, to situations when the right decisions are made. We identified COVIDx, an open-access data archive CXR data collection for COVID-Net which is composed of 13,800 CXR photographs in 13,725 patient instances. In addition, we explored how COVID-Net makes forecasts using an explainable approach to obtain greater understanding of essential variables for COVID situations, which will help clinicians, enhance their screening and raise their trust and clarity by utilizing COVID-Net for rapid device assistance screening.

References

1. Wong A, Shafiee MJ, Chwyl B, Li F (2018) Ferminets: learning generative machines to generate efficient neural networks via generative synthesis. arXiv preprint [arXiv:1809.05989](https://arxiv.org/abs/1809.05989)
2. Lin ZQ, Shafiee MJ, Bochkarev S, Jules MS, Wang XY, Wong A (2019) Explaining with impact: a machine-centric strategy to quantify the performance of explainability algorithms. arXiv preprint [arXiv:1910.07387](https://arxiv.org/abs/1910.07387)
3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223):497–506
4. Ng MY, Lee EY, Yang J, Yang F, Li X, Wang H, Lui MMS, Lo CSY, Leung B, Khong PL, Hui CKM (2020) Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiol Cardiothoracic Imaging 2(1):e200034
5. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L (2020) Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 200642

6. Lescure FX, Bouadma L, Nguyen D, Parisey M, Wicky PH, Behillil S, Gaymard A, Bouscambert-Duchamp M, Donati F, Le Hingrat Q, Enouf V (2020) Clinical and virological data of the first cases of COVID-19 in Europe: a case series. *Lancet Infect Dis*
7. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, Schluger NW, Volpi A, Yim JJ, Martin IB, Anderson DJ (2020) The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. *Chest*
8. Sendhilkumar NC, Ramesh GP (2020) Analysis of digital FIR filter using RLS and FT-RLS. In: *Advances in intelligent systems and computing*
9. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, Ji W (2020) Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology* 200432
10. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444
11. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, Bernheim A, Siegel E (2020) Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. *arXiv preprint arXiv:2003.05037*
12. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q, Cao K (2020) Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* 200905
13. Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, Tan W (2020) Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA*
14. Cohen JP, Morrison P, Dao L (2020) COVID-19 image data collection. *arXiv preprint arXiv: 2003.11597*
15. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition, June 2009. IEEE*, pp 248–255
16. Xu X, Jiang X, Ma C, Du P, Li X, Lv S, Yu L, Chen Y, Su J, Lang G, Li Y (2020) Deep learning system to screen coronavirus disease 2019 pneumonia. *arXiv preprint arXiv:2002.09334*
17. Pandey A, Prakash G (2019) Deduplication with attribute based encryption in E-health care systems. *Int J MC Square Sci Res* 11(4):16–24

Novel Approach to Monitor the Respiratory Rate for Asthma Patients



V. G. Sivakumar, S. P. Vimal, M. Vadivel, and V. Vijaya Baskar

Abstract It is found to be the fact that India has 18% of the world population and rising tension of chronic respiratory diseases. There is no proper understanding of the wide-spreading respiratory diseases, and their rapid updates are not available for states across the border of modern India. So, to hamper the death toll, I have proposed this system which continuously monitors Asthma patients using WSN. This proposed system is based on a flex sensor with a controller. This circuit is installed in a waist belt, which can monitor the breathing pattern of the patient continuously. Due to the variation in the flex sensor's value, the serial monitor displays the patients' live status. At times of any abnormalities, an SMS is triggered to the kin and friends of the patients. The suggested Respiratory Rate Monitoring System was checked and assessed on satisfactory findings.

Keywords Respiratory rate · Flex sensor · Arduino · Asthma patients · GSM

1 Introduction

Asthma affects up to 334 million people worldwide, and it has been a rising incidence for all the past three decades. It affects all genders, ethnic groups, but there is a broad difference inside the same nation in various countries and in different communities. This is the most prevalent genetic disease in kids and is more serious in un-affluent kids [1]. Drowsiness is considered to be expressed in the behavior not only of the nervous framework of the sympathetic nervous system. Therefore, by analyzing the ANS activity, it is considered to estimate somnolence prior to this; the PRC was used

V. G. Sivakumar (✉) · M. Vadivel
Vidya Jyothi Institute of Technology, Hyderabad, Telangana, India

S. P. Vimal
Sri Ramakrishna Engineering College, Coimbatore, India
e-mail: vimal.sp@srec.ac.in

V. Vijaya Baskar
Sathyabama Institute of Science and Technology, Chennai, India

to determine HRV [1]. A powerful HRV research method is used because it demonstrates changes in heart rate and rhythm during ventilation [2, 3]. The respiration rate (BR) is the cumulative breaths taken throughout a one-minute period are calculated when an individual is at rest. BR and HR are important signs to identify symptoms of cardiovascular diseases such as heart attack or Asthma [3]. The dissemination of breath analysis as a screening and monitoring method is decelerated by the device's expense, which experienced physicians can manage, and the shortage of systematic protocols for breath lab analysis. The International Organization for Breath Research works closely on the principle of standardized breath screening and analytical procedures [4]. Respiratory rate is among the main indicators of symptoms of a person's health. The respiratory rate of a human is measured in respirations [5]. The detector is centered on a plastic optical fiber (POF) in conjunction with the sufferer's torso and properly calibrated to be sensitive to normal breathing activity-induced malformations. This article describes a way to measure a person's respiration levels through thermography for Respiratory Alkalosis detection [6]. In the spectral images, we measure inhalation and exhalation to monitor the level of breathing [7–9]. In the above article, we have used both an embedded device and a camera to develop an embedded body breath detection monitoring system (EMSFBBD) that tracks the sufferer's breath and transmits the data on the Internet to a particular database [10]. In the UK, Asthma is a disease affecting about 5.4 million people. Alone in the UK every 10 s, someone has an Asthma attack, some of which are lifestyle-threatening. The dual focus on patient leadership Asthma is concerned with monitoring symptoms and the avoidance of Assaults of Asthma [11, 12]. The most common allergic signs are cough, gasping, stiffness of the chest, and shortness of breath. The bulk of the time, for most patients, the problem is secure, simple to handle, and these signs are either absent or moderate [13]. However, these signs could get worse after exposure to stimuli and leads to an attack [14]. And when patients feel fine, there is no solution for Asthma. They always have Asthma, and there could be flare-ups at any moment. Asthma ranges from moderate to extreme in intensity. In case of serious Asthma, problems in ventilation can be threatening [15]. The greatest hurdle in managing Asthma is regular compliance. Subsequently, this project aims to improve the information and use of Asthma action plans as a primary Asthma approach interference with routine symptom regulation.

2 Proposed System

The system architecture of proposed method is shown in Fig. 1, which constitutes a microcontroller which is the main processing unit in the proposed model, followed by flex sensors, a serial monitor, buzzer alarm, and a GSM model. Flex sensor is placed in the abdomen of a patient who is suffering from a respiratory problem like Asthma or immobile patients.

There are two sets of values obtained from the flex sensor. One set of values is obtained with the patients with a good breathing pattern and the other one with the

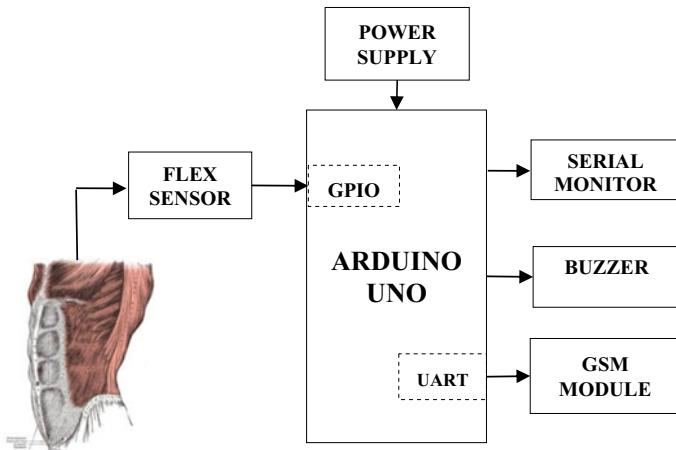
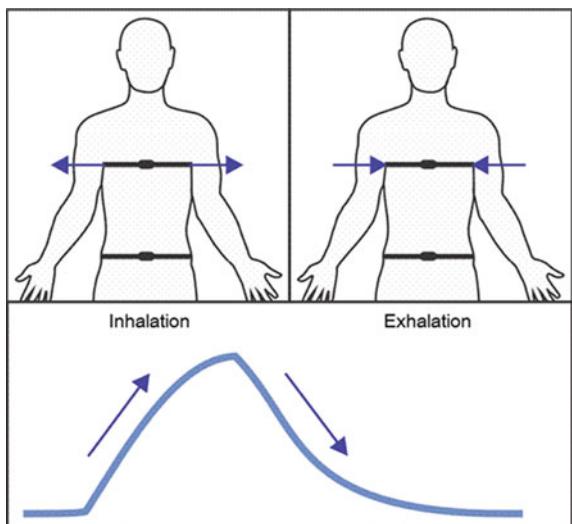


Fig. 1 Block diagram of the proposed method

patients who struggle to make out proper respiration. Figure 2 shows the breathing pattern of humans. So, once the proposed prototype is attached to the people with the problems mentioned above, it starts monitoring their breathing patterns continuously. If a person breathes normally, his monitored value will be printed in the serial screen, and again controller starts looking for the value. Whereas if a person does not breathe or his pattern looks somewhat similar to the second set of the pattern, then the value is monitored, and the controller triggers an SMS alert from the GSM module. Here, the controller commands the GSM module to go through a certain set of commands

Fig. 2 Breathing pattern of humans



which sends the message to the user. Finally, one can continuously monitor the respiratory rate of an Asthma patient using the proposed method.

3 Results and Discussions

Figure 3 shows the flex sensor placement in the body, which can be placed in the waist and the ribcage in which respiration movement can be widely seen.

Figure 4 shows the slow breathing pattern of the patient monitored through this flex sensor, and we can quite clearly see that breathe per minute is found to be 10. It is plotted between ribcage and volume.

Figure 5 shows the medium breathing pattern of the patient monitored through this flex sensor, and we can quite clearly see that breathe per minute is found to be 20. It is plotted between ribcage and volume.

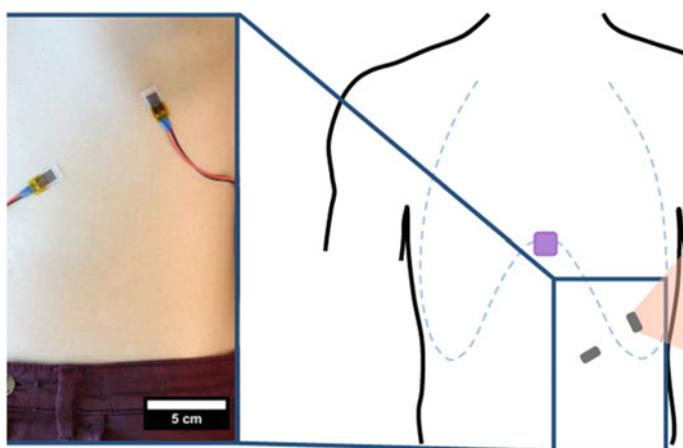


Fig. 3 Sensor placement

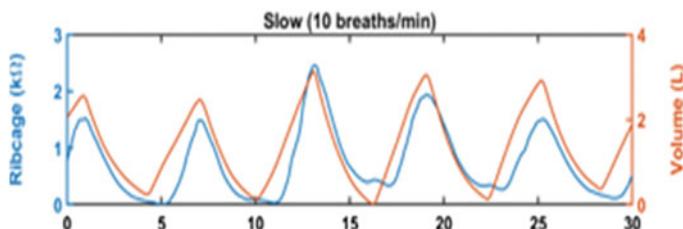


Fig. 4 Slow breathing pattern

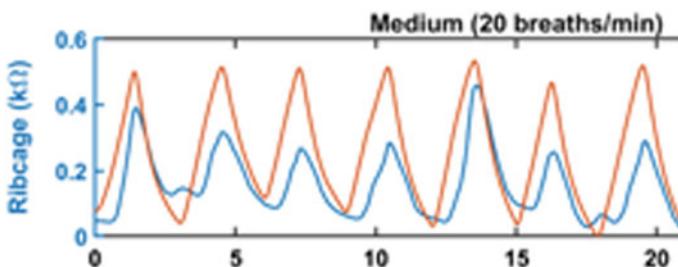


Fig. 5 Medium breathing pattern

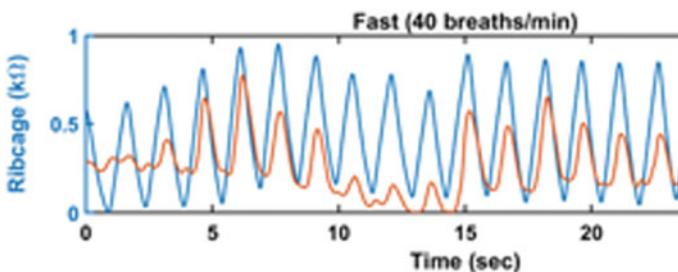


Fig. 6 Fast breathing pattern

Figure 6 shows the fast breathing pattern of the patient monitored through this flex sensor, and we can quite clearly see that breathe per minute is found to be 40. Graph is plotted between ribcage volume and time variation in seconds.

Figure 7 shows the hardware implementation of the proposed system. The figure shows the interfacing of a microcontroller with a flex sensor and LCD, which acts as a serial monitor in this model. In the shown picture, flex sensor is pushed backward, and based on the coding, the LCD shows that the flex sensor is bent backward. According to our necessity, we can code it as Patient is breathing. If no movement occurs for a few minutes, we can trigger the buzzer alarm so that the caretaker can interfere and take necessary actions accordingly. This system also gives an SMS alert via GSM to the caretaker to take precautions.

Figure 8 shows the voltage versus resistance graph of the flex sensor, which means the output voltage and resistance value when the flex sensor undergoes bend or tension over it. As we can see from the above graph, the maximum to minimum voltage ranges from 1.78 to 1.32 V. Similarly, resistance varies from 9.48 to 22.7 k Ω as inferred from Table 1.

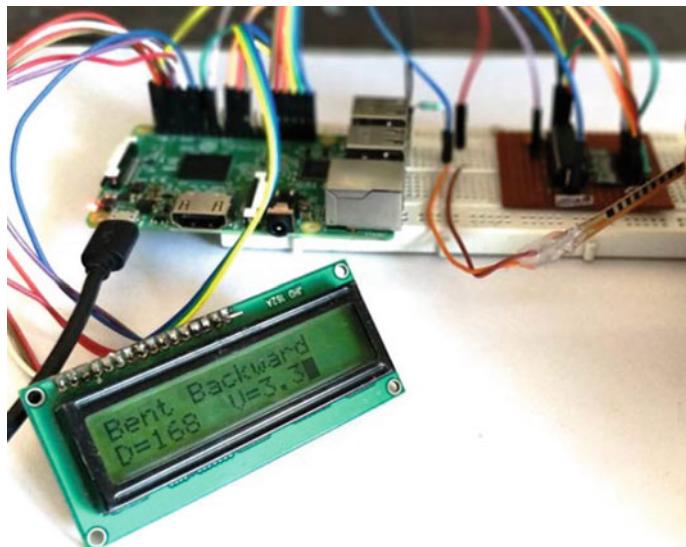


Fig. 7 Implementing the proposed hardware device

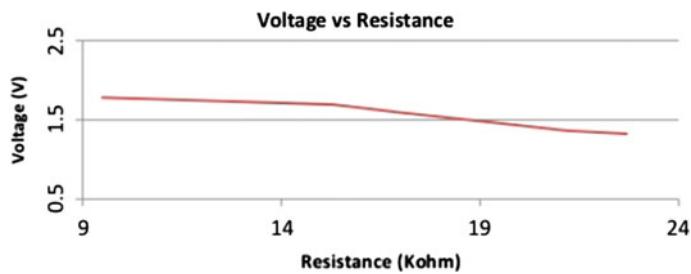


Fig. 8 Voltage versus resistance graph

Table 1 Value of resistance and output voltage of flex sensor for applied tension

Flex sensors (kΩ)	Vout (V)
9.48	1.78
15.3	1.69
17.0	1.59
21.2	1.36
22.7	1.32

4 Conclusion

This proposed prototype will reliably calculate the breathing pattern or respiratory rate of a person as opposed to portable respiratory rate monitoring devices based on the temperature probe and MEMS sensor. It is important to remember that medical gadgets are very costly, and these gadgets rarely have functions for monitoring the respiratory rate. And suggested systems implementation cost is very low when compared with gadgets and can be used in emergency care units in hospitals. As tested so far, gadgets or devices based on accelerometer and thermistor provide inaccurate readings, but this issue has been significantly reduced by the model examined in this article. Therefore, the proposed program meets all of the above-mentioned objectives. The proposed protocol is tested and verified in various conditions. This system gave output with high efficiency.

References

1. Igasaki T, Nagasawa K, Murayama N, Hu Z (2015) Dizziness prediction by cardiac output and/or respiratory rate variability in driving conditions with the logistic regression model. In: 2015 8th international conference on biomedical engineering and informatics (BMEI). IEEE, pp 189–193
2. Jaafar R, Rozali MAA (2017) Breathing rhythm & pulse rate estimate from a photoplethysmogram. In: 2017 6th international conference on electrical engineering and informatics (ICEEI). IEEE, pp 1–4
3. Lomonaco T, Salvo P, Ghimenti S, Biagini D, Bellagambi F, Fuoco R, Di Francesco F (2015) A breath sampling method that tests the effect of the breathing rate on the quality of the exhaled air. In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 7618–7621
4. Alam MM, Hussain M, Amin MA (2019) An innovative configuration of a push switch circuit & Arduino microcontroller breathing rate monitoring device. In: 2019 international conference on robotics, electrical and signal processing techniques (ICREST). IEEE, pp 470–473
5. Vallan A, Carullo A, Casalicchio ML, Penna A, Perronae G, De Vistro N, Milella A, Fracassi F (2014) For breath rate control, a plasma adapted fibre sensor. In: 2014 IEEE international symposium on medical measurements and applications (MeMeA). IEEE, pp 1–5
6. Hemanth K, Ramesh GP (2020) Energy efficiency and data packet security for wireless sensor networks using African buffalo optimization. IJCA 13:944–954
7. Swarnalatha A, Manikandan M (2020) Intravascular ultrasound image classification using wavelet energy features and random forest classifier. In: Advances in intelligent systems and computing
8. Bai Y-W, Li W-T (2010) Development & deployment of a body breath monitoring embedded display device using image processing techniques. IEEE. ISBN 978-1-4244-4316-1/10/
9. Kumar GH, Ramesh GP (2018) Novel gateway free device to a device communication technique for IoT to enable direct communication between homogeneous devices. Int J Pure Appl Math 118(16):565–578
10. Dam QB, Nguyen LT, Nguyen ST, Vu NH, Pham C (2019) e-Breath: breath detection and monitoring using frequency cepstral feature fusion. In: 2019 international conference on multimedia analysis and pattern recognition (MAPR). IEEE, pp 1–6
11. Ravichandran S (2017) High-end street lighting system with RTOS connecting to the Internet. Int J MC Square Sci Res 9(1):331–334

12. Shamsudheen S (2019) Smart agriculture using IoT. *Int J MC Square Sci Res* 11(4):25–33
13. Hemanth Kumar G, Ramesh GP (2019) Reducing power feasting and extend the network lifetime of IoT devices through localisation. *IJAST* 28(12):297–305
14. Zeleke B, Demissie M (2019) IoT based lawn cutter. *Int J MC Square Sci Res* 11(2):13–21
15. Fahad AAA (2019) Design and implementation of a blood bank system using web services in a cloud environment. *Int J MC Square Sci Res* 11(3):09-16

A Systematic Research on Identifying Mental Disorders in Social Networks Using Online Social Media Mining



S. Sai Jayanth, Shaik Nakarikanth Abja, and A. Mary Posonia

Abstract A novel tensor model that effectively integrates heterogeneous data from various stages was used to find social network mental disorder (SNMD) patients at an early stage based on their data. There have been studies of an increase in social network mental disorders such as cyber-dependency, information overload, and network limitation (social network mental disorder). Today, the symptoms of these psychiatric illnesses are often passive, resulting in a pause in clinical intervention. In this paper, we argue that online mining's social behavior enables us to effectively detect social network mental illness at an early stage. The mental variables used in the criteria (questions) cannot be viewed in social Internet behavior journals, making social network mental illness difficult to find. The diagnosis of social network mental illness in this groundbreaking new approach is not focused on self-disclosure of these mental factors by questionnaires. Instead, we deliver mental disorders finding (social network mental illness), a machine learning algorithm that uses features extracted from social network data to reliably identify potential social network mental disorder occurrences.

Keywords Mental disorder · Social network · Illness · Machine learning · Behavior

1 Introduction

Internet addiction disorder (IAD) is a type of behavior addiction with the patients dependent to the network just like those addicted to drug or alcohol [1]. Aroul et al. [2] investigate the issue, of simulated gambling via digital social media to inspect the connection of non-identical factors, e.g., grade and ethnicity. Mary and Baburaj [3] survey the risk factors related to Internet addiction. Vigneshwari and Aramudhan [4] look over the association of sleep quality and suicide attempt of Internet mental

S. Sai Jayanth · S. N. Abja · A. Mary Posonia (✉)

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

disorder addicts. Investigation shows that the teen age people with narcissistic tendencies and explicit item promoting to reserved shyness are particular unprotected to addiction with orbit showtime network (OSN) [5, 6]. Organizations and associations have consistently been worried about how they are seen by people in general. This worry results from an assortment of inspirations, including advertising and advertising. Saha et al. extract the social network current and linguist features from networked social media for depression patients to survey their patters. Usha et al. [7] survey sentimental and linguistic style of social media network data for critical depression disorder [MDD].

2 Literature Survey

Examine the issue of replicated betting using advanced and Internet-based technology to dissect the relationships between different variables, such as grade and ethnicity [8]. Examine the risk factors linked to Internet compulsion. Srilatha and Ulagamuthalvi [9] investigate the connection between Internet addicts' sleep quality and their desire to commit suicide [10]. Research demonstrates that youngsters with narcissistic inclinations and modesty are especially powerless against compulsion with OSNs [5, 6]. Reference [1] utilizes a based way to deal with gather and concentrate phonetic and content-based highlights from online Internet-based life to recognize borderline personality disorder and bipolar disorder patients. References [11, 12] extricate the topical and phonetic highlights from online Web-based life for discouragement patients to dissect their examples. Sathya Bama Krishna et al. [13] examine feeling and semantic styles of online networking information for major depressive disorder (MDD). The likelihood of mortality as a dormant state developing after some time. Srilatha and Ulagamuthalvi [9] propose a progressive learning strategy for occasion recognition and finding by first removing the highlights from various information sources and afterward learning through geological staggered model. Few more works using social network can be found in [14].

3 Proposed System

Currently, I am looking into data mining techniques to differentiate three forms of social media. Today, the symptoms of these psychiatric illnesses are often passive, resulting in a pause in clinical intervention. In this paper, we argue that online mining's social behavior enables us to effectively detect social network mental disorder at an early stage. We take the previous data from the social media for the classification of the data. The first phase of Fig. 1.

Project is to collect data from the various resources, and extraction process undergoes we extract the data from the database. We use MySQL software for the backend process. Feature extraction also done to get data more efficiently and fast. There

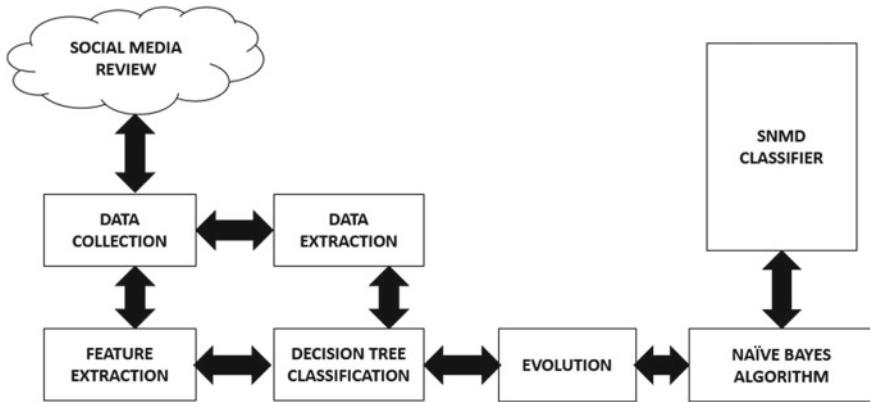


Fig. 1 Architecture diagram

will be login page containing many phases like data collection data, training and last phase is prediction, and these are all can be done by the user after logging into it. We use decision tree and Navies algorithm in which we take data from the database. Analyzing the data from database, we train the model and perform the prediction process based on training. The prediction results shows the percentage of disease occurred in system. A system architecture which is the conceptual model that defines to find the data collection and data extraction of the structure, behavior, and more of system. An architecture description is a formal to find the evolution to name and algorithm description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system (Fig. 1).

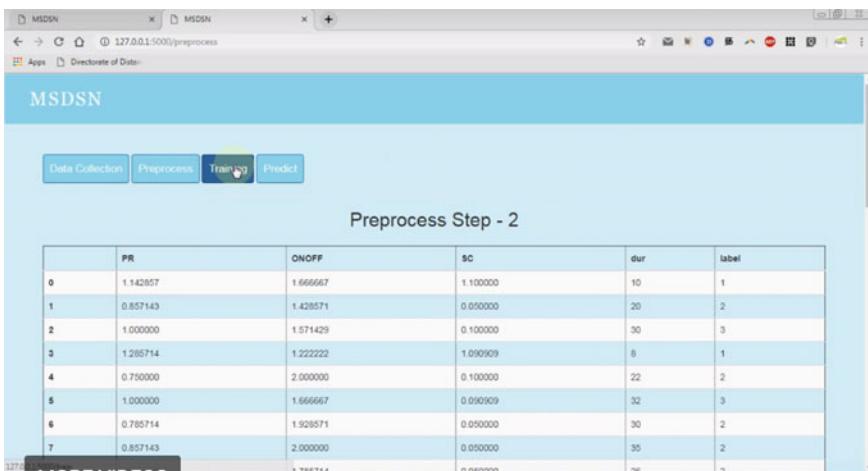


Fig. 2 User–login interface

In user-login interface used (Fig. 2), collect the number of customer data collection process. The step-by-step process like adding client name, age, gender, how many hours they spending their time on their mobiles, and how much data they are using in 24 h, in order to display the procedure for data collection in stepwise in the table.

The data collection is the process of gathering and measuring information of targeted 40 values variables to find the mental detection collecting all values of using other networking an established system, which then enables one to answer relevant questions and evaluate outcomes to find prediction of aim, arm strong, strength, illness, addiction as shown in Figs. 3 and 4.

In training step, we have to enter client range values and the range should be specific in all the entries. After entering the values, it shows the net compulsion

6	11	14	27	14	2	40	35	6	12	30	2
7	12	14	28	14	2	40	36	8	12	35	2
8	11	14	25	14	2	40	36	8	12	26	2
9	10	12	18	18	22	20	4	22	4	10	3
10	12	16	20	14	2	30	26	4	12	20	2
11	12	16	22	14	2	20	28	2	10	30	2
12	23	20	34	30	23	23	12	3	3	5	1
13	26	15	24	15	21	20	4	22	4	10	1
14	36	16	25	17	22	20	4	21	4	11	1
15	26	17	26	18	24	20	4	19	4	1	1
16	14	14	26	14	2	29	26	12	14	30	3
17	10	10	23	14	2	32	26	23	10	30	3
18	15	15	24	14	2	35	26	11	15	30	3
19	31	24	31	25	32	20	4	20	4	10	1

Fig. 3 Data collection

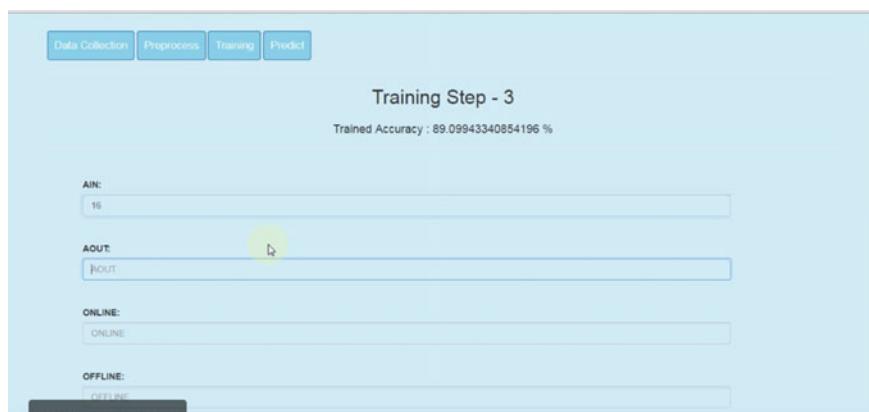


Fig. 4 Data training accuracy

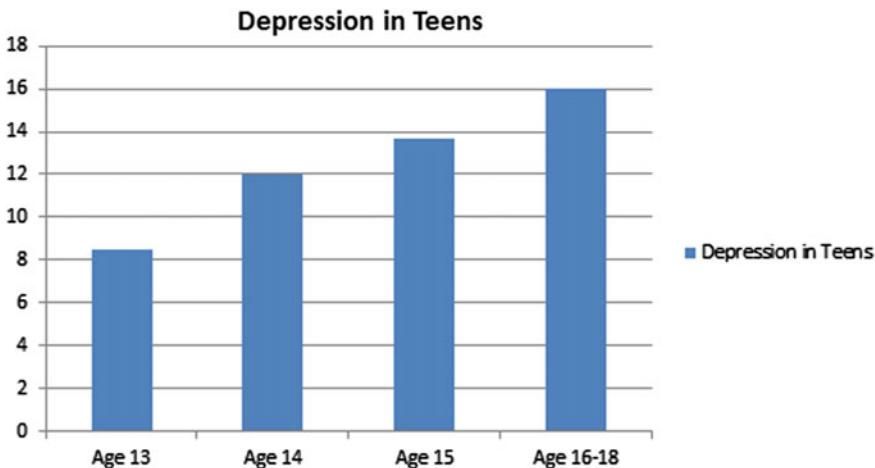


Fig. 5 Performance analysis

[NC] diagram. This step shows the mental disorders, if you want to change the graph prediction, change the range values and if we are getting a runtime error, we can play it from data collection.

4 Results and Discussion

The performance analysis of the project which is being developed is being inspected in processing with the previously used data. The visual representation of the study in graphical format of the performance with respect to the ease of user to automatically recognize SNMD patients at an early stage based on their OSN data using a novel tensor model that efficiently integrates heterogeneous data from various other social networks [OSNs] is shown in Fig. 5.

The SNMD data from different OSNs may be incomplete due to the heterogeneity (drawback). For example, the profiles of users may be empty due to the privacy issue, different functions on different OSNs (e.g., game, check-in, event), etc.

5 Conclusion

The proposed social network mental disorder structure investigates different highlights from information logs of other social network [OSNs] and a new constant approach for obtaining inactive characteristic among different other social network [OSNs] for social network mental disorder [SNMD] observation. This work means

collective attempt in the middle of computer scientist and mental healthcare research to address emerging in social network mental disorder [SNMDS].

References

1. Lin I-H, Ko C-H, Chang Y-P, Liu T-L, Wang P-W, Lin H-C, Huang M-F, Yeh Y-C, Chou W-J, Yen C-F (2014) The association between suicidality and Internet addiction and activities in Taiwanese adolescents. *Compr Psychiatry*
2. Aroul Canessane R, Dhanalakshmi R, Anu VM (2019) Implementation of tensor flow for real-time object detection. *Int J Recent Technol Eng* 8(2 Special Issue 11):2342–2345
3. Mary SP, Baburaj E (2013) Genetic based approach to improve E-commerce web site usability. In: 2013 5th international conference on advanced computing, ICoAC 2013, pp 395–399
4. Vigneshwari S, Aramudhan M (2015) Web information extraction on multiple ontologies based on concept relationships upon training the user profiles. In: Artificial intelligence and evolutionary algorithms in engineering systems. Springer, New Delhi, pp 1–8
5. Mary Posonia A, Kanmani Rajathi J (2015) Infrequent weighted item set mining using decision making approach algorithm. *Int J Appl Eng Res* 10(3):6817–6826
6. Chak K, Leung L (2004) Shyness and locus of control as predictors of internet addiction and internet use. *Cyberpsychol Behav*
7. Usha Nandini D, Sathyabama Krishna R, Nithya M, Pavithra R (2019) A resourceful information collecting system using smart black box. *J Comput Theor Nanosci* 16(8):3346–3350
8. Mary Posonia A, Vigneshwari S, Albert Mayan J, Jamunarani D (2019) Service direct: platform that incorporates service providers and consumers directly. *Int J Eng Adv Technol (IJEAT)* 8(6). ISSN: 2249 – 8958
9. Srilatha K, Ulagamuthalvi V (2019) Support vector machine and particle swarm optimization based classification of ovarian tumour. *Biosci Biotechnol Res Commun* 12(3):714–719
10. Young K (1998) Internet addiction: the emergence of a new clinical disorder. *Cyberpsychol Behav*
11. Young K, Pistner M, O'Mara J, Buchanan J (1999) Cyber-disorders: the mental health concern for the new millennium. *Cyberpsychol Behav*
12. Ankayarkanni B, Ezil Sam Leni A (2016) GABC based neuro-fuzzy classifier with multi kernel segmentation for satellite image classification. *Biomed Res. Special issue S158–S165*. ISSN 0970-938X
13. Sathya Bama Krishna R, Monica K, Aramudhan M (2015) Comparative analysis on spectral unsupervised feature selection: a review. *Int J Appl Eng Res* 10(2):2235–2240
14. Dey N, Borah S, Babo R, Ashour AS (2018) Social network analytics: computational research methods and techniques. Elsevier. ISBN: 9780128156414

Removal of Outliers and Missing Values in Diabetes Dataset Using Ensemble Method



M. D. Anto Praveena and B. Bharathi

Abstract Missing data imputation is an ongoing and crucial research topic in data mining. There may be many missing values in large dataset. However, there are few methods used solely for downstream analyses, with a few prediction tools, which definitely do need a full descriptor value matrix. We propose and assess a looping imputation method called ensemble based on few imputation methods. By calculating an average over a lot of regression trees which are unpruned, the ensemble method intrinsically constitutes a multiple imputation scheme. Using bagging estimate, boosting estimate and stacking estimate of the ensemble method, we are able to estimate the imputation error. Evaluation is finished on molecular descriptor datasets generated from a diverse choice of pharmaceutical fields with artificially delivered missing values ranging from 10 to 30%. The experimental end result demonstrate that missing values have an amazing impact at the effectiveness of imputation strategies and our approach ensemble is sturdier to missing values than the alternative ten imputation strategies used as benchmark. Additionally, the ensemble method exhibits appealing computational performance and can address high-dimensional data.

Keywords KNN · Stochastic regression imputation · Ensemble method · Simple regression · MICE · Multiple regression · Imputation

1 Introduction

The nature of molecular descriptor information complicates the improvement of highly accurate predictive models. Molecular descriptor records are commonly unevenly gathered as mentioned in [1]. These empty or unanswered values in statistics units are named missing values (information), and are of a trouble most researchers

M. D. Anto Praveena (✉) · B. Bharathi

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai 600119, India

B. Bharathi

e-mail: bharathi.cse@sathyabama.ac.in

face. Missing information may rise up for several reasons. For instance, by using twist of fate or a few molecules descriptor generator fails to offer descriptor data. Missing values imputation (MV) is usually an essential entry records assessment as mentioned in [2]. MV imputation stays an essential key step in facts preprocessing. Since many down-circulate analyses require a complete records set for implementation, MV imputation is a commonplace practice.

Many proven analytical procedures require completely discovered datasets of the molecular descriptors with none missing values. Even in today's pharmaceutical and agriculture science, this is rarely the case. The continuous development of new and improved measurement techniques in these fields creates challenges for data analysts not only due to high-dimensional multivariate descriptor data where the number of descriptors can greatly exceed the variety of observations where there are continuous descriptors.

Our aim is to implement a means of imputation that can cope with any kind of molecular descriptor data and make as few assumptions as possible about the statistics' structural elements. Random forests are ready to address specific valued-type information and allow interactive and nonlinear (regression) effects as a non-parametric technique. We soak up the lack of information by using an imputation scheme which loops in a first step by schooling an RF on observed values, followed by predicting the missing values after which iteratively moves on. We selected RF because under barren conditions such as excessive dimensions, complex iterations, and nonlinear record structures are known to be performing very well. Random forest is well suited for use in applied studies because of its precision and vigor, sometimes reminiscing of such circumstances.

Here, we compare our method with several other imputation methods. These imputation methodologies are applied on sub atomic descriptor datasets. Missing qualities are demonstrated by NaN in pandas as in [3]. We show that our methodology is serious to or outflanks the looked at strategies on the utilized datasets independently of the variable sort arrangement, the information dimensionality, the wellspring of the information or the measure of missing qualities.

In some samples of functions, error of imputation is reduced by up to 30%. The performance is usually achieved in just a few iterations which also makes our system computationally attractive. The ensemble method estimates give a very good approximation of the true imputation error having on average a proportional deviation of no more than 15–20%. In addition, our method needs no tuning parameter, and hence is easy to use as referred from [4].

2 Related Work

2.1 KNN

K Nearest Neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new data points, which means that a value will be assigned to the new data point based on how closely it matches the training set points. Using the following steps, we can understand its work.

Step 1: For implementation of any algorithm, we need dataset. Therefore, during the first phase of KNN, we load both the training and test dataset.

Step 2: Next, we’ll pick the value K , i.e., the nearest data points. K can be replaced as integer.

Step 3: Do the following for each of the test data points:

- 3.1 Using one of the methods namely: distance from Euclidean, Manhattan or Hamming measure the distance between test data and each training data row. Euclidean is the most widely used instrument for distance measurement.
- 3.2 Now sorting them in ascending order, based on the distance interest.
- 3.3 Next, select the top K rows from the sorted list.
- 3.4 Now, a class based on the most frequent class of those rows will be allocated the test point.

Step 4: End

2.2 Regression Imputation

Predicted value obtained by reverting the missing variable to other variables. So you are taking the expected value, based on other factors, instead of just taking the mean. This preserves relationships between variables that are involved in the model of imputation, but does not display variation around values which are predicted as referred from [5].

3 Proposed Work

Stochastic regression is a very effective type of supervised machine learning, as shown in Fig. 1 which tries to find the best possible straight line to describe main trend in underlying training data. It starts from a random position and then it adjusts or moves the line until and unless it finds the position that provides the minimum total average distance of data from the line. It uses mean squared error as a loss function

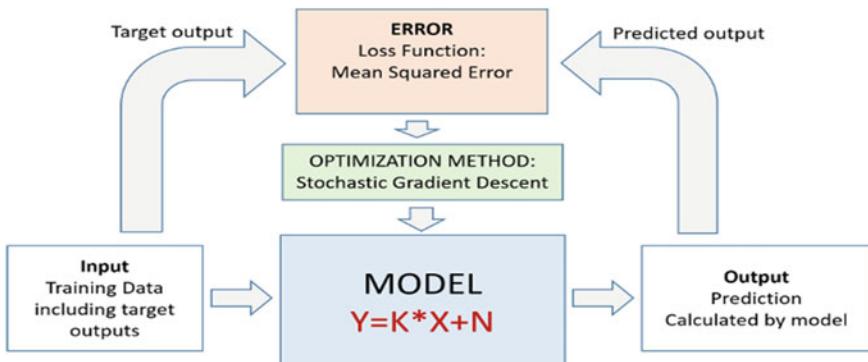


Fig. 1 Linear regression as supervised learning algorithm. Refer Zoran Sevarac [10]

and edge of stochastic gradient as an optimization algorithm. Stochastic method of optimization of the gradient descent knows how to move line in a direction that will probably reduce the error. In its general form, for single input x , and single output y , this algorithmic model is nothing short of a canonical mathematical formula used in a no longer coiled line $y = k * x + n$ and train it to find out values of k and n . The same technique can be used in multiple-input problems.

3.1 Architecture Diagram

See Fig. 1.

3.2 Stochastic Regression

Why would we need stochastic regression? We could just delete the entire row of data to help with the situation but since we are talking about medical dataset here, we have to ensure that we don't remove the data entirely instead of filling them, which is called common case analysis. If we replace the missing value with a common value, we could just under or overestimate the value. In case of diabetes dataset wherever we find a missing value we put **zero** value in there and further we replace the value with **NaN** calculate their numbers. We add some missing values with the values we got after prediction. Since the predicted values might lie along the regression hyper plane where we might encounter some actual noise. Then, we add some natural or normally distributed noise with a mean value of zero and a variance equal to the standard of the regression estimates to apply the uncertainty back to the imputed variable values. This method is called stochastic regression. In order to use regression to impute our values, first the dataset must have non empty values since regression can't run

without values so we insert random values in places where the values are missing and we do the following to impute the values.

Stochastic Regression Imputation algorithm in a for loop for all the variables of a dataset, while adding rows in R.

If the Train set have 1300 rows, the remaining 1700 rows are the Test set. First, fill in the unknown fields with random values in order to run regression on it.

1. For each column fit a linear model of this form: $Y \sim X_1$.
2. Use the model created to get the predicted value of the Y by using the first X_1 of the Test set.
3. After that take the first row of the Test set and rebind it to the Train set (now the Train set is 1301 rows).
4. Predict the value of Y using the 2nd row of X_1 from the Test set.
5. Repeat for the remaining 1699 rows of the Test set.
6. Apply it for all the remaining variables of the datasets (X_2, \dots, X_{14}).

Error concept formula decomposed as bias, uncertainty, and irreducible error.

The middle term is squared (Bias), and the last term is variance. You can see that Big E, which causes all of the hindrance in measuring the two terms. This is because you need to have the knowledge of the true population to measure the expectation value of some word.

So to decompose my prediction error in bias and variance, I shake hands with the old simulation and random data generator.

Below are the steps I have taken in doing so:

- (1) Self-defining a relationship that is to say defining the coefficients that we will call true population.
- (2) Use this to create a target variable, and add some noise.
- (3) Simulate n such different datasets, and based on them various linear as well decision tree.
- (4) Calculate the bias amount and variance that leverages the knowledge from phase 1.

3.3 Ensemble Method

Ensemble modeling is the method of running or extra linked but one-of-a-kind analytical fashions and then synthesizing the results into an unmarried source or spreading on the way to improving the accuracy of predictive analytics and facts mining [6]. For each modeling and statistical analysis, a single model can have biases, unnecessary uncertainty or outright inaccuracies that impair the reliability of its empirical results, mainly based on a single-statistical sample. Using special simulation methods can yield similar disadvantages. Through mixing extraordinary fashions or evaluating a few samples, scientists of facts and other analysts of facts will mitigate the impact

of those limitations and provide better statistics for business selection makers. One common example of modeling the ensemble is a random model of the forest. This approach to documenting mining leverages a few trees of choice, a kind of analytical model which is designed to predict results based entirely on exclusive variables and rules. A random version of the forest allows the decision trees that can look at one sample statistic of a kind, compare different factors or weight similar variables differently. The results of the various selection bushes are then either converted into a simple average or aggregated by additional weighting. Ensemble modeling has risen in popularity as the computational capabilities and superior analytics tool needed to run have been implemented by larger companies. However, the advent of Hadoop and other big data technologies has led companies to shop and evaluate additional quantities of information, gaining increased ability to walk analytical fashions on one of a kind data samples.

Ensemble approaches could be broken down into two classes:

Sequential ensemble methods in which the base learners (for example Adaboost) are generated consecutively. The fundamental motivation for consecutive approaches is the misuse of simple learner dependence [7]. Weighing already mislabeled models with a higher weight will help general execution.

Parallel ensemble approaches where the basic learners are generated simultaneously (e.g., Random Forest). The basic motivation is for fair strategies is to abuse independence among the base learners since the error can be significantly reduced by averaging. Many ensemble methods implement a single-basic learning algorithm that produces homogeneous base-learners, which is similar type learners, resulting in a homogeneous group [8].

There are also a few strategies that uses heterogeneous learners to activate heterogeneous ensembles, for example students of different kinds [9]. To be more precise than any of its individuals for example ensemble approaches, the base must be reliable as reasonably expected and as diverse as prudent.

3.4 Bagging

Bagging reflects the accumulation of bootstrap. One approach to decreasing an indicator fluctuation is to combine multiple measures together. For example, we can train M different trees on different subsets of the information (haphazardly picked with substitution) and register the ensemble:

$$f(x) = \frac{1}{M} \sum_{m=1}^M f_m(x)$$

Bagging uses bootstrap sampling to get the datasets to train the learners in the base.

Algorithm

```

1: Init data weights  $\{w_n\}$  to  $1/N$ 
2: for  $m = 1$  to  $M$  do
3:   fit a classifier  $y_m(x)$  by minimizing weighted error function  $J_m$ :
4:    $J_m = \sum_{n=1}^N w_n^{(m)} 1[y_m(x_n) \neq t_n]$ 
5:   compute  $\epsilon_m = \sum_{n=1}^N w_n^{(m)} 1[y_m(x_n) \neq t_n] / \sum_{n=1}^N w_n^{(m)}$ 
6:   evaluate  $\alpha_m = \log(\frac{1-\epsilon_m}{\epsilon_m})$ 
7:   update the data weights:  $w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m 1[y_m(x_n) \neq t_n]\}$ 
8: end for
9: Make predictions using the final model:  $Y_M(x) = \text{sign}\left(\sum_{m=1}^M \alpha_m y_m(x)\right)$ 

```

3.5 Adaptive Boosting Algorithm

Stacking algorithms are typically based on linear models, which, especially when predictions are highly correlated, may run into problems. In this review, we construct a greedy model stacking algorithm that overcomes this problem, while remaining very quick and easy to understand.

```

1: Input: training data  $D = \{x_i, y_i\}_{i=1}^m$ 
2: Output: ensemble classifier  $H$ 
3: Step 1: learn base-level classifiers
4: for  $t = 1$  to  $T$  do
5:   learn  $h_t$  based on  $D$ 
6: end for
7: Step 2: construct new data set of predictions
8: for  $i = 1$  to  $m$  do
9:    $D_h = \{x'_i, y_i\}$ , where  $x'_i = \{h_1(x_i), \dots, h_T(x_i)\}$ 
10: end for
11: Step 3: learn a meta-classifier
12: learn  $H$  based on  $D_h$ 
13: return  $H$ 

```

4 Result and Discussion

Various imputation methods are compared with the resultant datasets, and their accuracy is recorded.

Table 1 shows the various imputation and regression comparisons of the methods such as Simple Regression, KNN method, MICE regression, and stochastic regression method. It shows that the stochastic regression method has the most value when it comes to average in terms of precision in the field of medical dataset, keeping

Table 1 Comparison of imputation methods

Imputation method	Accuracy
Simple	75.4
KNN	77
MICE	76.1
Stochastic regression imputation	77.6

diabetes in mind. Hence, stochastic regression is the best when it comes to fill the places whose values were not present or missing.

5 Conclusion

Notwithstanding the techniques contemplated right now is regular to utilize ensembles in deep learning via training diverse and exact classifiers. To reach assorted variety, we will have to change models, hyper-parameter settings, and training strategies. We also henceforth see that our comparison directs us into concluding that stochastic regression brings in more accuracy when it comes to filling missing data or values in medical field.

References

- Christy A, Gandhi MG, Vaithyasubramanian S (2015) Cluster-based outer health care data detection algorithm. Procedia Comput Sci 50:209–215
- Roobini MS, Lakshmi M (2019) Classification of diabetes mellitus using soft computing and machine learning techniques. Int J Innov Technol Explor Eng 8(6S4). ISSN: 2278-3075
- Praveena MDA, Krupa JS, SaiPreethi S (2019) Statistical analysis of medical appointments using decision tree. In: 2019 fifth international conference on science technology engineering and mathematics (ICONSTEM), Chennai, pp 59–64
- Jabbar MA et al (2012) An evolutionary algorithm for heart disease prediction. In: Communications in computer and information science, vol 292. Springer Verlag, pp 378–389
- Aravind KRNVVD, Prayla Shyry S, Felix Y (2019) Classification of healthy and rot leaves of apple using gradient boosting and support vector classifier. Int J Innov Technol Explor Eng (IJITEE) 8(12). ISSN: 2278-3075
- Little RJ (1988) Missing-data adjustments in large surveys. J Bus Econ Stat 6:287–296
- Sariyar M, Borg A, Pommerening K (2011) Missing values in deduplication of electronic patient data. J Am Med Inform Assoc 19:76–82
- Kurz CF, Maier W, Rink C (2020) A greedy stacking algorithm for model ensembling and domain weighting. BMC Res Notes 13:70. <https://doi.org/10.1186/s13104-020-4931-7>
- Cox A, Rutter M, Yule B, Quinton D (1977) Bias resulting from missing information: some epidemiological findings. J Epidemiol Community Health 31(2):131–136
- Zoran Sevarac, From basic machine learning to deep learning in 5 minutes. Deepnetts. <https://www.deepnetts.com/blog/from-basic-machine-learning-to-deep-learning-in-5-minutes.html>

Literature Survey: Computational Models for Analyzing and Predicting the Spread of the Coronavirus Pandemic



Anubhav Soam, Kapeesh Kaul, and S. Ushasukhanya

Abstract Viral diseases are extremely widespread infections caused by viruses, which is a type of microorganism. Some of the common curable viral diseases are common cold, flu, pneumonia mumps, measles, etc. In addition to this, there are also some deadly viral diseases are human immunodeficiency virus (HIV), human papillomavirus (HPV), SARS, Ebola, etc., which is incurable. The recent coronavirus has also taken its place in this latter list for which the vaccine is yet to be discovered. As early diagnosis is the only option as of now which could control the death rate of this disease, several researchers are in the process of inventing drugs and vaccines for the same. At this stage, it is vital to develop some automated systems that could possibly detect the virus's presence at an early stage. Numerous scholarly articles concerning proposing computational models encompassing the spread of the coronavirus disease have been studied, analyzed, and juxtaposed with an aim to determine the optimality and accuracy of various models. This work aims to develop a collective study on the models developed so far for the prediction and spread of coronavirus.

Keywords Coronavirus · Deep learning · Convolutional neural network · Prediction

1 Introduction

A newly discovered, highly infectious coronavirus has spread like wildfire across the planet ever since its encounter with humanity in 2019. In the first few initial days after its discovery, the disease was limited to the city of Wuhan. However, it had spread worldwide within a quarter. Those who come in contact with the virus, firstly, show symptoms similar to that of the normal flu, i.e., majorly respiratory problems. In most cases, patients recover naturally without the need for any medical assistance. However, in some cases, patients with underlying medical conditions like heart diseases, diabetes, hypertension, etc., develop serious symptoms. The virus

A. Soam · K. Kaul (✉) · S. Ushasukhanya

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

spreads primarily through bodily fluids, namely—saliva and nasal mucus, which in turn means, spread through coughing and sneezing [1].

To the best of our knowledge, limited to what information is available to the public, currently, there is no readily available medical cure for the coronavirus disease. Considering recent technological developments with the aid of computer-assisted tools, various mathematical models have been tested against the actual curve to understand how the disease spreads [2]. For instance, models such as regression models and SEIR models have been all proposed by scholars to fit the curve. The possible implications of a system that would be able to predict the outburst of the disease accurately are endless. With its help, the government would be able to make informed decisions regarding the pandemic, such as determining the need for medical facilities based on the predicted numbers so that the country is already well-prepared for the huge influx of patients when the time comes. This exercise aims to review these aforementioned mathematical models and evaluate their feasibility in using them in a computational setting for such a prediction engine.

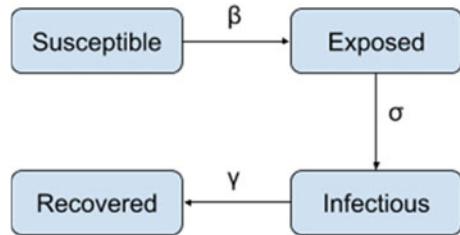
2 Related Work

All things considered using machine learning algorithms seem appropriate. Firstly, the idea of using pre-existing data to predict future outcomes is appropriate for this situation. In machine learning technical terms, the pre-existing data is considered as “training data” which in turn is used to train the model [3]. After this, the computer is tasked with performing calculations on the said training dataset. Now, as a result of these calculations, depending on the ML model used, its respective parameters are determined which will aid in the prediction of latter values [4]. Conclusively, the main idea behind this set of techniques is to “learn from experience.” Such a motto is fitting for this particular use case. Starting with regression analysis algorithms [5], these algorithms are based on the principle of estimating a relationship between a set of independent variables and a single dependent variable [6]. The dependent variable is sometimes termed as an outcome variable. In tasks involving forecasting and prediction, regression has proven to be efficient [7].

The SEIR model is one of the variations of the basic SIR model. Both of them, essentially, are compartmental models in epidemiology. These compartmental models aid in facilitating mathematical modeling of infectious diseases [8]. Here, labels like S, E, I, and R (susceptible, exposed, infectious, and recovered) [9] are assigned to the population which is under study. Over time of the study, a person may traverse through these states. For example, a person who was exposed at one point in time may recover from the disease and transition over to the recovered state. Figure 1 illustrates how an individual moves from one state to another.

The aforementioned “transitions” have been quantified using transition rates. For instance, from S to I, the transition rate can be formulated as

Fig. 1 SEIR transition model



$$\frac{ds}{dt} = \frac{-\rho SI}{N} \quad (1)$$

where N is the total population and ρ is the average number of contacts per person for a period, represents the probability of disease transmission from a susceptible individual to an infectious individual. σ is called incubation rate, represents the probability of exposed individuals becoming infectious. Similarly, γ is termed as recovery rate. Mathematically, these terms are formulated as follows:

$$\frac{dE}{dt} = \frac{-\rho SI}{N} - \sigma E \quad (2)$$

$$\frac{dI}{dt} = -\sigma E - \gamma I \quad (3)$$

Coming to the feasibility of this model, the idea it represents is mathematically and logically sound. The fact that a complex scenario of highly ambiguous data of disease spread can be boiled down to such a simple interpretation is exceptional. However, even though the model may lure mathematicians and epidemiologists into using this method using calculus, but, the scope of our study revolves around procedures that can easily be followed by a computer. Incorporating calculus in a machine-processed task can prove to be a difficult task. Moreover, SEIR is not a complete system. A complementing model like a conditional autoregressive (CAR) is needed to account for the multidimensional complexity of the system and to determine the transition rates which are quintessential for its underlying calculations. Also, even though SEIR accounts for, but does not explicitly incorporate individual differences. The most detrimental flaw is that it is not applicable for small datasets. So, in situations where a limited portion of the population is to be studied, SEIR is inapplicable.

Neural network is a class of models within machine learning literature that helps to solve problems that are very complex for humans to code. Inspired by biological, neural networks that function in animals, it functions in similar ways as that of an actual animal brain (Fig. 2).

For the scope of this study, it has been found that a recurrent neural network (RNN) is appropriate for this kind of implementation. Basically, it is a class of neural networks where the connection between nodes forms a graph along a temporal sequence. Moreover, RNNs can use their internal state to process variable-length

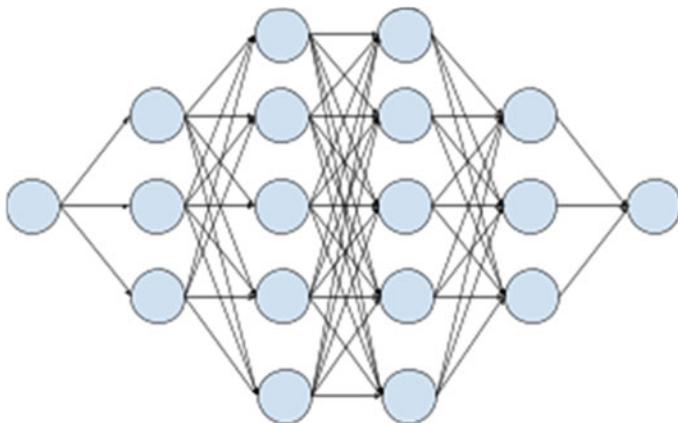


Fig. 2 Neural network

sequences of inputs. A pair of scholars namely Hochreiter and Schmidhuber tried to solve the problems related to RNN by building long short-term memory network (LSTM). Primarily, it tries to solve the vanishing/exploding gradient problem by incorporating gates and an explicitly defined memory cell, and these cells hold on to the previous values till the “forget” cell tells them to forget those values. Additionally, LSTM has an input gate that appends new stuff into the cell and the output gate decides when to pass the vectors one cell to a hidden state (Fig. 3).

A group of researchers has used a single model to predict the values of all the countries, and it was observed that such models performed very poorly due to their small dataset, while others who used more than three values such as death rate,

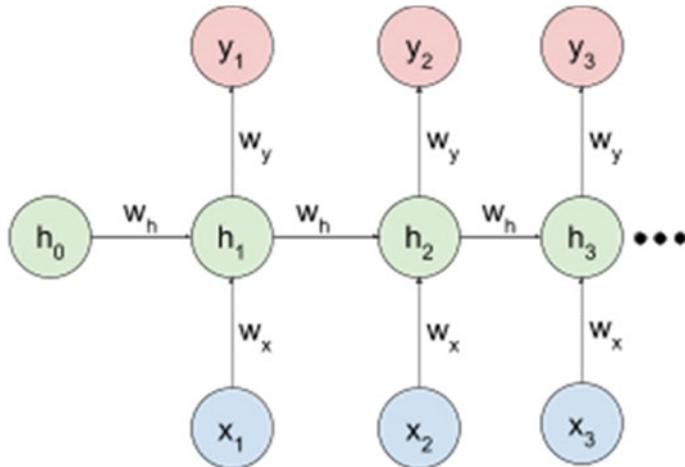


Fig. 3 Recurrent neural network

case rate, and recovery rate showed more accurate results as compared to the single variable model [10]. Evidently, applications of neural networks where the dataset is sizable have proven to be accurate [11]. However, one of the most prominent drawbacks of this system is that, inherently, neural networks require a large training dataset. Hence, it fails to produce accurate results when the data is not enough to accurately train the model. Moreover, neural networks are obsolete for single-variable datasets.

A hybrid model is created by the authors in [12] using CNN and RNN to diagnose COVID-19 using chest X-ray images. Many transfer learning methods like VGG, DenseNet, ResNet, and Inception are used in this diagnosis, and the performance in each and every case is compared to explicitly represent the best-suited model for the diagnosis of the disease. The model also uses gradient-weighted class activation map to pinpoint the region of interest (ROI) which helps in prediction the disease. Another model was developed by authors in [13] to predict the diagnosis of COVID-19 using optimized deep neural networks. The system also works well in the case of insufficient data. Unlike reverse-transcriptase polymerase chain reaction diagnosis, which is followed in current systems, the author has developed a tool which predicts based on weighted class loss function of the X-ray images. The model is also fine-tuned, and various transfer learning models performances are also compared.

3 Inferences

To begin with, using regression modeling for our purpose would seem to be suitable [15]. However, there is one obvious flaw. Linear regression is best suited for data that is mostly linear. A graph for the spread of an infectious disease tends to be highly unpredictable and in most cases, nonlinear [14]. It can be inferred that using regression systems would make our model have a marginal amount of overfitting.

Moving on to the SEIR model, it was found to be mathematically demanding, and even though it is logically sound, it does not incorporate individual differences within the population. Moreover, its implementation in small-sized datasets is not pragmatic.

Neural networks are accurate where we need highly accurate results for a situation where training data is plenty [16]. But, one notable flaw with neural network systems is that it needs quite large datasets to work upon. So, it fails to show accurate results when the data is limited and shows high levels of inaccuracy which demands optimization of the model before the declaration of the results.

4 Conclusion

After deep analysis and careful consideration of the aforementioned methods, it is safe to assume that retrofitting predefined elementary methods to such a complex

conundrum has, in most cases, proven to be a hit or miss. The conclusion obtained from this survey suggests formulation of a task-specific model that has the ability to comprehend the problem objectively. Thus, an abstract, unconventional algorithm is best suited to analyze such a situation.

References

1. WHO's information page on COVID-19. <https://www.who.int/health-topics/coronavirus>
2. Hamzah FAB, Lau CH, Nazri H, Ligot DV, Lee G, Tan CL, Shaib MKBM, Zaidon UHB, Abdulla AB, Chung MH, Ong CH, Chew PY, Salunga RE (2020) CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction
3. Jia L, Li K, Jiang Y, Guo X, Zhao T (2020) Prediction and analysis of coronavirus disease 2019
4. Yuan DF, Ying LY, Dong CZ (2012) Research progress on epidemic early warning model
5. Zhang F, Li L, Xuan HY (2011) Overview of infectious disease transmission models
6. Yang B, Pei H, Chen H (2016) Characterizing and discovering spatiotemporal social contact patterns for healthcare
7. Brownstein JS, Freifeld CC, Madoff LC (2009) Digital disease detection—harnessing the web for public health surveillance
8. Alessa A, Faezipour M (2018) A review of influenza detection and prediction through social networking sites
9. DeCaprio D, Gartner J, McCall CJ, Burgess T, Kothari S, Sayed S (2020) Building a COVID-19 vulnerability index
10. Sood N, Simon P, Ebner P et al (2020) Seroprevalence of SARS-CoV-2-specific antibodies among adults in Los Angeles County, California, on April 10–11, 2020. *JAMA* 323:2425–2427
11. Raissi M, Ramezani N, Seshaiyer P (2019) On parameter estimation approaches for predicting disease transmission through optimization, deep learning, and statistical inference methods
12. Islam MM, Islam MZ, Asraf A, Ding W (2020) Diagnosis of COVID-19 from X-rays using combined CNN-RNN architecture with transfer learning. <https://doi.org/10.1101/2020.08.24.20181339>
13. Punn NS, Agarwal S (2020) Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. *Appl Intell.* <https://doi.org/10.1007/s10489-020-01900-3>
14. Halloran ME, Ferguson NM, Eubank S et al (2008) Modeling targeted layered containment of an influenza pandemic in the United States. *Proc Natl Acad Sci* 105(12):4639–4644
15. Beyersmann J, Wolkewitz M, Allignol A, Grambauer N, Schumacher M (2011) Application of multistate models in hospital epidemiology: advances and challenges. *Biom J* 53(2):332–350
16. Eubank S, Kumar VSA, Marathe MV et al (2004) Structural and algorithmic aspects of massive social networks. In: SODA'04: proceedings of the ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp 718–727

Regression Analysis and Prediction of the COVID-19



Santosini Bhutia and Bichitrananda Patra

Abstract The corona virus disease is recognized as a global threat to the health industry and is a new challenge to the research area. To deal with this corona virus disease (COVID-19), which is currently sparked, all over the globe, machine learning (ML) plays a major role in variety of ways. This paper presents the analysis of the deadly COVID-19 outbreak to fight against this pandemic. This study is based on the dataset of confirmed cases, deaths, and recoveries worldwide as provided by the Johns Hopkins University. At first, we analyzed the pattern and characteristics of the growth of the pandemic by publicly available data. Secondly, we presented a comparative study and, finally, developed a future forecast model by taking three machine learning algorithms are support vector machine, linear regression, and Bayesian ridge regression.

Keywords Machine learning (ML) · Support vector machine · Polynomial regression and Bayesian ridge regression · Corona virus · Prediction model

1 Introduction

“Corona virus (CoV)” is prefixed with “novel,” as declared by the World Health Organization (WHO), it is a new strain of the virus. This virus runs from usual cold to hazardous disease, which started in Wuhan, the capital city of Hubei in China. It is perceived from two specific corona viruses, namely severe acute respiratory syndrome (SARS) and middle east respiratory syndrome (MERS). The SARS-CoV was reported in China in 2002 and was spread to twenty-four countries. It had been reported 8000 confirmed cases out of which 800 were dead leading to 10% mortality rate, while the MERS-CoV is reported in Saudi Arabia in 2013 with 2500 confirmed cases out of which 800 were dead leading to 34% mortality rate [1]. Now the new

S. Bhutia · B. Patra (✉)

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

strain of the virus “COVID-19” is sparked in Wuhan, China in December 2019 and has spread to around 220 countries worldwide [2].

Due to unavailability of proper medication, this COVID-19 is currently a dangerous disease which has created a high level of panic. Based on the clinical features, numerous sort of research in machine learning has been in progress for controlling this rapid spread of COVID-19. It is required to obtain a precise prediction model for the Government to recommend new policies to prevent the spread of the disease and also evaluate the efficiency of the imposed policies [3].

The corona virus has spread from Wuhan to its surrounding countries particularly to the large metropolitan regions, for example, Beijing, Shanghai, and Guangzhou. This disease transmitted due to the large population movement to celebrate spring festival before Chinese New Year (January 25). To stop this spreading, the Wuhan city declares a complete lockdown of the city on January 23. This epidemic is nonlinear and complex in nature and also various known and unknown factors associated with this spread [4]. This virus is able to increase its number twice consistently. At the beginning if the number is one, then it will increase to two thereafter four and so on. This increase in the number of infection in the growth is called exponential growth. This paper focuses to analyze, visualize and forecast the reported cases of corona virus by means of predictive modeling. It represents the reported cases from January 22, 2020 to December 18, 2020, followed by the prediction model using support vector machine, linear regression, and Bayesian ridge regression through time-series related data. The proposed method is outlined in Sect. 2 gives the representation of support vector machine, linear regression, and Bayesian ridge regression modeling. The dataset is briefly described in Sect. 3, and the experimental result of the comparison study is described in Sect. 4. The prediction models are plotted, and the conclusion is drawn in Sects. 5 and 6, respectively.

2 Proposed Method

This section describes the details of the proposed approach of COVID-19 analysis. This can be done by collecting the COVID-19 time-series dataset from the repository of Johns Hopkins University which is publicly available. It helps in understanding and extracting the knowledge about the disease. As a result, the spreading of the disease can be analyzed and aware the people to protect themselves from the disease. In this paper, the following task was performed. (i) How the corona virus spread across the world, (ii) Analysis of COVID-19 spread in the four most affected countries: USA, Brazil, India, Russia, (iii) How social distancing is important in reducing the growth rate, (iv) forecasting of COVID-19 for the next 30 days across the world.

In this paper, the forecast models are proposed using support vector machine (SVM), linear regression, and Bayesian ridge regression. These models fitted into the dataset containing the total number of COVID-19 confirmed cases worldwide considering with two performance metrics mean absolute error (MAE) and mean squared error (MSE). In the first step, the prediction for 15 days, i.e., from Dec 4,

2020 to Dec 18, 2020, is calculated and then is compared to actual confirmed cases which are followed by prediction for the future number of confirmed cases for the next 30 days.

2.1 Support Vector Machine

Support vector machine (SVM) [5] is supervised learning algorithm, which is recycled for mutually classification and regression problems. SVR classic depends on the components of SVM [6–8]. It is used for expectation and curve fitting in both linear and nonlinear regression type. SVR remains used to generate the hyper plane in n -dimensional article intergalactic where each data point is distinctly classified. The generalization equation for the hyper plane is signified as $y = wX + b$, wherever w is the weight and b is the intercept at $X = 0$. The tolerance margin is represented as epsilon ϵ . Margin is the width that the boundary could be increased by before hitting a data point, and support vectors are those data points that the margin pushes up against. The SVR regression model is imported from SVM class of Sklearn Python library [9]. The models parameters are: SVR ($C = 0.1$, $gamma = 0.01$, $epsilon = 1$, $kernel = poly$, $shrinking = True$, $degree = 3$). Basically, the SVR technique depends on the kernel function which is used to compute the dot product of two vectors.

2.2 Polynomial Regression

Polynomial regression is supervised machine learning algorithm. It models a relationship between the independent variable (x) and dependent variable (y) with n th degree polynomial. The dataset which is used for training is nonlinear in nature. This algorithm is trained according to the previous data and then tested on another dataset to validate its accuracy.

2.3 Bayesian Ridge

Bayesian regression is a machine learning algorithm used to predict continuous values [10]. Here Scikit-Learn library is used to implement Bayesian ridge regression. The important parameters of Bayesian ridge regression are:

`n_iter = 40`: number of iterations

`tot = 0.01`: when to stop the algorithm

`alpha_1 = 1e-05`: shape parameter of the regress line Gamma distribution over the alpha limitation

$\alpha_2 = 1e-05$: converse rule limit for the Gamma dissemination over the alpha limitation

$\lambda_1 = 0.001$: shape limitation of the Gamma distribution over the lambda limitation

$\lambda_2 = 0.0001$: inverse scale limitation of the Gamma distribution over lambda limitation.

3 Datasets

In this study, the datasets are collected from Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE) [11]. The dataset is accessible in time-series format with date, month, and year. It gives the information about the registered patient every day. In the dataset, everyday a new column is added at the end of the table. In this study, the dataset is accessed on December 18, 2020. This dataset consists of parameters such as Province/State, Country/Region, latitude, longitude, and dates. There are separate datasets for confirmed cases, deaths, and recoveries which consists of number of cases every day. Before analysis, pre-processing is required to transform the data. Here some irrelevant parameters like latitude and longitude were discarded for convenience, and dates were converted to date time object.

4 Experimental Results and Analysis

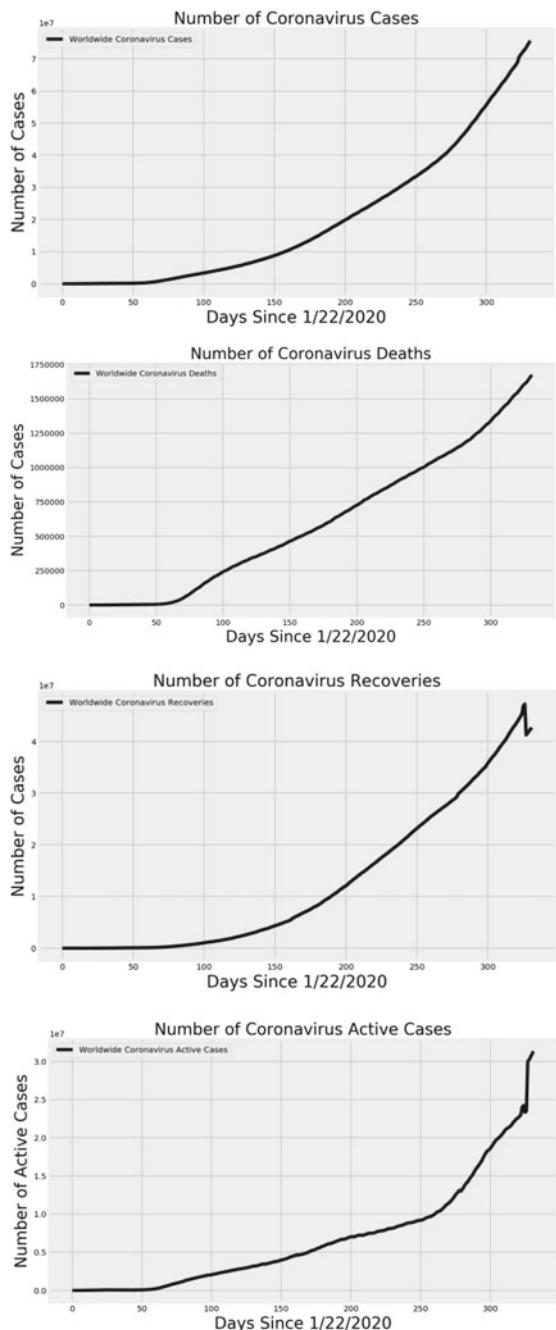
4.1 *The Corona Virus Spread Across the World*

In this section, we have presented the experimental result in detail in Fig. 1 across the world for the total number of confirmed cases, deaths, recoveries, and active cases from 22/01/2020 to 18/12/2020 in various countries/regions. From the graph, it is observed that the growth of the disease is in exponential form all over the world.

4.2 *Analysis of Corona Virus in the Four Most Affected Countries*

The outbreak of corona virus is an international emergency and is affecting our everyday life by lockdown in most of the countries. The four top most affected countries are USA, Brazil, India, and Russia in the world. The total number of confirmed cases for these four countries is plotted and shown. Also the comparison of these countries for confirmed cases, deaths, and recoveries is plotted from 01/03/2020

Fig. 1 Worldwide corona virus cases for total confirmed cases, deaths, recoveries and active cases



to 18/12/2020 and shown in Figs. 2, 3 and 4. From the graph, it is observed that the recovery rate is more in India, and the death rate is more in USA.

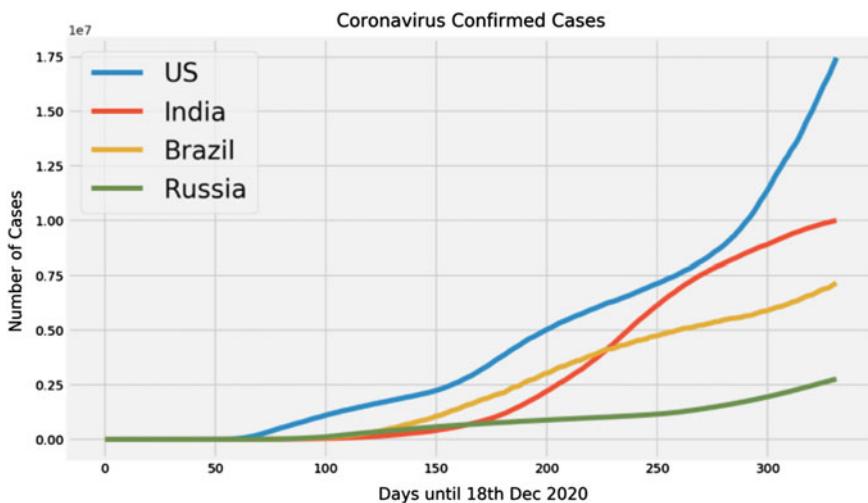


Fig. 2 Comparison of total number of confirmed cases in USA, Brazil, India, Russia

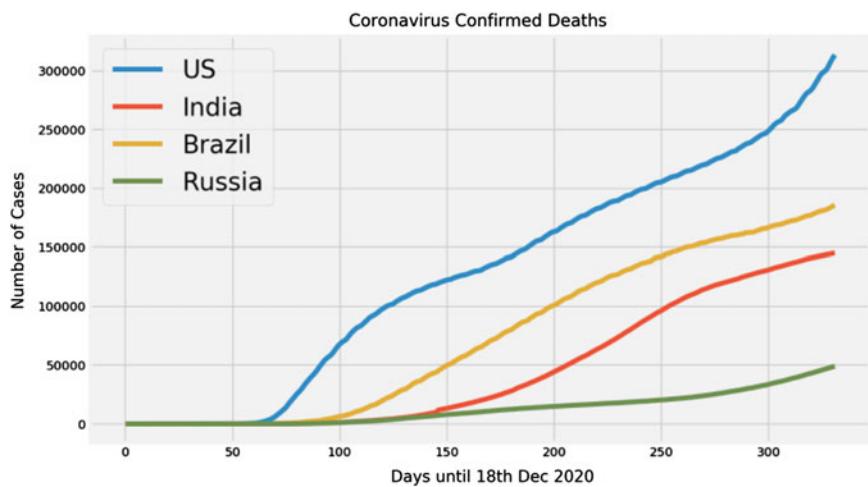


Fig. 3 Comparison of total number of deaths in USA, Brazil, India, Russia

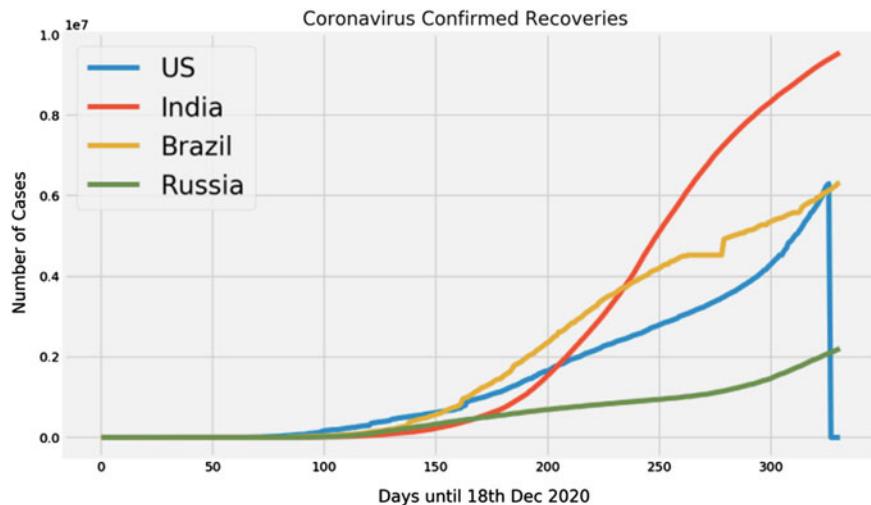


Fig. 4 Comparison of total number of recoveries in USA, Brazil, India, Russia

4.3 *Social Distancing Is Important in Reducing the Growth Rate*

Since it is a transmitted disease, if a positive corona patient comes in contact with other people, it will spread in others. To stop this spreading, Government imposed different rules and declared lockdown in their respective regions. Some corona positive patients and people having mild symptoms are quarantined to avoid meeting with other people. But there are some hidden infected people, who are asymptomatic or have mild symptoms, however, can still spread the disease. To break this spreading, World Health Organization (WHO) suggests social distancing is more important in reducing the growth rate of the disease. In addition to this, other activities are washing our hands frequently and not contacting our face.

The global economy and health have been affected by this pandemic. In this section, we represent the spread of COVID-19 across the world as of December 18, 2020. It spreads more rapidly than SARS and has similar symptoms like them. To regulator the extent of the disease and to understand the effect of the sickness, it is important to analyze the features of the sickness and expect the extent of the disease cross the globe. In this study, machine learning methods are used for data analysis to create the model which helps in forecasting for next 30 days across the world. First, support vector machine, polynomial regression, and Bayesian ridge algorithms were used to predict the number of confirmed cases for 15 days that is from Dt. 04/12/2020 to Dt. 18/12/2020. The MAE and MSE were also calculated and were shown in Fig. 5. Then the predicted values are compared with the actual values as shown in Table 1. Percentage of accuracy and error is calculated which is shown in Table 2. The accuracy of the models can be evaluated by comparing the

MAE: 2182776.690017684 **MAE:** 223475.63596688816
MSE: 4858999530630.158 **MSE:** 55783632985.30252

Date	SVM Predicted number of Confirmed Cases Worldwide
0	6.58172e+07
1	6.64285e+07
2	6.70436e+07
3	6.76626e+07
4	6.82855e+07
5	6.89123e+07
6	6.9543e+07
7	7.01777e+07
8	7.08162e+07
9	7.14587e+07
10	7.21051e+07
11	7.27555e+07
12	7.34099e+07
13	7.40683e+07
14	7.47307e+07

Date	Polynomial Predicted Number of Confirmed Cases Worldwide
0	6.29739e+07
1	6.35175e+07
2	6.40047e+07
3	6.46155e+07
4	6.51698e+07
5	6.57277e+07
6	6.62893e+07
7	6.68545e+07
8	6.74233e+07
9	6.79959e+07
10	6.85721e+07
11	6.9152e+07
12	6.97357e+07
13	7.03231e+07
14	7.09143e+07

Date	Bayesian Ridge Predicted Number of Confirmed Cases Worldwide
0	6.35505e+07
1	6.41274e+07
2	6.47091e+07
3	6.52956e+07
4	6.58869e+07
5	6.64832e+07
6	6.70844e+07
7	6.76906e+07
8	6.83019e+07
9	6.89183e+07
10	6.95398e+07
11	7.01666e+07
12	7.07987e+07
13	7.14361e+07
14	7.20788e+07

Fig. 5 Predicted confirmed cases across the world for 15 days

Table 1 Comparison of confirmed cases (actual) with the predicted result of the proposed models

Date	Actual value	Predicted value		Difference			
		SVM	Polynomial	Bayesian ridge	SVM	Polynomial	Bayesian ridge
04/12/2020	65,968,507	65,817,200	62,979,300	63,557,100	151,307	2,989,207	2,411,407
05/12/2020	66,610,808	66,428,500	63,523,300	64,134,400	182,308	3,087,508	2,476,408
06/12/2020	67,149,510	67,043,600	64,070,800	64,716,400	105,910	3,078,710	2,433,110
07/12/2020	67,663,843	67,662,600	64,621,800	65,303,300	1243	3,042,043	2,360,543
08/12/2020	68,302,251	68,285,500	65,176,500	65,895,000	16,751	3,125,751	2,407,251
09/12/2020	68,973,078	68,912,300	65,734,700	66,491,700	60,778	3,238,378	2,481,378
10/12/2020	70,466,211	69,543,000	66,296,600	67,093,300	923,211	4,169,611	3,372,911
11/12/2020	71,164,680	70,177,700	66,862,100	67,700,000	986,980	4,302,580	3,464,680
12/12/2020	71,787,629	70,816,200	67,431,300	68,311,700	971,429	4,356,329	3,475,929
13/12/2020	72,335,630	71,458,700	68,004,200	68,928,600	876,930	4,331,430	3,407,030
14/12/2020	72,859,287	72,105,100	68,580,800	69,550,600	754,187	4,278,487	3,308,687
15/12/2020	73,485,176	72,755,500	69,161,100	70,177,900	729,676	4,324,076	3,307,276
16/12/2020	74,219,546	73,409,900	69,745,200	70,810,500	809,646	4,474,346	3,409,046
17/12/2020	74,955,161	74,068,300	70,333,000	71,448,400	886,861	4,622,161	3,506,761
18/12/2020	75,672,814	74,730,700	70,924,600	72,091,700	942,114	4,748,214	3,581,114

Table 2 % of error and % of accuracy of the proposed models

Date	% of error			% of accuracy		
	SVM	Polynomial	Bayesian ridge	SVM	Polynomial	Bayesian ridge
04/12/2020	0.23	4.53	3.66	99.77	95.47	96.34
05/12/2020	0.27	4.64	3.72	99.73	95.36	96.28
06/12/2020	0.16	4.58	3.62	99.84	95.42	96.38
07/12/2020	0.00	4.50	3.49	100.00	95.50	96.51
08/12/2020	0.02	4.58	3.52	99.98	95.42	96.48
09/12/2020	0.09	4.70	3.60	99.91	95.30	96.40
10/12/2020	1.31	5.92	4.79	98.69	94.08	95.21
11/12/2020	1.39	6.05	4.87	98.61	93.95	95.13
12/12/2020	1.35	6.07	4.84	98.65	93.93	95.16
13/12/2020	1.21	5.99	4.71	98.79	94.01	95.29
14/12/2020	1.04	5.87	4.54	98.96	94.13	95.46
15/12/2020	0.99	5.88	4.50	99.01	94.12	95.50
16/12/2020	1.09	6.03	4.59	98.91	93.97	95.41
17/12/2020	1.18	6.17	4.68	98.82	93.83	95.32
18/12/2020	1.24	6.27	4.73	98.76	93.73	95.27

actual values of the confirmed cases with the predicted values of these 15 days. From this study, it is observed that SVM algorithm gives better accuracy that is 99.23%, whereas polynomial regression gives 94.55% of accuracy and Bayesian ridge gives 95.74% of accuracy.

5 Prediction

In this study, the COVID-19 global data were estimated by using three algorithms: support vector machine, polynomial regression, and Bayesian ridge. This problem was time series prediction problem. We have obtained a predicted time series for the next 30 days after December 18, 2020. The predicted values from 19/12/2020 to 3/1/2020 using support vector machine, polynomial regression, and bayesian ridge algorithm are shown in Fig. 6, and the corresponding graphical representation of the predicted values is shown in Figs. 7, 8, and 9.

Date	SVM Predicted Number of Confirmed Cases Worldwide
0 12/19/2020	7.53971e+07
1 12/20/2020	7.60675e+07
2 12/21/2020	7.67419e+07
3 12/22/2020	7.74205e+07
4 12/23/2020	7.8103e+07
5 12/24/2020	7.87897e+07
6 12/25/2020	7.94804e+07
7 12/26/2020	8.01752e+07
8 12/27/2020	8.08742e+07
9 12/28/2020	8.15773e+07
10 12/29/2020	8.22845e+07
11 12/30/2020	8.29958e+07
12 12/31/2020	8.37113e+07
13 01/01/2021	8.4431e+07
14 01/02/2021	8.51549e+07
15 01/03/2021	8.58829e+07

Date	Polynomial Predicted Number of Confirmed Cases Worldwide
0 12/19/2020	7.47925e+07
1 12/20/2020	7.54851e+07
2 12/21/2020	7.61834e+07
3 12/22/2020	7.68872e+07
4 12/23/2020	7.75966e+07
5 12/24/2020	7.83117e+07
6 12/25/2020	7.90326e+07
7 12/26/2020	7.97592e+07
8 12/27/2020	8.04917e+07
9 12/28/2020	8.12299e+07
10 12/29/2020	8.19741e+07
11 12/30/2020	8.27242e+07
12 12/31/2020	8.34802e+07
13 01/01/2021	8.42423e+07
14 01/02/2021	8.50104e+07
15 01/03/2021	8.57846e+07

Date	Bayesian Ridge Predicted Number of Confirmed Cases Worldwide
0 12/19/2020	7.58214e+07
1 12/20/2020	7.65711e+07
2 12/21/2020	7.73284e+07
3 12/22/2020	7.80933e+07
4 12/23/2020	7.8866e+07
5 12/24/2020	7.96465e+07
6 12/25/2020	8.0435e+07
7 12/26/2020	8.12314e+07
8 12/27/2020	8.20359e+07
9 12/28/2020	8.28486e+07
10 12/29/2020	8.36696e+07
11 12/30/2020	8.4499e+07
12 12/31/2020	8.53368e+07
13 01/01/2021	8.61831e+07
14 01/02/2021	8.70381e+07
15 01/03/2021	8.79018e+07

Fig. 6 Predicted confirmed cases across the world for 30 days

6 Conclusion

Corona virus causes extreme harm to the health in the entire world. In this paper, it gives the thought about the current trends of COVID-19 and the level of spread of the disease across the world. Anticipating the further spread can help the Government and the citizens to make proper arrangements to deal with the circumstance. Here three machine learning models were recycled to predict the promote extent using real-time data. Through this paper, it is observed that support vector machine gives better result as compared to polynomial regression and Bayesian ridge regression. The proposed methodologies predict the growth in total quantity of corona virus

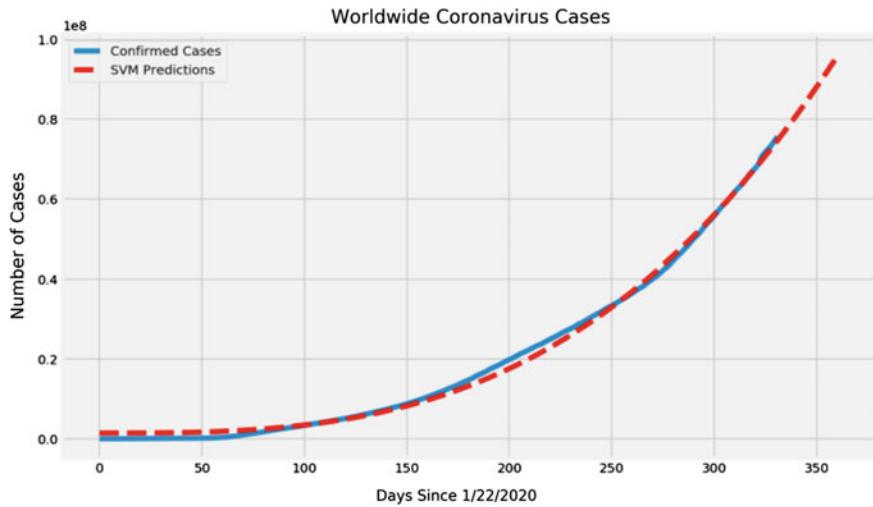


Fig. 7 Prediction of total number of confirmed cases using SVM

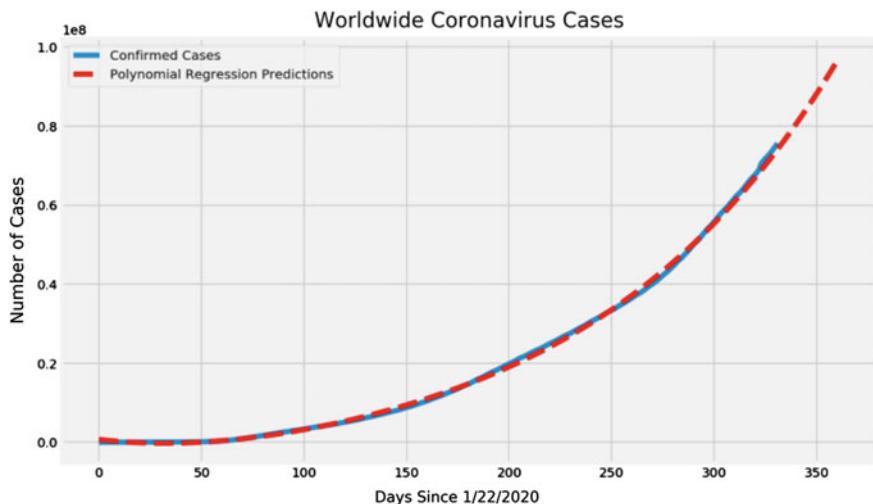


Fig. 8 Prediction of total number of confirmed cases using polynomial regression

infected confirmed cases for the next 30 days. In future study, we may be analyzed how to control measures with social distancing and proper hygienist.

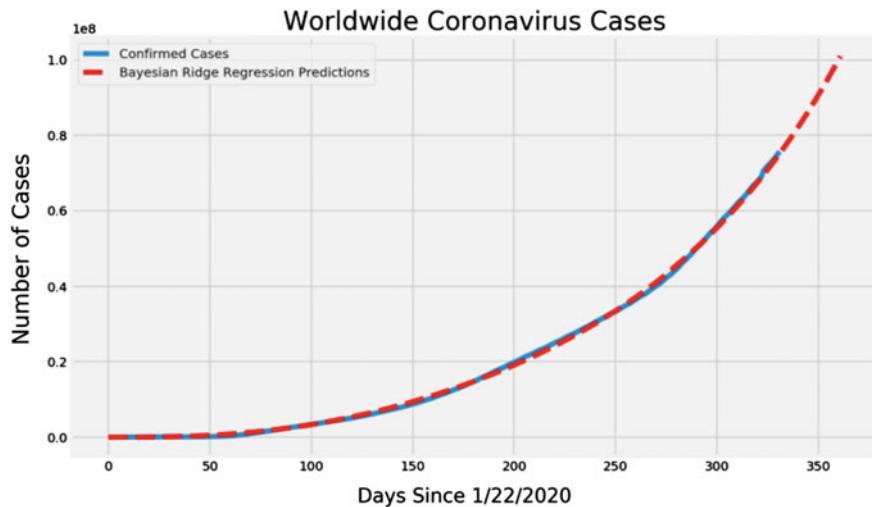


Fig. 9 Calculation of whole quantity of complete cases using Bayesian ridge regression

References

1. Sethy PK, Behera SK, Ratha PK, Biswas P (2020) Detection of coronavirus disease (COVID-19) based on deep features and support vector machine
2. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Zhang L (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 395(10223):507–513
3. Remuzzi A, Remuzzi G (2020) COVID-19 and Italy: what next? Lancet 395(10231):1225–1228
4. Gorbatenko AE, Baker SC, Baric R, Groot RJD, Drosten C, Gulyaeva AA et al (2020) Severe acute respiratory syndrome-related coronavirus: the species and its viruses—a statement of the Coronavirus Study Group
5. Müller KR, Smola AJ, Rätsch G, Schölkopf B, Kohlmorgen J, Vapnik V (1997) Predicting time series with support vector machines. In: International conference on artificial neural networks, Oct 1997. Springer, Berlin, Heidelberg, pp 999–1004
6. Ivanov D (2020) Predicting the impacts of epidemic outbreaks on global supply chains: a simulation-based analysis of the COVID-19/SARS-CoV2 case. Transp Res E. <https://doi.org/10.1016/j.tre>
7. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media
8. Sánchez A VD (2003) Advanced support vector machines and kernel methods. Neurocomputing 55(1–2):5–20
9. Sci-kit-learn (2020) https://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html
10. Zhang X, Ma R, Wang L (2020) Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. Chaos Solitons Fract 135:109829
11. https://raw.githubusercontent.com/CSSEGISandData/COVID19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

Deep Learning Neural Network-Based Pneumonia Classification



Anasua Banerjee and Minakhi Rout

Abstract Pneumonia is a serious critical infection found in human body. Pneumonia is broadly of two viral and bacterial. It causes ulcers in the affected lungs due to this reason many children and old persons die every year as reported by World Health Organization (WHO). Chest X-Ray is a way to detect whether a person is suffering from Pneumonia diseases or not. With a view to making the X-ray quality vivid, accurate and precise many augmentation and data pre-processing methods are used here. To expedite and ensure quality treatment in the remote areas, these automatic methods may be safely resorted to without depending on an expert radiologist's counseling. Recently, Convolution Neural Network (CNN) Deep learning algorithms are very much in use for medical imaging classification, precision and control. Here the proposed model is used to detect whether a person is having Pneumonia or not as well as this model may be applied in many areas like image classification, machine translation and image segmentation.

Keywords Deep learning · Convolution Neural Network (CNN) · Chest X-ray · Machine learning · Confusion matrix

1 Introduction

From times immemorial pneumonia has always been a bane of mankind. To protect and save mankind from these disastrous ailments, we must identify and detect these diseases in a very articulate manner and try to save all from little children to old-aged persons from this disaster. The tendency these days is to classify according to the specific bacterium or virus causing the infection.

Ascertaining whether a patient is suffering from pneumonia generated illness or not we have to go through some dataset of chest X-ray intensively to identify the real cause of illness and subsequently using the machine learning algorithm we have

A. Banerjee · M. Rout (✉)

School of Computer Engineering, Kalinga Institute of Industrial Technology Deemed to be University, Bhubaneswar, India

e-mail: minakhi.rout@kiit.ac.in

tried to identify and detect in the minutest and most subtle way the existence of this disease. Here, deep neural network and Convolution Neural Network (CNN) are used here which is acting as a binary classifier. As a result of this work, the workload of radiologists will be lessened without compromising the precision of their work.

The format of the paper goes thus: Sect. 2 describes the related work that has been reported earlier followed by Sect. 3 which provide the detailed overview of the deep learning network. Data set description provided in Sect. 4. Section 5 describes CNN model framework for classification, augmentation and pre-processing followed by model overview and result analysis in Sect. 6. Finally, Sect. 7 is about the concluding remarks of the work and future scope.

2 Literature Review

For the purpose of extraction of features from Chest X-Ray dataset, many techniques are used so that without taking any help from any trained radiologist to diagnosis whether a person having Pneumonia or not. These methods or techniques are mainly divided into two different categories, viz. handcrafted feature extraction techniques (Logistic regression) [1] and deep learning techniques (Convolution Neural Network). The research workers [2] proposed a multilane capsule network for the purpose of acquiring high accuracy at reduced cost. The authors [3–6] applied a CNN model to extract features from Chest X-Ray dataset. In this paper [7], they applied Alex Net, GoogleLeNet, VGGNet-16 and Res-Net50 on Chest X-Ray dataset for the purpose to diagnosis whether a person is suffering from Pneumonia disease or not. Researchers in [8] applied Mask R-CNN for the purpose of Pneumonia detection and obtained 79% accuracy. The research workers [5] applied their model so that it would be able to extract the features and classify on MRI image dataset to detect whether a person was having a brain tumor or not.

3 Overview of Deep Learning

Deep learning and machine learning being parts of artificial intelligence, here deep learning is being broached upon in detail. Deep learning is finding out the best solution in medical analysis. DL can help in detecting complicated and solving problems with the help of hidden layers, which have their existence between input and output layers. Convolution Neural network (CNN), a part of DL architecture, has convolution filters, pooling (Max pooling, Average pooling, Global average pooling) and fully connected layers, as its constituents. At first, the image of dataset is taken as input. On the image already taken as an input are applied Convolution Layers, Filters, Max pooling, Dense Layers, and Fully-connected Layers [9].

The above mentioned five layers performed as:

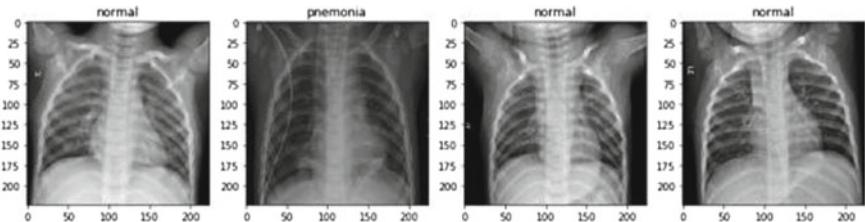


Fig. 1 Chest X-ray images

1. Convolution Layers are at first attached to original image which is already taken as input.
2. Filter, here the size being $[3 \times 3]$, in the second phase of operation is to be moved from top left to bottom right across the image. Mathematical calculation and operations will go on over the image according to the convolution filter. All these functioning will result in producing a feature map for each filter. A feature map is now being generated through activation function.
3. Max pooling's function is to draw out the maximum value revealed by filter according to the filters sliding over the image.
4. Dense Layer with its own weight and bias collects input from the preceding layers with all their activations.
5. Fully Connected Layer does the function of flattening the output after connecting all the neurons of all the layers.

A special note is to be taken that at the time of converting multidimensional tensor into single-dimensional Flatten layer has the most import function to play.

4 Dataset Description

In this paper, Chest X-Ray dataset was used here which is publicly available in the Kaggle website [10]. In this dataset, there are total 5856 images. Out of 5856 images, only 1583 are normal images and remaining are Pneumonia affected images. In this dataset, Chest X-Ray images have resolutions are ranging from 712×439 pixels to 2338×2025 pixels. In Fig. 1, sample images are shown.

5 CNN Classification Framework, Augmentation and Pre-processing

Keeping in view better augmentation method to improve the quality of dataset and that to by avoiding overfitting problem. We used various ways depicted here under.

Depiction is kindly to be noted like:

Fig. 2 Specification of image augmentation

Method	Values
<i>Rotation Range</i>	30
<i>Height- shift Range</i>	0.1
<i>Rescale</i>	1/255
<i>Weight- shift Range</i>	0.1
<i>Shear-Range</i>	0.2
<i>Fill Mode</i>	Nearest
<i>Horizontal-Flip</i>	True

Image Augmentation Settings (Figs. 2 and 3).

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 1339, 1339, 32)	896
conv2d_2 (Conv2D)	(None, 1337, 1337, 64)	18496
max_pooling2d_1 (MaxPooling2	(None, 668, 668, 64)	0
conv2d_3 (Conv2D)	(None, 666, 666, 64)	36928
max_pooling2d_2 (MaxPooling2	(None, 333, 333, 64)	0
flatten_1 (Flatten)	(None, 7096896)	0
dense_1 (Dense)	(None, 2)	14193794
dropout_1 (Dropout)	(None, 2)	0
activation_1 (Activation)	(None, 2)	0
Total params:	14,250,114	
Trainable params:	14,250,114	
Non-trainable params:	0	

Fig. 3 Model overview

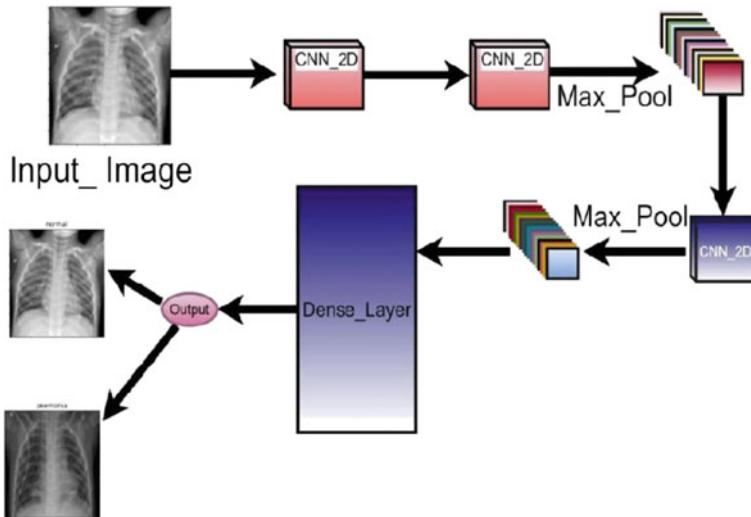


Fig. 4 Presence of pneumonia disease in chest X-ray image dataset detected by CNN model

6 Model Overview and Simulation Result Analysis

In this model at first we took Convolution layer whose filter size is 32, kernel size (3, 3) and input shape (1341, 1341, 3). Then we took Convolution size 64, kernel size (3, 3) with activation function as “Relu.” After that, we used Max pooling the size of which is (2, 2). Fourthly we used Convolution size 64, kernel size (3, 3) with activation function as “Relu.” After doing all these we used Max pooling the size of which is (2, 2). In order of sequence then, we used Flatten and subsequently Dense Layer with activation function as “Relu.”

An epoch is a number of times a machine is fed the whole training data set. To achieve precision, the full data set is to be impacted several times on neural network. Prior to the setting of updated model weights Batch size takes some samples of training data set and analyzes error gradients (Figs. 4 and 5).

7 ROC Curve

To measure accurately performance of model Receiver Operating Characteristic (ROC) curve is used. This curve ascertains the separability of different classes. Model is termed as perfectly depicted if ROC curve accuracy accredits it “1,” whereas 0.5 depicts it as the worst model. Whereas this model accuracy is computed as 0.828 which is shown in Fig. 6.

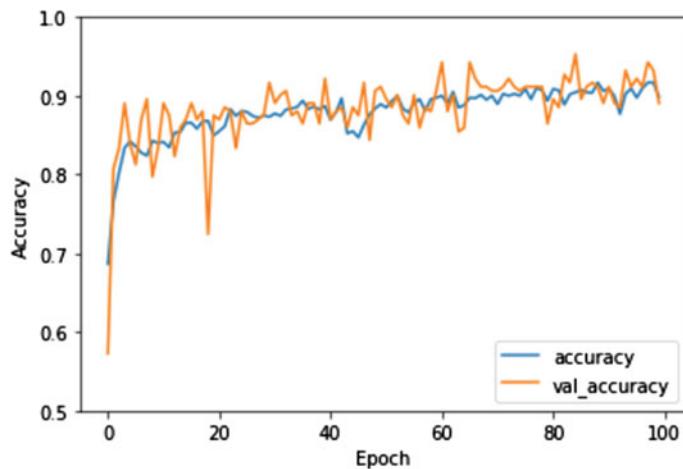


Fig. 5 Accuracy versus val accuracy

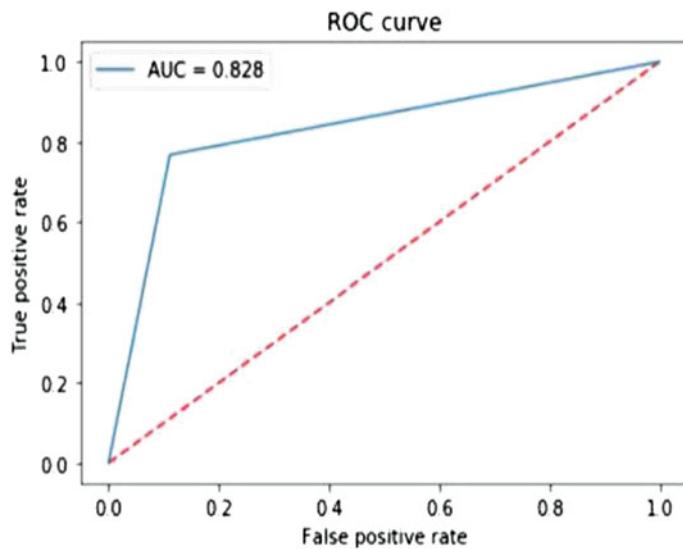


Fig. 6 ROC curve

8 Confusion Matrix

Confusion matrix has been used to evaluate the degree of performance in the machine learning classification problem. Confusion matrix can be used well when two or more classes are there which is shown in Fig. 7 and the confusion matrix obtained after applying the classification model is depicted in Fig. 8.

Fig. 7 Accuracy versus val accuracy

		Positive	Negative	
Positive	TP	FP		
	FN	TN		
Actual Values				

Fig. 8 Precision, recall and F1-score

	precision	recall	f1-score
0	0.86	0.75	0.80
1	0.77	0.88	0.82
accuracy			0.81
macro avg	0.82	0.81	0.81
weighted avg	0.82	0.81	0.81

To ascertain the existence of pneumonia and that to how much precisely this model can predict the performance quality confusion matrix with two classes is used here. In every confusion matrix, there are four different combinations of actual and predicted values which are shown in Fig. 7.

In True Positive sector, the true existence of pneumonia in chest X-ray and also the veracity of its existence are being noticed here.

In True Negative sector, the non-existence of the pneumonia disease and also the veracity of its non-existence noticed minutely here.

In False Positive sector, this model found the existence of pneumonia but the verified truth tells the otherwise that is pneumonia is non-existent.

In False Negative sector, this model found the non-existence of pneumonia but the verified truth suggests the otherwise that is pneumonia is exist.

9 Conclusion

In this paper, 2D Convolution Neural Network (CNN) model is used to predict whether a person is having Pneumonia infection or not. To expedite and ensure quality treatment in the remote areas, these automatic methods may be safely resorted to without depending on an expert radiologist's counseling. This model reached accuracy level nearly 95 Percentages and the ROC accuracy point is also around 82 percentages. As mentioned in the outset it may be stated again that this project may be of immense help to the physicians to determine their course of action. We are looking forward to applying advanced and new deep learning models like GAN and Auto encoder model in future on Chest X-Ray dataset.

References

1. Antin B, Kravitz J, Martayan E (2017) Detecting pneumonia in chest X-rays with supervised learning. *Semanticscholar.Org*
2. do Rosario VM, Borin E, Breternitz M (2019) The multi-lane capsule network. *IEEE Signal Process Lett* 26(7):1006–1010. <https://doi.org/10.1109/LSP.2019.2915661>
3. Stephen O, Sain M, Maduh UI, Jeong D-U (2019) An efficient deep learning approach to pneumonia classification in healthcare. *J Healthc Eng*
4. Al Mubarok AF, Dominique JAM, Thias AH (2019) Pneumonia detection with deep convolutional architecture. In: 2019 international conference of artificial intelligence and information technology (ICAIT), Yogyakarta, pp 486–489. <https://doi.org/10.1109/ICAIT.2019.8834476>
5. El-Dahshan E-SA, Mohsen HM, Revett K, Salem A-BM (2014) Computer-aided diagnosis of human brain tumor through MRI: a survey and a new algorithm. *Expert Syst Appl*. <https://doi.org/10.1016/j.eswa.2014.01.021>
6. Labhane G, Pansare R, Maheshwari S, Tiwari R, Shukla A (2020) Detection of pediatric pneumonia from chest X-ray images using CNN and transfer learning. In: 2020 3rd international conference on emerging technologies in computer engineering: machine learning and internet of things (ICETCE), Jaipur, pp 85–92. <https://doi.org/10.1109/ICETCE48199.2020.9091755>
7. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern reorganization
8. Siraztdinov I, Kholiavchenko M, Mustafaev T, Yixuan Y, Kuleev R, Ibragimov B (2019) Deep neural network ensemble for pneumonia localization from a large-scale chest X-ray database. *Comput Electr Eng*
9. Sun Y, Xue B, Zhang M, Yen GG (2020) Evolving deep convolutional neural networks for image classification. *IEEE Trans Evol Comput* 24(2):394–407. <https://doi.org/10.1109/TEVC.2019.2916183>
10. <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>

Fuzzy Time Series Model for Forecasting Agricultural Crop Production



K. Senthamarai Kannan, K. M. Karuppasamy, and R. Balasubramaniam

Abstract India is one of the largest producers of rice around the globe. In day-to-day life, forecasting techniques are used to predict weather, economy, population growth, agricultural crop production, and stock market. Recently, many researchers have tried fuzzy time series to predict future events. In agriculture, forecasting the rice production for a lead year is important for crop planning, agro-based resource utilization, and overall management of rice production. The main objective of this work is designing and implementing fuzzy time series model in the rice production. This proposed method yields better forecasting accuracy results in production, especially effective in the area of comparative study.

Keywords Fuzzy logic · Fuzzy time series · Mean square error · Rice production · Forecasting

1 Introduction

India is one of the largest rice producers in the world. Rice is the main staple food in the eastern, northeastern, and southern regions, while wheat is the main staple grains in the northern, northern hills, and western regions. Coarse cereals appear as important food in the western and southern regions [1]. Rice is one of the main food grains of India. Also, in India, that largest portion has been used for rice cultivation rice grains in tropical areas. Also, in India, that largest portion has been used for rice cultivation. Rice is cultivated in area that receives heavy rainfall annually. The soil and environment of India are in conducive to the production of rice. Next to China, India stands the largest producer of rice in the world. Rice contributes 40% of food grains of people from total food grain in countries. According to the annual report of

K. Senthamarai Kannan · K. M. Karuppasamy (✉) · R. Balasubramaniam
Department of Statistics, Manonmaniam Sundaranar University, Abishekappatti, Tirunelveli, Tamil Nadu 627012, India

K. Senthamarai Kannan
e-mail: senthamarai.kannan@msuniv.ac.in

Indian government regarding the production of rice during the last two years (2017–18 and 2018–19) in Andhra Pradesh (including Telangana), Karnataka, Kerala, and Tamil Nadu in as follows, Andhra Pradesh has produced more amount of rice for the two consecutive year's 2017–18 and 2018–19 in 3788 and 3733 yield (kg/ha), Karnataka has rice produces 3038 and 3011 yield (kg/ha), Kerala has rice produces 2757 and 2915 yield (kg/ha), and Tamil Nadu has rice produces 3630 and 3748 yield (kg/ha) [2].

A time series is an order of observation in use sequentially in time through an intrinsic feature that typically adjoining observations dependent in this analysis is concerned with technique used for analysis is dependence. Future estimations of time arrangement information of farming creation measure are neither precisely represented by a numerical capacity nor precisely represented by a likelihood dispersion [3]. The concept of fuzzy time series, capable of dealing with vague and imprecise data presented in terms of linguistic variables, was developed by Song and Chissom [4] by using the theory of sets and linguistic variable given by Zadeh [5, 6]. Yearly difference of the student enrollment number [7]. Developed to the computational method to high-order forecasting based on fuzzy time series using rice production [8]. A comparative study of fuzzy time series forecasting and Markov modeling [9]. Comparison of paddy prices in India from several states using time series model [10]. Forecasted to daily petrol price using DES, ARMIA, and fuzzy time series models [11]. Discussed to fuzzy time series, model based on forecasting accuracy used to terrorist attacks [12]. A comparative study has been carried out for Indian export data with two vast applications of forecasting, namely ARIMA time series model and fuzzy time series models [13]. Growth of population and financial success are the two main factors for growing rice interest in India. We have discussed with statewise comparison of growth in production and profitability of rice in India. Further estimates of the national commission on population, India's population will be 1340 million in 2021 [14]. It is estimated that the demand for rice will be 113.3 million tons by the year of 2021–22 [1].

This paper is comprising into seven sections as follows. Sect. 1 explains briefly about the topic and models dealt with. Section 2 is data preprocessing. Section 3 is basic concepts and definition of fuzzy time series model. Section 4 explain that are being discussed forecasting procedure of fuzzy time series model. Section 5 deals with the numerical results and discussion of the proposed method. And, in Sect. 6, the compared results are being exhibited. Finally, conclusions in Sect. 7.

2 Data Preprocessing Using Box Plot

The role of the Box plot is rice production yield in southern states of India (Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu). The upper and lower hinges of the box are indicated in third quartile and first quartile of the dataset. The line in the box indicates the median value of the data, and the open square represents the mean value of the dataset. The state Andhra Pradesh is very high spurious to the three states

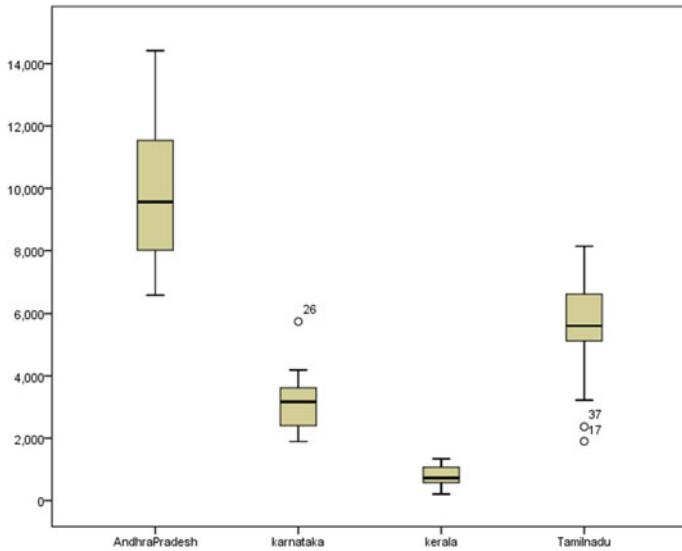


Fig. 1 Box plot for Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu

in Karnataka, Kerala, and Tamil Nadu. In particular, Karnataka states are very high production in 2016 and otherwise low from the rice production in remaining years. In similar, Tamil Nadu states very low rice production in 1997 and 2017 shown in Fig. 1.

3 Fuzzy Time Series Model and Definition

3.1 Fuzzy Time Series

This section discusses about some basic concepts of fuzzy time series. The combination of fuzzy technique and time series forecasting procedure opens up a new hopeful relevance progress in the research field [15]. Song and Chissom proposed fuzzy time series forecasting model described procedure as follows:

Let, U be the universe of discourse, where $U = \{u_1, u_2, \dots, u_n\}$ of the given historical data. The universe of discourse, $U = [D_{\min} - D_1, D_{\max} + D_2]$.

Where D_{\min} , D_{\max} , minimum data value, and maximum data value are denoted, and D_1, D_2 are two positive real numbers. A fuzzy set A_i of U is defined by,

$$A_i = f_{Ai}(u_1)/u_1 + f_{Ai}(u_2)/u_2 + \dots + f_{Ai}(u_n)/u_n,$$

where f_{Ai} is the membership function of fuzzy set A_i , $f_A: U \rightarrow [0, 1]$, and $f_A(u_i)$ indicates the grade of membership of u_i in A , where $f_A(u_i) \in [0, 1]$ and $1 \leq i \leq n$. The definition of fuzzy time series is reviewed as follows.

3.2 Definition of Fuzzy Time Series

This section for the time is being summarized basic fuzzy time series ideas provided for the following text.

Definition 1 Let, $Y(t)$, ($t = 0, 1, 2, \dots$) is a subset of R be the universe of discourse on which fuzzy sets $f_i(t)$, ($i = 1, 2, \dots$) are defined, and $F(t)$ is the collection of $f_i(t)$. $F(t)$ is called a fuzzy time series of $Y(t)$ [4, 11, 16–18].

Definition 2 If there exists a fuzzy relationship $R(t, t - 1)$, such that $F(t) = F(t - 1) * R(t, t - 1)$, where symbol $*$ is an operator, then $F(t)$ is said to be induced by $F(t - 1)$; the relationship can be denoted by $F(t - 1) \rightarrow F(t)$ [11, 16–18].

Definition 3 Suppose, $F(t - 1)$ by A_i and $F(t)$ by A_j fuzzy logical relationship can be defined by $A_i \rightarrow A_j$ where A_i and A_j are called the left hand side and right hand side of the fuzzy logical relationship [16–18].

Definition 4 If $F(t)$ is a time-invariant fuzzy time series, then the first-order fuzzy logical relationship is defined by $F(t - 1) \rightarrow F(t)$ [17, 18].

4 Step-by-Step Forecasting Procedure Will be Explained in Fuzzy Time Series

This model is based on historical data for yearly difference of the given dataset. First, no need to create a domain interval; second, no need to find the midpoint values of the each subinterval; third, do not use yearly change percentage in historical data while use the yearly difference in historical data [7]; fourth, greater adjustment for prediction formula of the inverse fuzzy numbers.

A model “a fuzzy time series forecasting model based on the yearly difference of the new creation records” is proposed. The following steps are given as.

- **Step 1** Let us define D is the discrete domain based on yearly difference of historical data.

$$D = E_i - E_{i-1} \quad (1)$$

where E_i and E_{i-1} are the historical data in i and $i - 1$ year.

- **Step 2** Compute the inverse fuzzy numbers of consecutive years based on domain v , [7] as given as

$$\left. \begin{aligned} v_\alpha &= \frac{1+0.0001}{\frac{d_\alpha}{d_\alpha} + \frac{d_{\alpha+1}}{d_\alpha}}, \\ v_\alpha &= \frac{0.0001+1+0.0001}{\frac{d_{\alpha-1}}{d_\alpha} + \frac{1}{d_\alpha} + \frac{0.0001}{d_{\alpha+1}}}, \quad 1983 \leq \alpha \leq 2018, \\ v_\alpha &= \frac{0.0001+1}{\frac{0.0001}{d_{\alpha-1}} + \frac{1}{d_\alpha}}. \end{aligned} \right\} \quad (2)$$

- **Step 3** Compute the forecasting formula of this model as

$$F\alpha = E\alpha - 1 + v\alpha \quad (3)$$

where $F\alpha$, $E\alpha - 1$, and $v\alpha$ are forecasting data, historical data, and inverse fuzzy numbers.

- **Step 4** Finally, compute the average forecasting error rate and mean square error using actual and estimated values.

1. Average forecasting error rate (AFER) [19] can be defined as

$$\text{AFER} = \frac{1}{n} \sum_{i=1}^n \frac{|E_i - F_i|}{E_i} \times 100\% \quad (4)$$

where n , E_i , and F_i are number of year, historical data, and forecasting data in i year.

2. Mean square error (MSE) [8] can be defined as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (E_i - F_i)^2 \quad (5)$$

where n , E_i , and F_i are number of year, historical data, and forecasting data in i year.

5 Computational Results and Discussions

The yearly rice productions data for thousand tons for the period from 1981 to 2019 for south states in India (Andhra Pradesh (AP), Karnataka (KA), Kerala (KL), and Tamil Nadu (TN)). This data collected from Ministry of Agriculture and Farmers Welfare, Government of India, were used for forecasting the future values using FTS model.

Step 1: To calculate D is the discrete domain based on yearly difference of historical data from Formula (1). For example, $D = E_i - E_{i-1}$. $D = d_{1982} = E_{1982} - E_{1981} = 856.5$.

Andhra Pradesh

$D = \{d_{1982} = 856.5\}$	$d_{1983} = -196.6$	$d_{1984} = 1119.4$	$d_{1985} = -1881.6$	$d_{1986} = 704.4$
$d_{1987} = -1022$	$d_{1988} = 495.6$	$d_{1989} = 3534$	$d_{1990} = -662$	$d_{1991} = -305.1$
$d_{1992} = -404.6$	$d_{1993} = -457.2$	$d_{1994} = 769.8$	$d_{1995} = -285.3$	$d_{1996} = -262.5$
$d_{1997} = 1671.8$	$d_{1998} = -2176$	$d_{1999} = 3368$	$d_{2000} = -1240.2$	$d_{2001} = 1820.2$
$d_{2002} = -1068.2$	$d_{2003} = -4062.8$	$d_{2004} = 1626$	$d_{2005} = 648$	$d_{2006} = -2103$
$d_{2007} = 168$	$d_{2008} = 1452$	$d_{2009} = 917$	$d_{2010} = -3703$	$d_{2011} = 3880$
$d_{2012} = -1523$	$d_{2013} = -1385$	$d_{2014} = 1214.7$	$d_{2015} = -1159.3$	$d_{2016} = -4076.7$
$d_{2017} = -36.3$	$d_{2018} = 713.8$	$d_{2019} = 80.5\}$		

Karnataka

$D = \{d_{1982} = 155.7\}$	$d_{1983} = -262.8$	$d_{1984} = 191.2$	$d_{1985} = 82.5$	$d_{1986} = -432.3$
$d_{1987} = 370.6$	$d_{1988} = -418.5$	$d_{1989} = 614.9$	$d_{1990} = -132.7$	$d_{1991} = 38.4$
$d_{1992} = 410.8$	$d_{1993} = 242.6$	$d_{1994} = 114.1$	$d_{1995} = -15.3$	$d_{1996} = -143.6$
$d_{1997} = 187.7$	$d_{1998} = 1.1$	$d_{1999} = 444.2$	$d_{2000} = 59.8$	$d_{2001} = 130$
$d_{2002} = -612.7$	$d_{2003} = -843.9$	$d_{2004} = 160.2$	$d_{2005} = 996.7$	$d_{2006} = 2197$
$d_{2007} = -2298$	$d_{2008} = 271$	$d_{2009} = 85$	$d_{2010} = -111$	$d_{2011} = 497$
$d_{2012} = -233$	$d_{2013} = -591$	$d_{2014} = 208.6$	$d_{2015} = 91.1$	$d_{2016} = -642.7$
$d_{2017} = -416.2$	$d_{2018} = 412.3$	$d_{2019} = 250.5\}$		

Kerala

$D = \{d_{1982} = 67.7\}$	$d_{1983} = -33.5$	$d_{1984} = -98.3$	$d_{1985} = -28.9$	$d_{1986} = -82.8$
$d_{1987} = -39.3$	$d_{1988} = -96.7$	$d_{1989} = -30.6$	$d_{1990} = 67.1$	$d_{1991} = 13$
$d_{1992} = -26.3$	$d_{1993} = 24.5$	$d_{1994} = -80.8$	$d_{1995} = -28.9$	$d_{1996} = -22.1$
$d_{1997} = -744.8$	$d_{1998} = 556.4$	$d_{1999} = -37.9$	$d_{2000} = 44.1$	$d_{2001} = -19.5$
$d_{2002} = -47.9$	$d_{2003} = -14.5$	$d_{2004} = -118.9$	$d_{2005} = 97.1$	$d_{2006} = -37.2$
$d_{2007} = 1.1$	$d_{2008} = -102.5$	$d_{2009} = 61.8$	$d_{2010} = 8$	$d_{2011} = -75.6$
$d_{2012} = 46.3$	$d_{2013} = -60.7$	$d_{2014} = 0.9$	$d_{2015} = 49.1$	$d_{2016} = -9$
$d_{2017} = -112.2$	$d_{2018} = 84.2$	$d_{2019} = 55.9\}$		

Tamil Nadu

$D = \{d_{1982} = 1448\}$	$d_{1983} = -2103$	$d_{1984} = 961.8$	$d_{1985} = 895.8$	$d_{1986} = 8.9$
$d_{1987} = -37.8$	$d_{1988} = 271.8$	$d_{1989} = -14.5$	$d_{1990} = 473.4$	$d_{1991} = -281$
$d_{1992} = 813.9$	$d_{1993} = 209.4$	$d_{1994} = -55.9$	$d_{1995} = 813$	$d_{1996} = -2272.8$
$d_{1997} = -3386.2$	$d_{1998} = 4989.9$	$d_{1999} = 1247.7$	$d_{2000} = -609.3$	$d_{2001} = -165.8$

(continued)

(continued)

$D = \{d_{1982} = 1448\}$	$d_{1983} = -2103$	$d_{1984} = 961.8$	$d_{1985} = 895.8$	$d_{1986} = 8.9$
$d_{2002} = -782.3$	$d_{2003} = -3006.9$	$d_{2004} = -354.3$	$d_{2005} = 1839.4$	$d_{2006} = 157.8$
$d_{2007} = 1390.6$	$d_{2008} = -1570.4$	$d_{2009} = 142.5$	$d_{2010} = 482.5$	$d_{2011} = 127.2$
$d_{2012} = 1666.3$	$d_{2013} = -3408.8$	$d_{2014} = 1299.9$	$d_{2015} = 489.2$	$d_{2016} = 1678.1$
$d_{2017} = -5147.7$	$d_{2018} = 4269.5$	$d_{2019} = -184.2\}$		

Step 2: Establish inverse fuzzy numbers of consecutive years based on domain v , as follows,

Andhra Pradesh

$$\left. \begin{aligned} v_\alpha &= \frac{1+0.0001}{\frac{1}{d_\alpha} + \frac{0.0001}{d_{\alpha+1}}}, \\ v_\alpha &= \frac{0.0001+1+0.0001}{\frac{0.0001}{d_{\alpha-1}} + \frac{1}{d_\alpha} + \frac{0.0001}{d_{\alpha+1}}}, \quad 1983 \leq \alpha \leq 2018, \\ v_\alpha &= \frac{0.0001+1}{\frac{0.0001}{d_{\alpha-1}} + \frac{1}{d_\alpha}}. \end{aligned} \right\} \quad (6)$$

Only for beginning value find out in the year of 1982.

$$v_{1982} = \frac{1 + 0.0001}{\frac{1}{d_{1982}} + \frac{0.0001}{d_{1983}}} = \frac{1.0001}{\frac{1}{856.5} + \frac{0.0001}{-196.6}} = \frac{1.0001}{0.001167034} = 856.96.$$

Next, except from find out the values of the year in 1982 and 2019, as given as,

$$v_\alpha = \frac{0.0001 + 1 + 0.0001}{\frac{0.0001}{d_{\alpha-1}} + \frac{1}{d_\alpha} + \frac{0.0001}{d_{\alpha+1}}}, \quad 1983 \leq \alpha \leq 2018,$$

$$v_{1983} = \frac{1.0002}{\frac{0.0001}{d_{1982}} + \frac{1}{d_{1983}} + \frac{0.0001}{d_{1984}}} = \frac{1.0002}{\frac{0.0001}{856.5} + \frac{1}{-196.6} + \frac{0.0001}{1119.4}} = -196.65,$$

.....

.....

$$v_{2018} = \frac{1.0002}{\frac{0.0001}{d_{2017}} + \frac{1}{d_{2018}} + \frac{0.0001}{d_{2019}}} = \frac{1.0002}{\frac{0.0001}{-36.3} + \frac{1}{713.8} + \frac{0.0001}{80.5}} = 714.71.$$

Finally, ending values find out in the year of 2019.

$$v_{2019} = \frac{0.0001 + 1}{\frac{0.0001}{d_{2018}} + \frac{1}{d_{2019}}} = \frac{1.0001}{\frac{0.0001}{713.8} + \frac{1}{80.5}} = \frac{1.0001}{0.0124225} = 80.51.$$

Karnataka

Only for beginning value find out in the year of 1982.

$$v_{1982} = \frac{1 + 0.0001}{\frac{1}{d_{1982}} + \frac{0.0001}{d_{1983}}} = \frac{1.0001}{\frac{1}{155.7} + \frac{0.0001}{-262.8}} = 155.725.$$

Next, except from find out the values of the year in 1982 and 2019, as given as,

$$v_{1983} = \frac{1.0002}{\frac{0.0001}{d_{1982}} + \frac{1}{d_{1983}} + \frac{0.0001}{d_{1984}}} = \frac{1.0002}{\frac{0.0001}{155.7} + \frac{1}{-262.8} + \frac{0.0001}{191.2}} = -262.933,$$

.....

.....

$$v_{2018} = \frac{1.0002}{\frac{0.0001}{d_{2017}} + \frac{1}{d_{2018}} + \frac{0.0001}{d_{2019}}} = \frac{1.0002}{\frac{0.0001}{-416.2} + \frac{1}{412.3} + \frac{0.0001}{250.5}} = 412.355.$$

Finally, ending values find out in the year of 2019.

$$v_{2019} = \frac{0.0001 + 1}{\frac{0.0001}{d_{2018}} + \frac{1}{d_{2019}}} = \frac{1.0001}{\frac{0.0001}{412.3} + \frac{1}{250.5}} = 250.509.$$

Kerala

Only for beginning value find out in the year of 1982.

$$v_{1982} = \frac{1 + 0.0001}{\frac{1}{d_{1982}} + \frac{0.0001}{d_{1983}}} = \frac{1.0001}{\frac{1}{67.7} + \frac{0.0001}{-33.5}} = 67.72.$$

Next, except from find out the values of the year in 1982 and 2019, as given as,

$$v_{1983} = \frac{1.0002}{\frac{0.0001}{d_{1982}} + \frac{1}{d_{1983}} + \frac{0.0001}{d_{1984}}} = \frac{1.0002}{\frac{0.0001}{67.7} + \frac{1}{-33.5} + \frac{0.0001}{-98.3}} = -33.507,$$

.....

.....

$$v_{2018} = \frac{1.0002}{\frac{0.0001}{d_{2017}} + \frac{1}{d_{2018}} + \frac{0.0001}{d_{2019}}} = \frac{1.0002}{\frac{0.0001}{-112.2} + \frac{1}{84.2} + \frac{0.0001}{55.9}} = 84.21.$$

Finally, ending values find out in the year of 2019.

$$v_{2019} = \frac{0.0001 + 1}{\frac{0.0001}{d_{2018}} + \frac{1}{d_{2019}}} = \frac{1.0001}{\frac{0.0001}{84.2} + \frac{1}{55.9}} = 55.901.$$

Tamil Nadu

Only for beginning value find out in the year of 1982.

$$v_{1982} = \frac{1 + 0.0001}{\frac{1}{d_{1982}} + \frac{0.0001}{d_{1983}}} = \frac{1.0001}{\frac{1}{1448} + \frac{0.0001}{-2103}} = 1448.244.$$

Next, except from find out the values of the year in 1982 and 2019, as given as,

$$v_{1983} = \frac{1.0002}{\frac{0.0001}{d_{1982}} + \frac{1}{d_{1983}} + \frac{0.0001}{d_{1984}}} = \frac{1.0002}{\frac{0.0001}{1448} + \frac{1}{-2103} + \frac{0.0001}{961.8}} = -2104.186,$$

.....

.....

$$\dots \dots \\ v_{2018} = \frac{1.0002}{\frac{0.0001}{d_{2017}} + \frac{1}{d_{2018}} + \frac{0.0001}{d_{2019}}} = \frac{1.0002}{\frac{0.0001}{-5147.7} + \frac{1}{4269.5} + \frac{0.0001}{-184.2}} = 4280.63. \\ \text{Finally, ending values find out in the year of 2019.}$$

$$v_{2019} = \frac{0.0001 + 1}{\frac{0.0001}{d_{2018}} + \frac{1}{d_{2019}}} = \frac{1.0001}{\frac{0.0001}{4269.5} + \frac{1}{-184.2}} = -0.018.$$

Step 3: A forecasting formula to forecast of this model as

$$F\alpha = E\alpha - 1 + v\alpha \quad (7)$$

Apply real number discrete domain and inverse fuzzy numbers Formula (6) and forecasting Formula (7) to forecast the rice production in AP, KA, KL, and TN from 1982 to 2019.

For example, $F1982 = E1981 + v1982 = 7011.4 + 856.96 = 7868.36$, similar from the years 1983 to 2019.

Fill in Table 1, the forecasting results on rice production in Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu from 1983 to 2019.

Step 4: Finally, it is compared with average forecasting error rate and mean square error using for actual and estimated values.

In this section, we assessed the forecasting efficacy of our proposed model for rice production data. Then, calculate average forecasting error (AFER) and mean square error (MSE) applying to Formulas (4) and (5). Forecasting accuracy measured in terms of mean forecasting error rate (AFER) and mean square error (MSE) is shown in Table 2.

6 Comparison of Four States: Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu

The forecasting results of rice production are shown in Table 1, and the forecast indicates that there are narrow variations between the actual and forecasted values of rice production in the selected states. The states Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu are compared with each other. The predicted value with actual value is computed using average forecasting error rate and mean square error [19]. The mean square error of AP, KA, KL, and TN values are 151,389.0003, 92,715.57221, 0.187049051, and 874.3084232. The AP state is very spurious to the other three states KA, KL, and TN. It exposes that AP state shows high forecast value comparing with other three states. The average forecasting error rate relationship of four states is AP > KA > TN > KL. The mean square error of AP, KA, KL, and TN values are 0.51, 1.96, 0.042, and 0.095%. The KA state is very spurious. The KA state shows high forecast value comparing with other three states. The mean square error relationship

Table 1 Actual and estimated values of AP, KA, KL, and TN

Year	Actual value Ei (AP)	Estimated value Fi (AP)	Actual value Ei (KA)	Estimated value Fi (KA)	Actual value Ei (KL)	Estimated value Fi (KL)	Actual value Ei (TN)	Estimated value Fi (TN)
1981	7011.4	–	2208.3	–	1272	–	4159	
1982	7867.9	7868.36	2364	2364.03	1339.7	1339.72	5607	5607.244
1983	7671.3	7671.25	2101.2	2101.07	1306.2	1306.19	3504	3502.814
1984	8790.7	8791.63	2292.4	2292.41	1207.9	1207.89	4465.8	4465.9333
1985	6909.1	6907.91	2374.9	2374.91	1255.9	1255.91	5361.6	5352.769
1986	7613.5	7613.72	1942.6	1942.24	1173.1	1173.09	5370.5	5370.501
1987	6591.5	6590.94	2313.2	2313.34	1133.8	1133.8	5332.7	5332.676
1988	7087.1	7087.22	1894.7	1894.56	1037.1	1037.14	5604.5	5605.261
1989	10,621.1	10,621.2	2509.6	4411.07	1006.5	1006.49	5590	5589.997
1990	9959.1	9959.1	2376.9	2376.83	1073.6	1073.59	6063.4	6065.125
1991	9654	9653.98	2415.3	2415.31	1086.6	1086.6	5782.4	5782.318
1992	9249.4	9249.37	2826.1	2825.67	1060.3	1060.29	6596.3	6596.382
1993	8792.2	8792.13	3068.7	3068.68	1084.8	1084.81	6805.7	6805.814
1994	9562	9562.49	3182.8	3182.9	1004	1003.98	6749.8	6749.787
1995	9276.7	9276.67	3167.5	3167.5	975.1	975.099	7562.8	7564.176
1996	9014.2	9014.17	3023.9	3024	953	952.997	5290	5289.063
1997	10,686	10,687.5	3211.6	3208.51	208.2	210.45	1903.8	1903.398
1998	8510	8509.14	3212.7	3212.7	764.6	765.571	6893.7	6893.437
1999	11,878	11,880.1	3656.9	3639.44	726.7	726.689	8141.4	8141.873
2000	10,637.8	10,637.4	3716.7	3716.71	770.8	770.823	7532.1	7532.173
2001	12,458	12,458.9	3846.7	3846.7	751.3	751.296	7366.3	7366.275
2002	11,389.8	11,389.6	3234	3233.63	703.4	703.418	6584	6584.233
2003	7327	7326.7	2390.1	2389.6	688.9	688.898	3577.1	3580.203
2004	8953	8952.98	2550.3	2550.13	570	570.06	3222.8	3222.727
2005	9601	9601.08	3547	3546.53	667.1	667.152	5062.2	5061.379
2006	11,704	11,701.1	5744	5744.17	629.9	629.765	5220	5220.028
2007	11,872	11,872	3446	3443.35	631	631	6610.6	6609.776
2008	13,324	13,322.8	3717	3716.97	528.5	527.5	5040.2	5037.976
2009	14,241	14,241.2	3802	3802.02	590.3	590.268	5182.7	5182.725
2010	10,538	10,535.4	3691	3690.96	598.3	598.301	5665.2	5664.95
2011	14,418	14,420.2	4188	4188.43	522.7	522.601	5792.4	5792.421
2012	12,895	12,894.8	3955	3954.95	569	569.015	7458.7	7456.934
2013	11,510	11,510	3364	3363.87	508.3	507.868	4049.9	4047.626
2014	12,724.7	10,295.3	3572.6	3572.6	509.2	509.2	5349.8	5349.764
2015	11,565.4	11,565.3	3663.7	3663.72	558.3	558.069	5839	5839.065
2016	7488.7	7534.57	3021	3020.52	549.3	549.299	7517.1	7516.914

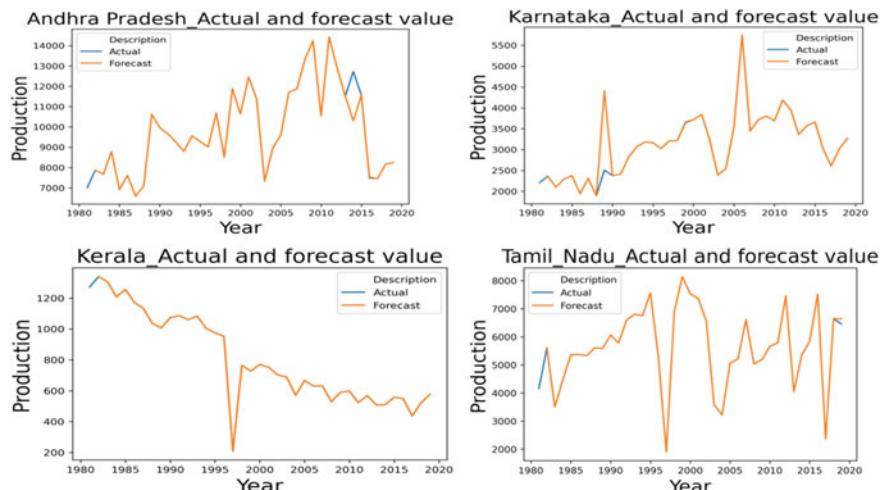
(continued)

Table 1 (continued)

Year	Actual value E_i (AP)	Estimated value F_i (AP)	Actual value E_i (KA)	Estimated value F_i (KA)	Actual value E_i (KL)	Estimated value F_i (KL)	Actual value E_i (TN)	Estimated value F_i (TN)
2017	7452.4	7452.39	2604.8	2604.7	437.1	437.2	2369.4	2366.17
2018	8166.2	8167.11	3017.1	3017.16	521.3	521.31	6638.9	6650.03
2019	8246.7	8246.71	3267.6	3267.61	577.2	577.201	6454.7	6638.882

Table 2 Average forecast error rate of AP, KA, KL, and TN

Error value	State			
	Andhra Pradesh (AP)	Karnataka (KA)	Kerala (KL)	Tamil Nadu (TN)
MSE	151,389.0003	92,715.57221	0.187049051	874.3084232
AFER (%)	0.51	1.96	0.042	0.095

**Fig. 2** Compared actual and estimated values of AP, KA, KL, and TN

of four states KA > TN > AP > KL is shown in Table 2. It has been studied that the predicted values are nearest to the actual values. And, some of the forecasting values are very close to the actual values shown in Fig. 2.

7 Conclusion

This paper discusses a new method of forecasting rice production employing fuzzy time series. The advantage of the method is its better accuracy mean square error and

average forecasting error rate comparing Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu states. The average forecasting error values which are shown in Table 2 are derived by using the proposed new method. The actual values and forecasted values are tabulated. It supports numerical and graphical representations of Fig. 2. This model forecasts production, an increase in the prices of rice in the upcoming years, and also a demand for the crop.

Acknowledgements The authors acknowledge the financial support of UGC under SAP (DRS-II) for carrying out this work.

References

1. Kumar P, Joshi PK, Birthal PS (2009) Demand projections for food grains in India. *Agric Econ Res Rev* 22(2):237–243
2. Agricultural statistics at a glance (2019) Government of India, Ministry of Agricultural and Farmers Welfare, Department of Agricultural Cooperation and Farmers Welfare, Directorate of Economics and Statistics
3. Rana AK (2018) Rice production forecasting through fuzzy time series. *Am Int J Res Sci Technol Eng Math* 23(1):158–162
4. Song Q, Chissom BS (1993) Fuzzy time series and its models. *Fuzzy Sets Syst* 54:269–277
5. Zadeh LA (1965) Fuzzy sets. *Inf Control* 8:338–353
6. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning—part I. *Inf Sci* 8:199–249
7. Wang H, Wang H, Guo J, Feng H (2014) A fuzzy time series forecasting model based on yearly difference of the student enrollment number. In: International conference on soft computing in information communication technology, pp 172–175
8. Abhishek, Kumar S (2017) A computational method for rice production forecasting based on high-order fuzzy time series. *Int J Fuzzy Math Arch* 13(2):145–157
9. Sullivan J, Woodall WH (1994) A comparison of fuzzy forecasting and Markov modeling. *Fuzzy Sets Syst* 64:279–293
10. Darekar A, Amarender Reddy A (2017) Forecasting of common paddy prices in India. *J Rice Res* 10(1):71–75
11. Sulaiga Beevi M, Senthamarai Kannan K, Syed Ali Fathima S (2020) Univariate time series models for fuel price. *Int J Sci Technol Res* 9(2):1490–1494
12. Manikandan M, Senthamarai Kannan K, Deneshkumar V (2013) Computational method based on distribution in fuzzy time series forecasting. *Int J Sci Res* 2(8):508–511
13. Senthamarai Kannan K, Sakthivel E (2014) Fuzzy time series model and ARIMA model—a comparative study. *Indian J Appl Res* 4(8):624–636
14. Samal P, Rout C, Repalli SK, Jambhulkar NN (2018) State-wise analysis of growth in production and profitability of rice in India. *Indian J Econ Dev* 14(3):399–409
15. Chen SM (1996) Forecasting enrollments based on fuzzy time series. *Fuzzy Sets Syst* 81:311–319
16. Arumugam P, Senthamarai Kannan K (2012) Computational algorithm of fuzzy stochastic model for forecasting. *J Algorithm Comput Technol* 6(3):375–383
17. Song Q, Chissom BS (1993) Forecasting enrollments with fuzzy time series—part I. *Fuzzy Sets Syst* 54:1–9
18. Song Q, Chissom BS (1994) Forecasting enrollments with fuzzy time series—part II. *Fuzzy Sets Syst* 54:1–8

19. Suresh S, Senthamarai Kannan K, Venkatesan P (2016) Higher order multivariate Markov chain model for fuzzy time series. *J Stat Manag Syst* 19(1):21–35

Data Visualization on Breast Phantom Mammogram Images Using Kernel Performance of SVM



A. R. Venmathi and L. Vanitha

Abstract Breast phantoms developed with nanotechnology are used as experimental models for evaluating and screening parameters while using mammographic screening. Nanotechnology is the emerging trend in manufacturing human organs to handle the difficult challenges of diseases like cancer. BI-RADS Atlas maintains a database for mammogram images and is used for analysis in this work. The affected area in breast phantoms is determined by using the region-growing technique and using the appropriate kernel function for categorizing cancer-affected areas. Data visualization is an attempt to understand data by putting it in a digital context in order to reveal models, patterns and associations that may not otherwise be identified. Python provides many large graphic libraries filled with several features. The classification accuracy for different types of kernels is determined by calculating sensitivity, specificity and classification accuracy. The analysis proves that the polynomial kernel of fourth degree has the highest classification accuracy compared to other types of kernels for this breast-phantom-based cancer classification.

Keywords Breast phantom · Data visualization · Micro-calcification · Nanodrug delivery · Region-based · Detection · Organ printing

1 Introduction

Among women, breast cancer is widespread globally, resulting in 25% of women diagnosed with cancer and a 15% death rate [1]. Mammogram, ultrasound and magnetic resonance imaging (MRI) are the prime diagnostic instruments for breasts. The significant limitations of these methods are either radiation or cost. To overcome these limitations, quality phantoms and efficient imaging techniques for assessing

A. R. Venmathi (✉)

Department of Electronics and Communication Engineering, Kings Engineering College, Chennai, India

L. Vanitha

Department of Electronics and Communication Engineering, Prathyusha Engineering College, Chennai, India

and testing these phantoms are required [2, 3]. The phantoms are numerical models used for clinical validation, which reduces the pain and risk to patients while they undergo medical diagnosis. The appropriateness of phantoms with human breasts gives the benefits of avoiding unwanted experiments on patients. Single phantoms fabricated with a single material urethane-based polymer for representing fibrogranular tissues and skin. Singlet uses a unique rigid material like bee wax to mimic granular tissue and skin. Commonly regular shapes of eggshells are used to model benign tumors, while irregular shapes are used to model malignant tumors. The micro-calcification clusters were included inside the phantom with the help of eggshells. These have similar X-ray attenuation characteristics as calcium oxalate. The eggshell contains enormous calcium content, which reflects calcification deposits inside the phantoms. Simulated micro-calcification clusters were of fine powders to represent benign tumors and irregular shape shells to show malignant tumors. The phantoms were imaged and characterized visually and quantitatively.

The virtual phantoms are categorized as rule-based and anatomy-based. These phantoms give a strong foundation for manufacturing physical structures with the derived definitions. Nanotechnology involves a significant role in processing these phantoms from a dedicated CAD system into tissue classes, glandular density and skin structure. Organ printing is the emerging method in nanospecialization to fulfill the requirements in the medical field, especially cancer patients, redeemed from tedious procedures. Printing of organs can be on Petri dishes instead of papers, while cells, chemicals and linkers replace the ink. 3D printing is the additive manufacturing technique for providing clear resolution in printing with reasonable cost to screen breast phantoms. PVC-based TMM material is used, and using 3D printing technology, breast phantoms are manufactured, which gives good-quality multi-modal images [4]. Opieliński et al. in their work proved ultrasound is capable of diagnosing breast phantom and identifying focal lesions [5]. Tuong and Gardiner claimed that phantom MRI images yielded good results in diagnosis [6] (Fig. 1).

Fig. 1 **a** Breast phantom.
b Mammogram image of the phantom

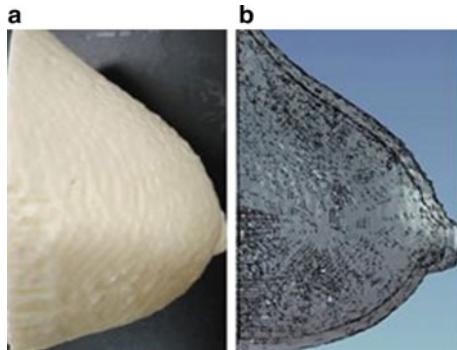


Table 1 Assessment categories of BI-RADS

Category	Impact
0	Need prior examinations
1	Negative
2	Benign
3	Probably benign
4	Suspicious
5	Highly suggestive of malignancy
6	Known biopsy proved

2 Literature Survey

Machine learning (ML) has changed the areas of medical image analyses and computer vision in recent years. However, there are still some questions about ML's use in clinical practice. Doctors may question the lack of classifier interpretability, or the point is that ML methods need vast quantities of data from preparation. Here we talk about some of the problems and display 1. As a generalization of the traditional algorithms automatically, it is possible to grasp how decision trees (a special class of ML models) and two are, how the dilemma of the labeled dataset (e.g. images) refer to manual and learning algorithms. The radiology department of the Academic Medical Centre in Amsterdam and the Rijnland Hospital in the Netherland maintain a database for breast phantom mammogram images. The organization of data divided for different breast compositions A, B, C, and D includes mass, asymmetry, architectural distortions, calcification and associated features. The final assessment is based on the findings shown in Table 1.

3 Region-Based Tumor Detection

Cells that persist in duplicating and failing to distinguish into specific cells that become immortal are called cancer cells. Extracting the image information is a vital task in image processing for image recognition, and the steps involved are twofold. The first step is detection, and the second step is segmentation. Based on specific criteria, the image divided into regions of the same nature extracts information from the image according to the area of interest. The gray-level discontinuities are used as a threshold level to detect the pixel dissimilarities regions. Region-based recognition is a method used for defining the area directly [7].

Image detection based on region-growing techniques involves a sequence of tasks, as shown in Fig. 2. Region growing can locate individual regions of the image and give information about the similar gray-level pixels, which can dissolve together to form a separate area that can define the area required to identify the affected cancerous region. The process starts with a single seed that is capable of adding similar pixels

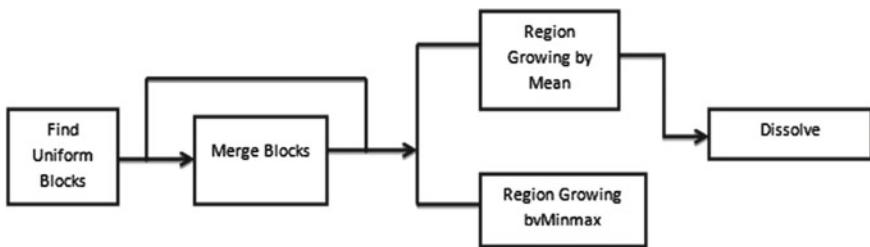


Fig. 2 Block diagram of region growing technique

slowly. This step is repeated for each newly added seed such that the exact area can be defined clearly. The process stopped if there are no more neighboring similar pixels. The property of each seed is determined for the coverage of the entire region of the image [8].

3.1 Finding Uniform Blocks

The initial step to define seed involves dividing the image into blocks. The selected 2×2 blocks enhance the region into 7×7 blocks. The single seed is chosen for each region to test the neighboring pixels for their gray levels and try to merge them into areas of the same gray levels. Since the digital image is a set of samples having continuously varying functions, the system checks whether the neighboring seed is less or greater than the threshold value for the origin seed. If the errors are small, then it is concluded that the pixel value nearing is related to the original value.

Partition the image into small seed regions.

The decision on region splitting or merging is done based on the max–min algorithm.

3.2 Region Growing with Gray Levels

As the defined threshold value determines which seed has to be added to form a merged region, the merged blocks are based on the minimum and maximum pixel intensities. If the difference in max–min values falls inside the threshold, then the seeds are combined into one block. If the difference in value exceeds the threshold, then the region is split into blocks forming an individual seed. Therefore, the merge split tree divides the complete image into regions carrying the same gray levels.

The properties of every spilt block are examined for the given criteria, and the algorithm is repeated for the entire region. The split and merge algorithm effectively detects the area of interest, and to avoid over-segmentation, the neighbors are verified [9].

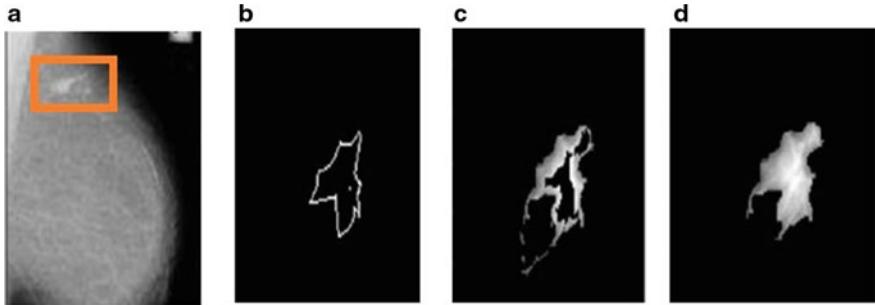


Fig. 3 **a** BI-RADS mammogram image composition D-calcification identified, **b** finding uniform blocks, **c** region growing, **d** merging

If $m = m + 1$, the region can be avoided and for the new region. And now compute. The difference is found out and decided whether the detection can be added or rejected, and Fig. 3 shows the sample output for region growing and merging algorithm [10].

4 SVM Classification

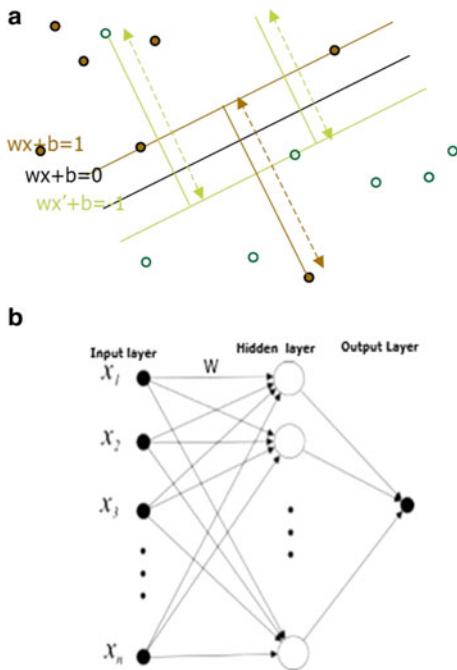
Support vector machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression challenges. It can solve linear and nonlinear problems and work well for many practical problems. The idea of SVM is simple as the algorithm creates a line or a hyperplane which separates the data into classes. An efficient method used for classifying unknown patterns is the kernel-based support vector machine [11, 12]. The input to the classifier is a feature vector, which organizes the data into different classes. Let x_i be the feature vector that belongs to training set X , mapped to one of two output categories ω_1 and ω_2 . Using this training dataset, SVM determines the best hyperplane.

$$g(x) = w^T x + w_0 = 0 \\ i = 1, 2, \dots, M$$

with a maximum margin, which can distinguish the input into two groups, as shown in Fig. 4a. The hyperplane determines a group that contains a high margin and gives improved organization performance. The architecture of the SVM structure is given in Fig. 4b.

The input layer of the network consists of neurons identical to the number of feature vectors. The intermediate layer, called the hidden layer, comprises h number of perceptions, and the value of ' h ' is fixed based on the experiment conducted. The output layer consists of single neuron that represents the output to anyone class, either benign or malignant [13]. In this work, SVM is proposed for the classification,

Fig. 4 **a** Hyperplane separating the class.
b Architecture of SVM



and the appropriate kernel function is determined from performance analysis. The input dataset is converted into a high-dimensional feature space, and the SVM kernel function makes the input linearly separable. Four generally used kernel functions are radial basis function, polynomial, sigmoid or hyperbolic function, and Mahalanobis, which are analyzed for breast cancer classification, and its performance is compared. The different kernel functions are described below:

- Radial basis function,
- Polynomial kernel function,
- Sigmoid kernel also called as the hyperbolic tangent kernel or multilayer perceptron,
- Mahalanobis kernel is the scaling factor.

5 SVM Classification

The performance of different kernels is compared based on skew, kurtosis, sensitivity, specificity and classification accuracy [14]. The goal of the SVM algorithm is to find an N -dimensional hyperplane that separates the data points. There are several potential hyperplanes to pick to differentiate the two types of data points. Maximizing the margin gap provides some reduction to maximize potential data points [15]. Hyperplanes are decision limits that help in the classification of data points. SVM

Table 2 State prediction with confusion matrix

States	Predicted state		
Actual states	True	TP	FN
	False	FP	TN

predicts the given scanned images into normal/abnormal. The SVM classifier is defined by as follows:

Evaluation Metrics

The performance measures used to compare the SLC system designed in the previous chapter with other schemes are accuracy, sensitivity and specificity [16]. They are defined below:

Accuracy: The classification accuracy metric gives the total number of correctly classified skin images. The formula to compute the accuracy is given below:

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

Sensitivity: TPR metric shows the ability of the system to identify skin cancer cases. The formula to compute the accuracy is given below:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity: TNR metric shows the ability of the system to identify normal or negative results. The formula to compute the accuracy is given below:

$$\text{sensitivity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

where

True positive (TP) = the number of positive trials correctly determined.

False negative (FN) = the number of positive trials wrongly determined as negative. False positive (FP) = the number of negative trials wrongly determined as positive. True negative (TN) = the number of negative trials correctly determined.

The state prediction with the confusion matrix is shown in Table 2.

6 Experimental Results

The extracted feature values and the performance measures were tabulated, and efficiency is compared for different kernels. The accuracy measure highlights the

kernel is having high sensitivity and low specificity. Table 3 shows the values obtained for mean, variance, standard deviation, skew, kurtosis, sensitivity, specificity and accuracy [17, 18] which are used for calculating the performance of various kernel functions, and the categorizing results obtained using SVM classifier with diverse functions of the kernel also tabulated. Figure 5 shows the graphical representation of the comparison between classification accuracy between different kernels [19, 20]. The accuracy of various means of kernel clustering for SVM classifiers shows the variation in bar chart.

7 Discussion and Conclusion

The proposed methodology for phantom images increases the convenience of the patient. The phantom was manufactured from original mammography, and the system techniques were applied to the phantom images. This can result from the severity of the calcification mass, and the physician can decide whether the patient needs a biopsy [21]. The results prove that the polynomial kernel with degree four performs better compared to another type of kernel in SVM classifier, which gives high classification accuracy. The classification accuracy can be increased by including more parameters or including a modified kernel in SVM.

Table 3 Experimental values for features and performance measure

Kernel type	Phantom mammo-composition	Mean	Variance	Standard deviation	Skew	Kurtosis	Sensitivity	Specificity	Accuracy %
RBF	A	0.5101	0.2497	0.4997	0.2	1.2	93.242	6.758	96.284
	B	0.6812	0.2170	0.469	0.3	1.6	92.178	7.822	96.546
	C	0.8661	0.1155	0.3399	0.3	1.2	95.321	4.679	96.228
D	A	0.6225	0.2350	0.4848	0.2	1.1	93.544	6.456	96.193
	B	0.6784	0.2182	0.4671	0.4	1.6	96.513	3.487	96.749
	C	0.8704	0.1128	0.3359	0.5	1.6	97.159	2.841	97.411
Polynomial	A	0.5612	0.2463	0.4952	0.4	1.7	95.601	4.399	96.656
	B	0.8448	0.1311	0.3621	0.5	1.8	97.732	2.268	96.824
	C	0.7542	0.1854	0.4306	0.3	1.5	96.120	3.88	96.635
Sigmoid	A	0.9288	0.0661	0.2570	0.4	1.6	97.174	2.826	96.753
	B	0.6904	0.2137	0.4623	0.5	2.5	95.396	4.604	97.213
	C	0.9813	0.0183	0.1323	0.6	1.9	91.626	8.374	95.146
Mahalanobis	A	0.6467	0.2285	0.4780	0.3	1.2	93.568	6.432	96.165
	B	0.6130	0.2378	0.4877	0.5	1.9	94.651	5.349	96.316
	C	0.5023	0.2500	0.5000	0.2	1.1	95.232	4.768	96.235

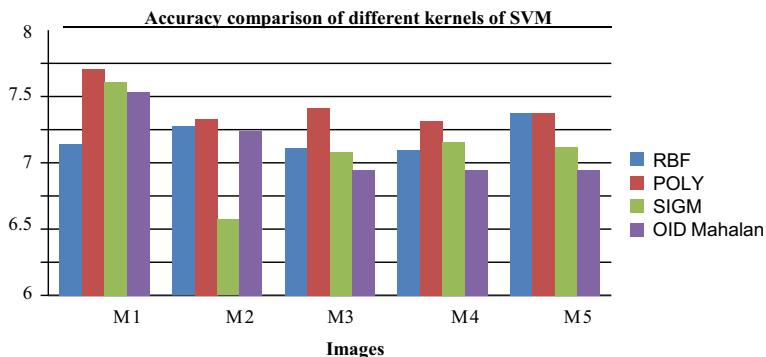


Fig. 5 Graph to highlight the best kernel of SVM

References

1. Carton AK, Bakic P, Ullberg C, Derand H, Maidment AD (2011) Development of a physical 3D anthropomorphic breast phantom. *Med Phys* 38(2):891–896
2. Wallace D, Ng J, Keall P, O'Brien R, Poulsen P, Juneja P, Booth J (2015) Determining appropriate imaging parameters for kilovoltage intrafraction monitoring: an experimental phantom study. *Phys Med Biol* 60(12):4835–4847
3. Jung J, Song S, Yoon S, Kwak J, Yoon K, Choi W, Jeong S, Choi E, Cho B (2015) Verification of accuracy of CyberKnife tumor-tracking radiation therapy using patient-specific lung phantoms. *Int J Radiat Oncol Biol Phys* 92(4):745–753
4. He Y, Liu Y, Dyer BA, Boone JM, Liu S, Chen T, Zheng F, Zhu Y, Sun Y, Rong Y, Qiu J (2019) 3D-printed breast phantom for multi-purpose and multi-modality imaging. *Quant Imaging Med Surg* 9(1):63–74
5. Opieński KJ, Pruchnicki P, Gudra T, Podgórski P, Kraśnicki T, Kurcz J, Sasiadek M (2014) Breast phantom imaging results from an ultrasound computer tomography research system. In: Information technologies in biomedicine, vol 3. Advances in intelligent systems and computing
6. Tuong B, Gardiner I (2013) Development of a novel breast MRI phantom for quality control. *Am J Roentgenol* 201(3)
7. Polak M, Zhang H, Pi M (2009) An evaluation metric for image segmentation of multiple objects. *Image Vis Comput* 27(8):1223–1227
8. Singh R, Singh J, Sharma P, Sharma S (2011) Edge based region growing. *Int J Comput Technol Appl* 2(4):1122–1126
9. Maitra IK, Nag S, Bandyopadhyay SK (2011) Automated digital mammogram segmentation for detection of abnormal masses using binary homogeneity enhancement algorithm. *Indian J Comput Sci Eng (IJCSE)*
10. Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med* 34:113–127
11. Vapnik VN (1995) The nature of statistical learning theory. Springer, Berlin
12. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
13. GokulRaj N, Umapathy K (2016) 5G wireless mesh network 802.11s load balancing architecture for 802.11 Bgn radio-PCI interface. *Procedia Comput Sci* 87:252–257
14. Umapathy K, Rajaveerappa D (2012) Implementation of low power and small area 128-point mixed radix 4-2 FFT processor for OFDM applications. *Eur J Sci Res (EJSR)* 84:112–118. ISSN: 1450-216X

15. Joseph F, Subbiah M (2018) Hybrid windowing adaptive FIR filter technique in underwater communication. *Int J MC Square Sci Res* 10(2):17–21
16. Kumarapandian S (2018) Melanoma classification using multiwavelet transform and support vector machine. *Int J MC Square Sci Res* 10(3):01–07
17. Babu VH, Balaji K (2020) Survey on modular multilevel inverter based on various switching modules for harmonic elimination. In: Intelligent computing in engineering. Springer, Singapore, pp 451–458
18. Lakshmi Narayanan K, Ramesh GP (2017) VLSI architecture for multi-band wavelet transform based image compression and image reconstruction. *J Eng Appl Sci* 2:6281–6285
19. Kumaran T (2020) Link quality and energy-aware metric-based routing strategy in WSNS. In: Intelligent computing in engineering. Springer, Singapore, pp 533–539
20. Satpathy RB, Ramesh GP (2020) Advance approach for effective EEG artefacts removal. In: Balas V, Kumar R, Srivastava R (eds) Recent trends and advances in artificial intelligence and internet of things. Intelligent systems reference library, vol 172. Springer, Cham
21. Manikandan G, Anand M (2020) Radix-2/4 FFT multiplierless architecture using MBSLS in OFDM applications. In: Intelligent computing in engineering. Springer, Singapore, pp 553–559

Role of Fog-Assisted Internet of Things-Enabled System for Managing the Impact of COVID-19



Upendra Verma , Mayank Sohani , Samarjeet Borah ,
Kapil Kumar Nagwanshi , and Sunil Pathak

Abstract The effect of pandemic COVID-19 outbreak has become a big matter of concern in the world. Healthcare industry demands for new technologies to fight against the pandemic. Fog-assisted IoT-enabled technology (Fog-IoT) is the alternative to cloud technology, which has potential strength to fulfill the requirements of patients as well as healthcare organization. In this paper, we explore and review the fog computing technology to mitigate the impact of COVID-19. Fog computing technology provides resources to IoT at proximity of network. This integrated technology is useful for dynamic monitoring of patients and provides rapid diagnosis to high risk patients. In healthcare industry, the delay sensitive patient information should be accessed in a fraction of seconds. So, fog computing could be a better solution for providing response intensive IoT application for medical emergencies.

Keywords Analytics · Fog computing · COVID-19 · Healthcare industry · IoT application · Medical emergencies

U. Verma · M. Sohani

SVKM's NMIMS University, MPSTME Shirpur Campus, Shirpur, Maharashtra 425405, India
e-mail: upendra.verma@nmims.edu

M. Sohani

e-mail: mayank.sohani@nmims.edu

S. Borah

Sikkim Manipal Institute of Technology, Majitar, Rangpo, East Sikkim, Sikkim 737132, India
e-mail: samarjeet.b@smit.smu.edu.in

K. K. Nagwanshi · S. Pathak

Department of Computer Science & Engineering, Amity School of Engineering & Technology,
Amity University Rajasthan, Jaipur, India
e-mail: dr.kapil@ieee.org

1 Introduction

The COVID-19 is an infection caused by severe respiratory syndrome Coronavirus-2 [12]. The COVID-19 has affected entire world that infection was first started from Wuhan, in December 2019 [7]. The epicenter of COVID-19 virus was connected to Wuhan's wet market [22]. Figure 1 shows the entire timeline of COVID-19 pandemic [1].

It is the fastest contagious virus attacks like influenza virus, causing ailments such as fever, breathlessness, and cough. The exact source of coronavirus is unidentified [4]. COVID virus is spherical positive sense RNA viruses ranging from 600 to 1400 Å in diameter [5, 19] as shown in Fig. 2.

The nature of virus is extremely contagious and required long incubation period (4–14 days). It is very difficult to find virus propagation mechanism. Firstly, virus is exposed through any person in close contact with an infected person. Secondly, virus is also transmitted through indirect contact with infected surfaces. After the corona outbreak came into the existence, several myths were also reported regarding to COVID-19. Table 1 shows the myths and actual facts surrounding the COVID-19 [25, 26].

This paper evaluated the use of fog computing technology that help to mitigate the adverse effect of this pandemic and expedite the recovery process. Fog computing can help us to fight against the virus without human intervention at any level. Fog computing was first proposed by CISCO in January 2014. Fog computing is the technology that distributes services and resources of computing, control, networking, and storage anywhere along the continuum from cloud of thing [2, 18]. Fog computing

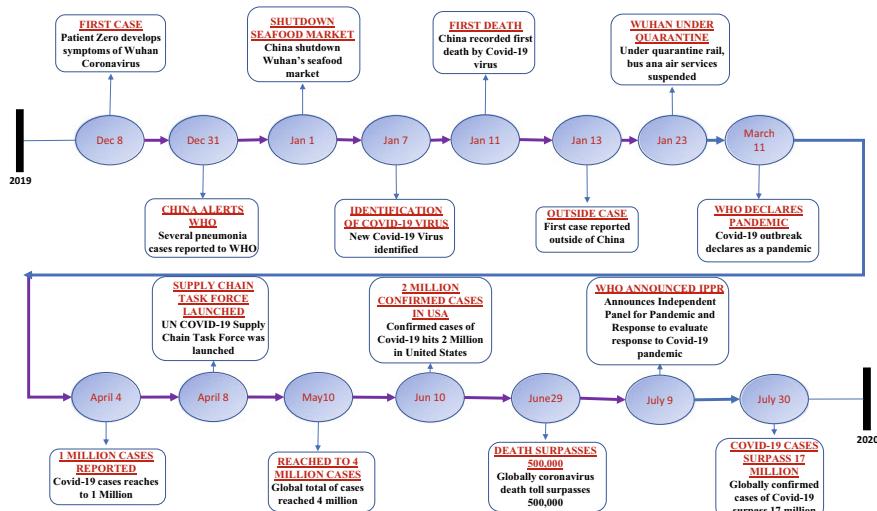


Fig. 1 Timeline of COVID-19 pandemic (figure is created by author as per literature available)

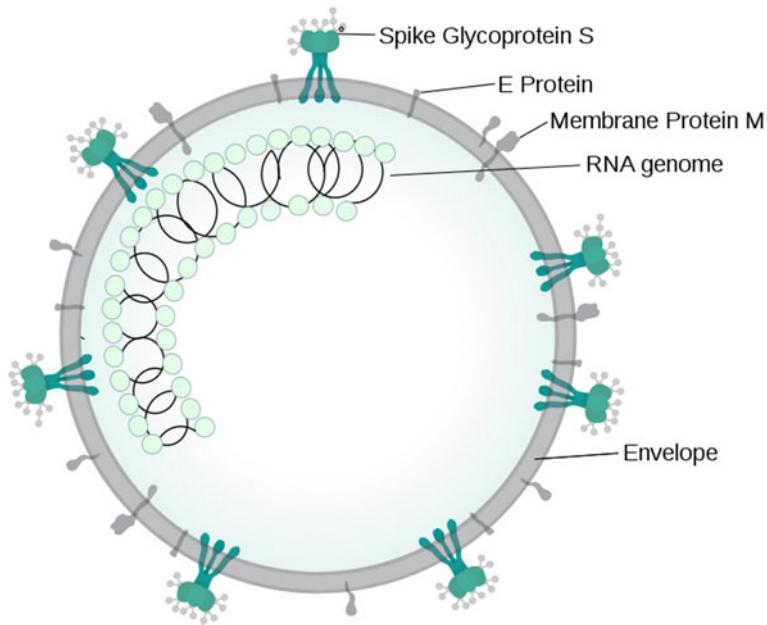


Fig. 2 COVID-19 virus virion structure (*Note* Image is licensed under CC license. *Image source* <https://commons.wikimedia.org>)

Table 1 Myths and facts surrounding COVID-19

Myths	Facts
Taking a hot bath can help to prevent COVID-19	Taking a hot bath does not prevent COVID-19
The Coronavirus can be spread via mosquito bites	Till now, there is no evidence to support this statement
COVID-19 disease have for life time	You can recover from the COVID-19
Thermal scanning can help to detect Coronavirus infection	Thermal scanners can only detect the presence of fever
Eating garlic can help to protect from COVID-19	Garlic cannot prevent from COVID-19
Only older people can infected from COVID-19	COVID-19 can affect people of all age groups
The coronavirus can be transmitted over 5G networks	The COVID-19 cannot be transmitted over 5G networks or radio waves transmissions
Warm weather can kill the COVID-19 virus	Warm weather cannot kill the COVID-19 virus
Hand dryers can be used to kill the corona virus	Hand dryers cannot be used to kill the corona virus

provides services closer to IoT devices [20]. The characteristic of fog computing is to process IoT data locally rather than globally. Fog computing brings the storage and processing load at the proximity of edge fog computing enables to provide faster decision making and processing of data, which supports healthcare authorities to take faster prevention action. The fog-enabled IoT is the cost effective and practical solution to fight against the problem with real-time monitoring at minimum response delay. So, fog and cloud technologies are working together to provide the solution for healthcare system.

2 Fog Computing Background

Fog computing provides applications and services at edge of the network [21]. In fog computing, fog nodes are placed proximity to IoT devices. So, fog provides real-time analysis of data. Fog computing is used when IoT devices require faster response time for applications such as smart traffic management and smart healthcare system. To fight against the coronavirus, patient data should be analyzed in fraction of seconds, and fog computing can be used to serve this purpose.

The first fog computing architecture was proposed by Bonomi et al. [3]. The characteristics preserved by fog computing are response time, mobility support, interoperability, wireless connectivity, distributed nature, real-time analysis of data, interconnectivity with cloud, supports large number of devices. These characteristic makes fog computing as a good technological solution to deal with the COVID-19 pandemic. Figure 3 shows the fog computing architecture.

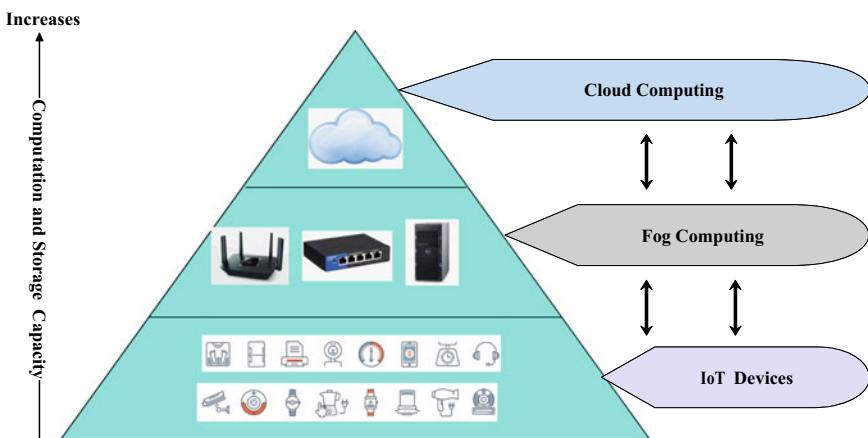


Fig. 3 Fog computing architecture

3 Need for Study

Coronavirus has affected across the world is revolving under the health crisis and crashing economies. Unfortunately, the numbers of infected patients are growing rapidly. However, fog-enabled IoT system is the emerging fore front approach to deal with the influences of COVID-19 pandemic. Fog computing plays a vital role to various domains such as smart transportations [10], smart city [17]. Fog computing has already employed to serve their purpose in smart healthcare monitoring system [16] and Internet of Medical Things (IoMT) [6]. According to WHO, the digital technologies can play indispensable role for handling the impact of COVID-19 [24].

4 Fog Computing for Mitigating the Impact of COVID-19 Pandemic

Fog computing is the new emerging and innovative technology, which provides services closer to the IoT devices. The real-time monitoring of patients is the major critical issues during this crisis. Several medical organization and government bodies are looking solution for real monitoring and processing of patient's data without delay. Fog computing can provide solution to the challenges of monitoring, contact tracing, surveillance, and patient's data processing. IoT/IoMT technology gives opportunities to solve various issues of COVID-19 [23] but still IoT facing various challenges. Fog computing can solve many IoT challenges [8] shown in Fig. 4.

So more particularly fog-IoT integrated technology can provide a solution to the challenges of real monitoring, rapid diagnosis, screening, and surveillance. Fog computing can be basis infrastructure for transforming the healthcare IoT from novelty to reality. The technology can be used to control the effect of SARS-CoV-2. The location of the fog closer to the end user that reduces the jitter, latency, and delay. The integration of fog computing and IoT for provisioning healthcare services to reduce the impact of COVID-19 pandemic. Figure 5 illustrates the use case of fog computing to mitigate the effect of COVID-19 outbreak.

To perform real monitoring of patients in quarantine center, the sensor network (sensors mounted on patient's body) generates patient health data (PHD) and PHD send to fog server for storage and processing. The fog server is capable to analyze patient's data in order to detected abnormal conditions. A decision framework located at fog server, which uses the machine learning techniques to identify the abnormal conditions that trigger an alarm system. If there is no abnormal case, then data is transferred to the cloud. In this approach, the intervention of healthcare authorities is not necessary in order to provide alarm notification service. Fog-assisted IoT-enabled system is the extension of conventional IoT and novel technology to fight against COVID-19 outbreak.

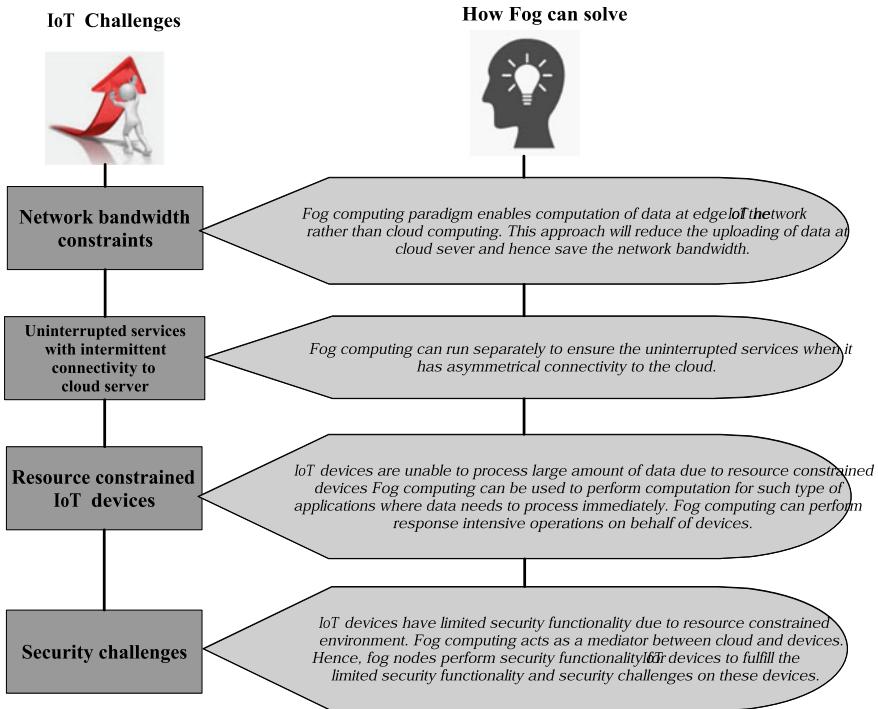


Fig. 4 Fog helps to address IoT challenges

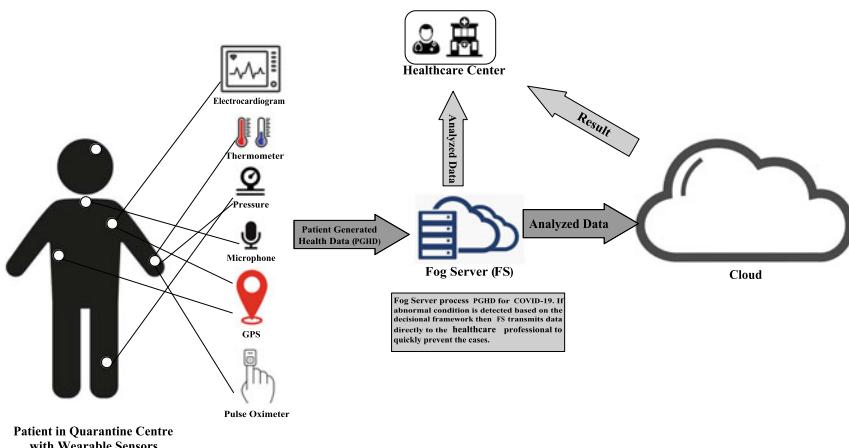


Fig. 5 Fog computing for mitigating the impact of COVID-19 outbreak

5 Applications of Fog Computing

5.1 Real-Time Processing of Patient Data

Fog computing can process the data and quickly alarm the patients and healthcare authorities. IoT devices are resource constrained to perform computing. Fog computing closer to users can be realized for real-time processing of patient data.

5.2 Remote Patient Health Monitoring

For many healthcare industries having a simple IoT-cloud architecture is not feasible due to fact that most hospital would not prefer patient data to be stored outside [11]. Using only cloud may cause delay during the real patient health monitoring. To deal with the COVID-19 situations, we have emergency response system that require real-time monitoring operations [14].

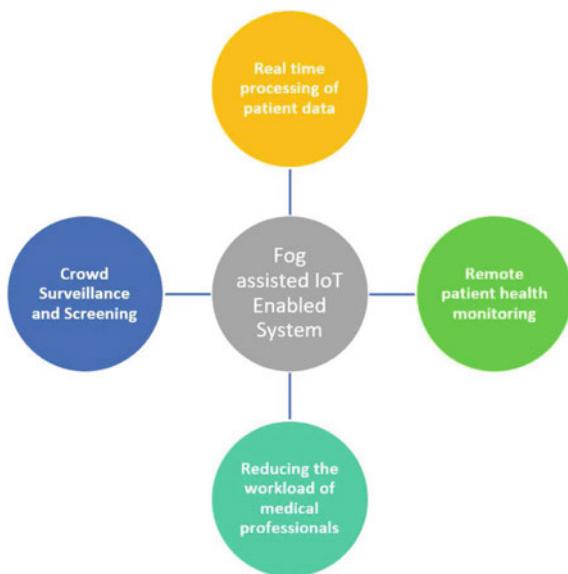
5.3 Crowd Surveillance and Screening

Social distancing is the necessary steps to stop the spread of COVID-19 outbreak. Fog computing devices can be deployed to various entry points of shopping complex outdoor, bus stations, railway stations, airports, etc., to access the facial identity of a person [9], and the data is useful for surveillance and observe the crowd efficiently [13, 15].

5.4 Reducing the Workload of Medical Professionals

Fog computing enables real time and remote monitoring of coronavirus infected patients without the intervention of doctors and healthcare professionals in order to provide healthcare services. The medical professionals need not to send alarm services to patients as discussed in Sect. 4. The fog computing assisted IoT system can provide response intensive decision making framework to reduce the work load of healthcare industry in order to deal with COVID-19 pandemic. Figure 6 illustrates the application domain of fog-enabled IoT system.

Fig. 6 Application domains of Fog-assisted IoT-enabled system



6 Conclusions

Fog computing provides computing and networking services at proximity of users to fight against COVID-19 pandemic. Fog computing is integrated with IoT called fog-assisted IoT-enabled architecture that provides effective use cases to identify coronavirus infected patients with minimal response time. In fog computing, patient's data are processed, and messages are conveyed to healthcare professionals without any type of intervention. Fog computing is not only useful to provides services closer to coronavirus patients, but also useful for real-time monitoring. In this review, we highlight a comprehensive outlook of COVID-19 outbreak. The myths and facts and timeline of COVID-19 are also discussed in a good detail. Fog computing technology is useful to prevent and manage the community spread of COVID-19 outbreak. Healthcare industries, professionals, and medical staff can create healthy environment to fight against COVID-19 pandemic with proper implementation of fog computing technology.

References

1. Anonymous (2020) Timeline of who's response to Covid-19. <https://www.who.int/news-room/detail/29-06-2020-covidtimeline>. Accessed 17 July 2020
2. Atlam H, Walters R, Wills G (2018) Fog computing and the internet of things: a review. *Big Data Cogn Comput* 2. <https://doi.org/10.3390/bdcc2020010>
3. Bonomi F, Milito R, Zhu J, Addepalli S (2012) Fog computing and its role in the internet of things. In: Proceedings of the first edition of the MCC workshop on mobile cloud computing, pp 13–16
4. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R (2020) Features, evaluation and treatment coronavirus (Covid-19). Statpearls [Internet]
5. Commons W (2020) File: 3d medical animation corona virus.jpg. Wikimedia commons, the free media repository (2020). https://commons.wikimedia.org/w/index.php?title=File:3D_medical_animation_corona_virus.jpg&oldid=479016449 (Online). Accessed 26 Feb 2021
6. Dubey H, Monteiro A, Constant N, Abtahi M, Borthakur D, Mahler L, Sun Y, Yang Q, Akbar U, Mankodiya K (2017) Fog computing in medical internet-of-things: architecture, implementation, and applications. In: *Handbook of large-scale distributed computing in smart healthcare*. Springer, Berlin, pp 281–321
7. Hiranya SA, Priya J, Lakshminarayanan Arivarasu VP (2020) Challenges faced by china on Covid-19. *Eur J Mol Clin Med* 7(1):2230–2235
8. Hu P, Dhelim S, Ning H, Qiu T (2017) Survey on fog computing: architecture, key technologies, applications and open issues. *J Network Comput Appl* 98:27–42
9. Hu P, Ning H, Qiu T, Zhang Y, Luo X (2016) Fog computing based face identification and resolution scheme in internet of things. *IEEE Trans Ind Inform* 13(4):1910–1920
10. Hussain MM, Alam MS, Beg MS (2019) Fog computing model for evolving smart transportation applications. *Fog Edge Comput Principles Paradigms* 22(4):347–372
11. Kraemer FA, Braten AE, Tamkittikhun N, Palma D (2017) Fog computing in healthcare—a review and discussion. *IEEE Access* 5:9206–9222
12. Vimal Kumar MN, Jaya R, Rubesh CM, Aakash Ram S (2020) Statistical analysis on novel corona virus: Covid-19. *Eur J Mol Clin Med* 7(1):95–103
13. Manikanth M, Swaminathan K et al (2020) MQTT based smart husbandry monitoring framework with live surveillance. *Eur J Mol Clin Med* 7(4):2412–2417
14. Mohammed AA, Burhanuddin M, Talib MS, Hameed ME, Ali MF (2020) A review on IoT-based healthcare monitoring systems for patient in remote environments. *Eur J Mol Clin Med* 7(3):2227–2235
15. Nasir M, Muhammad K, Lloret J, Sangaiah AK, Sajjad M (2019) Fog computing enabled cost-effective distributed summarization of surveillance videos for smart cities. *J Parallel Distrib Comput* 126:161–170
16. Paul A, Pinjari H, Hong WH, Seo HC, Rho S (2018) Fog computing-based IoT for health monitoring system. *J Sens* 2018
17. Perera C, Qin Y, Estrella JC, Reiff-Marganiec S, Vasilakos AV (2017) Fog computing for sustainable smart cities: a survey. *ACM Comput Surv (CSUR)* 50(3):1–43
18. Peter N (2015) Fog computing and its real time applications. *Int J Emerg Technol Adv Eng* 5(6):266–269
19. Singh B, Kumar V, Tripathi S (2020) A review of Covid-19 based on current evidences. *Int Res J Modernization Eng Technol Sci* 2(8):1449–1459
20. Singh RP, Javaid M, Haleem A, Suman R (2020) Internet of things (IoT) applications to fight against Covid-19 pandemic. *Diab Metab Syndr Clin Res Rev* 14(4):521–524
21. Singh SP, Nayyar A, Kumar R, Sharma A (2019) Fog computing: from architecture to edge computing and big data processing. *J Supercomput* 75(4):2070–2105
22. Sohrabi C, Alsaifi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R (2020) World health organization declares global emergency: a review of the 2019 novel coronavirus (Covid-19). *Int J Surg* 76:71–76

23. Swayamsiddha S, Mohanty C (2020) Application of cognitive internet of medical things for Covid-19 pandemic. *Diab Metab Syndr Clin Res Rev*
24. Whitelaw S, Mamas MA, Topol E, Van Spall HG (2020) Applications of digital technology in Covid-19 pandemic planning and response. *Lancet Digit Health*
25. WHO (2020) Coronavirus disease (Covid-19) advice for the public: mythbusters. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters> (Online). Accessed 26 Feb 2021
26. WHO (2020) Who situation report. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters> (Online). Accessed 26 Feb 2021

Intelligent Big Data Analytics: A Perspective for IoHT and HealthCare



Preetishree Patnaik, Brojo Kishor Mishra, Vivek Jaglan,
and Manoj Kumar Sahoo

Abstract Currently, we are in the COVID-19 pandemic situation, where the healthcare sector plays a major role to prevent the loss of life across the globe. During this COVID-19 period, the remote healthcare assistance and monitor systems show their impotence and existence. Science and technology are always associated with the healthcare sector, and it helps to diagnose the cause of bad health and helps to improve by proper treatment. The analysis of the causes plays a major role in treatment. With the advance of science and technology for the complex clinical solution in healthcare, IoHT (internet of health care things) and intelligent analytics in big data can play a bigger role in cancer detection, brain tumor, etc. The main focus of our work is to discuss a framework of intelligent analytics for big data utility to such a complex clinical problem on the IoHT platform. The digital imaging of medical imaging is playing a critical role in the clinical operation of a human in healthcare. The treatment of cancer or the brain is a very complex multistage treatment process; here, use of intelligent big data analytics and IoHT can help the doctors to provide better decision making which helps in treatment and recovery stages. So the remote healthcare sector, which is showing its utility during the COVID-19 situation, where intelligent data analytics play a major role to provide better healthcare solutions and monitoring to the human society. The intelligent medical systems for healthcare need proper data acquisition health condition, which need to analyze, store and process for each individual for clinical treatment and future reference. The past information periodically and present health condition can store and analyzed for better healthcare, this leads to an intelligent database system, where Intelligent data analytics play a big role, with the utility of the advanced platform it becomes robust and can be the backbone of the healthcare sector.

P. Patnaik · B. K. Mishra
GIET University, Gunupur, Odisha, India

V. Jaglan
GEH University, Dehradun, Uttarakhand, India

M. K. Sahoo (✉)
National Institute of Science & Technology (Autonomous), Berhampur, Odisha, India

Biju Patnaik University of Technology Rourkela, Rourkela, Odisha, India

Keywords Big data · Data analytics · IoHT · Intelligent healthcare system · Remote healthcare monitoring system

1 Introduction

Augmented analytics or intelligent big data analytics is an emerging platform utility achieved by integrating big data, data analytics with decision making or prediction is carried out by the use of machine learning or artificial intelligence. The main objective here to discuss a framework with intelligent data analytics for a medical utility where the data nature is complex to understand with a larger number. The augmented analytics or intelligent big data analytics for management and clinical taking into account with non-clinical managerial nature functions: planning, organizing, and clinical function: per history of the patient, clinical testing results, treatment phase, and recovery phase with the proper monitor.

Nowadays, there are more and more successful deep learning methods that have made remarkable progress in image classification, especially AI (artificial intelligence) and machine learning have beaten other traditional machine learning methods even human levels [1–3]. However, the deep learning model is only used in the natural image domain but no other domains, especially only a really small part is used in the medical image domain. Magnetic resonance imaging (MRI) is widely used in routine clinical diagnosis and treatments. In Japan, brain dock is being conducted for medical checkups [4]. We get brain dock MRI structured images with gender labels from clinics and using transfer learning methods to apply AI/ML on them to make a decision. We not only save a lot of doctors diagnose time, improves clinics, working efficiency, solve the insufficient health care resource problem [5], but also get a conclusion that there are some relationships between the MRI images. We propose an application of the transfer learning strategy to the brain dock MRI gender estimation and experiment to clarify the performance [6] (Figs. 1 and 2).

AI/ML is become a core of analytics, transforms business process, optimize and regroups the workforce, optimizes the infrastructure resources, and blends the operation and services. A 30% growth of new revenue in a cloud-based intelligent analytics solution is predicated by 2021 [7].

2 A Healthcare Framework by Utilizing Intelligent Big Data Analytics

Figure 3 is a graphical flow of the smart or intelligent framework for a healthcare system with the utilizing of augmented analytics or intelligent analytics in big data, which is built-in considerations of intelligent analytics for big data as a science and technology, system, service management for improving healthcare and hospital working decision making [8–10].

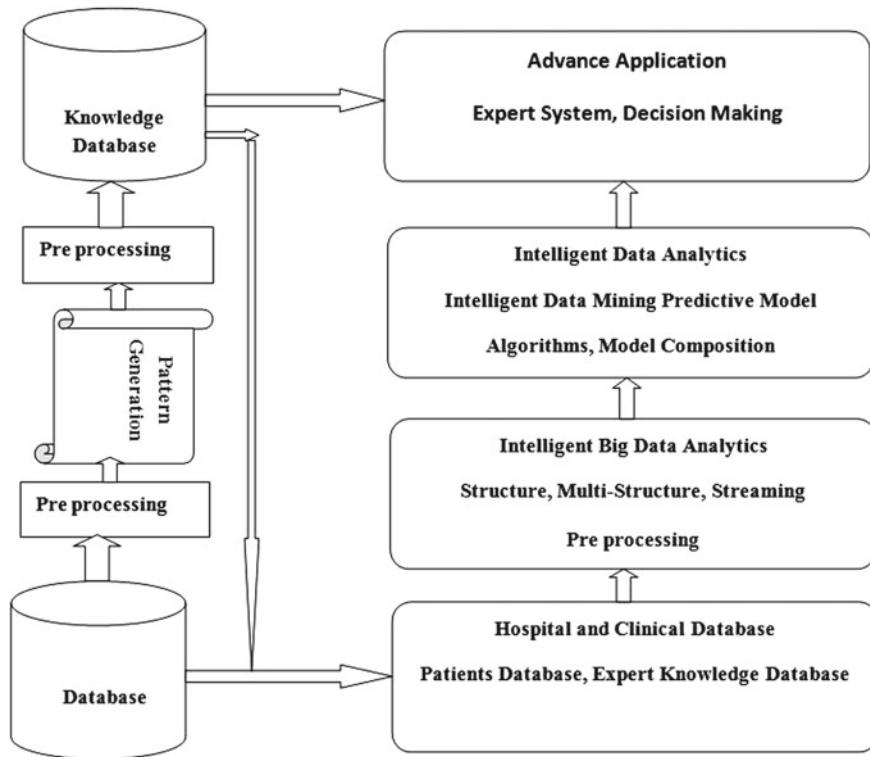


Fig. 1 Simple data flow healthcare framework model using intelligence big data analytics system

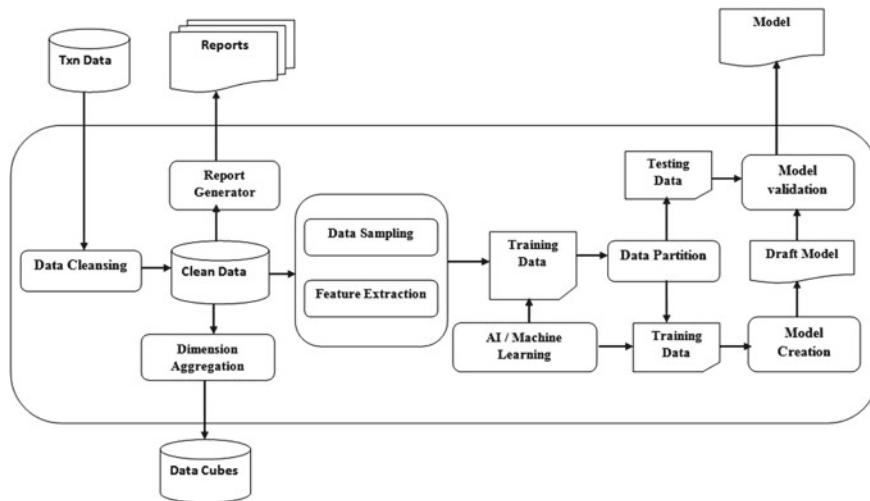


Fig. 2 The operational model of intelligence big data analytics system



Fig. 3 A health care framework utilizing intelligent big data analytics

A healthcare clinical systems are a human-made framework, in development biomedical equipment, the change in framework improved its performance with less time to taken for action in a better action [5]. In advance of medical imaging in the presence of smart and digital devices feel it the presence of it, the application of computing algorithm to understand and solving capability make it more effective, the nonlinear nature to understand heath related AI/ML gives a good platform to it. Like ANN, CNN, BAT algorithm, the PSO algorithm is used in digital imaging in the X-ray domain [11, 12]. A similar application is can be discussed in the case of TB, cancer in part of the human body, and brain related, where the data numbers are huge and data acquisition systems are complex by nature. Here, the AI or ML makes a significance help to analyzed as they can handle the nonlinear region as compared to a linear solution [13, 14].

2.1 The Components of Intelligent Big Data Analytics

Intelligent (Big) Database: The healthcare framework the input parameter, i.e., the data accusation process should be error-free process, it is the main parameter for the clinical process, and these data and their periodic analysis help to achieve to cure or treat a human being. These inputs initially handle by humans, then some biomedical instruments in analog nature, then digital, now programmable digital process. So that database changes its character according to input provided to it [1]. With the intelligent big data definition, the database has to store properly analyzed data with clinical details, and the unnecessary data has to lean up as shown in Fig. 2.

Intelligent Method: Artificial intelligence, machine learning, and deep learning: The intelligent in augmented analytics or intelligent big data analytics depends upon the

Fig. 4 The intelligent method used in big data analytics



process deploy to find a smarter output in the decision-making process; here, the intelligent is the degree of correctness of the decision by using a nonlinear process or algorithm: Artificial intelligence, machine learning, and deep learning [8, 12, 15]. Here, the developer should have properly understood the problem to identify and deploy these nature-inspired algorithms for nonlinear cases for solution (Fig. 4).

Artificial intelligence defined as a particular or goal orient solution for a task that exhibits nonlinear nature. Artificial intelligence is an algorithm that can learn, understand, and perform. Machine learning helps to build a system that can automatically learn and improve from experiences. The machine learning algorithms are trained with a huge number of data, and they are classified into three types supervised, unsupervised, and reinforcement learning. Machine learning method which trains and deploys with the human brain filters information method known as deep learning, which leads to a computing model to filter input data through layers for prediction and classification, an intelligent way of decision making. Apart from the AI/ML/DL and intelligent big database, intelligent analytics supported by mathematics and statistics methods as descriptive and predictive [12, 16, 17], their computing technology, web internet/data network technology, cloud technology. This leads to 50% of IT organization will be to use to improve application quality and deliver speed [7] is a key factor in the clinical framework.

Intelligent Big Data Analytics as a SaaS model (System and as Service): Intelligent data analytics can provide high-end solutions to a medical or clinical framework for complex serious diseases [7]. The clinical investigation involves expert examiners, researchers, prescient modelers, and analysts to break into the proper share of volume of data for organized exchange information for operation, in addition to this different co-related information are left out in a regular practice. Intelligent systems are created by integrating with the principles, methodologies, methods, and procedures of the intelligent Method (artificial intelligence, machine learning, and deep learning) for real-time clinical world problem solving [11]. This intelligent system is a system which should imitate, augment, and automate intelligent behaviors of human expert and solution with generating representations, and learning strategies [11] (Figs. 5 and 6).

Fig. 5 Intelligent big data analytics key benefits

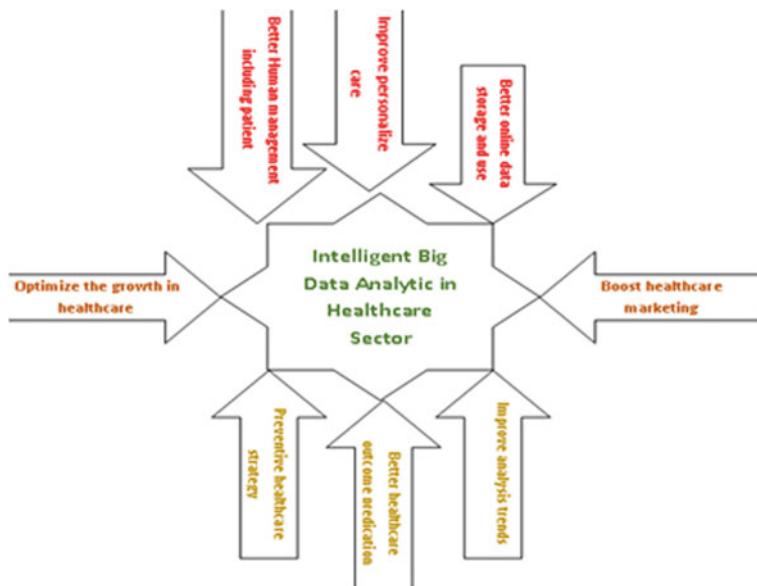
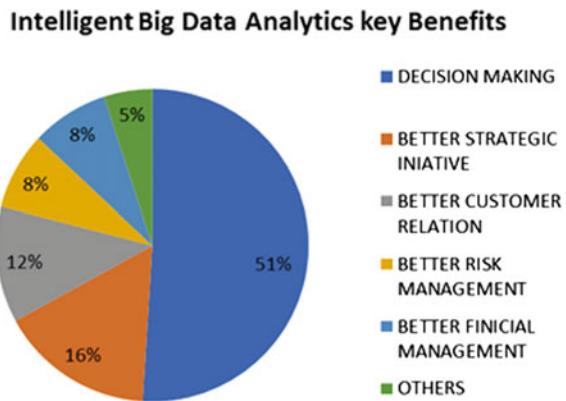


Fig. 6 Intelligent healthcare using analytics in big data platform

3 Medical Image Processing with Intelligent Analytics

With the advance in digital imaging in clinical performance increases rapidly, it provided significant information on the health of a human and organ work notwithstanding distinguishing disease states. It is quite useful for organ depiction, distinguishing tumors in lungs, spinal disfigurement finding, cancer detection, distinguishing tumors in the brain, brain injury in an accident, and so forth. Magnetic resonance imaging (MRI) or computed tomography (CT) is commonly used in routine

clinical diagnosis and treatment. AI or machine learning has a great revolution in medical research, for example, detect mitosis in breast cancer cells, predict the toxicity of new drugs, and understand gene mutation to prevent disease, computer-aided diagnosis (CADs) and so on [4, 5]. However, most of AI or machine model is only used in natural image domain but no other domains, especially only a really small part is used in the medical image domain. Magnetic resonance imaging (MRI) is widely used in routine clinical diagnosis and treatments. In Japan, brain dock is being conducted for medical checkups. We not only save a lot of doctors ‘diagnose time, improves clinics’ working efficiency, solve the insufficient health care resource problem, but also get a conclusion that there are some relationships between the MRI images [12, 14, 15]. The proposed methods can be extended for other medical imaging tasks. The proposed method and idea of the transfer learning for computer-aided diagnosis can be useful to make development in the medical diagnosis domain.

4 Intelligent Database for Healthcare Framework

As we discuss the need and intelligent system, here we take a 3-dimensional medical imaging for MRI image, here we consider an MRI data set from www.mathwork.com and simulated, generation and display of the Coronal Slices: Construction coronal slices almost similar to constructing sagittal slices. With the clinical standard, generation of 45 frames, starting 8 planes in with back to front moving into consideration, by ignoring other frames with 4-D array feature. This helps to diagnosis the brain leads to faster treatment (Figs. 7, 8, 9, 10, 11, and 12).

Simulation results as follow:

5 Conclusion

In this intelligent big data analytics study in the healthcare system, we discuss the importance of the integration of big data definitions for the medical healthcare sector and its utility, where medical imaging is one of them. The development of computing environment to make intelligence by using artificial intelligence, machine learning, and deep learning, for intelligent big database and their utility. It is quite important and critical for an intelligent data analytic process to use genuine data for proper assumption toward the illness. Such framework takes a smart choice to help the system as shown in Fig. 6.

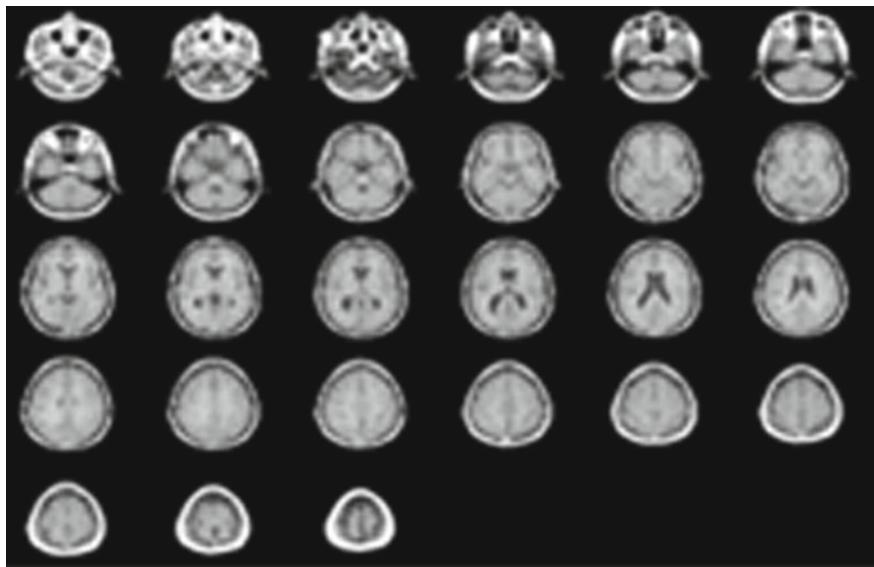


Fig. 7 The horizontal MRI data from (MRIDATA set)

Fig. 8 Raw sagittal slice
from horizontal slices

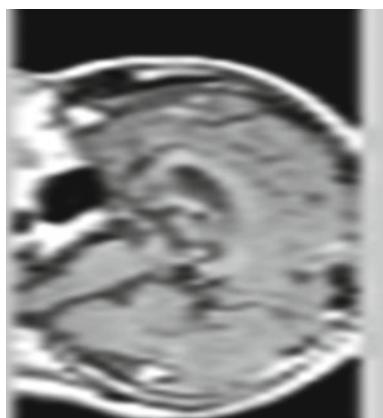


Fig. 9 Extraction sagittal
slice by using image
transformation, where
horizontal slices as inputs

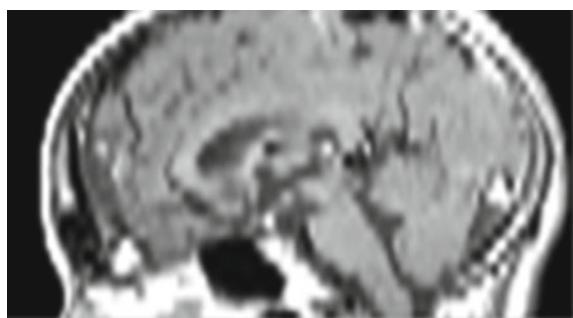


Fig. 10 Sagittal slice conversion from 3D input to 2D input

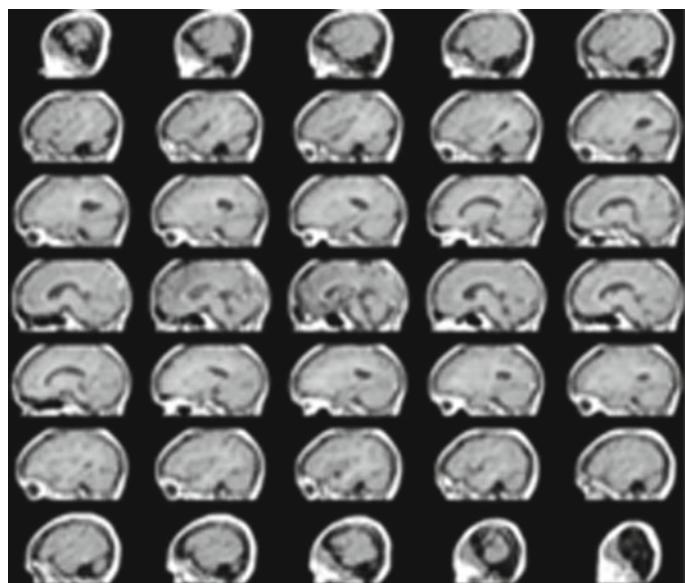
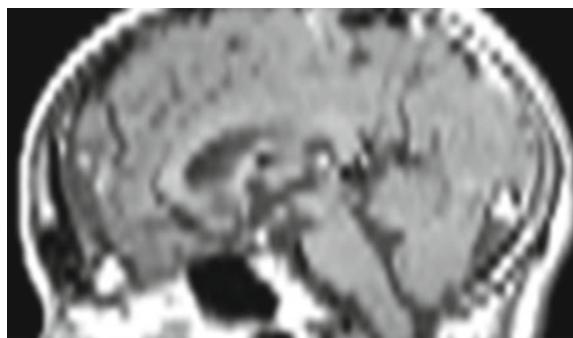


Fig. 11 Create and display sagittal slices (4-D array)

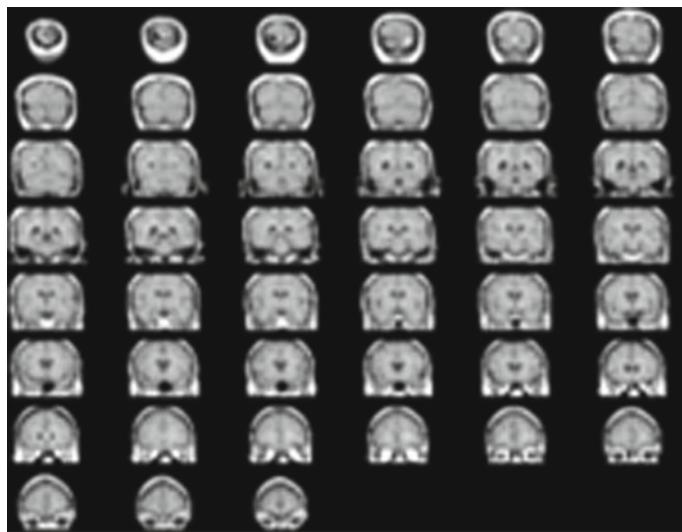


Fig. 12 Create and display coronal slices

References

1. McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D (2012) Huge data: the executives transformation. *Harvard Bus Rev* 90(10):60–68
2. Maseleno A, Huda M, Siregar M, Ahmad R, Hehsan A, Haron Z, Jasmi KA (2017) Combining the previous measure of evidence to educational entrance examination. *J Artif Intell* 10(3):85–90
3. Davis CK (2014) Viewpoint beyond data and analytics—why business analytics and big data really matter for modern business organizations. *CACM* 57(8):39–41
4. Maseleno A, Sabani N, Huda M, Ahmad R, Jasmi KA, Basiron B (2018) Demystifying learning analytics in personalized learning. *Int J Eng Technol* 7(3):1124–1129
5. Huda M, Maseleno A, Atmotiyoso P, Siregar M, Ahmad R, Jasmi K, Muhamad N (2018) Big data emerging technology: insights into innovative environment for online learning resources. *Int J Emerg Technol Learn (iJET)* 13(1):23–36
6. Al-Jarra O, Yoo PD, Muahidat S, Karagiannis GK (2015) Efficient machine learning for big data: a review. *Big Data Res* 2:87–93
7. Huda M, Maseleno A, Shahrill M, Jasmi KA, Mustari I, Basiron B (2017) Exploring adaptive teaching competencies in big data era. *Int J Emerg Technol Learn* 12(3)
8. Lynch C (2008) Huge information: how do your information develop? *Nature* 455(7209):28–29
9. Maseleno A, Hasan MM, Tuah N (2015) Combining fuzzy logic and Dempster-Shafer theory. *Indonesian J Electr Eng Comput Sci* 16(3):583–590
10. Sun Z, Wang P (2017) Big data, analytics and intelligence: an editorial perspective. *J Netw Math Nat Comput* 13(2):75–81
11. Moutinho L, Rita P, Li S (2006) Strategic diagnostics and management decision making: a hybrid knowledge-based approach. *Intell Sys Acc Fin Mgmt* 14:129–155
12. Russell S, Norvig P (2010) Artificial intelligence: a modern approach, 3rd edn. Prentice Hall
13. Schalkoff RJ (2011) Intelligent systems: principles, paradigms, and pragmatics. Jones and Bartlett Publishers, Boston
14. Sun Z, Finnie G (2004) Intelligent techniques in ecommerce: a case-based reasoning perspective. Springer, Heidelberg

15. Huda M, Maseleno A, Teh KSM, Don AG, Basiron B, Jasmi KA, et al (2018) Understanding modern learning environment (MLE) in big data era. *Int J Emerg Technol Learn* 13(5)
16. Sugiyarti E, Jasmi KA, Basiron B, Huda M, Shankar K, Maseleno A (2018) Decision support system of scholarship grantee selection using data mining. *Int J Pure Appl Math* 119(15):2239–2249
17. Kantardzic M (2011) Data mining: concepts, models, methods, and algorithms. Wiley & IEEE Press, Hoboken

Biomedical Data Classification Using Meta-learning: An Experimental Investigation



Tapas Ranjan Baitharu, P. K. Bharti, and Subhendu Kumar Pani

Abstract Supervised machine learning is a vast field, and meta-learning is one of the sub parts of that which is applied on data of previous years. Numerous data mining problems and techniques are available to take out unseen information from huge databases. Different input data are classified into target classes through classification assignment. We are using various machine learning algorithms in our model. Our overall classification contributed by output of base classifier which are under meta-learning. There are many challenges in the real world problem can be explained by combining machine learning algorithm. Meta-learning algorithms like bagging, dabbing, rotation forest, filtered classifier and decorate are widely used. Meta-learning is based on the principle from preceding experience and learns to enhance data mining work. In our paper, we investigate the performance of bagging, logitboost and filtered classifiers on a biomedical dataset by some popular learning algorithms for meta-data approaches.

Keywords Meta-learning · Meta-classifier · Classifications · Bagging · Logitboost and filtered classifiers

1 Introduction

Data mining uses the different techniques to construct the forecastation models through hidden patterns of the multi sourced databases. Particularly, healthcare system generates huge amount of data in high-velocity manner like gene expression data and medical imaging data. To analyze such huge velocity, data is very problematic through traditional techniques which may not predict correctly. This may lead the data mining become more and more imperative. Many associations generate the

T. R. Baitharu · P. K. Bharti
SVU, Gajraula, Uttar Pradesh, India
e-mail: vc@svu.edu.in

S. K. Pani (✉)
Krupajal Computer Academy, BPUT, Rourkela, Odisha, India

huge enormous data from its regular business operations and collected in modern data repositories frequently. Data mining field has been raised to address such demands. It merges database techniques and statistical approaches to find out unseen patterns in data and extract potentially helpful information for the development of business in competitive environment. Data mining techniques (like classification, clustering or association) have wide-ranging areas such as consumer behavior, market segmentation, banking, credit rating, fraud detection, crime analysis, genetic analysis, diseases prediction, traffic pattern detection in computer network, email classification, image and pattern recognition, etc.

These techniques were effectively applied on various healthcare tasks like disease prediction, prediction of cancer, prediction of heart attacks, etc. In biomedical, data mining that look for phenotypic data to construct the predictive models of disease detection and remedial response. In meta-learning, the different base learners will be merged and links their features to get together meta-knowledge which will be helpful to predict the accuracy of the ensemble learner [1, 2]. Generally, meta-learning targets at effective performance of the predictive system through initial experience. It is already proved that a meta learner performs better accuracy of traditional base learners [3, 4]. Field of machine learning uses, various meta-learning techniques such as voting, boosting, bagging and stacking.

Analyzing the effectiveness of a classification algorithm is significant to apply in practical problems and has considered as an exciting matter of research for machine learning and data mining society. Prediction done by the classification method in the classifier relates to the features present in the dataset. These characteristics also linked to this framework using different datasets. In this article, we study this part of meta-learning and account how heterogeneous dataset features have an effect on the classification performance of meta-learning techniques [5].

2 Methodology

On the following subsection, we investigated the data by obeying a methodical approach.

2.1 *Description of Datasets*

The dataset we are using has taken from UCI Repository [6]. The features associated with these selected samples are reflected with mixed statistical characteristics applied on various application domains. The characteristics of the datasets are summarized in Table 1.

Table 1 Dataset summarization

Name of the dataset	#No. of instances	#No. of classes	#No. of attributes	#Nominal/continuous	#Missing values (%)
Breast-Cancer	286	2	9	9/0	9 (0.35%)
Hepatitis	155	7	20	7/2	6 (0.15%)
Heart	294	8	14	7/6	6 (0.15%)

2.2 *Meta-learning System for Biomedical Data*

We have selected three meta classification methods in our study. They are of bagging, logitboost and filtered classifiers that integrate different base classifiers in its evaluation procedure [7–9]. The detail description of all used meta classifiers in this experiment is described here.

Bagging: Bagging or Bootstrap Aggregation is an easy and very prevailing ensemble technique. The role of ensemble technique is to construct a predictive model to forecast most accurately based on the predictions of the individual model which combines many machine learning algorithms. The major role of bagging is to reduce the variance of the high variance algorithm. So to reduce the variance, it uses Bootstrap procedure to machine learning methods. Bagging is used generally uses in the different decision trees. Once it is implemented, there is very less chance of over fitting of individual trees in the training data [10]. Due to this reason, the individual decision trees are developed deep and not able to be pruned. These grown trees contain both low bias and high variance. It is considered to be an important characterize of sub-models for forecastation using bagging. Bagging decision trees formations are based on the number of samples in its parameter. Samples can be selected on run. Once it is implemented the accuracy commences to stop producing improvement. In this case, it will take a long time to construct for very large number of models but never over fit in training data [11, 12].

LogitBoost: It comes under boosting category algorithm implemented by Jerome Friedman, Trevor Hastie and Robert Tibshirani. It is very effective category in machine learning and computational classifying theory. This algorithm is enhancement of AdaBoost algorithm into a new ensemble framework. It combines the cost methods of logistic regression with AdaBoost predictive model [13].

Filtered Classifier: several times, we execute a filter before apply of a classifier. The filters are essential for removing, transforming, minimizing misclassified instances and adding attributes. In our experiment, Weka's Filter class is used and then execute a number of filtering approached with the different class methods. We have executed filtered classifier class of the Weka library which is essential for implementing an arbitrary learner. Class for running an arbitrary classifier on data that has been passed through an arbitrary filter [14]. Arbitrary classifier requires an arbitrary filter for its implementation on training data.

Table 2 Bagging properties with REP decision tree base classifier

Properties	Description	Assigned values
Seeding	It uses the random number	01
representCopiesUsingWeights	It may not represents explicitly but may represents copies of instance using weights	False
storeOutOfBagPredictions	The required quantities of execution slots at the time of constructing the ensemble	02
numExecutionSlots	Each of bag size represent rep. similar to training set size percent	01
Percentage of bagSize	Each bag size will rep. as the training set percentage size	100
Count of numDecimalPlaces	It uses the number of decimal places as its output numbers in the predictive model	02
Number of batchSize	The chosen number of instances is used in prediction results	100
printClassifiers	Individual learner will be shown in the output	
numIterations execution	It needs iterations count at execution	10
Debug checking	Classifier may produce extra info to the console If it will be true	False
outputOutOfBagComplexityStatistics	It will get complexity-based statistics result if out-of-bag evaluation is executed	False

3 Experiment Design

We have used Weka¹ tool to perform the experiment [15]. For better accuracy purpose, we have taken e tenfold cross validation as the test mode. This procedure provides robustness to the classification. We have kept all three meta-learning algorithms classification accuracy on a dataset one after another.

We have set up the parameters according to the properties of three meta learners. First meta learner bagging uses REP decision tree base classifier. Second meta learner LogitBoost uses decision stump decision tree base classifier, and third meta learner filtered classifiers is internally implements the base classifier as J48 decision tree. Multivalued attribute based properties of discretize filters are also applied in our experiment. In this experiment, it uses discretize filter with multiple attributes. Properties of the bagging, LogitBoost and filtered classifier are used. Those properties have shown in Tables 2, 3 and 4. A snapshot of filter classifier tree view is shown in Fig. 1.

¹ www.cs.waikato.ac.nz/ml/weka/.

Table 3 LogitBoost properties with decision stump decision tree base classifier

Properties	Description	Assigned values
ZMax	Threshold for responses	3.0
numThreads	The number of threads to apply for batch prediction, that should be \Rightarrow size of thread pool	01
weightThreshold	Weight pruning of weight threshold (reduce to 90 for speeding up learning process)	100
Resume	Whether classifier can continue training after performing the requested number of iterations	False
useEstimatedPriors	Whether estimated priors are used rather than uniform ones	False
poolSize	The amount of the thread pool, for example, the quantity of cores in the CPU	02
useResampling	Whether resampling is used instead of reweighting	False
Shrinkage	Shrinkage limitation (use small value like 0.1 to lessen overfitting)	1.0
likelihoodThreshold	Threshold on improvement in likelihood	10

Table 4 Filtered classifiers properties with J48 decision tree base classifier

Properties	Description	Assigned values
Seeding	It uses the random number	01
Number of batchSize	The chosen number of instances is used in prediction results	100
Not to CheckCapabilities	If fixed, classifier ability is not tested before classifier is constructed (apply with care to decrease runtime)	False
Debug checking	If fix to true, classifier may produce extra info to the console	False
outputOutOfBagComplexityStatistics	It will get complexity-based statistics result if out-of-bag evaluation is executed	False

3.1 Performance Measure

We apply various metrics for evaluating the meta classifiers' predictive results in our experimentation. These are discussed below:

- Confusion Matrix: Predictions are obtained from the columns of the confusion matrix, and the actual class is represented by the rows. Accurate predictions always depend on the matrix diagonal. The common structure of confusion matrix is described below is

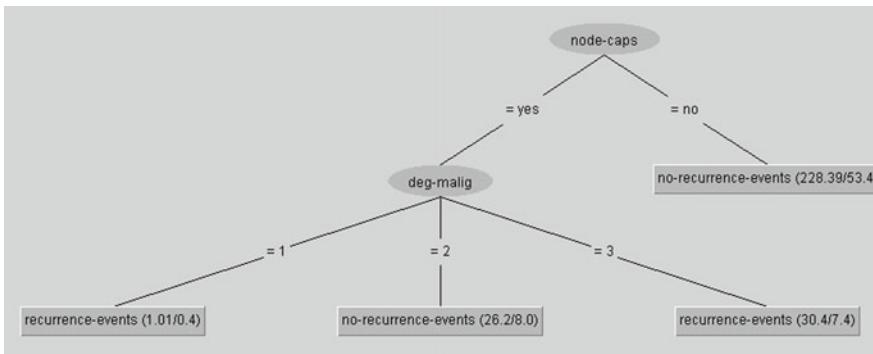


Fig. 1 Filter classifier tree view

TP	FN
FP	TN

in which, True Positives (TP) points to minority instance count that were accurately predicted, True Negatives (TN) points to the majority instance count that were accurately predicted. False Positives (FP) points to the majority instance count they were inaccurately predicted as minority class instances and False Negatives (FN) points to minority instance count that were inaccurately predicted as majority class instances.

Even if the confusion matrix presents an improved outlook on the classifier's performance comparing with accuracy, a more thorough investigation is preferable that are given by the more metrics.

- Recall: here classifiers could be recognized correct how much is shown in recall metric. $(TP + FN)$ corresponds to entire all minority members. Recall is shown below

$$\text{Recall} = \frac{TP}{TP + FN}$$

- Precision: This shows the percent of minority class instance comes under minority class through the classifier in the form of metric. $(TP + FP)$ signifies the total of positive forecastations by the learner.

Precision is specified by

$$\text{Precision} = \frac{TP}{TP + FP}$$

Overall it is understood that recall is a totality Measure and Precision is a meticulousness Measure. The value 1 of both recall and precision signifies that it is the ideal classifier but if the value changes nearer to one then it is very difficult to select the appropriate classifier. Those type cases different metrics as considered further are recommended in the literature.

- F-Measure: It consider as harmonic mean for both Precision and Recall. We consider it as amiddling between the two percentages. It truly shortens the evaluation among the classifiers. It is denoted by

$$F - \text{Measure} = 2/(1/\text{Recall} + 1/\text{Precision})$$

4 Result Analysis

We analyze the performance of mchosn meta classifiers in our experiment by considering Recall, Precision and F-Measure to quantitatively assess the classifiers [16, 17]. Various performance evaluations of the different classifiers in three datasets are described in Tables 5, 6, 7 and 8.

A comparative Meta classifiers performance with different datasets is displayed in Fig. 2.

The meta classifiers performance is not standardized across the datasets which is evident from Fig. 2. The predicted results show that filtered classifier and bagging produce the best accuracy in Hepatitis dataset. Similarly, the performance of filtered classifier and bagging is the second accuracy in heart dataset. Surprisingly, bagging which performs the least in breast cancer dataset. It is evaluated that filtered classifier performs best in both the datasets.

Table 5 Classifier performance of breast cancer dataset

Name of the classifier	Generated confusion matrix			Average precision	Average recall	Average F-mean	Accuracy (%)	Time taken (s)
Bagging	a	b	\leftarrow Classified as	0.641	0.692	0.639	69.23	0.05
	184	17	$ a = 1$					
	71	14	$ b = 2$					
LogitBoost	a	b	\leftarrow Classified as	0.702	0.724	0.705	72.37	0.04
	176	25	$ a = 1$					
	54	31	$ b = 2$					
Filtered Classifier	a	b	\leftarrow Classified as	752	755	713	75.52	0.07
	193	8	$ a = 1$					
	62	23	$ b = 2$					

Table 6 Classifier performance of hepatitis dataset

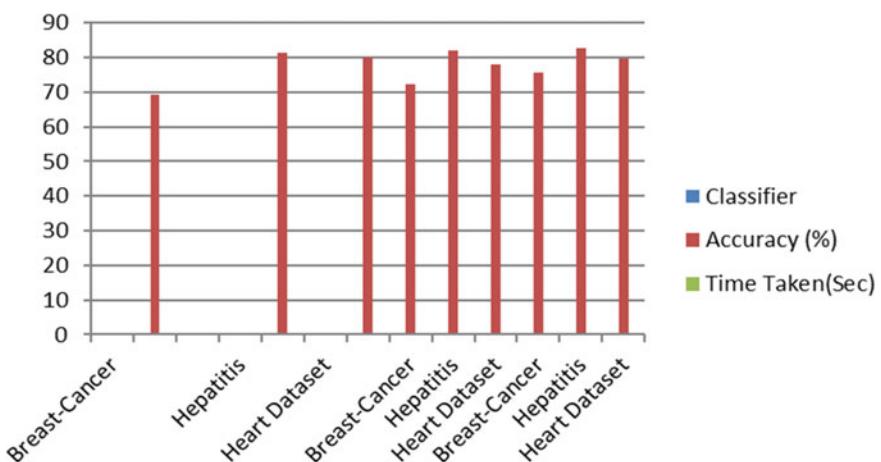
Name of the classifier	Generated confusion matrix			Average precision	Average recall	Average F-mean	Accuracy (%)	Time taken (s)
Bagging	a	b	← Classified as	0.786	0.813	0.780	81.2903	0.04
	8	24	$a = 1$					
	5	118	$b = 2$					
LogitBoost	a	b	← Classified as	0.808	0.819	0.812	81.93	0.02
	15	17	$a = 1$					
	11	112	$b = 2$					
Filtered Classifier	a	b	← Classified as	0.806	0.826	0.804	82.58	0.01
	11	21	$a = 1$					
	6	117	$b = 2$					

Table 7 Classifier performance of heart dataset

Name of the classifier	Generated confusion matrix			Average precision	Average recall	Average F-mean	Accuracy (%)	Time taken (s)
Bagging	a	b	← classified as	0.797	0.799	0.798	79.932	0.09
	162	26	$a = 1$					
	33	73	$b = 2$					
LogitBoost	a	b	← classified as	0.776	0.779	0.777	77.89	0.10
	160	28	$a = 1$					
	37	69	$b = 2$					
FilteredClassifier	a	b	← classified as	0.794	0.796	0.795	79.5918	0.01
	160	28	$a = 1$					
	32	34	$b = 2$					

Table 8 Performance of accuracy and time

Dataset	Classifier	Accuracy (%)	Time taken (s)
Breast-cancer	Bagging	69.23	0.05
Hepatitis		81.2903	0.04
Heart dataset		79.932	0.09
Breast-cancer		72.37	0.04
Hepatitis	LogitBoost	81.93	0.02
Heart dataset		77.89	0.1
Breast-cancer		75.52	0.07
Hepatitis	Filtered classifier	82.58	0.01
Heart dataset		79.5918	0.01

**Fig. 2** Performance of the meta classifiers

5 Conclusion and Future Directions

In our study, we have used three meta classifiers to evaluate their performance among multiple biomedical datasets to solve the formulated research problem. The performance shows that accuracy of the classification depends on the features of a dataset. Model formation by getting the interesting patterns discovers helpful knowledge acquired from huge data repositories. After analyzing the taken data produced by our model experiment, we concluded that there is no prove of consistency by classifiers in their predictive performances. Our work analyzes and asses the accuracy of the performance of three different meta classification algorithms such as bagging, LogitBoost and filtered classifier. Additionally, their result is more or less competitive displaying small difference. The experimental outputs illustrate that the maximum accuracy is established in filtered classifier with 82% in hepatitis dataset. Even though

the presented experiment used some practical algorithms of classification somehow, there is necessary to incorporate more types of algorithm related to art in future learning like gradient boosting machines, XGBOOST and LightGBM.

References

1. Abe H, Yamaguchi T (2002) Constructing inductive applications by meta-learning with method repositories. In: Progress in discovery science, Final report of the Japanese Discovery Science Project, pp 576–585. Springer, Berlin
2. Leyva E, Caises Y, González A, Pérez R (2014) On the use of meta learning for instance selection: an architecture and an experimental study. *Inf Sci* 266:16–30 (2014)
3. Lemke C, Gabrys B (2010) Meta-learning for time series forecasting and forecast combination. *Neurocomputing* 73:2006–2016 (Jun 2010)
4. Prudêncio RB, Ludermir TB (2004) Meta-learning approaches to selecting time series models. *Neurocomputing* 61:121–137 (2004)
5. Sun et al (2018) Meta-analysis of clinical trials. In: Principles and practice of clinical research, pp 317–327 (2018)
6. Dataset UH, UCI Machine Learning Repository [online]. <https://archive.ics.uci.edu/ml/machine-learning-databases/heartdisease/Heart>
7. Maudsley DB (1979) A theory of meta-learning and principles of facilitation: an organismic perspective. University of Toronto, vol 40, no 8, pp 4354–4355
8. RochaNeto AR, Barreto GA (2009) On the application of ensembles of classifiers to the diagnosis of pathologies of the vertebral column: a comparative analysis. *IEEE Lat Am Trans* 7(4):487–496
9. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the symposium on computer applications and medical care, pp 261–265
10. Dai W, Brisimi TS, Adams WG, Mela T, Saligrama V, Paschalidis IC (2015) Prediction of hospitalization due to heart diseases by supervised learning methods. *Int J Med Inform* 84(3):189–197
11. Parthiban G, Srivatsa SK (2012) Applying machine learning methods in diagnosing heart disease for diabetic patients. *Int J Appl Inform Syst* 3(7):2249–2868
12. Chaurasia V, Pal S (2014) Data mining approach to detect heart diseases. *Int J Adv Comput Sci Inform Technol* 2(4):56–66
13. Elter M, Schulz-Wendtland R, Wittenberg T (2007) The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Med Phys* 34(11):4164–4172
14. William HW, Mangasarian OL (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc Natl Acad Sci* 87:9193–9196
15. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. *SIGKDD Explor News* 11(1):10–18
16. Powers DM (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol* 2(1):37–63
17. Nithya R, Manikandan P, Ramyachitra D (2015) Performance analysis of meta classifiers algorithms using yeast dataset. *Int J Innov Res Comp Eng* 3(9):8062–8068

Computer Fraud and Security Data Analysis

Advanced Blockchain Security for Medical Data



G. S. R. P. Prasanth, G. V. Harish, and B. Bharathi

Abstract This project presents an efficient data security system for Medical Reports using BlockChain Technology. The exactness of conclusion and the viability of treatment can be improved by this proposed structure of social insurance frameworks dependent on the frameworks plus computing results plus parallel execution (ACP) approach. ACP approach is utilized to show and speak to patient's conditions, analysis, and treatment process, at that point break down and assess. Likewise, the developing square chain innovation with the human services framework makes simple the connection between the specialist and cloud, Doctor and patient and Patient and cloud for thorough social insurance information sharing, medicinal records survey, and care suitability. A frontend system is created utilizing Django structure for the patient module where the patient can enter the information for examining the sickness nearness without visiting the specialist. On the server-side, the information is gathered and decoded utilizing effective square chain innovation and a report is likewise produced for specialist confirmation which he can see utilizing his login made utilizing Django structure.

Keywords Healthcare · Blockchain · Dengue · Cloud · Encryption

1 Introduction

Medicinal consideration has been a fundamental piece of our lives thus the therapeutic information, for instance, remedies, past restorative records has likewise become an imperative part for patient's conclusion and for additional procedures. Customarily, therapeutic information was recorded on paper, which was inclined to

G. S. R. P. Prasanth (✉) · G. V. Harish · B. Bharathi

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

B. Bharathi

e-mail: bharathi.cse@sathyabama.ac.in

get harmed and changed. Along these lines, it was important to safeguard the information electronically. In any case, the restorative database could be altered or erased for all time.

At that point, there was likewise a worry on data leaking. Data leakage happens when an individual gets the patient's information knowingly or unknowingly without the acknowledgement of the patient this is called as data leakage. We need to give more security for the data and it should be hard for the person to access the information. Innovation consistently assumes a huge job on the off chance that it is tied in with upgrading the quality or about settling issues, for example, asset assignment alongside data hindering, here in medical-care information sharing innovation should have been advanced with time.

By and large, patients may have a great deal of specialist organizations as far as therapeutic human services that incorporate general doctors or experts or even advisors. Since an infection could be a result of the past ailment, so they all need to share wellbeing record safely with no control. Quiet need not be constantly an expert or to have a decent memory to recall every one of the information appropriately if every one of the information is put away and shared safely.

Patients need to continue refreshing their own medicinal information history. As indicated by the Fundamental Right to Life and Liberty under Article 21 and opportunity of articulation and development under Articles 19(1)(a) and (b) of the Constitution of India, a patient has right to counsel or get moved to another emergency clinic for his treatment. Presently, again it is patients' desire to share his information. Additionally, if an emergency clinic needs to share his information for look into there must be assent from the patient. Once more, in the event that assent is likewise there, at that point the information moving procedure takes a great deal of time. Additionally, if the information that is moved is in paper mode or even through email, there are time, speed, stockpiling, and security issues.

The storage of large databases always poses a security threat and cyber-attacks. Hackers try to recover the electronic health records and use it inappropriately. The patient's information stored is very delicate and should be safeguarded from external attacks.

One can likewise not depend upon a unified database in light of the fact that for all intents and purposes unique access controls for various clients, looking through technique over an encoded channel, huge memory for medicinal information stockpiling and so forth is troublesome.

Likewise, Al Omar et al. discuss about different issues that happen when information is put away in a scrambled configuration. Access to an appropriated record for sharing of restorative information in a straightforward way can ensure information security as bitcoin was believed to be verified for exchange. A patient reason driven usage model has to be designed and deployed because it involves therapeutic information of the patients. The patient has to be deliberated information in detail so that he/she clearly understands the amount of risk involved in sharing the information to the insurance agency. The patient should have liberty in deciding the receiver of information like blood donation centre or insurance agency, etc.

This innovation could protect information and subsequently ensure unwavering quality. What's more, if this innovation is utilized alongside distributed computing innovation, issues identified with capacity can likewise be evaporated, in light of the fact that cloud is trusted for putting away and overseeing information. The concept of blockchain has major application to update and reduce security concerns to the data stored on cloud. To be sure, restorative information sharing and putting away with blockchain-based cloud can address a great deal of issues of therapeutic information. Restorative thought has been an indispensable bit of our lives hence the helpful data, for example, arrangements, past remedial records has in like manner become a basic part for patients' finding and extra strategies.

Generally, the details of the patient are recorded on paper and then later updated on virtual records. Virtual Health Records (VHR) is maintained for permanent data storage and maintenance. Whenever data gets restored there is always a problem of data loss and denial of service. Information blocking happens when a substance, for example, an individual may be with or without his expect to find a workable pace that should not have been seen without patients or clinical centres concern.

Advancement reliably expect a very important activity if it is connected to improving the quality or about settling issues, for instance, resource dissemination close by information impeding, here in medical-care data sharing advancement ought to have been created with time. This endeavour has assessed every conceivable work on the remedial human administrations using blockchain development with a suitable close to examination.

The SHA-256 hashing figuring is a basic bit of the bitcoin show. It has seen execution in moving highlights of the advancement, for instance, bitcoin mining, Merkle trees and the creation of bitcoin addresses. Moreover, this can be used with any kind of therapeutic data security.

2 Related Work

The increasing volumes of social information partook in online interpersonal organizations, the foundation of dependable social connections over these stages, and the rise of innovations that permit fellowship systems to be induced from information traded in correspondence systems have persuaded scientists to assemble socially mindful confirmation plans. We direct the main examination that studies the writing identified with social verification. Right now, not just made a scientific classification for arranging all social validation plans sent in on the web or physical social settings and broadly examined their verification highlights, yet also fabricated a unique structure to assess the adequacy of all the social confirmation plans, recognized all the pragmatic and hypothetical assaults which might be put against such plans, tended to conceivable guard methodologies, and distinguished difficulties, open inquiries, and future research openings [1].

To improve the exactness of finding and the adequacy of treatment, a structure of equal medicinal services frameworks dependent on the fake frameworks and computational examinations and equal execution approach is proposed right now. PHS utilizes fake human services frameworks to demonstrate and speak to patients' conditions, conclusion, and treatment process, at that point, applies computational examinations to break down and assess different restorative regimens and actualizes equal execution for dynamic help and continuous advancement in both real and fake social insurance forms. Furthermore, we consolidate the developing blockchain innovation with PHS, through building a consortium blockchain connecting patients, emergency clinics, wellbeing authorities, and medicinal services networks for far-reaching social insurance information sharing, clinical records survey, and care suitability [2].

One specific pattern saw in medicinal services is the dynamic movement of information and administrations to the cloud, incompletely because of accommodation and reserve funds. There are, in any case, constraints to utilizing ordinary cryptographic natives and access control models to solve security and protection worries in undeniably virtual storage condition. Right now, study the possibility to utilize the blockchain innovation to ensure human services information is facilitated inside the cloud. We likewise depict the useful difficulties of such a suggestion and further research that is required [3].

Dengue is a significant overall arthropod-borne maladies. Dengue phenotypes depend on laboratorial and clinical tests that are off base. Objective: This paper presented an AI approach for the forecast of dengue fever seriousness dependent on human genome information [4].

Blockchain is a decentralized, trust less convention that joins straightforwardness, changelessness, and accord properties to empower secure, pseudo- unknown exchanges. Shrewd agreements are worked on a blockchain to help on-chain stockpiling and empower Decentralized Apps to interface with the blockchain programmatically. Programmable blockchains have created enthusiasm for the social insurance space as a significant answer to resolve key difficulties, for example, gapped interchanges, wasteful clinical report conveyance, and divided wellbeing records. This paper gives assessment measurements to survey blockchain-based DApps regarding their attainability, proposed ability, and consistency in the social insurance space [5].

Blockchain innovation has demonstrated its impressive flexibility as of late as an assortment of market parts looked for methods for fusing its capacities into their activities. While so far a large portion of the attention has been on the budgetary administration's industry, a few tasks in other help related territories, for example, human services show this is starting to change. Various beginning stages for blockchain innovation in the social insurance industry are the focal point of this report. With models for open human services the executives, client arranged clinical research and medication forging in the pharmaceutical area, this report intends to show potential impacts, objectives and possibilities associated with this troublesome innovation [6].

The ID of the persuasive clinical side effects and lab includes that help in the determination of dengue fever in the early period of the disease would help in structuring compelling general wellbeing the board and virological observation procedures. Keeping this as our primary target, we create right now a new computational knowledge-based strategy that predicts the determination progressively, limiting the number of bogus positives and bogus negatives [7].

The social insurance administrations industry is continually giving indications of progress and new opportunities. The transcendent necessities in the present medicinal services frameworks are to secure the patient's clinical report over potential hacks. Consequently, it is essential to have secure data that can simply affirm individuals can find a good pace clinical report. In this way, we have proposed blockchain innovation as a dispensed way to deal with award security in resulting to the clinical report of a patient [8].

Liu et al. [9] depict a novel methodology for execution of the propelled encryption standard calculation, which gives a fundamentally improved quality against first-request differential electromagnetic and force investigation with an insignificant extra overhead. Our technique depends on randomization in composite field number-crunching, which involves a low usage cost while doesn't modify the calculation, doesn't diminish the working recurrence, and keeps ideal similarity with the distributed standard [9].

The writing on side-channel investigation portrays various veiling plans intended to ensure square figures at the usage level. Such veiling plans commonly require the calculation of covered tables before the execution of an encryption work. Right now return to an assault which legitimately misuses this calculation to recoup all or a portion of the covers utilized. We show that safely executing concealing plans is just conceivable where one approaches a lot of arbitrary numbers [10].

Wellbeing Information is viewed as the most important data related to a person. Even though various reasonable approaches, rules, and consistency necessities are set up to defend wellbeing data, protection and security break stays key issues for electronic medicinal services frameworks. Right now centre of these issues and propose a security and protection model actualized in Methodist Environment for Translational and Outcomes Research. METEOR was created at Houston Methodist Hospital and comprises of two parts: the undertaking information distribution centre and a product insight and investigation (SIA) layer [11].

The ever-growing compromise of outstandingly varying enabled data making headways in clinical, biomedical and human administrations fields and the creating openness of data at the central territory that can be used requiring any relationship from pharmaceutical makers to clinical inclusion associations to crisis facilities have chiefly made social protection affiliations and all its sub-fragments in face of a flood of enormous data as at no other time experienced. While this data is being hailed as the best approach to improving prosperity results, increment significant encounters and cutting down costs, the security and assurance issues are overwhelming to such a degree, that social protection industry can't abuse it with its present resources. Directing and equipping the insightful power of enormous data, nevertheless, is basic to the accomplishment of every human help affiliation. It is correct now this paper

plans to show the forefront security and assurance issues in gigantic data as applied to restorative administrations industry and look at some available data security, data security, customers' finding a good pace strategy [12].

The chance to access on-request, unbounded calculation and capacity assets has progressively persuaded clients to move their wellbeing records from nearby server farms to the cloud condition. This change can lessen the expenses related with the administration of information sharing, correspondence overhead and improve Quality of Service. Preparing, putting away, facilitating and chronicling information identified with e-health frameworks without physical access and control can intensify verification and access control issues right now. Along these lines, persuading clients to move delicate clinical records to the cloud condition requires executing secure and solid validation and access control techniques to ensure the information. This paper proposes another data get to strategy that jam both validation and access control in cloud-based e-health frameworks. Our strategy depends on a zero-information convention joined with two-organize keyed access control. In each entrance demand, in light of the most extreme privileges of client, the base access is separated [13].

Non-Path Anonymous P2P convention is called Rumour Riding (RR). In RR convention, the first person shares key information and the figure content to various nearby nodes. The given key, figure content goes for arbitrary strolls independently in framework called gossip. The gossip consequently develops a mysterious way by way of irregular walks. The first person who initiates nor the last person who responds need not be worried about way structure and support. The talk riding convention utilizes irregular walk-based calculations which can be handily abused by uncooperative and malevolent hubs. Gossip riding convention concentrates just on mysterious looking and downloading on P2P frameworks. In any case, secrecy opens the entryway to potential abuses and misuses, misusing the P2P organize, allowing to assaults like answer assault. In the Rumour Riding convention, if the first person or who initiates goes about as malevolent hub, it might send counterfeit confirmation message to the recipient, resulting in system traffic and lessens the presentation of system. On the off chance that the responder goes about as a malevolent hub, at that point, there are conceivable outcomes for replay assault to happen [14].

This examination looks for is to explore the Peer to Peer organize topology and its custom conduct. The routine impact of node to node arrangement brings thinking about all the associated hubs as companions and these friends head-on with different friends to apply the undertakings appointed to every last one of them. The whole grasp of this system topology sticks on to the heterogeneous circle. Due to its complete system topology and inclination includes, the system is powerless against many rebuking assaults. Moreover, the ownership ascribed to the security issues of the system is relatively applied. Hence, this system topology is exposed to extensive threats [15].

3 Proposed System

3.1 Existing System

In the current framework, the paper displays the basic, productive, and reasonable plan of an our dengue occurrence expectation model consolidates support vector machine (SVM) calculation to gauge the exhibition of the model. No security of information. No web App. No earlier examinations of the sickness presence. Most of the preparation information is excess. No security gave about the information prepared. No application sort of thing for understanding, a medical clinic just as the specialist for a check.

Limitations

- Only the disease detection is taken care.
- No security for data transmitted over cloud.

3.2 Proposed System

Figure 1 describes about how the data is stored and encrypted using blockchain and how it is processed. At first, the patient details are obtained from a frontend application. Collecting the reports from the lab and encrypting the received data using AES and blockchain technology. All the data is uploaded in the cloud. When it is needed it is retrieved from the cloud. After retrieving the data is decrypted and collects the original reports. At the next step, the data is processed and validated.

Figure 1 shows the three modules in our project. The patient inserts the data collected from the test reports into the patient login. A frontend web application is developed in which the data is collected and before sending it to the cloud, we will encrypt the data using blockchain key to make the data secure. The Advanced Encryption Standard, or AES is a type of cypher that is being used in many organizations and companies to secure the sensitive data.

Figure 1 shows how the data travels from source to destination where it gets stored in the blockchain. There are lot of algorithms that we are going to implement to hide the original data from the hacker or the third person who is in need of the data in the middle of execution process. After going through all the process the data will be stored in the blockchain.

In hospital module, the data from the sensor is transmitted to the hospital. The patient data is analysed and a report is prepared and transferred to cloud. In the hospital side, we will receive the data and decrypt it using the AES decryption.

In doctor module, a separate application login is prepared for the doctor from which he can access the reports of the patients and analyse it. The AES decryption is being performed to decrypt the data transferred from the hospital module. Then, he

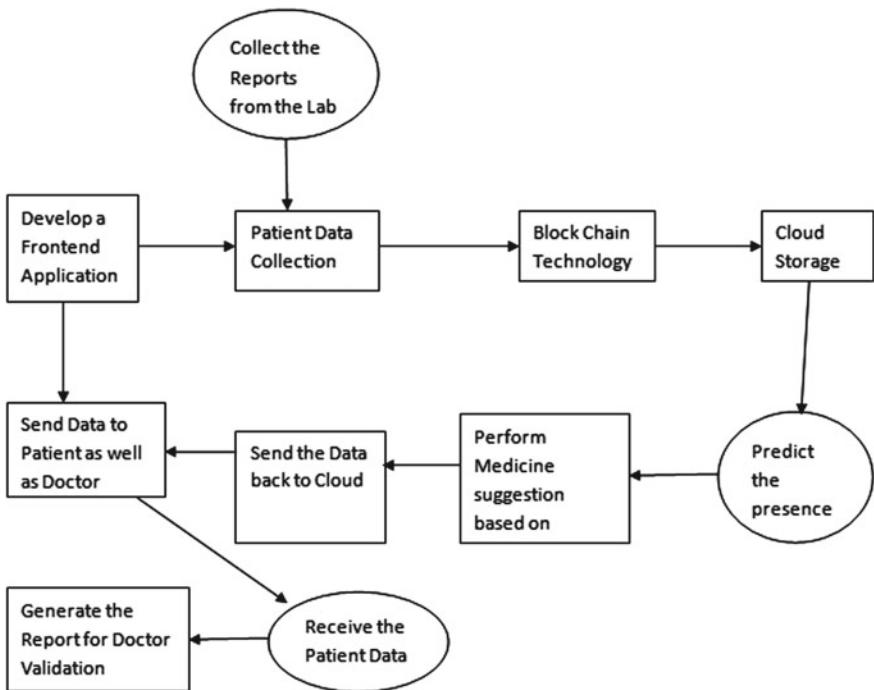


Fig. 1 System architecture diagram

can view using his login details about the patient condition and take necessary steps whenever needed.

In a proposed framework, a model for dengue discovery, controlling and treatment framework is manufactured and sent to check and show the viability and productivity of the square chain-fuelled social insurance structure. A web application is created to enter the wellbeing parameters to decide the infection nearness. Guarantees security to the information transmitted over the server to the specialist just as the patient. Programmed reportage for specialist check.

3.3 Advantages

- It allows secure data transmission over cloud using blockchain
- Automatic report generation of patient
- Web application for patient and doctor.

4 Conclusion

This project creates an efficient data security system in order to maintain the patient data, and this project creates an environment where we can store the data without getting it stolen or lost. We are using the blockchain technology for the efficient data security that this technology provides. The main purpose of using this blockchain technology in this project is to provide an efficient security for our patient data. We have used some other encryption and decryption methods to modify the data, so the third person or party will not be able to understand when they access them. We have used some of the machine learning algorithms to predict the patient condition using the data that the patient has submitted. We have used logistic regression for the prediction of dengue for a patient. By this process, it will be easy and efficient way to find out whether the patient is critical or not. Likewise, the developing square chain innovation with the human services framework makes simple the connection between the specialist and cloud, Doctor and patient and patient and cloud for thorough social insurance information sharing, medicinal records survey, and care auditability. Furthermore, this can be utilized with any sort of medicinal information security the previous user data classification.

References

1. Alomar N, Alsaleh M, Alarifi A (2017) Social authentication applications, attacks, defense strategies and future research directions: a systematic review. *IEEE Commun Surv Tutor* 19(no. 2):1080–1111
2. Wang S, Wang J, Wang X, Qiu T, Yuan Y, Ouyang L, Guo Y, Wang F-Y (2018) Blockchain-powered parallel healthcare systems based on the ACP approach. *IEEE Trans Comput Soc Syst* 5(4) (2018)
3. Jabeen F, Hamid Z, Akhunzada A, Abdul W, Ghouzali S, Esposito TC, Santis AD, Tortora G, Chang H, Choo KKR (2018) Blockchain: a panacea for healthcare cloud-based data security and privacy?" *IEEE Cloud Comput* 5(1):31–37 (2018)
4. Bingulac SP (1994) On the compatibility of adaptive controllers (Published conference proceedings style). In: Proceedings of 4th annual Allerton conference circuits and systems theory, New York, pp 8–16
5. Davi CCM, Pastor A, Oliveira T, Neto FBL, Braga-Neto U, Bigham A, Bamshad M, Marques ETA, Acioli-Santos B (2019) Severe Dengue prognosis using human genome data and machine learning. *IEEE Trans Biomed Eng* 2234–2341
6. Zhang P, Walker MA, White J, Schmidt DC, Lenz G (2017) Metrics for assessing blockchain-based healthcare decentralizedapps. In: 2017 IEEE 19th international conference one-health networking, applications and services (Healthcom), pp 1–4
7. Mettler M (2016) Blockchain technology in healthcare: the revolution starts here. In: 2016 IEEE 18th international conference on e-health networking, applications and services (Healthcom), pp 1–3; Juette GW, Zeffanella LE (1990) Radio noise currents n short sections on bundle conductors (presented conference paper style). In: Presented at the IEEE summer power meeting, Dallas, TX, June 22–27, Paper 90 SM 690-0 PWRS
8. Hari Rao VS, Naresh Kumar M (2012) A new intelligence-based approach for computer-aided diagnosis of Dengue fever. *IEEE Trans Inform Technol Biomed* 1089–7771

9. Liu W, Zhu S, Mundie T, Krieger U (2017) Advanced blockchain architecture for e-health systems. In: IEEE 19th international conference on e-health networking, applications and services (Healthcom). IEEE, pp 1–6
10. Masoumi M, Rezayati MH (2015) Novel approach to protect advanced encryption standard algorithm implementation against differential electromagnetic and power analysis. *IEEE Trans Inform Forensics Secur* 10(2):256–265
11. Tunstall MM, Whitnall C, Oswald E (2013) Masking tables—an underestimated security risk. In: International workshop on fast software encryption (FSE), pp 425–444
12. Puppala M, He T, Yu X, Chen S, Ogunti R, Wong STC (2016) Data security and privacy management in healthcare applications and clinical data warehouse environment. In: 2016 IEEE-EMBS international conference on biomedical and health informatics (BHI), pp 5–8
13. Abouelmehdi K, Beni-Hssane A, Khaloufi H, Saadi M (2017) Big data security and privacy in healthcare: a review. In: Procedia computer science, vol 113, pp 73–80 (2017). The 8th international conference on emerging ubiquitous systems and pervasive networks (EUSPN 2017)/The 7th international conference on current and future trends of information and communication technologies in healthcare (ICTH-2017)/affiliated workshops
14. Kahani N, Elgazzar K, Cordy JR (2016) Authentication and access control in e-health systems in the cloud. In: 2016 IEEE 2nd international conference on big data security on cloud (BigDataSecurity). IEEE international conference on high performance and smart computing (HPSC), and IEEE international conference on intelligent data and security 2016, pp 13–23
15. Christo MS, Meenakshi S (2016) Enhancing security properties of rumor riding protocol under various attacks scenario in P2P network. In: 2016 international conference on communication and signal processing (ICCPSP). 6 Apr 2016, pp 1130–1135

An IoT-Based Efficient Way of Monitoring Food Quality Management



Varsha Patil, Rajesh Kadu, Namrata Patel, and Kranti Bade

Abstract Ensuring an acceptable level of food quality and food safety is necessary to provide adequate protection for consumers. A food contamination can occur in the production process, but also a large part caused by the inefficient food handling because of inappropriate ambient conditions when the food is being transported and stored. There are many factors leading to food poisoning, typically changes in temperature and humidity are important factors. So the monitoring system capable of measuring temperature and humidity variability during transport and storage is of prime importance. Today almost everybody is getting effected by the food they consume, it's not only about the junk food, but all the packed foods, vegetables, products consumed and used in daily life, as all of them do not offer quality since their temperature, moisture, oxygen content vary from time to time. Majority of consumers only pay attention to the information provided on the packaging, i.e., the amount of ingredients used and their nutritional value but they forget that they are blindly risking their health by ignoring the environmental conditions to which these packets are subjected. Maintaining food freshness is the main criteria to protect human health. Yeasts and micro bacteria in the food result into organic acids, volatile acids and basic gases. Such by-products lowers pH of environment within and surrounding of the food, fish spoilage produces acetate, ammonia and CO₂. In this food quality monitoring system, our aim is reporting hazardous activities in the food spoilage by detecting acidic and volatile substances through absorption of these into volume of hydrogel in surrounding gases of food.

V. Patil · R. Kadu · N. Patel · K. Bade (✉)

Computer Engineering Department, SIESGST, Mumbai University, Mumbai, India

e-mail: bade.kranti@siesgst.ac.in

V. Patil

e-mail: varsha.patil@siesgst.ac.in

R. Kadu

e-mail: rajesh.kadu@siesgst.ac.in

N. Patel

e-mail: namrata.patel@siesgst.ac.in

Keywords FQM (food quality monitoring) • Sensors • Microcontroller

1 Introduction

Maintaining food security has become unconditional when it comes to food trade and customer demand. The food put on the market has to be of good quality and safe for consumption, as well as not be a source of disease and infection. For this reason, securing food safety and quality is a matter of international significance and a responsibility of food producers and governments.

In the country like us where food is abundant, people choose food product based on various quality factors such as texture, flavor, size, gloss and consistency. Food quality is important in food manufacturing requirement as it is directly related to the consumers' health issues. So some samples need to be routinely monitored at market food places.

In this paper, we are introducing food monitoring system designed on the basis of food sensors and microcontroller processing. The volatility in food substances, releasing of various gases or odor, texture of the food is being sensed continuously by food sensors from the time it has been started. This sensed data is then provided to the next processing unit of microcontroller, and alerts are sending to the consumers via mobile applications.

2 Review of Literature Survey

With the advancement in IoT technology, objects are interconnected and can be used in different smart home applications. Food quality monitoring system has become an important application from human health point of view. In paper, [1] proposed food quality monitoring system in which sensor senses state of food. It monitors the growth of yeasts and microbes from the spoilage of food placed in the fridge. In paper [2] the parameters like humidity, gas concentrations and ambient pressures are measured using IoT sensing layer. Here, IoT sensing layer uses FTIR [2] Fourier transform infrared to differentiate non-halal meat and from halal meat which is necessary for Muslims consumption.

Smart refrigerator system in paper [3] is easy to use and also economical for the user. In this system ARM microcontroller and Wi-Fi module transmits information to the android phone using by using IoT.

The IoT framework [4] designed for food monitoring at every stage of supply chain serves the purpose of preventive consumer health protection by maintaining required standard conditions. It involves continuous description of components of the food and nutrition system for the planning and program evaluation of the system.

According to the Food surveillance, monitoring and risk assessment department [5] of India various divisions are set up. Food surveillance and monitoring division

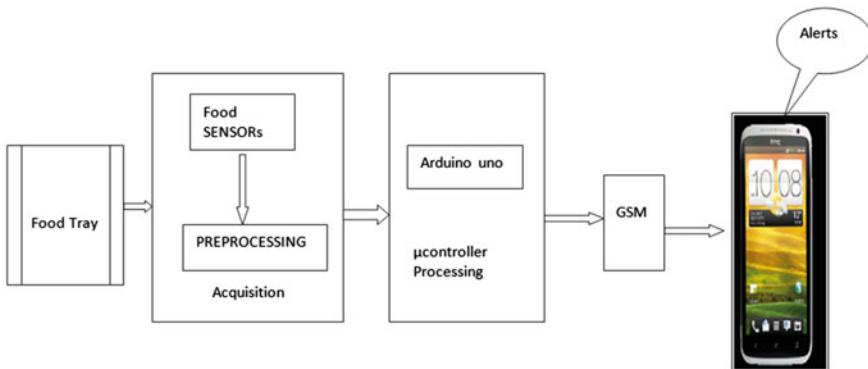


Fig. 1 The proposed method

which ensures that good agricultural practices, good manufacturing practices and good testing laboratory practices are followed by all stakeholders. Risk assessment division carry out risk assessment plan in line with the FHO/WHO guidance in this field.

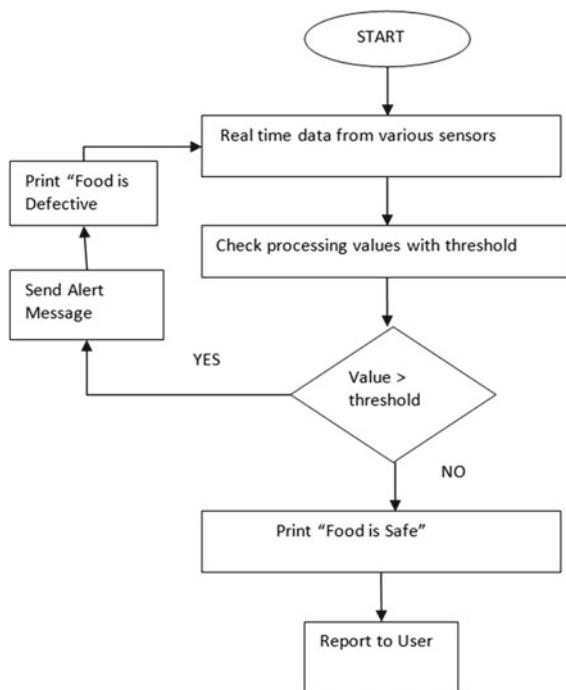
In paper [6], Shivaraj had developed gas analyzer for detecting gases like ammonia, NO₂, CO and CO₂, so this analyzer can be utilized in a variety of different roles and in food industries. Accorsi et al. [7] in this paper aims at contributing to set up general method to build an architecture of entities, food objects, flow of design and managing the interdependencies observed supply chain and input-output system. In this work, it monitors the temperature of the food for food safety.

In the present scenario, the work done in terms of sensed values that has been recorded and detailed analysis has been performed. In our project, we are trying to provide automated control alternatives for food quality monitoring by analyzing gases released from spoiled food (Figs. 1 and 2).

2.1 Sensors Subsystem

In the proposed system, sensors detects acidic and gaseous substances around food remotely. In the proposed system, sensors sense the state of freshness of food and send alert messages to user. Sensors monitor growth of yeasts and microbes and gases generating from spoiling food product. Electro sensors detect changes in the appearance and feel, changes in color and taste of food. This information given by sensors is transmitted from the device and signals alarm LED setup. The sensors will respond only smartphone readings if the frequencies are unchanged.

Fig. 2 Working of the system



2.2 Microcontroller Unit

The whole system is governed by the Arduino uno microcontroller. It is programmed using C language to receive the input from the mobile app over the Wi-Fi interface and depending on the input food item, it activates the particular sensors whose output values are compared against an existing database and algorithm. Depending on the levels of the output values, it is decided whether the food is safe for consumption or not. If yes, the microcontroller gives the signal for the green LED to glow and Wi-Fi transmits the all information to the android phone by using IoT. The items are out dated it is below the set threshold value to alert notification is send to the user's mobile to replace the food item.

In the working module, we have used easily manageable, stable and long life and simple drive circuit of Arduino Uno. These are key specifications of the Arduino Uno microcontroller unit.

Key Specifications

- Single-Chip RF Transceiver for 915-MHz ISM Band
- 902–928-MHz Operation
- FM/FSK Operation for Transmit and Receive
- Arduino Uno Operating Voltage – 5 V
- DC Current – 20 mA

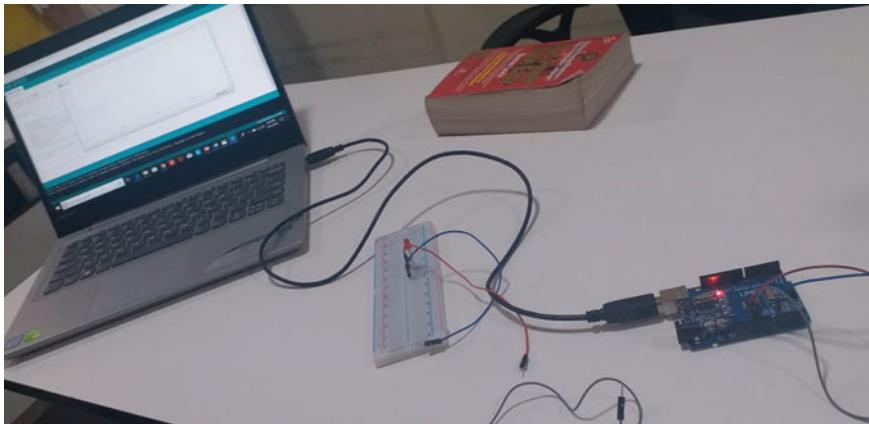


Fig. 3 Connection of Arduino Uno with MQ3 sensor

Main objective of the food quality monitoring system is the food safety and protection of consumer's health. The system has set some objectives as follows.

1. To ensure food safety by monitoring food storage and transportation at every level of supply chain.
2. Providing good quality food to the consumers, with the help of this technology which will further add to the reduction in the rate of food poisoning kind of diseases.
3. The performance and analysis of routine measurements, aimed at detecting changes in the nutritional or health status of the food doesn't guarantee that. Thus it increases the supplier standard.

2.3 Results

For experimental analysis liquid spirit has been taken because ripen fruits, spoiled onions also release alcoholic content in a small extent. Experiment results are represented as the range of RS gas 0.95. The results show that the alcohol content detected (Figs. 3, 4 and 5).

3 Conclusion

Our study contributes the increasing state of quality of food by continuous monitoring of gaseous contents released from food by MQ3 sensor. This information can be utilized on updating user from time to time basis.

```

void setup()
{
    Serial.begin(9600);
    pinMode(LED_BUILTIN, OUTPUT);
}

void loop()
{
    float sensor_volt;
    float RS_gas; // Get value of RS
    int sensorValue = analogRead(A0);
    sensor_volt=(float)sensorValue/1024.0;
    RS_gas = (5.0-sensor_volt)/sensor_volt;
    Serial.print("RS_ratio = ");
    Serial.println(RS_gas);
    if(RS_gas<0.95)
    {
        digitalWrite(LED_BUILTIN, HIGH);
    }
    else digitalWrite(LED_BUILTIN, LOW);
    delay(1000);
}

```

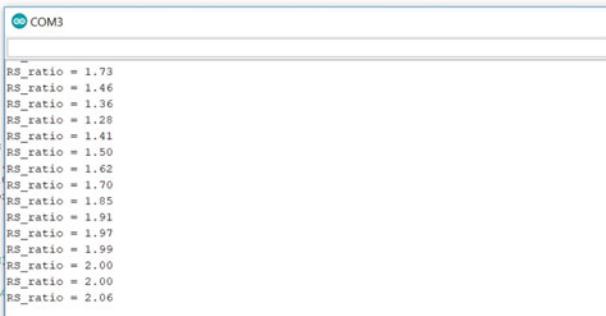


Fig. 4 Alcohol content measured by MQ3 sensor

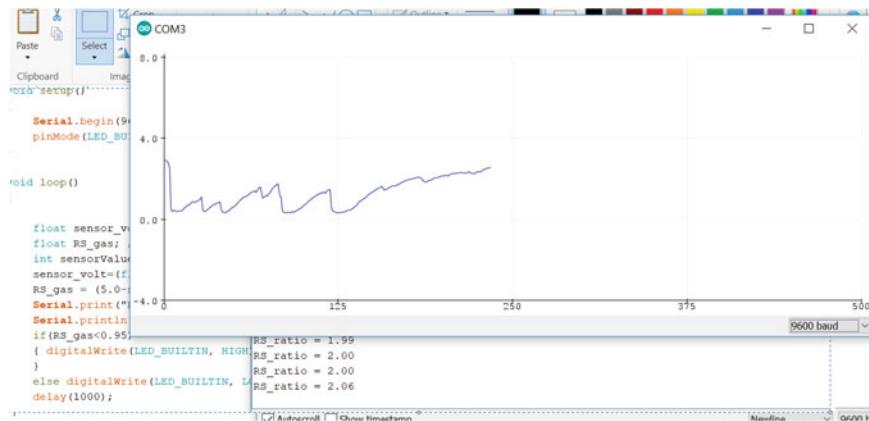


Fig. 5 Readings indicating alcohol content is above threshold

This proposed Food quality monitoring system can be deployed in the food storage area and food products are kept under continuous observations gives better results. The food monitoring system allows the authorities to get the status on real-time basis and take decisions accordingly. An option of smartphone with the monitoring system will simplify further the process of monitoring food spoilage. In future, this can be further expanded to determine the food contamination due to various pesticides, chemicals, etc. probed into the food.

References

1. Popa A, Hnatiuc M, Paun M, Geman O, Jude Hemanth D, Dorcea D, Son LH, Ghita S (2019) An intelligent IoT-based food quality monitoring approach using low-cost sensors, 13 Mar 2019. Published in Symmetry, 11, 374. <https://doi.org/10.3390/sym11030374>
2. Witjaksono G, Rabih AAS, Alva S, Yahya N (2018) IOT for agriculture: food quality and safety. In: IOP conference series materials science and engineering, April 2018, p. 343. <https://doi.org/10.1088/1757-899X/343/1/012023>
3. Shanmugapriya S, Kevinyaa A, Kavitha S, Manimegala K (2015) Detecting the food spoilage in refrigerator using food sensor. In: National conference on recent technologies for sustainable development 2015 [rechzig'15]—28th Aug 2015
4. Singh D, Jain P (2016) Iot based smart refrigerator system. Int J Adv Res Electron Commun Eng (Ijarece) 5(7) (2016)
5. Srivastava A, Gulati A (2016) iTrack: IoT framework for smart food monitoring system. Int J Comput Appl 148(No.12): 0975–8887
6. Shivaraj SN, Soumya P, Shriya D, Sourabh, Design and development of gas analyser for detecting ammonia, NO₂, CO and CO₂
7. Accorsi R, Bortolini M, Baruffaldi G, Pilati F, Ferrari E (2017) Internet of things paradigm in food supply chains control and management. Procedia Manuf 11:889–895

Host-Specific Outlier Detection Using Process Relation Semantics with Graph Mining



Binayak Panda and Satya Narayan Tripathy

Abstract Many small-scale organizations use computer system to run their business logic with a preferred list of application programs. They are open to threats in the form of an unknown or unrecognized program. The existing antivirus vendors use static as well as dynamic methods to protect a host from possible known threats, but they do not distinguish programs based on host-specific program list. It motivates to have a system which considers host-specific preference list of application and protects from any possible threat like unknown or unrecognized program. In this paper, a “Host Specific Outlier Detection Model” is proposed which learns all possible process relation semantics of a host using graph-based learning approach to detect an outlier program. The proposed system uses graph to represent the semantic relation among the processes running on a host in terms of predecessor and successor. Such a graph is named as process relation graph (PRG), and the paths of this graph are considered as key features representing the process relation semantics for a given process snapshot of the host. An optimized number of PRG’s are collected dynamically on a host to build a generalized process relation graph (GPRG). Such a GPRG enables detection of any outlier program with an accuracy of 96% for a suspected PRG at any time instance.

Keywords Host-Specific outlier detection · Process relation graph mining · Process relation semantics · Graph searching · Graph merging

1 Introduction

We are observing a regular increase in the number of computer hosts getting affected by unknown/unrecognized programs, specifically named as potentially unwanted programs or malwares. The two approaches named misuse detection and anomaly

B. Panda (✉) · S. N. Tripathy

PG. Department of Computer Science, Berhampur University, Berhampur, Odisha, India

S. N. Tripathy

e-mail: snt.cs@buodisha.edu.in

detection are popular to deal with such threats. Misuse detection is an approach used to protect a host from malicious activities by using static or dynamic signature comparison methods. Most of the antivirus programs rely on signature verification for the detection of known threats, but the signature database needs to be updated regularly. But the code obfuscation techniques like polymorphism and metamorphism diminishes the performance of signature-based detection techniques. On the contrary, in anomaly detection, a host is monitored to detect a significant deviation in its normal operation. Novelty detection and outlier detection are two words representing anomaly detection. Novelty detection is about distinguishing a new observation to be normal or abnormal from a known list of observations. Novelty detection is about finding that newly observed data differ in some way from the earlier observations considered during the knowledge gaining process. Outlier detection aims to detect the abnormal observations present in the total number of observations. It is applied to problems where descriptions of abnormalities, i.e., negative observations are insufficient, but the descriptions of normality's, i.e., positive observations are sufficiently available.

Rapid increase in the number of unknown programs makes threat level to be the same irrespective of number of detection techniques available. This justifies the unavailability of descriptions for unknown programs which are considered as negative samples for a host. But looking into the usual operation on a host, the list of known programs is very restricted, which justifies the availability of descriptions of positive samples for a host. This motivates to apply outlier detection approach on a host to protect it from unknown or unrecognized programs. The proposed system dynamically collects process snapshots which represents the process relation semantic, and it helps in finding outlier process on a host. In this work, the set of paths of the graph PRG are considered as key features representing the process relation semantics for a given process snapshot of the host. Construction of PRG with optimized paths is the key idea and the estimated time complexity is found as $O(n^2)$ with n as the number of processes in that process snapshot. A GPRG is constructed using a list of optimized PRG's which represents all possible process relation semantics of the host. Graph merging is the key mechanism in building GPRG and the estimated time complexity is found as $O(n^3)$ with n as the number of PRGs. Any suspected PRG gets searched on the GPRG to detect existence of outlier. Graph searching is the key logic here and works in $O(n^3)$ time which is quite impressive in graph searching.

The remaining sections of this paper are as follows: Section 2 mentions related works on anomaly or outlier detection techniques and the use of graph mining as a dynamic program analysis technique. Section 3 speaks about the proposed host-specific outlier detection system using graph. Section 4 mentions the experimental work, results, and complexity analysis of the system. Section 5 speaks the concluding remark.

2 Literature Review

Protecting a host from known threats as well as unknown threats is attracting researchers to apply various methods on attributes of a host or program. Referring the system calls and kernel modules of a host, an anomaly detection approach is designed which resulted with very less false alarms [1]. Utilizing the runtime features gathered from processes and network operations of a node, an anomaly detection system is designed using a one-class support vector machine for a node of cloud infrastructure [2]. This anomaly detection system successfully detected new malware for a node in cloud infrastructure. Using runtime metrics with a two-class support vector machine, an online adaptive anomaly detection (AAD) framework is designed to detect abnormalities in the system [3]. This framework suffered from misclassification on new anomalous instances because of data imbalance problems [4]. A naive anomaly detector (NAD) learns n-grams referring to a normal dictionary to detect anomalies. The dictionary is expected to have all possible occurring n-grams from system traces. Any unseen n-gram is considered as a case of anomaly. It uses a tuneable detection threshold to conclude the unseen n-gram as normal or anomalous. NAD achieved 100% detection rate with almost zero false alarms [5]. Online analysis of multiple components such as memory usage, I/O operation count and CPU time usage, etc., are considered for entropy-based anomaly testing (EbAT) in cloud infrastructure to do anomaly detection [6]. Runtime behaviors of programs are found from the application programming interface (API) used and the control flow graph (CFG) of the executable. A combination of API and CFG information as API-CFG graphs are used to represent the semantic aspect of programs. Graph mining API-CFG graphs are represented as feature vectors that are used to classify unseen benign and malicious codes [7]. A graph constructed from instruction traces of executable and graph kernel is used to create a similarity matrix between graphs [8]. Graph mining approaches are new areas being investigated for malware analysis. A compression-based graph mining approach with graph matching mechanism to detect unknown benign as well as malware programs became very efficient [9]. Graph metrics are used rather than other patterns for malware detection [10]. Similarity matrices are used for the classification of malware and benign programs using a support vector machine. This graph-based malware detection shows significant improvement in malware detection rate. Graphs are powerful data structures that capture the semantic relationship between related objects of a system [11]. Graph-based methods are also computationally powerful enough to detect abnormalities in a system like other anomaly detection methods. Finding graph similarity is very popular to solve the problem of similarity in two structural objects, estimation of maximum sub-graph (MSG) and graph edit distance (GED) is the key concept behind graph similarity [12]. These are NP-complete problems. With advanced computing capabilities, finding similarity between graphs with GED is not reasonable for graphs with more than 16 nodes [13]. So graph similarity search is very costly in terms of computation time. Two or more linearly ordered directed graphs can be merged to represent a single generalized

linearly ordered directed graph with preserving the successor and predecessor relation [14]. The proposal of SimGNN provides an approach to solve graph similarity problems by mapping graphs into vectors and pairwise node comparison to conclude [15]. It has been claimed that the computation of similarity is done in quadratic time with respect to the number of nodes in two graphs.

This paper shows dynamic analysis on process relation semantics, as a host-specific outlier detection method for real-time protection against possible threats. Use of graph to represent process relation semantics and graph mining on various instances helps in finding a generalized process relation graph to detect outliers in the host.

3 System Design

3.1 System Overview

The system has three modules like PRG construction, merging of PRG's to find the GPRG, and searching a PRG at GPRG to detect occurrence of an outlier. The modular design of the system is depicted in Fig. 1. The first two stages can be viewed as the learning phase, and the last phase can be viewed as the testing phase of the system.

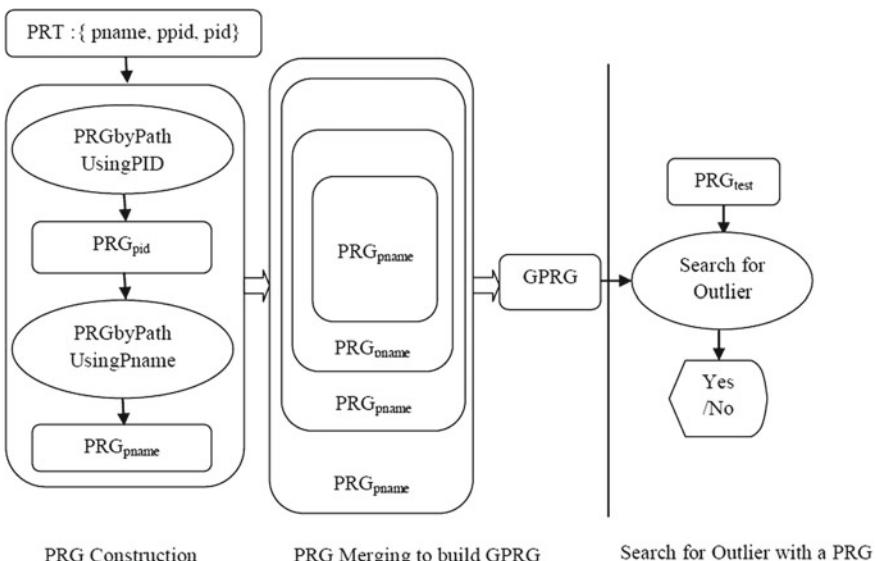


Fig. 1 Modular division of the system

3.2 PRG Construction

Windows management interface command (WMIC) is used to find the process relation table (PRT) as shown in Table 1 for a particular timestamp. Each entry of PRT is a 3-tuple (pname, ppid, pid) where pname is the process name, ppid is the parent process ID, and pid is the processID. This PRT shows process relation semantic as successor-predecessor relation, which is a snapshot of the PRG of a host at some instance. Construction of PRG uses the path as an attribute to represent the graph and uses path pruning as the important mechanism involved. A process is assigned with a unique pid and a unique pname in an operating system. The pname remains same but except few, the pid changes every time the operating system boots. There exist several hierarchical predecessors (a process who creates another process) and successor (a process which got created) relation among processes. Table 1 shows a sample PRT of a host from which three sample hierarchical predecessor and successor relations can be observed as “System Idle (pid 0) → System (pid 4) → smss.exe (pid 276),” “parent_wininit.exe (pid 360) → wininit.exe (pid 436) → services.exe (pid 484) → svchost.exe (pid 1008),” and “parent_explorer.exe (pid 1096) → explorer.exe (pid 2072) → chrome.exe (pid 3776).” The conclusion from the observation is that services.exe and svchost.exe hold relational semantic between them as parent process and child process. Such a process relation semantic will always hold. Conceptually in windows, an instance of smss.exe creates wininit.exe and terminates immediately; hence, the parent process for wininit.exe is named here as parent_wininit.exe. Similar observations are found for winlogon.exe and explorer.exe and considered similarly.

A PRT can be viewed as a directed acyclic graph PRG = (V, E). V is a finite set of nodes, i.e., $V = \{PID_0, PID_1, PID_2, \dots, PID_n\}$ and each PID_i represents a pid, E is a finite set of edges, i.e., $E = \{PRS_0, PRS_1, PRS_2, \dots, PRS_m\}$ and each PRS_i represent process relation semantic between two processes, hence, $E \subseteq V \times V$. A node is a source node if its in-degree is 0 and a sink node if its out-degree is 0. A path in a graph is an ordered sequence of nodes where predecessor and successor relations are preserved among them. A simple path is a path with no repetition of

Table 1 A sample PRT of a host

c:\> wmic get parentprocessid, processid, name

pname	ppid	pid	pname	ppid	pid
System_Idle	0	0	svchost.exe	484	1008
System	0	4	svchost.exe	484	1040
smss.exe	4	276	taskhost.exe	484	1424
csrss.exe	360	384	explorer.exe	1096	2072
wininit.exe	360	436	winlogon.exe	428	576
services.exe	436	484	chrome.exe	2072	3776
lsass.exe	436	508	AcroRd32.exe	2072	4736
lsm.exe	436	516	dwm.exe	1040	1036

the node. In this work, a set of optimized paths are used to represent a graph, which is a novel approach compared to the traditional adjacency matrix or adjacency list approach.

A PRG_{pid} is a graph representing a set of simple paths, where each path is a linear ordering of PID_i 's starting with a source $\text{PID}_{\text{source}}$ and ending with some PID_m preserving the process relation semantics among them as shown in Eq. (1). Table 2 shows the set of paths representing PRG_{pid} corresponding to PRT_i .

$$\text{PRG}_{\text{pid}} = \{\text{PRG}_{\text{pid}} \cup \{\text{PATH}_k\} | \text{PATH}_k \in \{\text{PID}_{\text{source}}, \dots, \text{PID}_m\}\} \quad (1)$$

In PRG_{pid} , it is observed that there exist longer paths from some source node to some sink node which contains other paths starting with the same source node as a sub-path as highlighted in Table 2. Hence, such sub-paths are pruned out to reconstruct the PRG_{pid} as a set of simple paths, where each path is a linear ordering of PID_i 's starting with a source $\text{PID}_{\text{source}}$ and ending with a sink PID_{sink} preserving the process relation semantics among them as shown in Eq. (2). The Algorithm 1 PRGbyPathUsingPID explains the use of PRT_i to find all paths of PRG_{pid} with pruning out sub-paths present and leaving all the possible paths from the source node to sink node only.

$$\text{PRG}_{\text{pid}} = \{\text{PRG}_{\text{pid}} \cup \{\text{PATH}_k\} | \text{PATH}_k \in \{\text{PID}_{\text{source}}, \dots, \text{PID}_{\text{sink}}\}\} \quad (2)$$

Table 2 A sample PRG_{pid} (set of paths)

0 → 4	360 → 436 → 508	360 → 436 → 484 → 1424
0 → 4 → 276	360 → 436 → 516	1096 → 2072
360 → 384	428 → 576	1096 → 2072 → 3776
360 → 436	360 → 436 → 484 → 1008	1096 → 2072 → 4736
360 → 436 → 484	360 → 436 → 484 → 1040	360 → 436 → 484 → 1040 → 1036

Algorithm 1: PRGbyPathUsingPID

Input: A PRT as a dictionary $ppid_dict$ {pid: ppid} where the pid is the key and $ppid$ is the value and a list $ppid_list = []$ containing only $ppid$.

Output: A set of paths representing PRG_{pid} as explained in Eq. (2) and as shown in Table 3.

1. $PRG_{pid} = \emptyset$
2. For each ($pid \in \text{key}(ppid_dict)$ and $pid \notin ppid_list$)
 - a. $PATH = \{pid\}$
 - b. While (True)
 - If ($ppid_dict[pid] \neq \epsilon$)
 $PATH = PATH \cup ppid_dict[pid]$
 $pid = ppid_dict[pid]$
 - Else
 Break While
 - c. $PRG_{pid} = PRG_{pid} \cup PATH$
3. Return PRG_{pid}

PRG_{pid} as shown in Table 3 is the set of paths representing all the processes running on a host at one time instance preserving the process relation semantic. As pid changes every time operating system boots, pid needs to be mapped to pname. PRG_{pname} is the graph with set of paths by pname found from PRG_{pid} as shown in Table 4.

From the highlighted portion of Table 4, it is observed that path pruning operation is essential to find the optimized PRG_{pname} as shown in Table 5. A path by name from PRG_{pid} is to be included in optimized PRG_{pname} or not is decided as given in Eq. (3). The Algorithm 2 PRGbyPathUsingPname explains the use of PRG_{pid} to find

Table 3 A sample PRG_{pid} (set of paths after path pruning)

$0 \rightarrow 4 \rightarrow 276$	$428 \rightarrow 576$	$1096 \rightarrow 2072 \rightarrow 3776$
$360 \rightarrow 384$	$360 \rightarrow 436 \rightarrow 484 \rightarrow 1008$	$1096 \rightarrow 2072 \rightarrow 4736$
$360 \rightarrow 436 \rightarrow 508$	$360 \rightarrow 436 \rightarrow 484 \rightarrow 1424$	
$360 \rightarrow 436 \rightarrow 516$	$360 \rightarrow 436 \rightarrow 484 \rightarrow 1040 \rightarrow 1036$	

Table 4 A sample PRG_{pname} (set of paths)

parent_System → System → smss.exe	parent_wininit.exe → wininit.exe → services.exe → svchost.exe
parent_csrss.exe → csrss.exe	parent_wininit.exe → wininit.exe → services.exe → taskhost.exe
parent_wininit.exe → wininit.exe → lsass.exe	parent_wininit.exe → wininit.exe → services.exe → svchost.exe → dwm.exe
parent_wininit.exe → wininit.exe → lsm.exe	parent_explorer.exe → explorer.exe → chrome.exe
parent_winlogon.exe → winlogon.exe	parent_explorer.exe → explorer.exe → AcroRd32.exe

Table 5 A sample PRG_{pname} (after path pruning)

parent_System → System → smss.exe	parent_wininit.exe → wininit.exe → services.exe → taskhost.exe
parent_csrss.exe → csrss.exe	parent_wininit.exe → wininit.exe → services.exe → svchost.exe → dwm.exe
parent_wininit.exe → wininit.exe → lsass.exe	parent_explorer.exe → explorer.exe → chrome.exe
parent_wininit.exe → wininit.exe → lsm.exe	parent_explorer.exe → explorer.exe → AcroRd32.exe
parent_winlogon.exe → winlogon.exe	

optimized PRG_{pname} by pruning out sub-paths and leaving all the possible paths from the source node to sink node only. For each PATH_{byname} ∈ PRG_{pid},

$$\text{PRG}_{\text{pname}} = \begin{cases} \text{PRG}_{\text{pname}} \cup \{\text{PATH}_{\text{byname}}\}, \text{ If } (\text{PATH}_{\text{byname}} \notin \text{PRG}_{\text{pname}}) \\ \text{and } (\#P \in \text{PRG}_{\text{pname}} | \text{PATH}_{\text{byname}} \subset P) \\ \text{PRG}_{\text{pname}} \cup \emptyset, \text{ Otherwise} \end{cases} \quad (3)$$

Algorithm 2: PRGbyPathUsingPname

Input: A PRG_{pid} by pid's and a dictionary pname_dict {pid: pname} where pid is the key and pname is the value.

Output: A set of optimized paths for PRG_{pname} by pname as shown in Table 5.

1. PRG_{pname} = \emptyset
2. For each (PATH ∈ PRG_{pid})
 - a. PATHbyname = \emptyset
 - b. For each (node ∈ PATH)
 - i. PATHbyname = PATHbyname ∪ pname_dict[node]
 - c. If (PATHbyname \notin PRG_{pname}) & ($\#P \in \text{PRG}_{\text{pname}} | \text{PATHbyname} \subset P$)
 - i. PRG_{pname} = PRG_{pname} ∪ PATHbyname
 3. Return PRG_{pname}

3.3 PRG Merging

A collected set of PRGs are merged to construct a GPRG, which will have all possible process relation semantics. GPRG is represented as a set of all possible paths of a specific host from some trivial source node to some trivial sink node. Let PRG_{list} = {PRG₁, PRG₂, ..., PRG_m} be the set of PRGs, where each PRG_i is computed at various time intervals at various levels of use of the host. A path in GPRG using paths of PRG_i's is evaluated as shown in Eq. (4), value of optimal **m** to be found from the experiment with respect to the host. Algorithm 3 BuildGPRG explains the

```

parent_System-->System-->sms.exe , parent_wininit.exe-->wininit.exe-->lsass.exe , parent_wininit.exe-->wininit.exe-->lsm.exe ,
logon.exe , parent_wininit.exe-->Wininit.exe-->services.exe-->MspEng.exe , parent_wininit.exe-->Wininit.exe-->services.exe-->spc
t.exe-->wininit.exe-->services.exe-->armsvc.exe , parent_wininit.exe-->Wininit.exe-->services.exe-->mosquitto.exe , parent_wini
vices.exe-->taskhost.exe , parent_explorer.exe-->explorer.exe-->mseces.exe , parent_explorer.exe-->explorer.exe-->NitroPDFPrint
er.exe-->explorer.exe-->igfxtray.exe , parent_explorer.exe-->explorer.exe-->hkcmd.exe , parent_explorer.exe-->explorer.exe-->
explorer.exe-->explorer.exe-->GrooveMonitor.exe , parent_explorer.exe-->explorer.exe-->jusched.exe , parent_wininit.exe-->Wininit.e
st.exe-->igfxsrvc.exe , parent_explorer.exe-->explorer.exe-->ONENOTEM.EXE , parent_GoogleCrashHandler.exe-->GoogleCrashHandler.e
>wininit.exe-->services.exe-->NisSrv.exe , parent_explorer.exe-->explorer.exe-->Apoint.exe-->ApMsgFwd.exe , parent_wininit.exe-->
-->svchost.exe-->WUDFHost.exe , parent_ApmEx.exe-->pnpEx.exe , parent_crss.exe-->crss.exe-->conhost.exe , parent_explorer.e
e.exe-->chrom.exe , parent_wininit.exe-->wininit.exe-->services.exe-->svchost.exe-->wsauclt.exe , parent_RdrCEF.exe-->RdrCEF.exe
explorer.exe-->explorer.exe-->AcroRd32.exe-->parent_explorer.exe-->explorer.exe-->spyder3.exe-->pythonw.exe-->pythonw.exe-->cm
xe-->explorer.exe-->calc.exe , parent_wininit.exe-->wininit.exe-->svchost.exe-->dm.exe , parent_wininit.exe-->vir
>SearchIndexer.exe-->SearchProtocolHost.exe , parent_wininit.exe-->wininit.exe-->services.exe-->SearchIndexer.exe-->SearchFilterHc
.exe-->explorer.exe-->WINWORD.EXE , parent_explorer.exe-->explorer.exe-->EXCEL.EXE , parent_wininit.exe-->wininit.exe-->services.
, parent_wininit.exe-->wininit.exe-->services.exe-->svchost.exe-->audiiodg.exe , parent_explorer.exe-->explorer.exe-->cmd.exe-->pyt
ht_wininit.exe-->wininit.exe-->sppsvc.exe , parent_wininit.exe-->wininit.exe-->services.exe-->svchost.exe-->taskeng
xe-->explorer.exe-->WINWORD.EXE-->AcroRd32.exe-->RdrCEF.exe-->RdrCEF.exe , parent_explorer.exe-->explorer.exe-->pythonw.exe-->pyt
h.exe-->explorer.exe-->SnippingTool.exe , parent_wininit.exe-->wininit.exe-->services.exe-->svchost.exe-->wispris.exe , parent_e
-->mspaint.exe , parent_explorer.exe-->explorer.exe-->vlc.exe , parent_explorer.exe-->explorer.exe-->devcpp.exe , parent_explor
erType.exe-->MathTypeLib.exe , parent_wininit.exe-->wininit.exe-->services.exe-->svchost.exe-->MniPrvSE.exe , parent_explorer.exe
exe , parent_explorer.exe-->explorer.exe-->POWERPNT.EXE , parent_explorer.exe-->explorer.exe-->AcroRd32.exe-->RdrCEF.exe-->RdrCE
f.exe-->explorer.exe-->FreeCell.exe , parent_wininit.exe-->wininit.exe-->services.exe-->svchost.exe-->dilhost.exe , parent_firefox.e
x.exe , parent_explorer.exe-->explorer.exe-->notepad.exe , parent_explorer.exe-->explorer.exe-->Hearts.exe , parent_explorer.exe
re.exe , parent_explorer.exe-->explorer.exe-->chrom.exe-->software_reporter_tool.exe-->software_reporter_tool.exe , parent_expl
exampp-control.exe-->httpd.exe-->httpd.exe , parent_explorer.exe-->explorer.exe-->putty.exe , parent_explorer.exe-->explorer.exe
-->explorer.exe-->explorer.exe-->dns3.exe-->dns3server.exe . parent_explorer.exe-->explorer.exe-->Mahjong.exe .

```

Fig. 2 A sample GPRG

use of PRG_i 's in the construction of GPRG. Figure 2 shows some sample paths of GPRG. For all $\text{PATH} \in \text{PRG}_i$, $1 \leq i \leq m$,

$$\text{GPRG} = \begin{cases} \text{GPRG} \cup \{\text{PATH}\}, \text{If } (\text{PATH} \notin \text{GPRG}) \\ \text{and } (\#\text{P} \in \text{GPRG} \mid \text{PATH} \subset \text{P}) \\ \text{GPRG} \cup \emptyset, \text{Otherwise} \end{cases} \quad (4)$$

Algorithm 3: BuildGPRG

Input: $\text{PRG}_{\text{list}} = \{\text{PRG}_1, \text{PRG}_2, \dots, \text{PRG}_m\}$, set of m PRGs found at several time instances.

Output: A GPRG as to be built using Eq. (4) and found as shown in Fig. 2.

1. $\text{GPRG} = \emptyset$
2. For each $\text{PRG} \in \text{PRG}_{\text{list}}$
 - a. For each $\text{PATH} \in \text{PRG}$
 - i. If $(\text{PATH} \notin \text{GPRG})$ and $(\#\text{P} \in \text{GPRG} \mid \text{PATH} \subset \text{P})$ $\text{GPRG} = \text{GPRG} \cup \text{PATH}$
3. Return GPRG

3.4 Searching PRG in GPRG

For identification of outlier, a PRG_{test} is to be constructed from a PRT_{test} collected at some specific time instant. The PRG_{test} is to be searched in the GPRG for outlier

in terms of path. It is a problem of searching a graph as sub-graph in another. A graph $G_1 = (V_1, E_1)$ will be a sub-graph of graph $G = (V, E)$, if $V_1 \subset V$ and $E_1 \subset E$. Equation (5) specifies the steps involved in deciding whether PRG_{test} is a sub-graph of the GPRG or not. Algorithm 4 SearchPRG explains the steps of searching PRG_{test} in GPRG. If PRG_{test} is not found to be a sub-graph, it returns the set of paths $OutlierInPath$ which represent the presence of outlier in the host. For each $PATH \in PRG_{test}$,

$$OutlierInPath = \begin{cases} OutlierInPath \cup \{PATH\}, & \text{If(PATH } \notin \text{GPRG)} \\ \text{and } (\nexists P \in \text{GPRG} | PATH \subset P) \\ OutlierInPath \cup \emptyset, & \text{Otherwise} \end{cases} \quad (5)$$

Algorithm 4: SearchPRG
Input: PRG_{test} and GPRG.
Output: Yes or No with $OutlierInPath$: Which is a set of PATHS containing some outlier program for the host.

1. $OutlierInPath = \emptyset$
2. For each $PATH \in PRG_{test}$
 - a. If ($PATH \notin \text{GPRG}$) and ($\nexists P \in \text{GPRG} | PATH \subset P$)
 $OutlierInPath = OutlierInPath \cup PATH$
3. If $| OutlierInPath | == 0$
YES, PRG_{test} is a subgraph of GPRG
Else
NO, PRG_{test} is not a subgraph of GPRG
4. Return $OutlierInPath$

4 Experiment and Complexity Analysis of the System

4.1 Experiment and Result Analysis

With a questionnaire, a preferred list of applications of a host is chosen based on user's need or interest.

- The host being monitored is created from a known to be a clean disk image, and a list of trusted third-party programs are installed based on the user's preference.
- The host is used for normal activities and applications are used with varying loads from light to heavy with respect to the number of applications. At regular interval of 10 second, the PRT's are collected. With this schedule, 1000 instances of PRT's are collected from a continuous run of 3 hour. In a week cycle, 7000 instances of PRT's are collected.
- On every collected PRT, the PRG_{pname} is constructed by invoking the algorithms $PRGbyPathUsingPID$ and $PRGbyPathUsingPname$ one after another.

BuildGPRG is invoked on the list of $\text{PRG}_{\text{pname}}$ and the GPRG is constructed. For higher accuracy $|\text{PRG}_{\text{list}}|$, i.e., an optimal \mathbf{m} is to be finalized from the experiment.

In the training phase, for building GPRG, 80% of the total collected PRT's are considered with a random selection. For the testing phase, a total of 1000 PRG's are considered, out of which 300 PRG's are randomly chosen from earlier collected set, and 700 PRG's are collected from the host running some additional programs beyond the preferred list of programs. For each of the PRG's, the SearchPRG is invoked to reach a conclusion. The performance metrics recall, precision, F1-Score, and accuracy of the system are evaluated using the following fundamental parameters. Result is a case of true positive (TP): when an unknown PRG is detected as a case of outlier, true negative (TN): when a known PRG is not detected as a case of outlier, false positive (FP): when a known PRG is detected as a case of outlier and false negative (FN): when an unknown PRG is not detected as a case of outlier. Equation (6) illustrates the calculation of recall of the model which is the ability of the model to identify unknown PRG correctly, i.e., the probability of an unknown PRG being detected correctly. Equation (7) illustrates the calculation of precision of the model which says how many unknown PRG's are identified correctly, i.e., the probability that an unknown PRG will be detected correctly. F1 Score is the harmonic mean of precision and recall, which is as illustrated in Eq. (8). Equation (9) illustrates the overall accuracy of the system, i.e., the degree to which it detects a newly tested PRG correctly.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (9)$$

Evaluation of optimized \mathbf{m} with several iterations during construction of GPRG and the respective change in performance metrics are mentioned in Table 6. Figure 3

Table 6 Performance comparison in finding optimized \mathbf{m} for GPRG

	$m = 3000$	$m = 4000$	$m = 5000$	$m = 6000$
Recall	0.92	0.95	0.97	0.97
Precision	0.93	0.96	0.97	0.98
F1	0.93	0.95	0.97	0.97
Accuracy	0.90	0.94	0.96	0.96

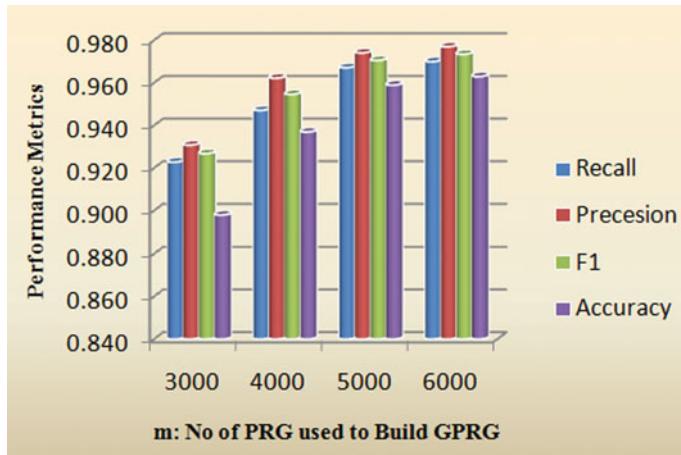


Fig. 3 Number of PRG used to build GPRG versus performance metrics

depicts graphically the change in performance metrics to build the optimized GPRG. It is observed that a change of m from 3000 to 4000 results in a significant improvement in performance, a change of m from 4000 to 5000 results in a little improvement in performance, and a change of m from 5000 to 6000 the performance became consistent. Hence, the optimal m is accepted as 5000 in the experiment with accuracy, recall, precision, and F1 Score above 96%.

4.2 Complexity Analysis of the System

The system uses graph mechanism; hence, time analysis becomes essential. The estimation of time complexity for each stage of the model follows.

- PRG Construction: One time traversal on a PRT to generate ppid_dict, pname_dict and ppid_list needs constant time as $\theta(n)$ assuming n records in the PRT. Statement-2 of *PRGbyPathUsingPID* iterates for each key of dictionary ppid_dict, and it backtracks to find the path from sink node to source node. Hence, the total time will be in order of $O(n^2)$. Statement-2 of *PRGbyPathUsingPname* iterates for each path in PRG_{pid} and all nodes on the path. Hence, the total time will be in the order of $O(n^2)$. Assuming m PRT's are taken into consideration, the total time will be $m * (O(n^2) + O(n^2)) \cong O(mn^2)$.
- GPRG Construction: Statement-2 of *BuildGPRG* iterates through every PRG, for each PRG the statement-2.a goes through all the paths, and for each path, it considers all the nodes. Hence, the total time will be in the order of $O(n^3)$. Hence overall time for the training phase will be $O(mn^2) + O(n^3) \cong O(n^3)$.
- Search PRG: Statement-2 of *SearchPRG* iterates through all the paths of PRG_{test} , and for each path, all the paths of GPRG are processed node by node. Hence,

the total time will be in the order of $O(n^3)$. In a best scenario, the time can be evaluated as $O(kn^2)$, where the longest path will have k nodes with k sufficiently small compared to n the number of paths. In testing phase, the best case time is $\theta(n) + O(kn^2) \cong O(kn^2) \cong O(n^2)$ and the worst case time is $\theta(n) + O(n^3) = O(n^3)$.

5 Conclusion

This work shows process relation semantic as a key feature to do dynamic analysis on a specific host for detection of outlier, with a list of known programs to be used on that host. Process relation semantic is well described using a graph PRG, which is represented by a novel approach of set of paths. A time-efficient path pruning approach to find an optimized PRG and a time-efficient graph merging approach to build optimized GPRG are important findings of this work. An optimized m , i.e., the number of PRG to be used for building optimized GPRG also plays a vital role in accuracy of system. The experimental result shows 96% accuracy in detecting outliers using the GPRG which contains all possible process relation semantic as set of paths with an optimized m as 5000. It is found that use of path to represent the graph rather than the adjacency matrix or adjacency list in this problem domain resulted in lesser complexity. The complexity of searching for PRG_{test} on GPRG is found as $O(n^2)$ in best case and $O(n^3)$ in worst case, which is very promising.

References

1. Murtaza SS, Khreich W, Hamou-Lhdj A, Couture M (2013) A host-based anomaly detection approach by representing system calls as states of kernel modules. In: IEEE 24th international symposium on software reliability engineering (ISSRE), Pasadena, CA, 2013, pp 431–440. <https://doi.org/10.1109/ISSRE.2013.6698896>
2. Watson M, Shirazi SN, Marnerides A, Mauthe A, Hutchison D (2015) Malware detection in cloud computing infrastructures. IEEE Trans Dependable Secure Comput 13:1–1. <https://doi.org/10.1109/TDSC.2015.2457918>
3. Pannu HS, Liu J, Fu S (2012) Aad: adaptive anomaly detection system for cloud computing infrastructures. In: IEEE symposium on reliable distributed systems, pp 396–397
4. Marnerides A, Malinowski S, Morla R, Kim H (2015) Fault diagnosis in DSL networks using support vectormachines. Comput Commun. <http://www.sciencedirect.com/science/article/pii/S0140366415000080>
5. An N, Duff A, Naik G, Faloutsos M, Weber S, Mancoridis S (2017) Behavioural anomaly detection of malware on home routers. 12th international conference on malicious and unwanted software (MALWARE), Fajardo, 2017, pp 47–54. <https://doi.org/10.1109/MALWARE.2017.8323956>
6. Wang C (2009) Ebaf: online methods for detecting utility cloud anomalies. In: Proceedings of the 6th middleware doctoral symposium. ACM, p 4. <https://doi.org/10.1145/1659753.1659757>
7. Eskandari M, Hashemi S (2012) A graphmining approach for detecting unknown malwares. J Vis Lang Comput 23:154–162. <https://doi.org/10.1016/j.jvlc.2012.02.002>

8. Anderson B, Quist D, Neil J, Storlie C, Lane T (2011) Graph-based malware detection using dynamic analysis. *Comput Virol* 7:247–258. <https://doi.org/10.1007/s11416-011-0152-x>
9. Wüchner T, Cisłak A, Ochoa M, Pretschner A (2019) Leveraging compression-based graph mining for behaviour-based malware detection. *IEEE Trans Depend Secure Comput* 16(1):99–112. <https://doi.org/10.1109/TDSC.2017.2675881>
10. Jang J-W, Woo J, Mohaisen A, Yun J, Kim HK (2015) Mal-Netminer: malware classification approach based on social network analysis of system call graph. *Math Probl Eng* 2015:20. <https://doi.org/10.1155/2015/769624>
11. Akoglu L, Tong H, Koutra D (2014) Graph-based anomaly detection and description: a survey. <https://arxiv.org/abs/1404.4679v2>
12. Zeng Z, Tung AKH, Wang J, Feng J, Zhou L (2009) Comparing stars: on approximating graph edit distance. *PVLDB* 2(1):25–36
13. Blumenthal DB, Gamper J (2018) On the exact computation of the graph edit distance. *Pattern Recognition Letters*
14. Schwägerl F, Uhrig S, Westfechtel B (2015) A graph-based algorithm for three-way merging of ordered collections in EMF models. *Sci Comput Program* 113:51–81. <https://doi.org/10.1016/j.scico.2015.02.008>
15. Bai Y, Ding H, Bian S, Chen T, Sun Y, Wang W (2019) SimGNN: a neural network approach to fast graph similarity computation. In: The twelfth ACM international conference on web search and data mining (WSDM'19), February 11–15, 2019, Melbourne, VIC, Australia. ACM, New York, NY, USA, 9 p. <https://doi.org/10.1145/3289600.3290967>

Intrusion (Hybrid) Detection System for Cloud Computing Environments



G. Nagarajan, R. I. Minu, and T. Sasikala

Abstract Cloud computing is the technology where we can store information somewhere using the Internet. We can upload our information into it. We can work in many computing environments, as it is serving well. Some intruders can steal the data from it. Virtualization helps us in forming many systems which helps us in server building. The intrusion detection system is used to detect the intruders who are trying to manipulate the information in the cloud, which is a bad scenario. We are trying to develop an algorithm that helps us in finding the intruders in cloud computing environments. We should take care of security as the data is uttermost important, so the safety of it is also important. We have sections divided where each section would help us our research better understand it. Many technologies are used for protecting the data in these environments. It is very much great that protecting data in cloud areas is not an easy task. Many researchers are finding new algorithms to protect the data.

Keywords Cloud computing · Intrusion detection · Cloud security · Anomaly · Hybrid

1 Introduction

Cloud computing is a demanding technology where we can do all our computations in the cloud by avoiding the use of physical devices. It helps us a lot in storing the data and doing the computations online. The Internet is the key requirement for

G. Nagarajan (✉) · R. I. Minu · T. Sasikala

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

R. I. Minu

e-mail: minur@srmist.edu.in

T. Sasikala

e-mail: dean.computing@sathyabama.ac.in

R. I. Minu

SRM Institute of Science and Technology, Kattankulatur, India

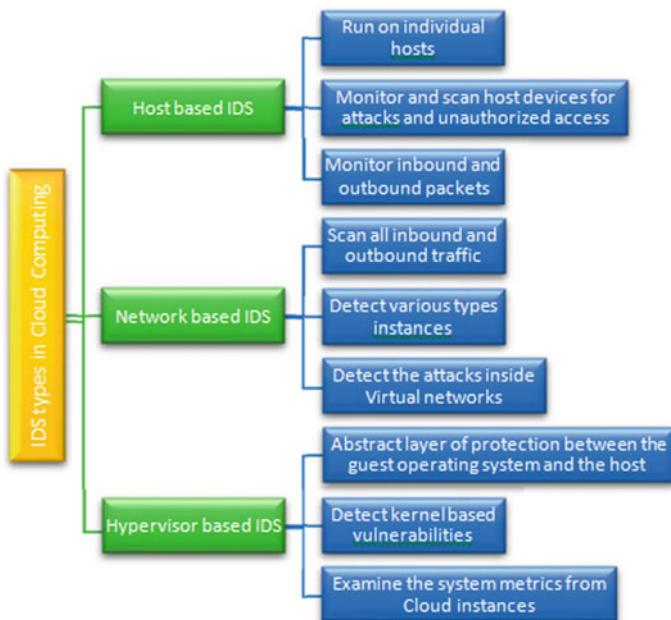


Fig. 1 Types of intrusion detection system (IDS) in cloud computing [2]

doing all the cloud operations. Virtualization helps us in sharing the single system into multiple systems we call virtual machines which helps us in computing online. In these days, everyone is using the cloud because it provides great features to work on from uploading the data to doing all the computations. To secure this information, we are coming up with algorithms like SHA, MD5, AES these algorithms are very much useful in protecting the data in the area. We used these algorithms for the safety of data because we have SHA. For suppose, it is a very powerful algorithm that helps us keep the data unmodified it has patterns to secure the data this is how we protect the data in cloud environments. They are the family of cryptography hash functions it is released by NIST. National Institute of standards and Technology. In the form of Fig. 1, [1] a detailed description of each type of intrusion detection system (IDS) in cloud computing.

2 Related Work

Intrusion detection system or preventing of intruders from the cloud computing environments is a difficult task everyone has contributed their own contribution in order to protect the cloud computing environments. some of them are approached many hashing algorithms to protect the data in the environments we referred all their research papers so that we will get an idea about the approaches they followed to

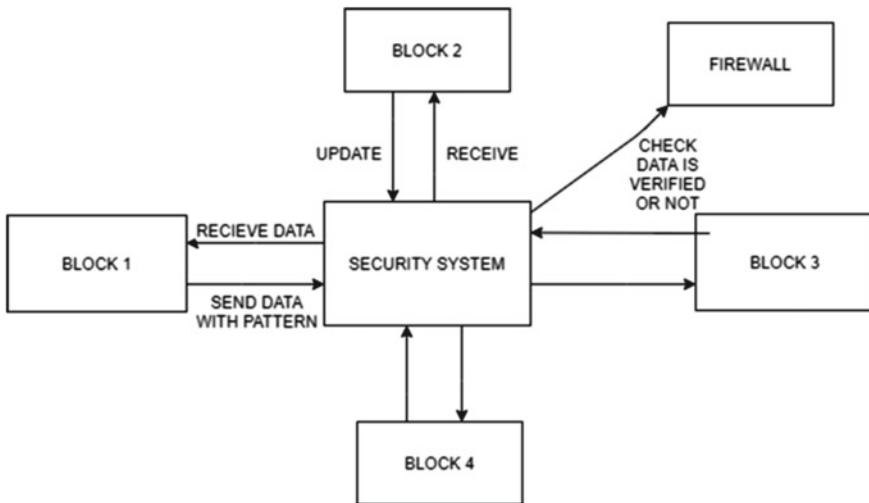


Fig. 1 Architecture diagram

detect the intruder and the way they approached to detect the intruders in the cloud computing environments some used alarm management techniques and also machine learning techniques for detecting the intrusion.

Atallah et al. [3] developed an evolutionary design for intrusion detection program, he used genetic programming techniques helps us in detecting the intruders who are trying to modify the data in the cloud so that the data or information stored in it will be protected in their approach they have used linear genetic programming, multi expression programming and gene expression programming. Used to analyse the intrusion detection prevention system, they thought using this approach which is lightweight and easy to use and which is better than approaching machine learning models used for detection programs. There are several traditional intrusion detection systems available in literature [4–11] which may be adopted to apply in cloud environment.

3 Existing System

The cloud data sharing system has evolved so much where we can store a large amount of data in the system where it is not physically stored, but it feels like it is stored in our real system. The main important thing in cloud environments is securing the cloud data. Researchers are trying to secure the cloud data by implementing many algorithms and implementations for only protecting the data in the cloud it is a challenging process all the existing systems include the intrusion detection system using anomaly detection, network intrusion detection systems, host-based intrusion detection system and signature-based detection system. All these systems secure the

cloud in different ways by taking the different inputs for the better security of the cloud as the existing systems uses some downgraded implementations which are not sufficient for the storing the cloud in a large size, so the existing systems are protecting the resources only to the minimal content which is going to store in the cloud. The Internet is the key thing for the running of the cloud sharing systems.

In paper [12], the author addresses the problem faced on verifiability of outsourcing data in cloud. Their finding efficiently detects the misbehavior of the cloud server using their own Yao's garbled circuit with homomorphic encryption. In the Atallah and Li [13] finding, they tried to address the problem in sequence comparison in secure outsourcing. They are implemented a secured protocol for this kind of problem (Table 1).

4 Proposed System

The main objective of this paper is to provide a schematic of secure data sharing scheme model with secure key distribution scheme. Using this mode, data sharing is possible in a secure way in a dynamic group. The key distribution scheme specified in this proposed system uses a secure communication channel. In the proposed system model, the user will receive the private key from the group manager without any certification using this secure communication channel. The verification of the public key of the user was authorized by the group manager. The durability of this scheme was analyzed using a predefined security analysis. By implementing this method for data sharing, collusion attacks can be prevented. Only those user who are in the list of group user can send data in the cloud. A fine-grained control of data in cloud can be achieved using this methodology. The user with abnormal activity were revoked from the group list. They cannot access the data in the cloud once they got revoked. This user revocation was achieved using a polynomial function. The other salient feature of this methodology is that if a new user joint the group with an existing user private key, and they need not to generate a key again which support the dynamic properties of this scheme.

A. Login Module

This is the first activity module in our project where we use our credentials to login into the system. We have a user account and Admin account separately because Admin has more privileges than the user as Admin has to create the account for every user when creating an account has been assigned to some group. If the details provided by the user or admin did not match it will show error where we have to validate the details correctly in order to use the system.

B. Registration Module

A new user who wants to use the application needs to be registered first before he uses all the features of the application. In order to use the application, a user must create an account with the help of Admin. The user has to set a unique strong password which

Table 1 Analysis of cloud based IDS [1]

Features references	Detection technique	IDS type	Positioning	Detection time	Data source	Attacks covered	Limitations/challenges
CIDS for cloud computing networks, 2010 [14]	Signature based	Distributed	Each cloud region	Real time	Network traffic, signatures of known attacks	Protects system from single point of failure, DoS and DDoS	Can't detect unknown attacks, high-computational overhead
Securing cloud from DDOS attacks using IDS in VMs, 2010 [15]	Network based	Virtual Switch		Real time	Network packets, signatures of known intrusions	Secures VMs from DDoS attacks	Detects only known attacks
Integrating a NIDS into an open source cloud computing environment, 2010 [16]	Network based	At each node		Real time	Network traffic, normal usage of resources like CPU	Only Known attacks particularly SIP flooding	Can't detect unknown attacks,
Autonomic agent-based self-managed IDPS, 2010 [17]	Anomaly based	Host based	N/A	Real time	Network traffic, System activities (system calls etc.)	Can detect all types of attacks in real-time	Implementation details are not given
Multi-level IDS and log management in CC, 2011 [18]	Host based	At each guest OS		Real time	User behaviors, known attack patterns	Can detect both known and unknown attacks at a fast rate	Consumes more resources for high level users
Distributed intrusion detection in clouds using MAs, 2009 [19]	Distributed	At each VM		Real time	Audit data, known intrusion patterns, system logs	Can detect both known and unknown attacks	There is a limit on the number of VMs to be visited

(continued)

Table 1 (continued)

Features references	Detection technique	IDS type	Positioning	Detection time	Data source	Attacks covered	Limitations/challenges
Collabrat: Xen Hypervisor based collaborative IDS, 2011 [20]	VMM based, distributed	At each VMM	Real time	Audit data, anomaly database	Can detect hyper-call based attacks on VMM and host OS	Cannot detect other types of attacks	
IDS for cloud computing, 2012 [21]	Hybrid	Distributed	At the processing server	Real time	Audit data, user profiles, signatures of known intrusions	Can help CSP to improve its quality of service, detects unknown attacks	The proposed idea is theoretical, no implementation provided
Bayesian classifier and snort based NIDS in cloud computing, 2012 [22]	Network based	At the processing servers	Real time	Network packets, known attack signatures, prior events	Detects all types of attacks	Complexity increased due to integration of both, signatures and anomalies	
IDS in cloud computing environment, 2011 [23]	Host based and network based	At each node	Real time	Logs of user activities, signatures of known attacks	Can detect all known attacks, may detect unknown attacks using ANN	Experimental results are not given	
GCCIDS, 2010 [2]	Host based	At each node	Real time	Audit data, user profiles	Known attacks, Unknown attacks using ANN	Accurate detection requires more training time, there is a limit on number of rules	

contains all the alphabets and some special characters which helps the password be strong so that it will be hard for the intruders to crack the password. There will be some validation for the successful registration of the user. This will help.

C. Creating Instance

In this type of data securing, the data gets secured first by encrypting the data before it gets stored in the cloud so that we can know that the data in the cloud is modified or not as the data has no control over the internet because the owner itself can't say that whether the data in the cloud systems are secured or not. In this module, the data gets protected before it gets stored in the cloud environment it is encrypted using the cryptographic algorithms.

D. Find Collusion Module

This module will help in finding the collusion which is intruder who is trying to modify the data which is present in the cloud platform.

5 Results and Discussion

By this research paper, we are here by to discuss the output of our project which helps in protecting the cloud environments, here the data which is going to store in cloud environments will be undergone to these methods like AES, MD5, SHA so that the file will be set some pattern whenever someone is trying to steal the information or updating of the information the pattern gets changed and we will capture the identity of the intruder who tried to steal or modify our information. We use the key for the group so that whenever we upload the file or send the file to receiver in between we may have intruder who tries to steal our data so we should take care of the data and finding the intruder is much needful for the protection of the data this method of finding the intruder in the cloud computing environments will help in finding the intruder for the protection of data.

6 Conclusions

The problems in the cloud environments mostly are security issues which are very problematic for the people who are storing the data in the cloud environments which they thought like safe but protecting the data in the cloud environments is a challenging task we have three algorithms namely SHA, AES, MD5 which helps in protecting the data by encrypting the data which are going to store in the cloud which helps us protecting our data which is being stored in the cloud environments there is key generation using the secure hashing algorithm. So there is in need for secure outsourcing sequence comparison algorithm to address this collusion attack.

This paper illustrated a methodology for authentic user data access on cloud. In this paper for key generation and sharing SHA algorithm were used by the trusted authorities. The encryption of the file is done by using AES algorithm in trusted authority module. The data owner can compute the hash value using MD 5 algorithm. The detail of the generated key was stored in a database which will be used dynamically. If there is any abnormality, the keys will be verified by the trusted authority. The trusted authority will securely saved this data on cloud through CSP module. The confidentiality of the communicated data was secured against collusion attacks.

References

1. Mehmood Y, Shibli MA, Habiba U, Masood R (2013) Intrusion detection system in cloud computing: challenges and opportunities. In: 2013 2nd national conference on information assurance (NCIA), pp 59–66. IEEE (2013)
2. Mikail A, Pranggono B (2019) Securing infrastructure-as-a-service public clouds using security onion. *Appl Syst Innov* 2(1):6
3. Atallah MJ, Kerschbaum F, Du W (2003) Secure and private sequence comparisons. In: Proceedings of ACM Workshop Privacy Electronic Society (WPES), Washington, DC, USA (2003), pp 39–44
4. Borah S, Panigrahi R, Chakraborty A (2018) An enhanced intrusion detection system based on clustering. In: Saeed K, Chaki N, Pati B, Bakshi S, Mohapatra D (eds) Progress in advanced computing and intelligent engineering. Advances in intelligent systems and computing, vol 564. Springer, Singapore (2018)
5. Dutt I, Borah S, Maitra IK (2020) Immune system based intrusion detection system (IS-IDS): a proposed. *IEEE Access* 8(2020):34929–34941
6. Dutt I, Borah S (2015) Some studies in intrusion detection using data mining techniques. *Int J Innov Res Sci Eng Technol* 4(7):5500–5511
7. Prasad KK, Borah S (2013) Use of genetic algorithms in intrusion detection systems: an analysis. *Int J Appl Res Stud (iJARS)*. 2(8) (2013). ISSN: 2278-9480
8. Panigrahi R, Borah S (2019) Dual-stage intrusion detection for class imbalance scenarios. *Comput Fraud Secur* 2019:12–19
9. Dutt I, Borah S, Maitra IK, Bhowmik K, Maity A, Das S (2018) Real-time hybrid intrusion detection system using machine learning techniques. Advances in Communication, Devices and Networking, Springer, pp 885–894
10. Borah S, Chakraborty D, Chawhan C, Saha A (2011) Advanced clustering based intrusion detection (ACID) algorithm. Advances in Computing and Communications, Springer CCIS series, vol 192, Part 1, pp 35–43 (2011). ISSN: 1865-0929. https://doi.org/10.1007/978-3-642-22720-2_4
11. Borah S, Chakraborty A (2014) Towards the development of an efficient intrusion detection system. *Int J Comput Appl* 90(8):15–20
12. Feng Y, Ma H, Chen X (2015) Efficient and verifiable outsourcing scheme of sequence comparisons'. *Intell Autom Soft Comput* 21(1):51–63
13. Atallah MJ, Li J (2004) Secure outsourcing of sequence comparisons. In: Proceedings of International Workshop Privacy Enhancing Technol. (PET), Toronto, ON, Canada (2004), pp 63–78
14. Bakshi A, Yogesh B (2010) Securing cloud from DDOS attacks using intrusion detection system in virtual machine. In: 2010 second international conference on communication software and networks, pp 260–264

15. Mazzariello C, Bifulco R, Canonico R (2010) Integrating a network IDS into an open source cloud computing environment. In: 2010 sixth international conference on information assurance and security, pp 265–270
16. Patel A, Qassim Q, Shukor Z, Nogueira J, Júnior J, Wills C (2010) Autonomic agent-based self-managed intrusion detection and prevention system. In: Proceedings of the South African information security multi-conference (SAISMC 2010), pp 223–234
17. Lee JH, Park MW, Eom JH, Chung TM (2011) Multi-level intrusion detection system and log management in cloud computing. ICACT pp 552–555
18. Dastjerdi AV, Bakar KA, Tabatabaei SGH (2009) Distributed intrusion detection in clouds using mobile agents. In: Third international conference on advanced engineering computing and applications in sciences (2009), pp 175–180
19. Bharadwaja S, Sun W, Niamat M, Shen F (2011) Collabra: a Xen Hypervisor based collaborative intrusion detection system. In: Eighth international conference on information technology: new generations (2011), pp 695–700
20. Shelke PK, Sontakke S, Gawande AD (2012) Intrusion detection system for cloud computing. Int J Sci Technol Res 1(4):67–71
21. Modi CN, Patel DR, Patel A, Muttukrishnan R (2012) Bayesian classifier and snort based network intrusion detection system in cloud computing. In: Third international conference on computing, communication and networking technologies, 26th–28th July 2012
22. Dhage SN, Meshram BB, Rawat R, Padawe S, Misra MP (2011) Intrusion detection system in cloud computing environment. In: International conference and workshop on emerging trends in technology (ICWET 2011), pp 235–239
23. Vieira K, Schulter A, Westphall CB, Westphall CM (2010) Intrusion detection for grid and cloud computing. IEEE Computer Society (July/August 2010), pp 38–43

Intrusion Detection and Prevention Systems Using Snort



Shubham Sharma, Parma Nand, and Pankaj Sharma

Abstract We live in a digital era, where everything is digital and online, ranging from private data to information on any topic. As this advancement in information technology is a boon, there is nothing stopping it from becoming a big thing for everyone. Day-by-day cybercrimes are increasing, and every other day, we hear about major data breaches and data leaks of million and millions of individuals. Hence, it is necessary to protect our digital self. A sure and dependable way to do so is by using an IDS. This paper provides an insight on IDS, its working and functions. So, in this paper, we will provide all necessary information regarding IDS, and this can also help us to create our very rule set to protect our network environment.

Keywords IDS · IPS · Firewall and machine learning

1 Introduction to IDS

An interface for hardware or applications that identify malicious activity or policy breaches of a network or networks is an intruder. Any intrusion activity or violation is typically reported directly to an admin or captured remotely and uses a protection data and event control system. Detection of intrusion is the method of monitoring and analyzing the events that exist on a computing system or network for signs of possible incidents that are breaches of computer security protocol abuses, fair use rules or standard security procedures or imminent threats [1, 2].

S. Sharma (✉) · P. Nand · P. Sharma

Department of Computer Science Engineering, School of Engineering and Technologies, Sharda University, Greater-Noida, India

e-mail: 2019007684.shubham@pg.sharda.ac.in

P. Nand

e-mail: parma.nand@sharda.ac.in

P. Sharma

e-mail: pankaj.sharma1@sharda.ac.in

Platforms are tracked by intrusion detection for malicious purposes, and incorrect warnings are always disposed of. Therefore, businesses need to perfect their IDS items when they first download these. As compared to malicious behavior, this involves setting up the intrusion prevention mechanisms appropriately to realize what normal traffic flow looks like. Intrusion monitoring systems also monitor the incoming data packets of the device to search the involved suspicious operations and send warning warnings simultaneously. An intrusion detection is classified into the following five classes.

1.1 Network Intrusion Detection Systems (NIDS)

Inside the network, a fixed point is established which acts as a review unit for all the network packets flowing through the devices. It consists of a document with the subnets that are considered to be safe and then that particular set of subnets are cross-referenced with the original set of traffic generating subnets to restrict the malicious incidents.

1.2 Host Intrusion Detection Systems (HIDS)

It can run on a set of individual hosting network pointers. The tracing of inward and outward packages comes under the scope of HIDS. The unusual activities are reported to the authority.

1.3 Protocol-Based Intrusion Detection System (PIDS)

It consists of the machine or agent which can periodically inoculate at the monitoring end of the server, detecting the protocols between the two usual users such as operator/expedient and the server and interpreting that.

1.4 Application Protocol-Based Intrusion Detection System (APIDS)

It remains as scheme or agent that typically resides inside a community of servers. It detects the intrusions by monitoring and interpreting the communication on application-specific protocols. For starters, this would explicitly track the SQL procedure to the middle ware as it manages with the file of the Web server.

1.5 Hybrid Intrusion Detection System (HIDS)

It stands generated by the mixture of two or more methods to the intrusion detection. In order to grow a whole vision of the net environment, host agent or application information is joint with net knowledge in the mixture of intrusion detection scheme [3–6].

2 IDPS Principles

Intrusion detection is a mechanism by which events that arise on a computing system or network are monitored and analyzed for signs of possible injuries that are compromises or immediate chances of violations of computer security legislation, fair use rules or standard security practices. There are numerous explanations for problems, including ransomware (e.g., viruses, malware), criminals who get unwanted Internet network access, and registered computer operators who misappropriate or threaten to improve extra rights under which they are illegal. An intrusion prevention system (IPS) is a software that has all the features of an intrusion detection and will help to avoid possible incidents as well. IDS and IPS technologies have many of the same capabilities, and administrators can typically disable preventive features in IPS items, enabling them to function as IDSs.

The intrusion detection framework and intrusion prevention framework would then be referred to as IDS and IPS, respectively. IDS and IPS platforms have many of the capabilities, and administrators can typically disable preventive functionality in IPS items, enabling them to function as IDSs. Therefore, in the remainder of this guide, the word intrusion detection and prevention systems (IDPS) are briefly used to refer to both IDS and IPS technologies. Certain cases are specifically mentioned [7].

3 Comparison of IDS, IPS and Firewall

The key difference is that the firewall conducts acts such as traffic blocking and filtering when a system administrator is detected and warned by an IPS/IDS or stops the attack according to configuration. A firewall permits traffic based on a collection of configured rules [8] (Table 1).

Table 1 Difference between IDS versus IPS versus firewall

Parameter	IDS	IPS	Firewall
Philosophy	An intrusion detection system (IDS) is an application for devices or software that monitors traffic for malicious behavior or breaches of policies and sends detection warnings	IPS is a framework that inspects, detects, classifies and then proactively prevents malicious traffic from attacking traffic	Firewall is a network protection system based on default rules that filter incoming and outgoing network traffic
Principle of working	Recognizes traffic in real time, scans for traffic trends or attack signatures and generates warnings	Inspects traffic in real time, searches for traffic patterns or attack signatures and then avoids detection attacks	Filters IP address-based traffic and port numbers
Configuration mode	For monitoring and detection, inline or as end host (via span)	Inline mode, generally being in layer 2	Layer 3 mode or transparent mode
Placement	Via port span, non-inline (or via tap)	Inline after firewall usually	At the network perimeter inline
Traffic patterns	Analyzed	Analyzed	Not analyzed
Action on unauthorized traffic detection	Alerts/alarms on detection of anomaly	Preventing the traffic on detection of anomaly	Block the traffic

4 Working of IDPS

An intrusion detection (IDS) is a scheme that controls Internet circulation for doubtful activity including warnings since this action is observed. While the major purposes are the recognition and warning of abnormalities, by identifying malicious activity or anomalous traffic, such intrusion detection may bring charges, like obstructing traffic received from questionable Internet Protocol (IP) accounts.

There are also server-based intrusion detection services that are used to protect data and are programmed in cloud deployments. We may classify their research in distinct and separate ways.

4.1 Behavior-Based Working

Detection and avoidance of behavior-based intrusion analyze traffic patterns and behavior. While this may appear to be technological jargon, it is not, and it is

extremely effective at detecting abnormal or deviant behaviour that poses a security risk to your network and, more significantly, instantly reducing the danger.

4.2 Anomaly-Based Working

An intrusion detection focused on anomaly is also an intrusion detection device for tracking system activity and categorizing it as either normal or abnormal to identify intrusions and abuse of both network and device. Systems were used for neural nets to considerable effect.

4.3 Rule-Based Working

In a law intrusion detection scheme, an assault will either be identified if a rule is located in the rule base or it remains hidden when it is not identified. If this is combined with FIDS, further infringements that have remained hidden by RIDS can be found [9].

5 Machine Learning for IDPS

Machine learning methods have evolved exponentially in the last decade, allowing automation and forecasts on scales never dreamed before. This inspires researchers and developers to come up with new uses for these lovely techniques. Machine learning methods were soon being used to reinforce network security schemes.

An attack, such as brute force, denial of service or even penetration from inside a network, is the most frequent vulnerability to a network's security. With the evolving trends of network activity, a complex approach to detecting and preventing such intrusions is expected. To learn and battle advanced attackers who can easily circumvent simple intrusion detection mechanisms, we need a dataset that is modifiable, reproducible and extensible (IDS). This is where artificial intelligence (AI) comes in. We will look at the different machine learning approaches that can be used to generate accurate IDS.

Machine learning is concerned with the development of programmers that can learn from results. The learning process starts with observations or data in order to identify trends in the data and make better decisions based on the examples given. The key objective is for machines to understand without human intervention and change their behaviors accordingly. Algorithms classified into several variations.

5.1 Supervised Learning

It may use labeled examples to apply what they have observed in the past to forecast future events. A training dataset is used to generate an estimated function that can be used to make assumptions about the output values. The device will have goals for new inputs after adequate testing.

5.2 Unsupervised Learning

When training details aren't labelled or classified, they're exploited. Unattended learning explores how structures can infer a function from unlabeled data to explain a hidden structure.

5.3 Semi-supervised Learning

It uses unlabeled data for instruction, with a combination of labeled and unlabeled data.

5.4 Semi-supervised Learning Falls Between Unsupervised Learning and Supervised Learning

When you do not have enough labeled data to create an effective model or you do not have the opportunity or money to have more, semi-supervised methods may be used to extend the training data collection.

6 Comparison Between Different Research Work

On the basis of above categories of intrusion detection prevention systems working. We have compared some famous research work. In this comparison, we have summarized their key feature of their paper like its research methodology, results of their work, parameter they use in this research, conclusion of work, description of the research paper, authors name, journal/conference/publication and at last their techniques in this paper.

After analyzing all these parameters and techniques of IDPS, we may conclude that the Snort is simply the greatest strategy out there for intrusion detection/prevention systems after evaluating all these research studies and their strategies.

So, why is Snort selected? Snort has the potential to do actual traffic monitoring and Internet Protocol (IP) network packet recording since it is a permitted software network-based intrusion detection/prevention (IDS/IPS).

Snort performs protocol review, scanning and material fitting. But with Snort, the key benefit is that we can apply our very own set of rules by using this intrusion prevention strategy as per our need for network protection.

7 All We Need to Know About Snort

Snort is a permitted device that is open foundation for intrusion detection. It is a very popular and powerful multi-packet instrument that is operated by many various individuals and companies. It is one of the intrusion detection/prevention schemes focused on identity. The beauty of this instrument resides in the formulation of laws.

You can create/design traffic blocking rules or simply send notifications, log notifications to a log file, send them to the console or view them on the screen.

It can be run from the command line as a packet sniffer mode that simply looks at the information in the header and prints the details on the screen. It can be used as a packet logging mode that takes each packet and logs it into the root directory's log files.

Later on, you can view the file via Snort or tcpdump.

For the study and capture of real-time raw packet data in NIDS format, Snort uses promiscuous-mode NICs. Snort can perform real-time packet logging, content search/matching and protocol analysis and can also detect a number of attacks with known loopholes.

8 Basic Architecture of Snort

We all know that, as a packet sniffer, Snort first began. The tcpdump, or its broad graphical brother Wireshark, is another famous example of a packet sniffer. Snort applied a few things to its architecture in order to develop into the IDS program it is today (Fig. 1).

9 Rules in Snort

That Snort principles are divided into two smaller modules, the rule heading as well as the policy choices. The rule system displays the policy's process, configuration, source and destination e-mail accounts and Web masks, as well as destination sequence data. It has several uses as shown in (Fig. 2).

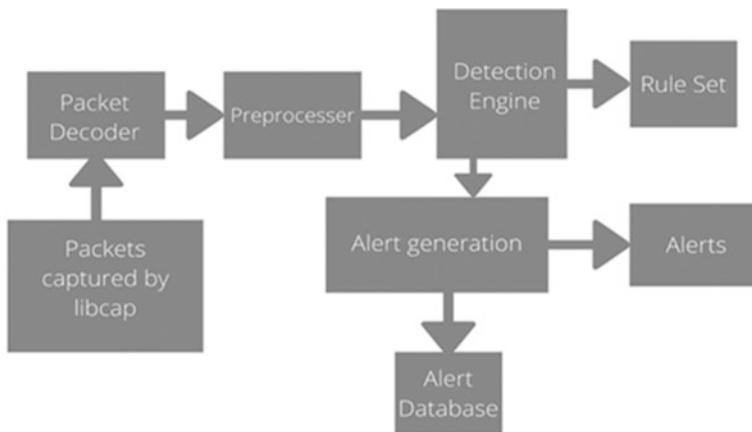


Fig. 1 Architecture of Snort

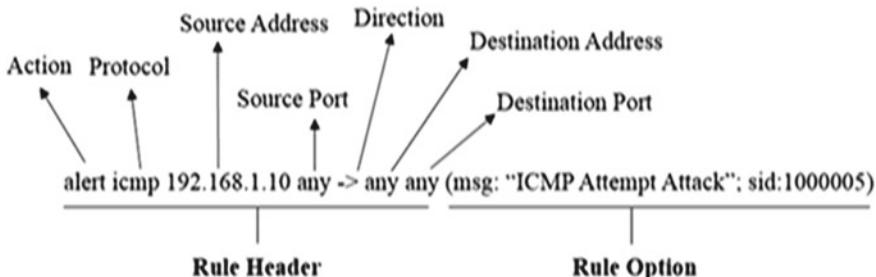


Fig. 2 Rules format in Snort

10 Methodology of Snort

Impact of digital use different methods in classifying incidents. The three identification technique classes will be mentioned: signature-based, anomaly-based and full state protocols.

10.1 Focused on Signing

A sign is a sequence alluded to by an identified hazard. Signature-based detection is the practice of comparing signs to known incidents in classifying possible injuries. Samples of signs would be as obeys.

A telnet effort by such a “root” account would be a desecration of an establishment’s welfare schemes.

A logging file with a serial monitor quality of 645 and for operating system, suggesting that now the server inspection was removed.

10.2 Focus on Anomaly

Anomaly-based identification is the procedure of matching explanations of what behavior remains deemed acceptable to observable phenomena to discern important abnormalities.

An IDPS has profiles that use anomaly-based identification to represent the usual activity of objects, such as participants, guests, network links or requests.

The patterns remain established by nursing the features of acceptable behavior ended a long historical.

10.3 State Full Protocol Analysis

The method of comparing predetermined profiles of commonly agreed meanings of benign protocol operation for and protocol state against observed events to identify anomalies is state complete protocol analysis.

11 Setup and Results

Since we all recognize, the executive summary of the analysis paper is where you report the outcomes of your thesis based on the available evidence gathered as a result of the methods [or techniques] you have implemented. The final report would explicitly state the outcomes, excluding bias or understanding, and be presented in a logical order.

11.1 Set Up the Snort

Requirements: Ubuntu, Windows, Snort, Wireshark.

Firstly, we need the Snort software in our Ubuntu server. Now, let us, to enable Snort, just use command.

```
sudo apt-get install snort
```

Because when setup begins, the GUI we checked before will ask you about it. Give your name here and click enter. Then, that will ask you regarding your network IP. Also, you can provide either a single IP or an IP array.

It would then query you anyway for the description of both the interfaces, re-include it and press enter. To make any improvements inside, use nano or any text editor to open the file type as Snort is enabled. Use this order below:

```
sudo nano /etc/snort/snort.conf
```

Read down the thumb drive below position number 45 to identify a network for protection. Configure the networking addresses which you are protecting.

ipvar HOME_NET (Our IP)

To make Snort: IDS function, operate below from button. This command publishes the full file, and the assigned tasks are checked manually.

```
sudo snort -A console -i (our interface) -c /etc/snort/snort.conf
```

11.2 Catch the ICMP Packets

ICMP is one of the network communication suites, as defined in RFC 792. ICMP results are usually used and then generated for diagnostic or control purposes of answer to failures in equipment called (as specified in RFC 1122).

ICMP failures are redirected to the IP address of the initiating message. When the Snort is designed and enabled, we will continue to make improvements to its rules based with our own request and want. Because we need first look first and brightest for our Mac address and wireless connection. To proceed presently,

ifconfig

As we have got our details, ens33 is our network interface, and 192.168.0.12 is the IP address. We can now build the rules according to us within this location /etc./snort/rules/. Our rule to catch the ICMP packets is

```
alert icmp any any -> 192.168.0.12 any (msg: "ICMP Packet found"; sid:10,000,001;)
```

All we need to do after setting up the rule is checking the IP address of the Windows system with the aid of this command:

ipconfig

We have got Windows IP address which is **192.168.0.6** and follow the process by sending ICMP packets to victim with the help of ping command.

ping 192.168.0.12

Now, we need to activate the listener for conformation that our packets are heading toward right direction. By listening to the network operation with the aid of Wireshark, we can also verify if they send those packets or not.

As we can know, both request and response packets were caught in the Wireshark. Now, we are testing the Snort and Bravo!! Earlier rules we created worked perfectly fine, and Snort captured all the packets that came for our secured IP address.

We have successfully captured all the ICMP packets in our Snort settings our protected by IP address. In this experiment, in 26,000 attack attempts, we have achieved success rate of about 99%, and failure is about 1.01%.

11.3 Catch FTP Packets

That file transfer is a popular way of transferring data files between a communication protocol on a Web server. FTP is constructed on a network–server network model using independent plane and information relations between the sender and receiver.

alert tcp any any -> any 21 (msg: "FTP Packet found"; sid:10,000,002;)

alert tcp any any -> any 22 (msg: "SSH Packet found"; sid:10,000,003;)

alert tcp any any -> any 80 (msg: "HTTP Packet found"; sid:10,000,004;)

This rule will help us get these packets, and it will give us an idea for enforcing another rule as well after using this rule, as per our need. We have just tried find out our Kali IP address which is **192.168.0.9**. We are using the FTP rule this time. Check the IP address and start sharing the FTP packets with the help of NMAP tool.

**ifconfig
nmap -p 21 192.168.0.12**

Now, we need to validate this with the help of network listener like Wireshark. We can see that we captured successfully all request and reply packets of this conversation.

Like earlier, we have checked our packets similarly we checked them now. We can now check the packets that have been captured and yes!! We did it successfully captured them. In this experiment, in 25,000 attack attempts, we have achieved success rate of about 98.89%, and failure is about 1.112%.

12 Conclusion

IDS is now the natural next step for several companies before introducing gateway technology at the edge of the network. IDS will provide security against inner attackers and outsiders if there is no traffic coming through the barrier whatsoever. But the main statements are very important to note at all times. If these same points really aren't conformed to an IDS deployment together with a firewall unaccompanied will not create an extremely secure system.

12.1 Durable Verification

To encrypt data or possible abuse, an IDS uses very efficient excel spreadsheet methods; organizations must also maintain, but they provide clear customers with an in-place authentication and encryption framework.

12.2 Not Because All IPS Are Really an Answer to Safety Risks

IDS has an excellent job at checking the incidents at intrusion are monitored and documented. Businesses must use a process of staff training, device testing and debugging and commitment to a decent safekeeping strategy in instruction to diminish the risk of invasion.

12.3 An IDS Is No Substitution for a Sound Information Security

An IDS, as is the case for most protection and tracking systems, serves as one component of an organizational information security. It is another well approach to safeguard that intrusions besides bugs, malware infections, etc.

12.4 Personal Intervention Is Required

The system administrators or system administrator will examine an incident after event is identified and documented, control how it happened, address the problem and take the appropriate action to prevent the incidence of the very equal intrusion with fashionable potential.

13 Future Scope

The IDS is here to stay, I have no doubt, while future systems will certainly take a different shape than our current models. While hopefully, the concepts discussed here are attainable. The concepts of mathematics and artificial intelligence (AI) needed for success are already being developed, tested and enhanced. With the NIDES and EMERALD ventures, SRI has a great beginning, using a distributed model similar to what is described above.

References

1. Kim JW, Bentley P (2002) Artificial Immune Model
2. Dutt I, Borah S (2015) Some studies in intrusion detection using data mining techniques. *Int J Innov Res Sci Eng Technol* 4(7):5500–5511
3. Das N, Sarkar T (2014) Survey on host and network-based intrusion detection system. *Int J Adv Netw Appl*
4. Panigrahi R, Borah S (2019) Dual-stage intrusion detection for class imbalance scenarios. *Comput Fraud Secur* 2019:12–19
5. Dutt I, Borah S, Maitra IK, Bhowmik K, Maity A, Das S (2018) Real-time hybrid intrusion detection system using machine learning techniques. *Advances in Communication, Devices and Networking*, Springer, Berlin, pp 885–894
6. Dutt I, Borah S, Maitra IK (2020) Immune system based intrusion detection system (IS-IDS): a proposed. *IEEE Access* 8(2020):34929–34941
7. Cheung S, Duterte B, Fong MW, Lindqvist U, Skinner KM, Valdes A (2006) Using model-based intrusion detection for SCADA networks. Computer Science Laboratory SRI International (2006)
8. Bivens A, Palagiri C, Smith R, Szymanski B, Embrechts M (2002) Network based intrusion detection using neural networks. ANNIE
9. Lawal OB, Bitola AI, Lunge OB (2013) Analysis and evaluation of network-based intrusion detection using snort freeware. *Afr J Comput & ICT*
10. Jyotsna V, Rama Prasad VV, Munawara Prasad K (2011) A review of anomaly-based intrusion detection systems. *Int J Comput Appl* 28:26–35
11. Basha N, Bharani I, Shanmugam D, Ahmed AM (2007) Hybrid intelligent intrusion detection system. World Academy of Science, Engineering, and Technology, p 11
12. Devi Krishna KS, Ramakrishna BB (2013) An artificial neural network-based intrusion detection system and classification of attacks. *Int J Eng Res Appl*
13. Onashoga SA, Akinde AD, Sodiya AS (2009) A strategic review of existing mobile agent based intrusion detection systems. *Iss Inform Sci Inform Technol* 6
14. Shah B, Trivedi BH (2012–13) Artificial neural network-based intrusion detection system: a survey. *Int J Comput Appl* 6:13–18
15. Sodiya AS, Ojesanmi OA, Akinola OC, Aborisade O (2014) Neural network-based intrusion detection systems. *Int J Comput Appl* 106(18)
16. Moradi M, Zulkernine M, A neural network-based system for intrusion detection and classification of attacks. *Nat Sci Eng Res Council Canada*
17. Kanalayasiri U, Sangwanpong S (2015) Network-based intrusion detection model for detecting TCP SYN flooding
18. Staniford-Chen S, Cheung S, Crawford R, Dilger M, Frank J, Hoagland J, Levitt K, Wee C, Yip R, Zerkle D, GrIDS (2017) A graph based intrusion detection system for large networks. DARPA

19. Jirapummin C, Wattanapongsakorn N, Kanthamanon P (2018) Hybrid neural networks for intrusion detection system
20. Fuchsberger A (2005) Intrusion detection systems and intrusion prevention systems, information security technical report
21. Zeng J, Guo D (2009) Agent-based intrusion detection for network-based application. *Int J Netw Secur* 8(3)
22. Qassim QS, Zin AM, Ab Aziz MJ (2016) Anomalies classification approach for network-based intrusion detection system. *Int J Netw Secur* 18
23. Singh AP, Singh MD (2014) Analysis of host-based and network-based intrusion detection system. *I.J. Comput Netw Inform Secur*
24. Amini M, Jalili R (2018) Network-based intrusion detection using unsupervised adaptive resonance theory (ART). Iran Telecommunication Research Center (ITRC)
25. Labib K, Vemuri R (2018) NSOM: a real-time network-based intrusion detection system using self-organizing maps
26. Meftah S, Rachidi T, Assem N (2019) Network-based intrusion detection using the UNSW-NB15 dataset. *Int J Comput Digit Syst*
27. Branch JW, Bivens A, Chan CY, Lee TK, Szymanski BK (2002) Denial of service intrusion detection using time dependent deterministic finite automata, Research conference
28. Palagiri C (2017) Network-based intrusion detection using neural networks
29. Verma A, Ranga V (2019) Machine learning-based intrusion detection systems for IoT applications. Springer Science Business Media, LLC, part of Springer Nature

Performance Assessment of End-to-End Routing Protocols in Cognitive Radio Ad-Hoc Networks



Debabrata Dansana and Prafulla Kumar Behera

Abstract Routing in cognitive radio ad-hoc network is one of the important aspects in wireless communication. There are different routing protocols in cognitive ad-hoc network (CRAHN), and a comparative analysis of such protocols is required to find out the best possible routing protocol. In this paper, two data-driven routing protocols, namely ad-hoc on-demand distance vector (AODV) and weighted cumulative expected transmission time (WCETT), are evaluated and compared. For end-to-end routing, data processing is done using NS-2 simulator. CRCN patch analysis has been provided to validate the theoretical claims. The performance of WCETT and AODV is evaluated in terms of throughput, routing overhead, packet delivery ratio (PDR) and normalized routing overhead (NRL). The simulation result is found better in favor of WCETT as compared to AODV because of its better route selection strategy in CRAHN.

Keywords Routing protocol · Cognitive radio ad-hoc network (CRAHN) · Performance assessment · AODV · WCETT · NS-2

1 Introduction

In the field of communication technology, the advanced use of wireless computing and communicating devices has led to the overcrowding of the radio spectrum. On the authority of a study by the FCC, it has been disclosed that the spectrum utilizes only 15–85% of the time, and for the rest of the time, it remains unused. Many researchers have proposed cognitive radio networks (CRNs) as a way to deal with this. Channels for transmission are selected expediently in case of cognitive radio networks. This can be called as cognitive radio ad-hoc networks (CRAHNS) [1], where reliability in data dispersion is hard to achieve. One of the biggest challenges that a cognitive user faces is how to find a suitable spectrum without any interference from the primary node the spectrum belongs to. Also, CR nodes had to communicate

D. Dansana (✉) · P. K. Behera

Department of Computer Science and Application, Utkal University, Bhubaneswar, Odisha, India

along with the non-disturbing lines of PR nodes by limiting CR to PR interference. The main intention is to create a new channel selection strategy which will cause less interference with PR nodes than other strategies, which will maximize odds that the message is delivered, minimize the overhead and normalize routing overhead which will increase the data dispersal reachability.

Routing in CRAHN remains an important area of research. The routing blueprint in CRAHNS is bifurcated into two pigeonholes [2] based on the association linking the routing algorithm spectrum management and the spectrum management. In another proposal, the routing selection and the spectrum decision are executed independently by the MAC and network layers which are disassociated with each other. A routing protocol RACONA is presented, which suggests the inclusion of spectrum availability over time in a link which cost metric for each of the links [3]. We had created a new design CRAHN, which takes into account two routing metrics, i.e., spectrum decision and the routing selection, and they are merged together with network layers. The end-to-end route and the spectrum to be used are decided simultaneously for each node by the source node. In spite of the fact that the collaborative design is reactive to the mutation of spectrum, it yields better end-to-end execution and hold up quality of service (QoS)—hard and fast applications. The researchers had presented the SEARCH protocol, which blends geographical/topological routing and spectrum allocation to provide help in avoiding regions designated by PU recreation and also determines the foremost track-channel amalgamation to reduce the end-to-end delay [4]. A new path is suggested by knitting on-demand routing and spectrum band selection, and the amount of data to be transported is responsible for selecting the end-to-end route [5]. Within sight of the preceding strategy depend on the demand of routing strategy (like AODV [6]), which works by discovering a solitary/isolated pathway between a source/spring and a terminus node. In paper [789], the author focused on succeeding problem by taking different routing metrics on end-to-end performance. In paper [10], author done analysis on secondary isolated receiver and transmitted probability as well as path connectivity between different nodes. Channel quality indicator (CQI) [10] is proposed which is a MAC protocol in CRAHNs. A quality of services (QoS) and quality of experience (QoE) survey has been done in multimedia communication over cognitive radio network (MCRNs) [11]. Simulation helps us to build a mathematical model of the world and run it several times on a system. Data analysis helps us to analyze the real-time data using various statistical methods. Statistical analysis is required for the estimating performance properties and compares performance characteristics for the different approaches by analyzing the performance of the data, and we will identify the essential insights for making better decision.

In this research paper, a comparative analysis of two data-driven different routing protocols, namely AODV and WCETT, has been carried out for CRAHNs. Data processing and data validation for theoretical claims have been done by NS-2 simulating environment and using CRCN patch analysis. The research work is arranged as follows. Section 2 furnished theoretical background of different routing networks (AODV and WCETT), Sect. 3 illustrates the simulated results, and in the end, Sect. 4 concludes the research work and the future work.

2 Routing Scheme in CRAHNs

There are different types of routing protocols available in cognitive radio ad-hoc networks. A detailed discussion of different routing protocol is provided in [7–10, 12].

2.1 AODV

AODV is an on-demand, solitary path, loop-free interspaced vector protocol. It merges the on-demand route discovery mechanism into dynamic source routing with the concept of terminus sequence number from dynamic terminus sequenced to distance vector. AODV route update rules are as follows.

1. If $(seq_num_i^d < seq_num_j^d)$ or $(seq_num_i^d = seq_num_j^d \text{ and } hop_count_i^d > hop_count_j^d)$
2. $seq_num_i^d := seq_num_j^d$
3. $hop_count_i^d := hop_count_j^d$
4. $next_hop_i^d := j$
5. End.

Flowchart of RREQ has been shown in Fig. 1.

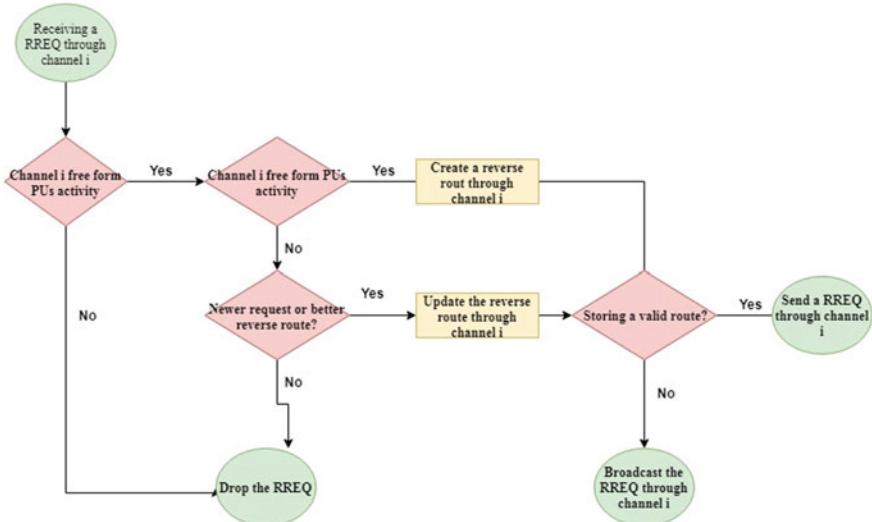


Fig. 1 Flowchart for RREQ

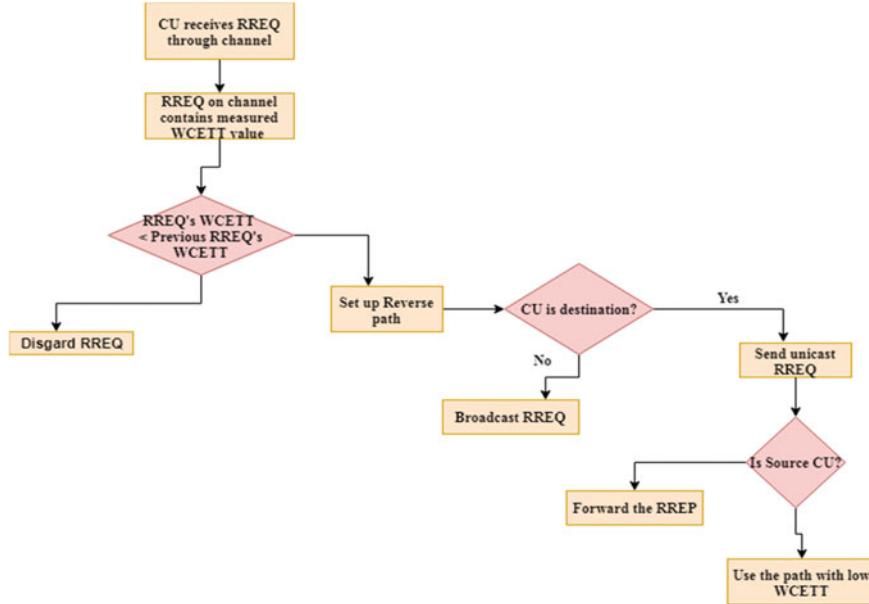


Fig. 2 Flowchart for WCETT routing

2.2 Weighted Cumulative Expected Transmission Time

This protocol works by assigning weight to each of the links that had provided time in the transference linked out expected transmission time (ETT). Therefore, the node must stand by for the channel resource's availability before data channeling because this might lead to longer detain and lower performance. A flowchart diagram of WCETT is represented in Fig. 2 for RREP and RREQ.

Let us assume a link denoted as ' i ' from node 'x' to 'y'. Now, here we have to compute ETT of the packet transmitted on this link whose efficacy is denoted by ETT (we are going to explain the computation of ETT in the upcoming subsections.). Next, we have to figure out a way to amalgamate the solitary ETT link weights of hops along a track into a metric that will reflect given optimum path.

The metric we have taken into consideration is described as weighted cumulative ETT (WCETT). Keeping the objective in mind, what we want for WCETT is to increment in worth as we keep on adding more and more links to an existing route. This can be done by setting WCETT as the total no of ETT's hops. Excluding that, the summation of ETTs also has a physical denotation, i.e., a packet which travels along the path; it experiences some end-to-end delay.

So, for path of having n numbers of hops,

$$\text{WCETT} = \sum_{i=1}^n \text{ETT}_i \quad (1)$$

Regarding an n -hop path, which consists of total k channels, X_j is represented as

$$X_j = \sum_i \text{ETT}_i \quad 1 \leq j \leq k \quad (2)$$

Hence, X_j is taken as the summation of the total transmission times of the hops on channel j . The bottleneck channel which has the largest X_j will dominate the throughput.

So,

$$\text{WCETT} = X_j \quad (3)$$

It is noticeable that the paths which are more channel-diverse will be favored by this metric. Therefore, increment in number of the hops to the paths will not always increase the value of the metrics. It is because the value of the metric is not affected by the additional hops, which do not use bottleneck channels.

Thus, the advantageous properties of the two metrics, expressed in Eqs. (1) and (3), can be merged by taking their weighted average:

$$\text{WCETT} = (1 - \beta) \sum_{i=1}^n \text{ETT}_i + \beta X_j$$

where β is a tunable parameter subject to $0 \leq \beta \leq 1$.

3 Simulation and Result Analysis

Network Simulator-2.31 (NS-2.31) with CRCN [12] patch is used for simulation results. Table 1 provides relevant simulation parameters.

In this section, IEEE. 802.11 with FTP as a traffic source has been used, and the evaluation of the performing routing protocol AODV and WCETT is done by using multiple random topologies named as *random.tcl*. Whereas, secondary users placed

Table 1 Simulation parameters

Topology	$500 * 500 \text{ m}^2$
Traffic type	File transfer protocol
Packet size	512 bytes
Simulation time	50 s
Node speed	20 m/s
MAC layer	IEEE 802.11
Transport layer	Transfer control protocol

in $500 \times 500 \text{ m}^2$, and we can run each simulation for 50 s by adding a number of nodes in the subsequent simulation.

Figure 3 shows the throughput of two different routing protocols, i.e., WCETT and AODV over time. Analyzing the various data of different protocols, we can see the trends at different time interval. We can do future prediction by looking to the past trends. It is observed that throughput of WCETT protocol is better than AODV protocol.

Various performance metrics such as throughput, PDR, NRL, and routing overhead have been considered for the performance evaluation of routing protocols through data processing. The processed data of AODV and WCETT protocols has

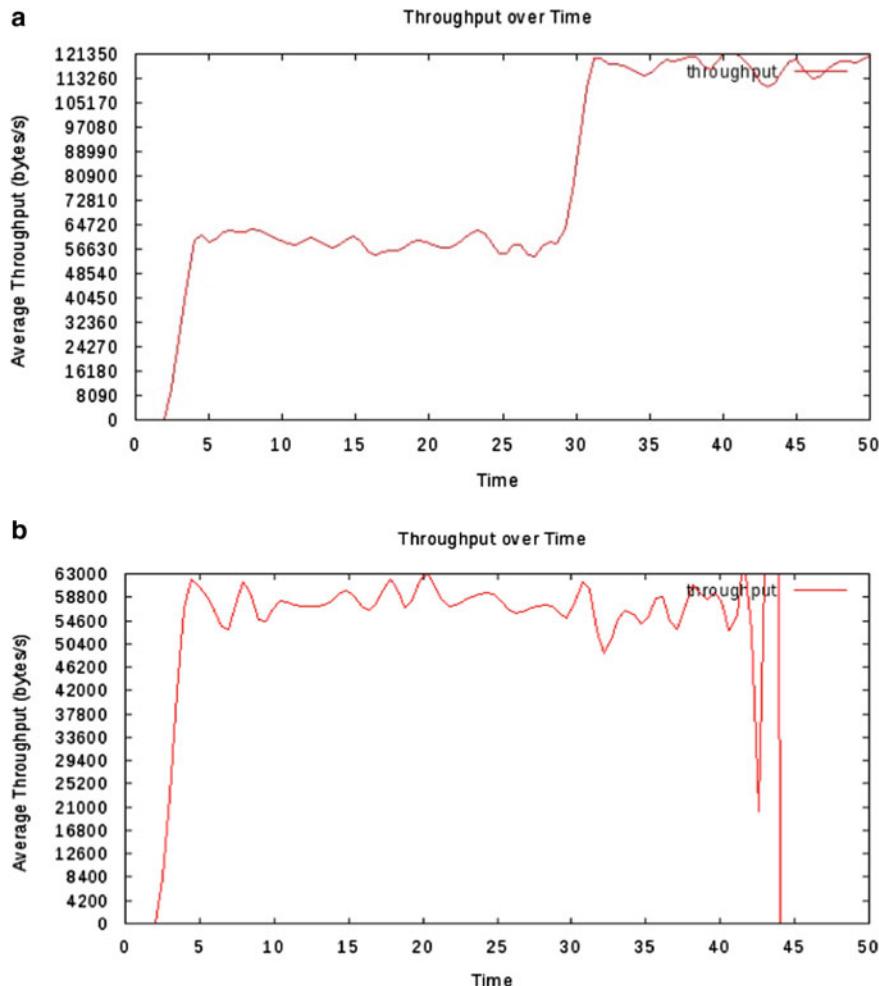


Fig. 3 Throughput of **a** WCETT and **b** AODV

been represented in Tables 2 and 3, respectively. Interference and channel have been taken as constant, and overhead, NRL, PDR and delay have been varied.

Throughput is the ratio between numbers of sent packet to the total time, whereas PDR is the ratio between number of packets delivered and number of packets sent. NRL is the ratio between total numbers of routing packet to packet received. Figure 4 shows different comparative analysis of AODV and WCETT protocols. Different parameters such as overhead, PDR and NRL have been compared with respect to the number of nodes. From Fig. 4a, it is observed that for different number of nodes, routing overhead for AODV is higher than WCETT. From Fig. 4b, it is observed that the PDR of the WCETT-based routing is constantly higher than AODV routing. The network which has multiple channel support shows the highest number of throughput and lower delay than a single channel network. Hence, in terms of average throughput,

Table 2 Data representation of AODV protocol

AODV	Node varies from 10 to 100							
nodes	Send	Received	Overhead	NRL	PDR	Interface	Channel	Delay
10	5123	5110	3128	0.61	99.74	1	5	0.01
20	4075	4067	8390	2.06	99.8	1	5	0.012
30	2486	2443	17,085	6.99	98.27	1	5	0.027
40	1838	1813	24,472	13.49	98.63	1	5	0.031
50	1822	1813	28,896	15.93	99.5	1	5	0.033
60	1819	1802	34,434	19.1	99.06	1	5	0.028
70	1634	1598	39,488	24.71	97.79	1	5	0.02
80	1696	1650	44,738	27.11	97.28	1	5	0.024
90	1897	1884	47,483	25.2	99.31	1	5	0.024
100	1557	1520	54,294	35.71	97.62	1	5	0.028

Table 3 Data representation of WCETT protocol

WCETT	nodes varies from 10 to 100							
Nodes	Send	Received	Overhead	NRL	PDR	Interface	Channel	Delay
10	5587	5564	530	0.095	99.58	1	5	0.00924
20	5448	5444	1060	0.194	99.92	1	5	0.00967
30	5342	5331	1592	0.298	99.79	1	5	0.00985
40	5275	5255	2120	0.403	99.62	1	5	0.00998
50	5194	5186	2651	0.511	99.84	1	5	0.01013
60	5176	5151	3180	0.617	99.51	1	5	0.01008
70	5127	5117	3713	0.725	99.8	1	5	0.01015
80	5100	5100	4240	0.831	100	1	5	0.01018
90	5095	5069	4772	0.941	99.48	1	5	0.01019
100	5076	5046	5303	1.05	99.4	1	5	0.01015

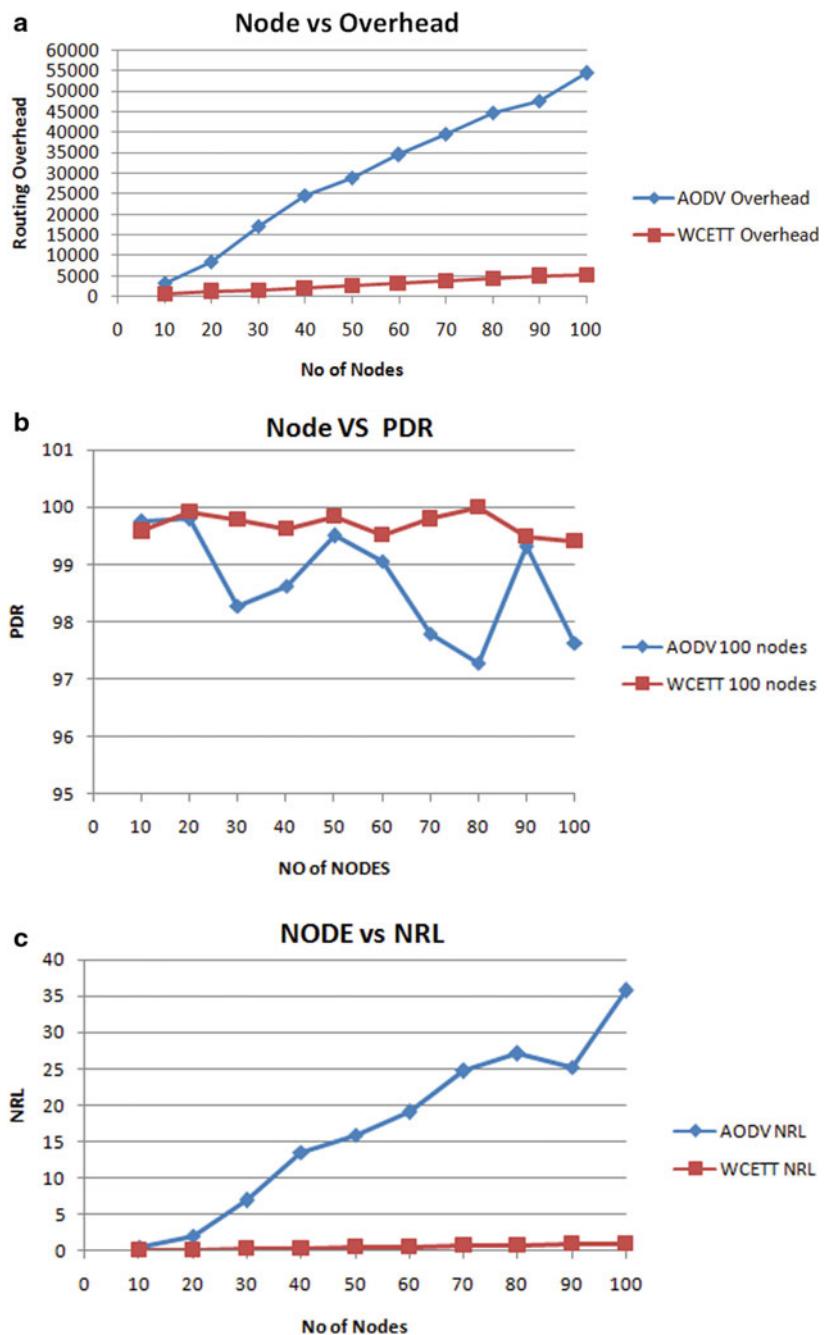


Fig. 4 Performance comparison. **a** Overhead. **b** PDR. **c** NRL

the overall performance of the network and packet delivery ratio are being increased. From Fig. 4c, it is observed that normalized routing overhead (NRL) of the WCETT throughout the running time of the simulation is based on the routing and remains higher than AODV routing throughout the running time of the simulation.

4 Conclusion

In this research work, performance of data-driven end-to-end routing in CRAHNs is evaluated between AODV and WCETT protocols. CRCN is used with NS-2.31 simulator for performance analysis. By analyzing behavior of different protocol on various data, it can be observed that the throughput of WCETT is higher than AODV. With the increase in node numbers from ten to hundred, WCETT protocol performed better as compared to AODV protocol in routing overhead, PDR and NRL, and this is because of optimum route selection process of the WCETT. As security plays a vital role in wireless communication, the future work of this research work is to give the security to improve the QoS and secure the communication.

References

1. Akyildiz IF, Lee W-Y, Chowdhury KR (2009) CRAHNs: cognitive radio ad hoc networks. *Ad Hoc Netw* 7(5):810–836
2. Wang Q, Zheng H (2006) Route and spectrum selection in dynamic spectrum networks. *Proc IEEE CCNC 2006*:625–629
3. Talay AC, Altilar DT (2009) RACON: a routing protocol for mobile cognitive radio networks. In: *Proceedings of ACM CORONET 2009*, Bejing, China, pp 73–78
4. Marina MK, Das SR (2006) Ad hoc on-demand multipath distance vector routing. *Wirel Commun Mob Comput* 6:969–988
5. Perkins CE, Belding-Royer EM (1999) Ad hoc on demand distance vector (AODV) routing. In: *Proceedings of IEEE WMCSA 1999*, New-Orleans, USA, pp 90–100
6. Floyd S, Henderson T (1999) The NewReno modification to TCP's fast recovery algorithm. In: Internet engineering task force, request for comments, experimental 2582, April 1999
7. Kulkarni S, Markande S (2015) Comparative study of routing protocols in cognitive radio networks. In: 2015 international conference pervasive computing, pp 1–5 (2015)
8. Chowdhury KR, Akyildiz IF (2011) CRP: a routing protocol for cognitive radio ad hoc networks. *IEEE J Select Areas Commun* 29(4):794–804
9. Salim S, Moh S (2013) On-demand routing protocols for cognitive radio ad hoc networks. *EURASIP J Wirel Commun Netw* 1–10
10. Dynamic Channel Assignment MAC Protocol for Cognitive Radio Ad Hoc Networks (CRAHNs) (2019) *Int J Recent Technol Eng* 8(4):10446–10452. <https://doi.org/10.35940/ijrte.d9170.118419>
11. Jalil Piran M, Pham Q-V, Islam SMR, Cho S, Bae B, Suh DY, Han Z (2020) Multimedia communication over cognitive radio networks from QoS/QoE perspective: a comprehensive survey. *J Netw Comput Appl* 172:102759. <https://doi.org/10.1016/j.jnca.2020.102759>
12. Singh K, Moh S (2016) Routing protocols in cognitive radio ad hoc networks: a comprehensive review. *J Netw Comput Appl*

13. Chowdhury KR, Di Felice M, Akyildiz I (2009) TP-CRAHN: a transport protocol for mobile cognitive radio ad hoc networks. Proc IEEE INOCOM 2009:2482–2490

Unstructured Log Analysis for System Anomaly Detection—A Study



Anukampa Behera, Chhabi Rani Panigrahi, and Bibudhendu Pati

Abstract Nowadays, with the rapid pace of innovation, a typical production infrastructure is getting huge, complicated and difficult to manage. Hence, incident detection and action have become a challenge to the operations and information security (InfoSec) teams. As we are moving toward deployments of complicated or complex large-scale micro-service architectures, the kind of data generated from all those systems is huge. So, it becomes very difficult to identify if anything goes wrong in underlying systems, i.e., the system is vulnerable to various attacks. Keeping track of the flow of traffic and user activities on a large scale, complicated environment is becoming very costly and unmanageable. Traditional systems and ways are becoming inefficient for zero-day security issues. So it is highly recommended to develop a system that is capable of raising an alarm for any detected anomaly after performing an automatic analysis of the generated logs. This study is conducted to review the research work on unstructured log analysis for the purpose of monitoring the system and anomaly detection. We have identified the datasets used for this purpose and also pointed out the challenges involved in unstructured data analysis.

Keywords Log analysis · Micro-service · Log parsing · word2vec · temp2vec · Natural language processing

1 Introduction

Recently, development of Web applications is preferably done using micro-service design for better scalability, flexibility and reliability. This approach has a paradigm shift from monolithic application architectures to micro-service architectures with the purpose to improve scalability and have better reliability features in applications.

A. Behera

Department of Computer Science and Engineering, ITER, S'O'A (Deemed to be), Bhubaneswar, India

A. Behera · C. R. Panigrahi (✉) · B. Pati

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

Here, there are numerous small (micro) services that are interconnected with many other micro-services to provide complex services like Web applications [1]. Many of the world Web business leaders have migrated most of their infrastructure to micro-services. A set of services that are secluded, failure supple and scalable leads to an application built on micro-service architecture. Here, each of the services is a full-fledged application which communicates with other similar service [2]. It has become quite a profitable venture for companies those have implemented micro-service architecture in terms of faster release of software, thus making the team size smaller and enabling them to concentrate on the unit work they have been assigned with [3]. Distribution of applications over numerous systems that may be implemented virtually or physically has resulted in better scalability and reliability. The very much in-practice agile DevOps style can easily adopt this because of its low vulnerability to error while upgrading. As it is a collection of services, the testing also needs to be conducted only for the upgraded components instead of the entire application which yields a higher rate of automation in production units, enabling them for shorter product release cycle and rapid recurrent deployments [4].

In spite of being so advantageous, micro-service architecture has put forth many challenges especially in the area of performance monitoring where the problems must be detected before it puts the entire system under risk [5]. As the applications are distributed over numerous virtual machines, the monitoring of nodes becomes a humongous task. Due to the frequent upgradation of services and multiple instances of the same application with different version deployed simultaneously, it becomes quite challenging to detect and forecast anomalies based on historical data. Moreover, detecting anomalies manually becomes quite unmanageable and does not result in much efficiency [6]. First of all, the system administrators need to manually go through a thorough examination of these huge logs generated every minute by a system that is distributed at a large scale, looking for deviations. Lack of detailed system knowledge comes next. The developer who is associated or the administrator in-charge of the system may not have adequate knowledge about the designing and functioning of the system, as commercial-off-the-shelf components are often used by many big enterprises. As distributed systems become more complex, they contribute to the deterioration of the efficiency of diagnosing problem manually [7]. Thus, the need for development of artificial intelligent-based automatic anomaly detection and monitoring system becomes a standard and compulsory requirement of many distributed systems so that the quality of service can be ensured. In this paper, a study has been done for all the works that have been done till date for unstructured log analysis and automatic anomaly detection. A summarization of various methods researchers have suggested, are presented for the same. The remaining of the paper is organized as follows: In Sect. 2, the importance of unstructured log analysis is discussed, Sect. 3 presents the approaches used in the literature for log analysis, and in Sect. 4, the datasets used for this purpose are listed. In Sect. 5, some specific challenges that may arise during the log analysis process have been identified, and Sect. 6 represents the conclusion.

2 Background

“If data had mass, the earth would be a black hole.”—told by Stephen Marslan—hits in the right direction toward which researchers must orient our research. By the term big data, we normally visualize a large portion of the huge structured as well as unstructured data, and it is nothing but machine-originated log data. In large applications, from security to debugging and troubleshooting, every phase lots of logs are generated. Within these logs, lots of useful information are available for which proper action must be taken to ensure system security and smooth functioning. But it is only possible if the logs are properly stored, managed and analyzed.

According to the survey done [8], it shows that above 50% of the digital companies that have successfully deployed big data projects are using some kind of log management. And for majority of them, log management is done in a priority basis as for the purpose of network-related observations, and log-files are considered as the prime source of data. As per DevOps glossary, the data file that contains information about various blueprints in which service is used, all the actions performed and jobs done when a system is operational, is termed as log-file. A log-file is generated by the computer.

These logs can be recorded in various forms—structured, semi-structured and unstructured [9].

Structured Data: These are formatted and organized data. Example: Content of a table in a relational database system, spreadsheets where one entity is represented by each row and related records.

Semi-Structured Data: These are generally tagged data. Here, unlike relational database management system structured representation does not exist, the data are either tagged or marked by any other pointers that are available. These tagging or pointers separate semantic units, and the related fields and related information within the data are hierarchically organized. Example: json or xml files.

Unstructured Data: Here, none of the data are organized in any pre-defined manner, and mostly data are in the continuous text format. It might contain so much of important information, but as they are not organized, it is very difficult to extract them.

When the need of the system is to process peta bytes of events generated related to security every day, if traditional approach is followed, it could definitely take a very long time even turning to days or a week to compile and analyze all the data that are available in logs. The source of the logs may be countless endpoints on various sensor grids where each such sensor grid might be having different log structures. In order to handle the volume, variety and velocity (3 vs), the only affordable and scalable approach is to introduce it to an efficient log management system.

Some applications of efficient log analysis

- If logs can be properly monitored, they can provide the insight of the system’s hardware and software health, real-time statistics on the load bearing capacity of applications and servers—which is helpful in improvement of overall business

intelligence [10]. Organizations, especially IT, can improve their system reliability for the end users by analyzing the log-file about their system performance. It helps them to understand when capacity enhancement is needed to optimize the user experience. One can reach to the root cause of the problems like slow processing of queries, transaction errors or some application bugs that is having a reverse impact on the service provided. This information can be used to generate alert for hardware upgradations and software updates.

- In a networked or cloud environment as network logs [11] are available, efficient monitoring can help in keeping a watch on data breech and can take care of the security posture of cloud computing environments which is a very serious aspect. In these types of logs, one can trace things like unsuccessful log-in attempts, failed user authentication or unexpected server overloads, all of which can generate an alert for a cyber-attack to be in progress. So, any type of deviation found in the system can generate an alert for a possibility of a security breech.
- Log analysis adds a new dimension in opening new opportunities for business analytics [12]. For example, when any customer makes a transaction in an e-commerce site, his/her entire footprint is captured which when analyzed helps in understanding the behavior of the customer, his/her likings and disliking, their life styles and quality of the products that they purchase. These all help the organizations to give better and more customized service to their customer, hence contributing a lot in the all dimensional growth of their business.

Depending on the various services provided, organizations need to adhere to standards set by the government and requirements from the industries to give assurance on the functionality as well as safety. Recording the data log and analyzing them further, help the organizations to take a safeguard against any type of internal as well as external threat. Besides playing a vital role to ensure cyber security, forensic investigations in case of any fraudulent activities, various audit requirements can also be taken care of by using log-files analysis [13].

3 Anomaly Detection Approaches

Xu et al. [14] proposed an approach to detect problem by using console log analysis and the built-in monitoring service provided by software systems. They claimed that by a better preprocessing even using simple algorithm better results can be yielded. To raise the accuracy in log parsing, they used source code as a reference for better understanding of unstructured log data. It helped to extract identifiers and state variables which are normally not considered while log parsing making the process more difficult. With this, the complexity was reduced and noise in any form was removed. Hence, it solved many challenges posed in machine learning related to constructing of features. In this approach, frequent pattern was used to filter general events, and PCA detection is used for anomaly detection.

Guang et al. [15] in their work focused on detecting the program invariants. Invariants generally refer to a condition or logical rule which should be satisfied, while a code is executed. These invariants may be used along with some given pre- and post-condition for assertion. They used following pre-processing before proceeding toward detection of anomalies. Firstly, when console log was received, the log messages received in unstructured format were passed through the log parser in order to convert them to tuple form. Next, these log messages were grouped, and a count vector was calculated for each such group. Then, every message counter vector was compared with the learnt invariants, and for those who deviate from them, anomalies were raised. They claimed that their suggested method performs better than PCA-based methods and detected anomalies with finer details which could provide insight cues for problem diagnosis.

Breiber et al. [16] used some popular data mining methods to detect anomaly. Their work was based on the heuristic scanning, activity monitoring and integrity check methods suggested by earlier researchers. Heuristic scanning includes searching for any unfamiliar or atypical command in an execution of a regular program. Activity monitoring focuses on the sequence of files executions, monitoring the process – typically behavior of APIs, various blended data sources and system calls, to point out if anything unusual is found. Integrity checker checks the stored files or network packets for alteration by making a hash computation each time matching with the earlier stored checksum. To detect anomalies, they used an anomaly profile by creating rules from transactions blocks which were collected across various log sources. They extracted session start time, duration of the session and service type from log to create a transaction block which was converted to binary information using standard algorithms. An identifier was used for recognizing the log record spatially. Here, Hadoop technology was used to reduce the log analysis time. They have tested the suggested algorithm to perform better on Hadoop by showing an error rate below 0.1% in Hadoop than that was implemented in Java.

Tuor et al. [17] developed a language modeling framework for detection of cyber anomaly by exploring the use of recurrent neural network (RNN)-based language models. The network language model was updated on the run based on the events that have occurred the day before. If it found any event that had a low probability of occurrence, that event was flagged as an anomaly. This model reduced the deployment time, gave an optimized solution for analysts by providing them related and specific event to evaluate and also was capable of processing the real-time events.

After this, natural language processing was introduced for anomaly detection in unstructured logs. Du et al. [18] proposed DeepLog, a deep neural network which was highly influenced by natural language processing. It used a long short-term memory (LSTM) to model the sequences of log entries based on a certain pattern and syntax. So, DeepLog could learn the pattern from executions that were termed as normal and raised anomaly for all deviations captured. They had also proposed to upgrade the model incrementally to learn and handle new log patterns. Unlike many methods that have been suggested, DeepLog performed anomaly detection for each log entry rather than grouping them to sessions.

Wang et al. [19] used the Thunderbird generated logs for giving a comparative analysis of LSTM, GBDT and Naïve Bayes algorithms. Here, first oversampling was applied to the cleaned logs to get an equilibrium sample. The logs were divided to two classes—exception and normal class, based on the whether they contained typical exception-based keywords. In this work, for extraction of features, skip-gram model of word2vec and TF-IDF were used. If word2vec was used to extract features, then GDBT and Naïve Bayes gave better result. LSTM works in the detected relationship between contexts and was not affected by any type of feature selection method adopted. This study gave more importance on feature reduction algorithm. The study concluded that for semantic information extraction from logs, word2vec was much more effective than other methods.

Lu et al. [20] presented a comparative study of CNN, LSTM and MLP. Their experiment shows that convolutional neural network gives better result in terms of higher and faster detection of anomalies on big data logs. In order to obtain the key and a data router based on observations for log key sequences as per the execution order, on raw log data, a two-level parsing mechanism was applied.

Brown et al. [21] established relational mapping between features drawing important features using attention mechanism. They proposed various attention modes such as fixed, syntax, semantic and tiered attention model.

Farzad et al. [22], for the purpose of detecting anomalies and further classifying them, proposed Auto-LSTM, Auto-BLSTM and Auto-GRU models. They had also given a comparative analysis between earlier known models and their proposed models when applied on data sets obtained from various high-performance servers and super computers. In the proposed architecture, they converted the log message that is in text form to numeric form by using word frequency. To train the encoder, they had labeled the entire dataset into two parts like the normal data are labeled as positive and those with abnormal behavior are labeled as negative. Next, the positive and negative encoder's outputs were concatenated which after duplicate and noise removal were used for proposed deep learning algorithm.

Meng et al. [23] proposed a unified data-driven deep learning framework inspired by word embedding. They used an LSTM model that was attention-based, for the sequential and quantitative anomaly detection simultaneously. To extract the semantic for an effective detection of anomaly, template2vec—a synonym–antonym-based novel-word representation method was proposed. As instead of using only template index they also captured on the semantic information, the problem of false alarm was reduced. For extracting temporary templates from a new log, they used FT-Tree which calculated the template vector. It was then matched with the existing template vectors to find similarity. It found out whether the logs captured were an entire new log or were existing log with a small variation. This approach took care of the challenge of new log templates being generated between two adjacent periodic re-training.

Zhang et al. [24] proposed an anomaly detection approach, namely LogRobust. It searched the log events for extracting semantic information and presented them as semantic vectors. They used Bi-LSTM model that was attention-based for detecting anomalies. This model learned the significance of various log events by capturing

contextual information in the log sequence and thus got the capability to mark out unstable log events and sequences. The challenge of log instability and handling ever changing real-world data could be handled by this approach.

Wang et al. [25] proposed a deep neural network model. It predicted the number of level-1 logs currently available by using long-term memory (LSTM). As it compared the actual logs with predicted number of logs, any surge could be reported. They analyzed the router logs and list information about attributes and their status for training their proposed model which was based on unsupervised learning algorithms. Primarily, they used isolation forest [26], OneClassSVM [27, 28] and density-based algorithm LocalOutlierFactor [29] to identify the cause for the surge in logs.

In Table 1, the summary of the works done till now in the field of anomaly detection by unstructured log analysis is presented.

4 Datasets Used

As most of the logs are generated by various applications running across the globe, from security and privacy point of view these logs cannot be made public. Thus public datasets are unavailable for research purpose. So, before discussing the anomaly detection part, first the challenge of finding a proper dataset needs to be taken care of. Oliner and Stearley [32] have first published five system logs obtained from the Top500 Supercomputer list with the log features. A gist of the log characteristics is shown in Table 2.

Zhu et al. [33] had used sixteen datasets for log analysis from various sources like distributed system logs, operating system logs, logs generated from mobile systems, various applications running on server as well as applications running stand alone. Hadoop distributed file system (HDFS) was used as a standard benchmark by Breier et al. [16], Lou et al. [15] and Xu et al.[14] in their works. In total, 24,396,061 log messages were generated from 29 log events. Zhang et al. [24] had used data produced from the industry (Microsoft) and a hug log collection from Loghub [34] where 440 millions of log messages are available which are around 77 GB in size.

5 Challenges to Deal with

In this study, various challenges were identified that were faced during log analysis areas of applications.

Debugging: While logs help a lot in debugging an application, the researchers have found out statistical anomaly detection to be challenging. Even after a log message is found to be anomalous, it might be a single instance but at the same time contributing a lot to the security threat. But as the logs are unable to provide any further evidence, sometimes it is not possible to distinguish whether the message is the cause of the problem, or it is the symptom that is shown for a further bigger problem or simply

Table 1 Summary of the works on anomaly detection by unstructured log analysis

Year	Author(s)	Solution/model proposed	Dataset used for experimentation	Method used for pre-processing/parsing	Method(s) used
2019	Wang et al. [25]	LSTM	Netengine40E Router Log	Parsing based on attribute behavior	Directed graph
2019	Meng et al. [23]	LSTM	BGL HDFS	FT-Tree [30]	Template2Vec
2019	Farzad [22]	Auto-LSTM Auto-BLSTM Auto-GRU	BGL IMDB Openstack Thunderbird	NA	Counting word frequency
2019	Meng et al. [23]	LSTM	BGL HDFS	FT-Tree [30]	Template2Vec
2019	Farzad [22]	Auto-LSTM Auto-BLSTM Auto-GRU	BGL IMDB Openstack Thunderbird	NA	Counting word frequency
2019	Meng et al. [23]	LSTM	BGL HDFS	FT-Tree [30]	Template2Vec
2019	Zhang et al. [24]	Attention-based Bi-LSTM—LogRobust	HDFS Industry data (Microsoft) Data from loghub	Drain [31]	FastText
2018	Brown [21]	attention mechanism on LSTM	LANL Cyber-security dataset	Tokenization	NA
2018	Wang [19]	LSTM	Thunderbird	Noise removal	Word2vec, TF-IDF
2018	Lu [20]	Model based on CNN	HDFS	Logs-key sequences, session key	LogKey2Vec
2017	Tuor [17]	LSTM model based on RNN	Los Alamos National Laboratory (LANL) Kent 2016	Log-line tokenization	Language models
2017	Du [18]	DeepLog	HDFS Openstack-log	Spell [30]	Value for parameter work-flow
2015	Breier [16]	Heuristic scanning, integrity check	DARPA intrusion detection evaluation set, Snort logs	MapReduce	NA
2010	Lou [15]	Invariant	Hadoop CloudDB	Log parsing	NA
2009	Xu [14]	PCA	Darkstar Hadoop (HDFS)	Log parsing	Frequent pattern-based filtering

Table 2 Gist of log characteristics for system logs generated from Top500 Supercomputer

System name	Supplier	Days active	Produced log size (GB)	Production rate (bytes/s)	No. of messages produced	No. of alerts produced
Blue Gene/L	IBM	215	1.207	64.976	4,747,963	348,460
Thunderbird	Dell	244	27.367	1298.146	211,212,192	3,248,239
Red Storm	Cray	104	29.990	3337.562	219,096,168	1,665,744
Spirit (ICC2)	HP	558	30.289	628.257	272,298,969	172,816,564
Liberty	HP	315	22.820	835.824	265,569,231	2,452

harmless. Moreover, as statistical methods are heavily dependent on the quality of logs; they count a lot on the logging of events those are termed as important. But at the same time, the problem remains in finding the important events as the methods cannot self tag themselves as “important” on their own.

Performance: In the process of debugging or optimizing system performance, log analysis is quite helpful. If the way resources of the system are used can be tracked, then one can get an insight of the performance of the system. According to Oliner et al. [10], two major contributing factors are the system environment and how much workload the system has. The component interaction, most of the times, raises to the performance problem. In order to trace these interactions, many types of heterogeneous logs generated from various resources need to be synthesized. By heterogeneous logs, we mean logs in different formats and may be recorded in different time zones across the world. Hence, while synthesizing the ordering of events on the basis of time when they have occurred, becomes very difficult and challenging. Sometimes, a general harmless routine activity, performed on one component, which is not logged, may cause serious trouble for the adjoining component such as flushing a log to the disk. As the log is not generated, it is impossible to trace to the root of the problem.

Oliner et al. in their another work in 2010 [35] have suggested a method of computing influence. It finds out interrelated components in a group of components by tracking the anomalous behavior exhibited within a single time slot. But, in order to implement this method, the logs need to be maintained in a very detailed manner; which will impact adversely the performance of the system.

As a solution to the above-mentioned situation, instead of maintaining detailed log, a sample log may be used, which again has a underlying risk of missing out important logs! Erlingsson et al. [36] have suggested Fay, which collects, processes and analyzes software execution traces. Here, the user can specify the events that need to be measured which are formulated as queries based on which instrumentation is inserted to the running system, measurements are aggregated and analysis mechanism is provided. Oliner et al. have recommended writing dynamic programs with sample-based logging method to cater to challenge posed by maintenance of detailed logs at a scale.

Security: In the regular intrusion detection, system security is checked based on the signatures. Signatures uniquely represent malicious activities, anomalous behavior and the complete footprints, where it differs from one trusted user to another. But this type of system works fine with known attacks where it can match with the existing database of signatures. But in case it is a new attack signature method may not be useful. In such a situation, anomaly detection can be used, but it will report all types of deviations recorded. Setting a mark for an activity to be termed from anomalous to suspicious is difficult. So, it might end up in raising too many true-negative cases. Yuan et al. [37] have mentioned about “adversary” as a big challenge. Here, an adversary will follow the same footprints as that of a trusted operation, ending up generating almost the same log that is generated by a regular normal process. They have suggested that during development of an application the developers can design to create more detailed and elaborative logs so that it will not be possible for an adversary to follow the same log without showing up the deviation in the user behavior.

Prediction: As per the analysis given by Oliner et al. [10], by analyzing the logs, one can get to know the increasing/decreasing load on the system in terms of traffic, resource sharing requirement and how the system is dealing with it. Using the logs, predictive models can be built that will help an organization in planning its resources, in taking decision like whether to enhance capacity or not etc. It can also help business organization for making better marketing strategies, fund distribution as well as help them to manage the inventories in a profitable way. But the challenge that is faced here is that, in order to use this, one must have enough domain knowledge about the specific system. If the system upgrades to a new one or becomes complex, the model will not be able to predict as per the upgraded standard. Ganapati et al. [38] have suggested to use statistical learning techniques for performance prediction. Still they pose some challenges like statistical methods can process numeric data, but many a times, the logs contain lots of categorical data. Conversion from categorical data to meaningful numeric information requires proper domain knowledge and is a tedious process. Thus, taking a correct course of action from prediction given becomes challenging.

All these years despite many techniques have evolved for efficient log analysis, still some challenges remain. In the greater use of the large-scale micro-service architecture where a system comprises many smaller systems, use of a single log-file to monitor events from all these sub-systems is not possible. Moreover, the sub-systems may generate heterogeneous logs. In this scenario, cross-correlation becomes very difficult.

The logging process itself needs some scrutiny in order to control the long-windedness of logging in case of spikes or adversarial behavior. Some malicious activities may be propagated using logging as a tool. So, it poses a challenge to extract maximum information from the log content while minimizing instrumentation.

The problem of log instability is mentioned by Zhang et al. [24]. They have found the source of this problem as the generation of various logs by modifying the source code and while trying to handle the impurities in big data. Whenever a

task is performed, to generate the logs, generally no rules are followed which create non-standard logs. So, a generic approach to handle logs is very difficult to propose.

From the above discussion, a conclusion can be established that it is almost impossible for humans to comprehend the log data manually and detect any anomaly if found.

6 Conclusion

Despite of several challenges posed in unstructured log analysis, continuous research and brainstorming are going on since several years. Researchers are adopting various machine learning algorithms to handle this problem. Recently, DeepLog analysis using natural language processing has taken the lead. Various datasets that researchers have used have been listed for further use, as getting a dataset for experimentation in case of unstructured log analysis is a big challenge as mostly they are industry generated and are not shared due to privacy policies. This work is a summarization work done in the field of anomaly detection in unstructured log analysis. Here, a study has been done on various different methods researchers have proposed with a comparison and the challenged they have handled. In our study of various publications done till date, it is observed that natural language processing methods using deep learning are producing more accurate results than the other methods.

References

1. Available <https://dzone.com/articles/what-is-microservices-an-introduction-to-microserv>. [Online]. Last accessed on 03/12/2020
2. Dragoni N, Giallorenzo S, Lafuente AL et al (2017) Microservices: yesterday, today, and tomorrow. In: CCIS, editors. Present and ulterior software engineering. Nizwa. Springer, pp 273–278
3. Balalaie A, Heydarnoori A, Jamshidi P (2016) Microservices architecture enables DevOps: migration to a cloud-native architecture. IEEE Softw 33(3):42–52
4. Aderaldo CM, Mendonça NC, Pahl C, Jamshidi P (2017) Benchmark requirements for microservices architecture research. In: IEEE/ACM 1st international workshop on establishing the community-wide infrastructure for architecture-based software engineering (ECASE), Buenos Aires, pp 8–13
5. Du Q, Xie T, He Y (2018) Anomaly detection and diagnosis for container-based microservices with performance monitoring: 18th international conference, ICA3PP 2018, Guangzhou, China, Nov 15–17, proceedings, part IV
6. Fu Q, Lou J, Wang Y, Li J (2009) Execution anomaly detection in distributed systems through unstructured log analysis. In: Ninth IEEE international conference on data mining, Miami, FL, 2009, pp 149–158
7. Jayathilaka H, Krantz C, Wolski R (2017) Performance monitoring and root cause analysis for cloud-hosted web applications. In: Proceedings of the 26th international conference on World Wide Web—WWW’17
8. Available <https://www.csoonline.com/article/2935362/log-management-is-leading-use-case-for-big-data.html>. [Online]. Last accessed on 09/12/2020

9. Available <https://www.graylog.org/post/turning-unstructured-data-into-structured-data-with-log-management-tools>. [Online]. Last accessed on 03/12/2020
10. Oliner A, Ganapathi A, Xu W (2011) Advances and challenges in log analysis: logs contain a wealth of information for help in managing systems. Queue 9, 12:30–40
11. Grace LKJ, Maheswari V, Nagamalai D (2011) Web log data analysis and mining. Advanced computing, pp 459–469
12. Available <https://www.graylog.org/post/how-big-data-and-log-management-work-hand-in-hand>. [Online]. Last accessed on 05/12/2020
13. Available <https://semantext.com/blog/log-analysis/>. [Online]. Last accessed on 20/12/2020
14. Xu W, Huang L, Fox A, Patterson D, Jordan MI (2009) Detecting large-scale system problems by mining console logs. In: Proceedings of the ACM SIGOPS 22nd symposium on operating systems principles—SOSP'09
15. Lou J-G, Fu Q, Yang S, Xu Y, Li J (2010) Mining invariants from console logs for system problem detection. In: USENIX annual technical conference, pp 23–25
16. Breier J, Branišová J (2015) Anomaly detection from log files using data mining techniques. In: Information science and applications. Springer, pp 449–457
17. Tuor A, Baerwolf R, Knowles N, Hutchinson B, Nichols N, Jasper R (2017) Recurrent neural network language models for open vocabulary event-level cyber anomaly detection. [arXiv: 1712.00557](https://arxiv.org/abs/1712.00557)
18. Du M, Li F, Zheng G, Srikumar V (2017) DeepLog: anomaly detection and diagnosis from system logs through deep learning. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security (CCS'17). Association for Computing Machinery, New York, NY, USA, pp 1285–1298
19. Wang M, Xu L, Guo L (2018) Anomaly detection of system logs based on natural language processing and deep learning. In: 4th international conference on frontiers of signal processing (ICFSP), pp 140–144
20. Lu S, Wei X, Li Y, Wang L (2018) Detecting anomaly in big data system logs using convolutional neural network. In: IEEE 16th international conference on dependable, autonomic and secure computing, 16th international conference on pervasive intelligence and computing, 4th international conference on big data intelligence and computing and cyber science and technology congress (DASC/PiCom/DataCom/CyberSciTech). IEEE, pp 151–158
21. Brown A, Tuor A, Hutchinson B, Nichols N (2018) Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In: Proceedings of the first workshop on machine learning for computing systems, pp 1–8
22. Farzad A, Gulliver TA (2019) Log message anomaly detection and classification using auto-b/lstm and auto-gru. [arXiv:1911.08744](https://arxiv.org/abs/1911.08744)
23. Meng W, Liu Y, Zhu Y, Zhang S, Pei D, Liu Y, Chen Y, Zhang R, Tao S, Sun P et al (2019) Loganomaly: unsupervised detection of sequential and quantitative anomalies in unstructured logs. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19. International joint conferences on artificial intelligence organization, vol 7, pp 4739–4745
24. Zhang X, Xu Y, Lin Q, Qiao B, Zhang H, Dang Y, Xie C, Yang X, Cheng Q, Li Z et al (2019) Robust log-based anomaly detection on unstable log data. In: Proceedings of the 27th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering, pp 807–817
25. Wang X, Wang D, Zhang Y, Jin L, Song M (2019) Unsupervised learning for log data analysis based on behavior and attribute features. In: Proceedings of the international conference on artificial intelligence and computer science, pp 510–518
26. He P, Zhu J, Zheng Z, Lyu MR (2017) Drain: an online log parsing approach with fixed depth tree. In: IEEE international conference on web services (ICWS). IEEE, pp 33–40
27. Available <https://towardsdatascience.com/outlier-detection-with-one-class-svms-5403a1a1878c>. [Online]
28. Nguyen TBT, Liao TL, Vu TA (2019) Anomaly detection using one-class SVM for logs of juniper router devices. In: Duong T, Vo NS, Nguyen L, Vien QT, Nguyen VD (eds) Industrial

- networks and intelligent systems. INISCOM 2019. Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, vol 293. Springer, Cham
- 29. Chepenko D (2018) A density-based algorithm for outlier detection, Sep 16. <https://towardsdatascience.com/density-based-algorithm-for-outlier-detection-8f278d2f7983>
 - 30. Du M, Li F (2016) Spell: streaming parsing of system event logs. In: IEEE 16th international conference on data mining (ICDM). IEEE, pp 859–864
 - 31. Zhang S, Meng W, Bu J, Yang S, Liu Y, Pei D, Xu J, Chen Y, Dong H, Qu X et al (2017) Syslog processing for switch failure diagnosis and prediction in data center networks. In: IEEE/ACM 25th international symposium on quality of service (IWQoS). IEEE, pp 1–10
 - 32. Oliner A, Stearley J (2007) What supercomputers say: a study of five system logs, pp 575–584. <https://doi.org/10.1109/DSN.2007.103>
 - 33. Zhu J, He S, Liu J, He P, Xie Q, Zheng Z, Lyu MR (2019) Tools and benchmarks for automated log parsing. In: IEEE/ACM 41st international conference on software engineering: software engineering in practice (ICSE-SEIP). <https://doi.org/10.1109/icse-seip.2019.00021>
 - 34. Available <https://github.com/logpai/loghub>. [Online]
 - 35. Oliner AJ, Kulkarni AV, Aiken A (2010) Using correlated surprise to infer shared influence. In: Proceedings of the international conference on dependable systems and networks, Chicago, IL, pp 191–200
 - 36. Erlingsson Ú, Peinado M, Peter S, Budiu M (2011) Fay: extensible distributed tracing from kernels to clusters. In: Proceedings of the 23rd ACM symposium on operating systems principles, Cascais, Portugal
 - 37. Yuan D, Zheng J, Park S, Zhou Y, Savage S (2011) Improving software diagnosability via log enhancement. In: Proceedings of architectural support for programming languages and operating systems, Newport Beach, CA
 - 38. Ganapathi A, Chen Y, Fox A, Katz RH, Patterson DA (2010) Statistics-driven workload modeling for the cloud. In: Workshop on self-managing database systems at ICDE, pp 87–92

Malware Detection System Using API-Decision Tree



D. Anil Kumar, Susanta Kumar Das, and Manoj Kumar Sahoo

Abstract Recent growth in the Internet shows the utility of the information technologies utilities to the common man, and it plays a major role and changes the information exchange speed and quantity. The 5G technologies platform gives a better platform for this at cheaper rate than 4G. The data science application over the Internet also increased, which leads to the further developed and need to development in the platform, namely big data, cloud computing, Internet of things (IoT), artificial intelligence, machine learning, etc. So the data traffic is more; hence, the security of these information is a key issue to address. The cyber-attack is a common threat associated with Internet. The cyber-attack is carried out with virus or malware program, antivirus, and other security measure are deploy appropriate stage in the stand alone system or over the network for safety of the data and information exchange. The virus and malware program are smart program enter in to data exchange point of communication. So, smart and intelligent decision taking program have to deploy to stop and delete them. The malware detection is a key issue, malware is slow poison type harmful to the system, and it reduces the computing efficiency and sometime leads to finical loss, due to the user information by malware attack. So in this paper, we are addressing the malware detection using decision tree (DT) and using C4.5 (J48) classification algorithm find a suitable solution method form malware attack in a API routine.

Keywords Malware · Decision tree · Classification · API calls · Support vector machine

D. Anil Kumar (✉) · S. K. Das · M. K. Sahoo
Berhampur University, Berhampur, Odisha, India

National Institute of Science and Technology, Berhampur, Odisha, India

S. K. Das
e-mail: skd.cs@buodisha.edu.in

1 Introduction

The recent trend shows arise of the use and utilities of the Internet in every sector of utility, with the support of the development and the technology in data communication. The ability of the 5G technology in data communication makes the internet service faster and cost-effective to the end-user than the 4G. The cloud platform and Internet of things (IoT) devices make things more robust on 5G platform with higher data processing and communication speed. The need for high-speed processing and communication depends upon the software and corresponding hardware environment with the consideration of the processing element, i.e., the microprocessor or microcontroller.

The software environment processing speed is depending upon the design of the software and deployment environment. During the communication of the data or information, there are some of the unwanted processes that may be attacked or attach to the software environment, and the unwanted software which is associated to destroy or make harmful to the process is known as a virus, while some of this unwanted software will reduce the speed and performance primarily known as malware. The malware is the unwanted software component present in the process of computing without the permission of the user. The malware can lead to bigger issue if stay in the software environment at the user end.

In today's world of data computing, which involves a huge amount of data, malware is a spoilsport or threat to the process and the system at the user end. The number and types of this malware are increasing rapidly and exponentially in the computing world, with the increasing complexity of malware. So, there is a need to stop or prevent this malware, the antivirus and security product making segment of the IT (Information Technology) sector key a close eye and address the problem [1]. The malware is well available over the Internet, which makes a smart entry to any system, while downloading information from the Internet. The entry of malware to any system makes its place by scanning vulnerabilities in the operating system and leads to unwanted execution process to slow down the process by adding unauthorized and unrequired association affecting the system performance or underperformance have to carry out.

The malware can be characterized as below [2]. The malware can be active in the process as a standalone type or it creates a group as a similar type element to create a larger group. The malware either is in standalone or groups it reduces the user system performance. The Internet is the prime source of malware, which can be infected or enter into the host user during its communication over the Internet. In the present time, the utility of the Internet is increasing, with the utility of increase on the Internet the malware type, complexity and number also increase to attack or infect the host system. Malware is user-friendly for the cyber attackers and its attacking and harmfulness depend upon the skill level of the attacker, and it is also difficult to detect, stop or remove the malware at the user system end. It can easily access the host user system even if the user end deploy the basic security measurement, even it found that from the previous malware attack as a cyber-security issue multiple can easily

bypass or penetrate authentications deployment architecture. So malware is a good cyber weapon for cyber-attack with more effective and easy communication, due to the presence and utilization of the Internet. The malware provides a good amount of financial gain to the attacker or the group of attackers by performing different cybercrime. The detection of the malware is a key issue in cyber-security, so the malware detector can be defined as a mathematical model as a function with domain and proper range, where the domain is defined as a program or set of program for execution with the range definition malicious or benign [3]. The detector should have the capacity to determine correctly for a program is either benign or malicious by nature. The detector generally detects the malware from its signature, the machine code pattern generated by the malware program. The system security program uses this signature to verify from the system allowed signature list for storage and memory element of the system. This type of malware detection at the user end can be classified as statics, dynamic, or hybrid (a combination of both) [4].

2 Related Work

Faraz Ahmed, Haider Hameed, Zubair Shafiq, and Muddassar Farooq proposed the windows API method, where they use a machine-learning algorithm to detect the malware or benign from the windows API, and an efficient classifier is deploy for the detections [5]. This method or tool shows good efficiency whenever the API categories are less or minimum. Yi-Dong Shen, Zhong Zhang, and Qiang Yang use an objective-oriented utility-based association (OOA) mining method to detect the malware by using the signature, which is related to the host system. Here the main focus is to determine the degree of usefulness of the signature before stopping or delete malware. Yanfang Ye, Dingding Wang, Tao Li, Dongyi Ye, and Qingshan discuss malware detection by utilization of association mining-based classification [6]. Here the analysis is carried out on the windows API executable files list table, the reference file list is provided by the KingSoft Corporation antivirus laboratory, and this method is an alternative process for the detection of malware against the process which is used to deploy Naive Bayes, support vector machine (SVM), and decision tree techniques [7]. The data mining association mining-based classification is widely used by the cyber and system security product developers like KingSoft Corporation, Norton Antivirus, and McAfee Virus Scan. This method is adopted by of King Soft Corporation for their antivirus package in scanning tool to detect malware. The malware detection process is depending upon the API call routine and well defines the execution file format. The decision needs to more precise; otherwise, it stops the execution of the needed file. This decision can be dependent upon the file format and the format database which leads to the precision of malware detection. Similarly, the techniques discussed in [8–10] can be adopted to apply for the same.

3 API-Decision Tree (APIDT)

The decision-making process for detection should be robust, and the decision needs to be well defined with high precision-based computing, which needs to stop or detect the malware into the user system. Different method is deployed for the purpose, and decision tree (DT) is one of the efficient and popular methods. So, DT is defined, as a support tool with tree type architecture to process all conditions of the algorithm and take the best possible decision from all well-defined possible consequences, considering the event outcomes, cost of the resources, and their utility. The DT is more effective with the presence of conditional statements for the control, well defined in the applied algorithms. So, DT is a well construct predictive architectural process with consideration of the output feature space or subspace, and this leads to a robust and proper correct prediction of decision. DT, as a classifier set proper hierarchical decisions to depend on the features. A smaller deviation in the training set leads to a different DT, so the training set needs to be well-defining input set to the DT. The input dataset is needed lesser preparation for a smaller dataset or need scale normalization before the DT model design or applied at the user end. The DT structure remains unaffected from the transformation, hence not sensitive to outliers. DT remains unaffected even if the parameters relation is nonlinear by nature. The malware detector developer needs to take care of the top-level node division and the tree expansion with proper domain knowledge and updated skill to handle the attack on data and information [6, 11].

The DT uses the popular algorithms mainly Iterative Dichotomiser-3 (ID3), C4.5, C5.0, and classification and regression trees (CART). ID3 algorithm is based on a multi-way tree, and the node finds is depend on the largest feature from the feature category to consider, which can be grown to the maximum size. C4.5 is the next to improve algorithm than ID3, which removed the restriction on the feature category for searching in a discrete interval. C5.0 is more accurate with the lesser memory utility and smaller well defines rule sets than C4.5. CART (Classification and Regression Trees) is very similar to C4.5, and it supports the numerical target variables (regression) without computing rule sets. CART is used to construct binary trees considering the feature and threshold, responsible for the largest information gain at every node.

The DT can be the approach for simplification can be discussed as size control of the tree, robust test space for modification, test search and Modify test search, restrictions utility condition of Database, well define adaptable alternative data structures. The control tree size can be check and measure in Pre- or post-pruning process. The modification of the test space needs to measured or verified by test data driven or any known hypothesis-driven for data have to carry out. The change in search process can be modify the selection measures deploy, by changing the continuous features or look-ahead search method need to use. The database restrictions can be used by proper addition of, selection of case or feature to take proper decisions. There should be a flexible in the data structure to support to the malware detection architecture, which is based on decision graphs or well-defined rules.

The DT uses the popular algorithms, mainly Iterative Dichotomiser-3 (ID3), C4.5, C5.0, and classification and regression trees (CART). ID3 algorithm is based on a multi-way tree, and the node finds is depend on the largest feature from the feature category to consider, which can be grown to the maximum size. C4.5 is the next to improve algorithm than ID3, which removed the restriction on the feature category for searching in a discrete interval. C5.0 is more accurate with the lesser memory utility and smaller well defines rule sets than C4.5. CART (Classification and Regression Trees) is very similar to C4.5, it supports the numerical target variables (regression) without computing rule sets. CART is used to construct binary trees considering the feature and threshold, responsible for the largest information gain at every node. In this study, we deploy C.4.5 improved version, i.e., J48 algorithm with more accuracy to achieve for detection of malware [12, 13]. J48 algorithm is introduced by Quinalan in 1993.

I. Algorithm

Pseudo code for J48 algorithm

```

1: Creation of a root node with number of “ N”;-- petal from this to create or
mapped
2: If (Class”T” belongs to same category feature” C”)
    { petal node=N; mark N as class C; return N; }
3: Calculate Information gain(Ai) of For all event class, i=1 to n
    —calculation for greatest value
4: Ta= testing attribute;
5: N.Ta= attribute having highest information gain;
6: if (N.Ta==continuous) { find minimum energy i.e threshold; }
7: For (Each T’ in the splitting of T)
8: if (T’ is empty) { child of N is a petal node;}
9: else {petal of N= dtree (T')}
10: calculate classification error rate of node N;
11: return N;
```

II. Entropy and Information Gain Formula

The formula for the information gain and entropy can be expressed as

$$\text{Gain} = \text{info}(T) - \sum_{i=1}^s \left(\frac{|T_i|}{|T|} * \text{info}(T_i) \right)$$

and

$$\text{entropy} = \text{info}(T) - \sum_{i=1}^{N_{\text{class}}} \left(\frac{\text{freq}(Cj, T)}{|T|} * \log_2 \frac{\text{freq}(Cj, T)}{|T|} \right)$$

III. Confusion Matrix

Table 1 Confusion matrix

	Predicted class positive	Predicted class negative
Actual class-positive	TP	FP
Actual class-negative	FN	TN

The performance of the algorithm can be measured by using the confusion matrix method and represented in Table 1.

IV. Accuracy

The accuracy of the DT can express as a fellow and calculated from the confusion matrix condition.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\sum(\text{ALL CLASS CASES})} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

V. Experimental setup

The malware detector is used on an Intel Core i7 3rd generation processor with 8 GB RAM and UBUNTU as environmental OS. The sample training data consist of 134,500 instances, normal 44,341, known attack 86,749, and unknown source attack 3410. For accuracy, we train and test this data on ninefold cross-validation. We deploy the WEKA is an open-source tool. The datasets used for the training of the model are, s instances.arff, normal.arff, known_attack.arff, and unknown_source_attack.arff. The *arff* format dataset is define format, “Attribute Relation file Format” is a standard used in WEKA.

VI. Result

The database is used different data attributes as define above. The J48 DT is used as the decision-making classifier as discuss above. The experimental explained in this paper includes thousands of malware executable files with about 200 benign files. In this experiment, the classification accuracy is observed as 87.05% in WEKA environment with a kappa statistic of 0.822.

4 Conclusions

In this paper, we find the DT is used fully to protect the data and information at the user end. The result shows more dataset need to check its robustness and accuracy to make it more useful. The C.4.5 improved version, i.e., J48 algorithm is a good solution to prevent malware.

References

1. Rehman RR, Hazarika GC, Chetia G (2011) Malware threats and mitigation strategies: a survey. *J Theor Appl Inf Technol* 29(2):69–73
2. OECD Ministerial Meeting Report (2007) Malicious software (malware): a security threat to the internet economy. Korean Communication Commision, Final draft, May 2007
3. Vinod P, Laxmi V, Gaur MS (2009) Survey on malware detection methods. *Proc Hacker* 2009:74–79
4. NwokediIdika, Mathur AP (2007) A survey of malware detection techniques. SERC Library
5. Ahmed F, Hameed H, Zubair Shafiq M, Farooq M (2009) Using spatio-temporal information in API calls with machine learning algorithms for malware detection. In: Proceedings of the 2nd ACM workshop on artificial intelligence and security (AISeC 2009), pp 55–62
6. Ye Y, Wang D, Li T, Ye D (2007) IMDS: intelligent malware detection system. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining (KDD’07), pp 1043–1047
7. Ye Y, Wang D, Li T, Ye D, Jiang Q (2008) An intelligent pe-malware detection system based on association mining. *J Comput Virol* 4:323–334
8. Dutt I, Borah S, Maitra IK (2020) Immune system based intrusion detection system (IS-IDS): a proposed. *IEEE Access* 8(2020):34929–34941
9. Panigrahi R, Borah S (2019) Dual-stage intrusion detection for class imbalance scenarios. *Comput Fraud Secur* 2019:12–19
10. Dutt I, Borah S, Maitra IK, Bhowmik K, Maity A, Das S (2018) Real-time hybrid intrusion detection system using machine learning techniques advances in communication, devices and networking. Springer, Berlin, pp 885–894
11. Shen Y-D, Zhang Z, Yang Q (2002) Objective-oriented utility-based association mining. In: Proceedings of the IEEE international conference on data mining (ICDM 2003), pp 426–433
12. Jain YK (2012) Upendra, an efficient intrusion detection based on decision tree classifier using feature reduction. *Int J Sci Res Publ* 2(1). ISSN 2250-3153
13. Sahu S, Mehtre BM (2015) Network intrusion detection system using J48 decision tree. In: International conference on advances in computing, communications and informatics (ICACCI)
14. Leonard LC (2017) Web-based behavioral modeling for continuous user authentication. <https://www.sciencedirect.com/topics/computer-science/decision-trees>

ANFIS for Fraud Automobile Insurance Detection System



Gopikrishna Panda, Sunil Kumar Dhal, Rabinarayan Satpathy,
and Subhendu Kumar Pani

Abstract Fraud claims of automobile insurance are great loss for insurance companies as well as client with insurance policy. The main motive of this work is to devise a mechanism first as predictive model to classify whether a policy claim is classified as genuine or not and secondly what different kinds of parameters should be goaled to find fraud claims. To accomplish this motive, greatly precise prediction models are made by finding key features set through feature selection methods that are necessary for avoiding loss in future. The study of parametric and non-parametric algorithms that are statistical learning in nature reflects on to diminish unpredictability and escalates the possibilities of finding the accurate claims. The required feature set that is important for a framework is investigated by calculating feature relevant formed on the perceived specifics of a policy claim through a cross-validation. This is further tested for improvement of the efficiency through which claims of automobile fraudulent are precisely distinguished using adaptive neuro-fuzzy inference system. The accuracy of the model based on adaptive neuro-fuzzy inference system can reach more than 98% with an exact feature set chosen through a cross-validation. The endeavour attempted here would surely benefit not only insurance sector but also other fraud detection in other sectors like credit card, financial communication and many more in a way to get rid of increasing fraud in financial sector.

Keywords Automobile insurance fraud prediction system (AIFPS) · Adaptive neuro-fuzzy inference system (ANFIS) · Discrete wavelet transform · Artificial neural network (ANN)

G. Panda (✉) · S. K. Dhal · R. Satpathy
Sri Sri University, Cuttack, Odisha, India

S. K. Pani
Krupajal Computer Academy, Bhubaneswar, Odisha, India

1 Introduction

Insurance fraud is common in recent time, and human effort to detect this fraud is proving time-consuming as well as costly. Fraud takes place in various sectors such as insurance [1–3], credit card [4, 5], telecommunications [6] and financial communications [7–9]. Out of all sectors, fraud in insurance is very common and frequently done. Automobile insurance fraud in insurance sector is most frequent fraud that happens to give loss to insurance companies as well as for policyholders. Information related to such frauds where money acquiring unethically is mentioned in [10].

Car insurance domain specially attracts such fraud to be carried out easily which gives rise to false policyholder to set or plan traffic accidents and register false insurance claims (inflationary pressure) to gain illegal profit for their insurance policy [11]. It has been observed and revealed that nearly 21–36% of the car insurance claims are fake and 3% of claims are arraigned [12].

Two distinctive sorts of fraud occur that include devious and non-amateur fraud, and other types are performed by arranged batch of people. As the arranged group of fraud has done less fake insurance cases than the devious fraud, financial loss (outflow of revenue) was more due to arrange group of fraud as in [13]. As discussed in the paper [14], detection of the fraud would be challenging because of various reasons. One of the reasons is high voluminous data that is changing constantly with time. In real world for analysing these huge data sets, the speedy, the innovative and excellent performance algorithms are required. As there is well-known fact dealing with huge data set, elaborate analysis execution is costly affair. Presence of such issues demands requirement of exact entry of real policy records for each person should be updated regularly to avoid car insurance fraud and entry should be analysed properly and exactly.

Recently, one of the dynamic research fields is machine learning (ML) with the advancement of hardware and software computing environment related to various application fields with highly vague problem issue. Insurance sector is related to one of the problems in recent year. ML is capable of process of automation through less consuming time and personalized by nature. Successive results are produced in the case of ML and ANN processes where the objective of ML and ANN is achieved. The larger voluminous nature of data in insurance fraud should be analysed and interpenetrated which are results of advanced ML and ANN learning techniques to less sampled cases for exact prediction. The fraud insurance detection as well as interpretation is largely depending on the group of people proficient in insurance and efficiency of these people in the insurance issue that is not sufficient to solve issue. Therefore, the data analysis entails expert people with expertise of high efficiency with high exactness degree, though it is vulnerable to do error, and learning method based on ML and ANN can boost the accuracy with graded estimation for automation-based prediction-making model and mechanism with specialist behaviour.

Bayesian network, Gaussian mixture model and hidden Markov model are some traditional methods, or the common methods evaluated and executed for recognition

include species and diseases of human, animal, birds, etc. Many researchers used these traditional methods executed have faced failure to achieve the high accuracy and efficiency. Various neurocomputing systems were conceptualized for recognition of insurance frauds like probabilistic neural network, decision tree, linear discriminant analysis and support vector machine. Adaptive neuro-fuzzy inference system based on machine learning for fraud classification and prediction is next level of evolution in artificial neural network. In this paper, we discuss about the necessity of ANFIS based on machine learning utility towards the issues of fraud in automobile insurance sector.

The paper is organized after introduction as follows; the methodology of the research ANFIS is discussed elaborately in Sect. 2. Section 3 focuses on the observation of the proposed system. Section 3 analyses the techniques which are being used for feature selection by using discrete wavelet transforms and discussed briefly. This Sect. 3 elaborates on the approach by the use of adaptive neuro-fuzzy inference system. With the help of the flow diagram, the proposed approach is described in detail. In Sect. 4, the results and discussions are then discussed, and those are found from the experiments carried out. Finally, Sect. 5 discusses conclusions from the existing work and provides works for possible future extensions.

2 Methodology

Adaptive system is a multi-layer feed-forward network in which there are different nodes. Each node executes a specific function (node function) on input signals as well as the parameters set related to this node. The functions of the node may change from node to node, and the option of each node function depends on the overall input–output function through which this is required to be carried out by the adaptive network. The thing to be noticed is the links in an adaptive network that only indicates the flow direction of signals between nodes. There are no weights related to the links. To express different adaptive capabilities, the use of both circle and square nodes in an adaptive network is performed. While a circle node (fixed node) has none parameter, but square node (adaptive node) has parameters. An adaptive network has parameters which are the mixture of the parameter sets of each adaptive node. According to provided training data and a procedure of gradient-based learning, the updating of parameters is done to reach an optimum input–output mapping.

Soft computing primarily consists of modules, i.e. these are ANN and fuzzy system (FS). Universally, these two components have been acquired for their benefaction in the areas of known and unknown possibility. Abundant system free evaluation data is highly optimum for ANNs, but ample work data is given, and knowledge rendered by expert is important for fuzzy systems to perform as shown Fig. 1. So, both ANN and FS can give both numeral quantitative potential and subjective-qualitative potential, respectively, and can be executed to resolve the problem that arises where excellent accuracy of implementation is not realizable. Both ANN and fuzzy make a hybrid system that can be built to render potential or properties like adaptability,

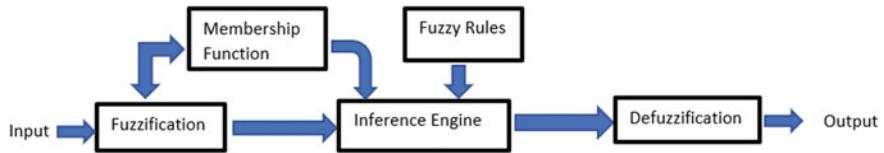


Fig. 1 Fuzzy system block diagram

parallelism, nonlinear processing, robustness and learning in data voluminous fields for familiar and non-familiar system [15]. Numerous fields are embracing ANFIS for their various applications. Out of two fuzzy models, i.e. Mamdani fuzzy model or TSK fuzzy model any one can be used to particular issue. Although, TSK model is universally chosen due to its smooth and computationally proficient results than Mamdani's model. In this section, there is a proposal for a class of adaptive networks which functions same as fuzzy inference systems. The proposed system here is evidently named as ANFIS, which stands for adaptive network-based fuzzy inference. For the application of the hybrid learning rule, the description of system deals with depiction of the parameter set to the node sequentially. Apart from this, there is representation of application of the Stone–Weierstrass theorem to ANFIS with simplified fuzzy if–then rules in addition to relation of the radial basis function network to this kind of simplified ANFIS.

2.1 Architecture of ANFIS

For the assumption of the adaptive network-based fuzzy inference system, a, b are two inputs and c as one output of the system. For the basis of this system, two fuzzy if rules of Takagi and Sugeno rule are taken for the system [16] (Fig. 2).

Rule I: If a is a_1 and b is b_1 , then $f_1 = P_1a + Q_1b + R_1$.

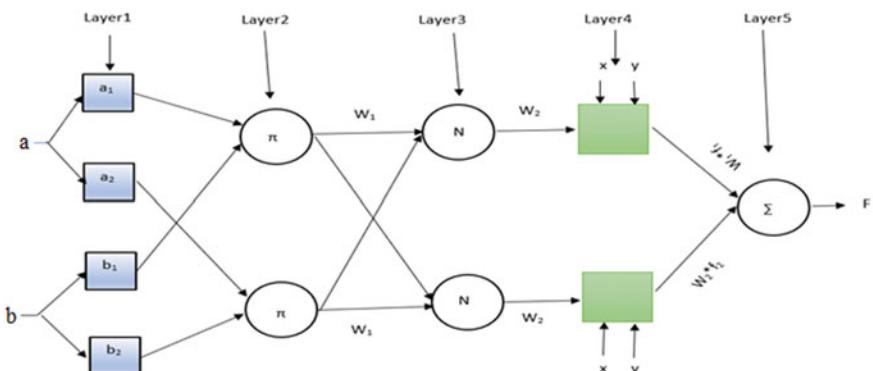


Fig. 2 ANFIS algorithm-based structural design

Rule II: If a is a_2 and b is b_2 , then $f_2 = P_2a + Q_2b + R_2$.

Where aj and bj = fuzzy variable related to antecedent inputs.

Pj, Qj, Rj = values of constant.

Layer 1: Node variables are numerically evaluated for layer of node j as

$$\begin{aligned} M1Pj &= \mu aj(x), & j &= 1, 2 \\ M1Pj &= \mu bj(x), & j &= 3, 4 \end{aligned}$$

where member functions are $\mu aj(x)$ and $\mu bj(x)$.

Layer 2: Firing function is evaluated as

$$MP2, j = Wj = \mu aj(x) * \mu bj(y) \quad j = 1, 2$$

Layer 3: Normalized firing strength is the ratio of j th firing strength to total firing strength as

$$MP3, j = \overline{W_j} = \frac{W_j}{W_j + W_2} \quad j = 1, 2$$

where $\overline{W_j}$ = normalized firing strength.

Layer 4: each node of j , there is node function with square node

$$M_j^4 = \overline{W_j} f_j = \overline{W_j} (P_j x + Q_j y + R_j),$$

where the output of layer 3 is $\overline{W_j}$, and $\{P_j, Q_j, R_j\}$ is the parameter set. Parameters in this layer will be referred to as consequent parameters.

Layer 5: The single node in this layer is a circle node labelled as \sum that computes the overall output as the summation of all incoming signals

$$M_1^5 = \text{overall output} = \sum_j \overline{W_j} f_j = \frac{\sum_j w_j f_j}{\sum_j w_j}$$

This ANFIS algorithm is used for prediction of fraud in automobile insurance for aiding the insurance company for timely detection of fraud of the client [17–19]. This process is examined, evaluated and analysed in relation to result in the next section of this work, i.e. observation. The model using algorithm is implemented in automobile insurance fraud prediction system (AIFPS).

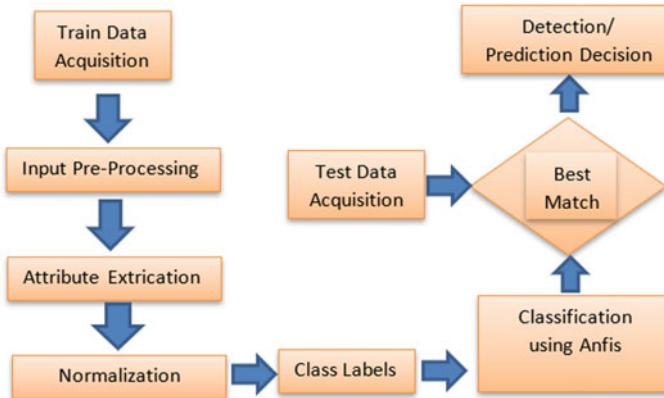


Fig. 3 Depicts the design flow of the proposed system

2.2 Observation (Simulation)

Simulation of the proposed system is described here in this section. MATLAB 2016b software is used for executing the experiments and simulating model. The test platform is Intel core 5i 8th generation, 2.2-GHz CPU, 8-GB RAM processor with Windows 10 operating system. An automobile insurance fraud prediction system (AIFPS) designed using ANFIS is reflected in Fig. 3.

The design flow begins with data acquisition, pre-processing, feature extraction, classification and prediction decision. The work points to the study of the execution of the automobile insurance fraud prediction system that will provide reliable, accurate and minimized speed overall. The response created by system is formed on decision brought by the system. For both training and testing phases of classification, the data acquisition task is collected separately. Data from data repository is pre-processed for eventual manipulation. In this case, the data set consists of 3451 participants involved in 1561 collisions in Slovenia between the years 1999 and 2008. The set was made by merging two data sets, one labelled and one unlabelled taken from data repository.

Data pre-processing constitutes of noise removing or filtering. In this work, zero first-order FIR filter with a 6 dB/octave gain is used to remove the noise from the input data. FIR filter is used because its capability is to reduce to zero artefacts along with other associated noise like background noise. The next stage is the feature extraction. Data matrix features include temporal features as well as spectral features. In DWT, the approximate coefficient of feature vector is 44 in dimension with the use of Daubechies 4 and decomposition level of 10 which gives an optimum performance. The subsequent step of the proposed model is utilization of method of principal component analysis (PCA) which helps to extract attributes dimension reduced depending on the dimension of these attributes. PCA technique is used for variable reduction so correlated variables are removed and uncorrelated variables are taken. The

Table 1 ANFIS requirements

Input data size	Data matrix
SNR	0–3 Db
ANFIS type	ANFIS with five layers
ANFIS training method	Back propagation gradient descent method together with least squares
Average training epochs	10–200
Total number of membership functions	10
Type of membership function	Gbellmf, linear
RMSE with five linear at 200 epochs	1.304×10^{-3}

results of first principal component differ as much as the difference can be observed as with approaching components as possible way as in data as well. PCA is a universal approach for related patterns and reducing dimension of voluminous data dimension. The data having attributes vectors majorly consists of special number of information based to the data. For this fraud prediction system, the attribute length is minimized due to the principal components number which gives the best possible performance during training. The classifier in this work used for the recognition/prediction is ANFIS.

The group of the least-squares (LS) method and the back propagation gradient descent (BPGD) method is put in application for training FIS membership function parameters to copy a training data set provided. The crucial requirement of ANFIS along with five layers is reflected in Table 1. Data matrix feature of minimized length is utilized for the training of the ANFIS for 10–200 epochs. The results are evaluated on the accordance of ten trials of epoch on average. Gbellmf MF is used for performing fuzzification. Because of which the Gbellmf MF provides the very realistic results to fuzzy transformation which are achieved from these experiments. The sample of ten such MFs to be considered which give the best results is executed. For training the ANFIS, the root mean squared error (RMSE) is the measure with which cost function parameter is calculated for training data.

3 Experimental Details and Results

In terms of computational time and accuracy, the performance of AIFPS is analysed. The computational time as a whole is minimized to big level after using proposed system. 100 data matrix size in total is been provided to the system for training, validation and testing, and from them, a small set of samples with artefacts are also made for testing. The training sample trained in the system is made elaborately, and the system is related to minute variation in signal-to-noise ratio (SNR) range between 0 and 3 dB to achieve the dependable, powerful and accurate prediction. The results obtained is compared with [20] and inferred that the ANFIS produces

relatively RMSE values, but the computation delay is specifically less. As a result, the RMSE convergence is reached to best in a handful of epochs. The minimum RMSE and time for training in 200 epochs in accordance with the ANFIS are shown in Table 1. The prediction accuracy rates and time delay achieved out on a small number of training epochs with ANFIS and ANN for automobile insurance fraud prediction are shown in Table 2. Thus, for optimum identification and authorization of design of prediction system, the ANFIS approach is used. At midst of training of ANFIS, plot of the average RMSE convergence is depicted in Fig. 4. The edge of ANFIS-based approach over reflects the result on an average 30% less epochs than the ANN-aided method and yields between 5 and 18% better winning rates. Each

Table 2 Comparison of average rate accuracy in percentage and computational time delay in seconds for both ANFIS and ANN based on prediction of automobile insurance fraud

No. of epoch	ANN with wavelet		ANFIS with wavelet	
	Accuracy %	computation time in seconds	Accuracy %	Computation time in seconds
10	85	0.92	90	0.61
20	85	0.92	91	0.62
50	85	0.95	91	0.55
100	82	0.98	97	0.51
150	81	0.89	95	0.53
200	80	0.91	96	0.54

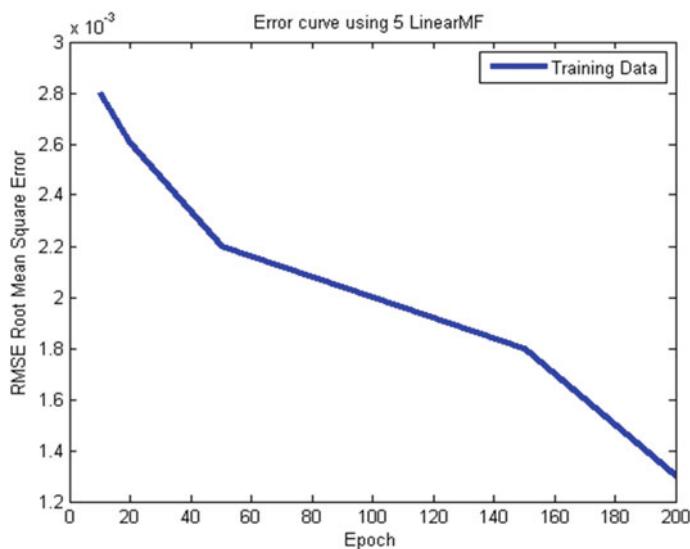


Fig. 4 ANFIS average RMSE convergence

data specimen needs training delay time at midst from 1.1 to 2.3 s. The ten trials were executed on the specimen sets on which the results are inferred and the inference on average is reflected. The root mean squared error (RMSE) equation is defined by formulae given below:

$$\text{RMSE} = \sqrt{\frac{1}{D} \sum_{t=1}^D (A_t - F_t)^2}$$

where A_t is actual value, F_t is fitted value subsequently, and the list of training or testing sample is D . The RMSE is majorly the fundamental calculation medium to actuate the level of learning executed by the ANFIS.

4 Results

The outcomes reflect test results found in Table 2 in this section of the paper, while testing the proposed system. The testing begins with ANN along with wavelet and compared the performance of ANFIS with wavelet.

After observation from Table 2, ANN with wavelet performed on time delay of 0.8–0.9 s gives an accuracy of 80–85%. Table 2 reflects ANFIS along with wavelet performed at time delay of 0.5–0.6 s rendering an accuracy of prediction of 90–98%.

Using ANFIS along with wavelet, the classification accuracy increases from 90 to 98% with time delay between 0.5 and 0.6 s which is shown in representation in Figs. 5 and 6. In Fig. 5, the black line depicts the classification accuracy of ANFIS along with wavelet which lies between 90 and 98% which depends on number of epoch. As the epoch increments, the black line becomes straight horizontal line during (epoch 20–epoch 45), then suddenly increases to some extent after 50–100 epochs and decreases steadily from epoch 100 to epoch 150. Lastly, the curve increases steadily after epoch 150–epoch 200. This black line depicts that ANFIS with wavelet is better than other technique used on the matter of classification accuracy. The blue line denotes that ANN along with wavelet is lower than former method.

Figure 5 depicts time delay of both methods with respect to number of epoch where the time delay of ANFIS along with wavelet is better than ANN along with wavelet. The black line with Δ on the curve depicts time delay of ANFIS along with MFCC, and green line with * on the curve depicts time delay ANN along with wavelet in Fig. 6. As the classification accuracy is best in ANFIS with wavelet, the computation time delay is perfect in comparison with other method used.

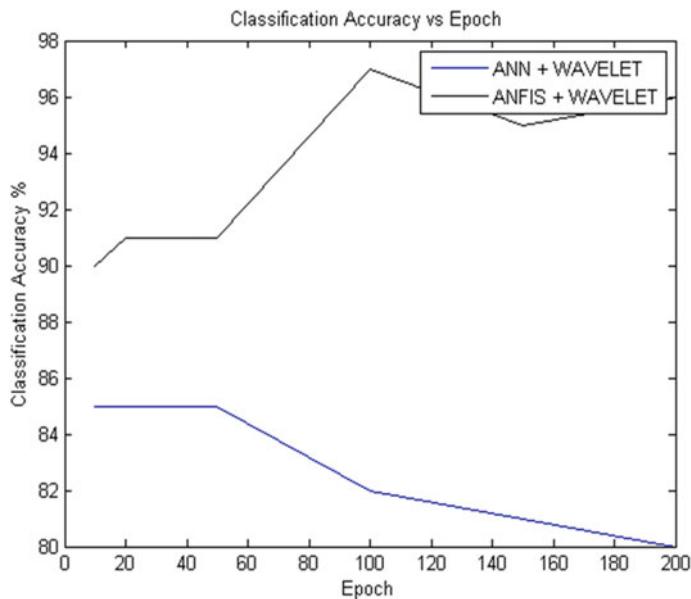


Fig. 5 Graphical depiction of ANN and ANFIS in accordance with accuracy percentage versus epoch

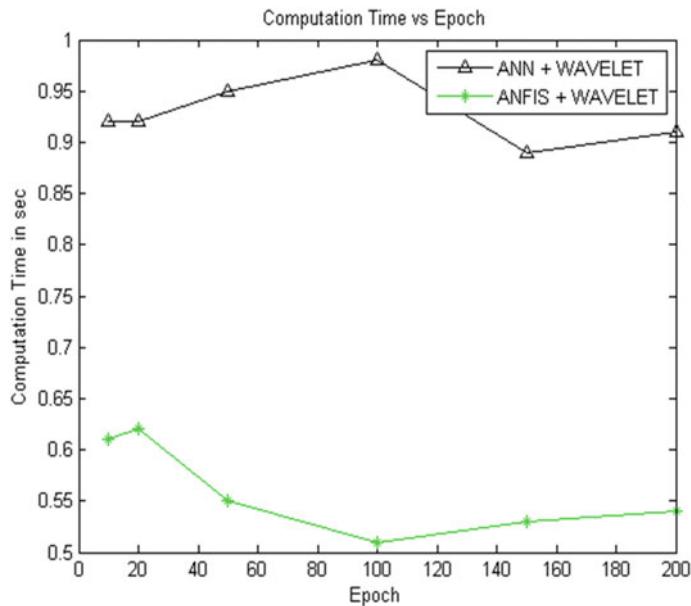


Fig. 6 Graphical depiction of ANN and ANFIS in accordance with computation time versus epoch

5 Conclusion

The concept of ANFIS is implemented in the paper for Automobile fraud detection prediction system to detect frauds. Efforts are performed to work out the insurance fraud prediction with the neural network of fuzzy logic-based. The system is regularly combined and associated with sensory encoding, learning and decoding functionalities. The system works in a time-based structure, where the possibilities of neuro-fuzzy concept are considered for information processing and cognitive computing. The prediction performance of the system was examined and evaluated with two methods, i.e. ANFIS and ANN used where ANFIS along with wavelet resulted in the best method with high accuracy and low delay time. Implementation of software in the office with the help of hardware implants would be beneficial for automobile fraud prediction system. This type of adaptive neuro-fuzzy generation neural network would make wave for further research in advanced machine learning in near future and will also bring to a great revolution in various areas of financial crisis yielding advantageous as a whole.

References

1. Ormerod TC, Ball LJ, Morley NJ (2010) Informing the development of a fraud prevention toolset through a situated analysis of fraud investigation expertise. *Behaviour Inf Technol* 31(4):371–381. <https://doi.org/10.1080/01449291003752906>
2. Li J, Huang K-Y, Jin J, Shi J (2008) A survey on statistical methods for health care fraud detection. *Health Care Manage Sci* 11(3):275–287. <https://link.springer.com/article/10.1007/s10729-007-9045-4>
3. Estevez PA, Held CM, Perez CA (2006) Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Syst Appl* 31(2):337–344. <https://doi.org/10.1016/j.eswa.2005.09.028>
4. Kirkos K, Spassis C, Manolopoulos Y (2007) Data mining techniques for the detection of Fraudulent financial statements. *Expert Syst Appl* 32(4):995–1003. <https://doi.org/10.1016/j.eswa.2006.02.016>
5. Kotsiantis S, Koumanakos E, Tzelepis D, Tampakas V (2006) Forecasting fraudulent financial statements using data mining. *Int J Comput Intell* 3(2):104–110
6. Holton C (2009) Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multi-billion dollar problem. *Dec Support Syst* 46(4):853–864. <https://doi.org/10.1016/j.dss.2008.11.013>
7. Almeida MPS (2009) Classification for fraud detection with social network analysis. MD diss., Technical University of Lisbon
8. Ayuso M, Guillen M, Bolancé C (2011) Loss risk through fraud in car insurance. XARXA DE REFERENCIA EN ECONOMIA APLICADA (XREAP). <https://ssrn.com/abstract=1857007>
9. Nian K, Zhang H, Tayal A, Coleman T, Li Y (2016) Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *Fin Data Sci* 2(1):58–75. <https://doi.org/10.1016/j.jfds.2016.03.001>
10. White Paper (2012) Combating insurance claims fraud. how to recognize and reduce opportunistic and organized claims fraud. SAS® Fraud Framework for Insurance. Retrieved from <http://www.sas.com/reg/wp/no/42477>
11. Bolton RJ, Hand DJ (2002) Statistical fraud detection: a review. *Stat Sci* 17(3):235–249. <http://www.jstor.org/stable/318278>

12. Wang Y, Xu W (2018) Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Dec Support Syst* 115:87–95
13. Subudhi S, Panigrahi S (2020) Use of optimized fuzzy C means clustering and supervised classifier for automobile insurance fraud detection. *J King Saud Univ Comput Inf Sci* 32(5):568–575
14. Rodrigues LA, Omar N (2014) Auto claim fraud detection using multi classifier system, pp 37–44
15. Shin H, Kim KH, Song C et al (2010) Electrodiagnosis support system for localizing neural injury in an upper limb. *J Am Med Inform Assoc* 17:345–347
16. Rondina JM, Filippone M, Girolami M et al (2016) Decoding post-stroke motor function from structural brain imaging. *Neuroimage Clin* 12:372–380
17. Goleiji L, Tarokh MJ (2015) Identification of influential features and fraud detection in the insurance industry using the data mining techniques. *Majlesi J Multimedia Process* 4(3)
18. Li Y, Yan C, Liu W, Li M (2017) A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud detection identification. *J Appl Soft Comput*
19. Li Y, Yan C, Liu W (2016) Research and application of random forest model in mining automobile insurance fraud. *IEEE*
20. Asadi H, Kok HK, Looby S et al (2016) Outcomes and complications after endovascular treatment of brain arteriovenous malformations: a prognostication attempt using artificial intelligence. *World Neurosurg* 96:562–569

Image and Video Data Analysis

Computer Vision-Based Alert System to Detect Fatigue in Vehicle Drivers



Jyotsna Rani Thota, B. J. Jaidhan, Mukkamala S. N. V. Jitendra, A. Shanmuk Srinivas, and A. S. Venkata Praneel

Abstract The most prevailing problem around the world is the increasing number of road mishaps. The number of lives lost on road is much more than any other disease and disaster combined every year. The main cause of accidents to occur is improper and inattentive driving. With proper research on driver drowsiness and the behavioral pattern, the driver exhibits we can reduce accidents. This paper aims to implement the non-intrusive approach to detect the fatigue of the driver and warn the driver immediately to prevent the accident. This paper uses dlib and OPENCV libraries to implement the proposed system. This system uses a dlib face landmark detector to identify 68 distinct spots to apply various face forecast techniques. The live video stream is taken from the camera and is decomposed into continuous frames. The dlib library identifies the eye landmarks and is used to calculate the eye aspect ratio (EAR). The alarm will start given the EAR falls below the threshold value set for some consecutive frames. EAR is obtained by calculating the Euclidean distance between measured eye co-ordinates. OpenCV is used as a primary image processing tool. Python language is used as the main coding language. This drowsiness detection mechanism monitors EAR continuously for drowsiness and gives the alarm sounds if EAR falls below the threshold. Our experimental results show that the proposed system works at a good pace works well in real time. The hardware required for the implementation of this paper is a decent camera that can capture at least 15 frames per second. This paper will help with a significant decrease in the number of accidents.

Keywords Facial landmark identifiers · Fatigue recognition · Eye aspect ratio · Live frames extraction · EAR threshold

J. R. Thota · B. J. Jaidhan · M. S. N. V. Jitendra (✉) · A. Shanmuk Srinivas ·
A. S. Venkata Praneel

Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam, AP 530045, India

1 Introduction

This part traces the requirement for the drowsiness discovery framework in vehicles and the review of the ongoing picture handling framework structured in this system. Driver exhaustion is the reason for vehicle mishaps. Recent studies gage that every year 1200 passings and 76,000 wounds can be linked to drowsiness-related accidents. There has been abundant advancement in the field of well-being if there should arise an occurrence of mishaps as airbags, and so forth. Notwithstanding, the improvement of advancements for distinguishing or forestalling fatigue in the driver's seat is a significant test in the field of mishap prevention frameworks. Due to the peril that tiredness presents out and about, techniques should be produced for neutralizing its effects. The point of this task is to build up a model drowsiness detection framework. This framework will essentially quantify and record the ongoing highlights of the driver or the driving example and persistently assess them based on the levels predetermined to demonstrate fatigue. The driver may give indications of exhaustion from various perspectives. The task reviews various techniques for eye recognition and checking whether the recognized eyes are open or shut. Based on the occasions, the eyes are seen as shut, and it very well may be resolved whether the individual is drowsy or not. Another strategy for weakness discovery is the assessment of the highlights of the mouth. On the off chance that the driver yawns, giving indications of exhaustion, it very well may be utilized to trigger the caution framework [1]. Articulations showing outrageous indignation, frustration, stun, energy, and so forth show that the driver is not in the best state to drive and may vacillate at it. This ought to be utilized to remind the driver as a suitable alert to either deliberately leave the vehicle or deliberately recover the self-control to drive securely [2]. Fatigue recognition in the non-intrusive structure might be done effectively by expanding the number of parameters on which the driver is being observed. This will prompt a calculation that can be effectively adaptable to identify exhaustion. Such a driver exhaustion location framework has a colossal extension in the vehicle business. There are a huge number of vehicles being made each year, and weariness among drivers is a strong mishap factor. Subsequently, vehicle fabricating organizations are the greatest market for such a framework [3]. A car manufacturer has presented a weariness recognition framework in specific models of very good-quality vehicles. Be that as it may, the proposed model that this task targets making will be unmistakably increasingly the modest when contrasted with that model, subsequently making a greater market for this item.

Indian streets are inclined to weakness-related mishaps basically among truck drivers, significant distance transport drivers, and BPO representatives during the late hours [4]. The presentation of such a framework in significant distance vehicles and out in the open vehicle would diminish the number of mishaps. Need for the task several examinations has explored the connection between driver weariness and crash chance and has endeavored to evaluate the hazard increment [5]. For a situation control investigation on vehicle drivers, a study contrasted 571 accidents-included drivers and 588 non-crash-included drivers driving in a similar territory

and on similar occasions. Driver factors are taken from mishap enrollment and extra meetings. Considering conceivable puzzling factors (sex, age, financial status, yearly kilometers, speed, street type), they found a solid connection between intense weariness (in the light of the loss of rest the prior night) and crash inclusion [11]. Crash hazard was many times higher with a score of 4 on the Stanford sleepiness scale (95% certainty interim); multiple times higher for driving somewhere in the range of 2 and 5 am (95% interim); and just about multiple times higher when drivers had rested for under 5 h in the previous 24-h time frame (95% certainty interim). For a situation control study, the study contrasted crash-included drivers, and a comparable gathering of non-crash-included drivers at a similar area, heading, time, and day [12]. The chance of accident was multiple times higher for drivers who had answered to have nearly nodded off in the driver's seat (95% certainty interim). The information gathered in the 100 Car Naturalistic Driving Study shows that driving while exhausted expands a driver's danger of association in an accident or close accident by about multiple times. In investigations of expert drivers (transport, lorry, truck), an organization found that following 11 h of work length, the accident hazard pairs. The impact of the undertaking term is constantly entrapped with the impacts of the hour of the day and here and there additionally with the time allotment conscious and past absence of rest [13]. The length of a trip might be of lesser significance contrasted with these different variables—many exhaustion-related mishaps happen after driving for just a couple of hours. A short trip can likewise wind up in exhaustion-related accidents since the time of day and long, and unpredictable working hours are more grounded indicators of weariness than time spent driving [14]. The relationship of non-clinical (way of life) determinants of weariness with a crash has not been the subject of careful research. There is as yet an absence of information concerning the commitment of expanding complete long periods of work and moves timetables to driver weakness [15]. While examination into fatigue and rest apnea in truck drivers has prompted familiarity with these issues and some change of work conditions, occupationally instigated weariness inconceivably a lot of bigger quantities of suburbanites has gotten little consideration [16]. The Motor Administration (FMCSA), the trucking firm, expressway prosperity supporters, and transportation authorities have all recognized driver fatigue as a significant cause for vehicle safety issues [17]. Drowsiness impacts mental sharpness, decreasing an individual's ability to work a vehicle safely, and growing the threat of human bumble that could incite fatalities and injuries. Besides, it has been seemed to slow reaction time, reduces care, and debilitates judgment [18]. Broadened time frames in the driver's seat in troubling driving circumstances make truck drivers particularly slanted to drowsy driving mishaps. Successfully watching out for the issue of driver lethargy in the business motor vehicle industry is a forcing and multi-faceted test [19]. Operational requirements are contrasting, and factors, for instance, work plans, commitment times, rest periods, recovery openings, and response to customer needs can vary extensively. Moreover, the collaboration of the indispensable physiological segments that underlie the advancement of laziness, specifically the homeostatic drive for rest and rhythms, is mind-boggling [20].

Meanwhile, the troubles hinder a singular, clear response to the problem, and there is an inspiration to acknowledge that driver tiredness can be suitably regulated taking all facts into account, in this manner achieving an imperative decline in related hazard and improved prosperity. Keeping in mind, the necessity for a lessening in crashes related to driver fatigue in transportation will require some innovative thoughts and propelling methodologies. Vehicle mechanical procedures, both open and creating, have fantastic potential as huge and practical devices to address exhaustion. Inside any total and fruitful depletion, the administrator's program, an on-board device that screens driver state dynamically may have a certifiable motivating force as a security net. Tired drivers show some unmistakable lead, including eye gaze, eyelid advancement, understudy improvement, head positions, and appearance. Non-obtrusive methods are as of now being utilized to assess a driver's preparation level through the visual impression of his/her condition using a remote camera and top-tier headways in PC vision. Continuous headway in research and advances in PC gear advancements have made it possible to measure head present, eye gaze, and eyelid improvement exactly and logically. The readiness checking advances screen generally on-line and continuously—bio-direct pieces of the overseer; for example, eye gaze, eye end, hindrance, head state and improvement, and heartbeat. To be sensible and significant as driver notice systems, these contraptions must get decode and input information to the head in genuine driving conditions. Taking everything into account, there exists a need, and thus, advancing undertakings are in progress, to endorse overseer-based, on-board weariness watching developments in a genuine naturalistic driving condition. Considering the requirement for non-intrusive moderately reasonable modules for exhaustion discovery in vehicles, this undertaking has been detailed to screen the driver's activities and responses, check for weariness and on the identification of weakness or laziness, animate the fitting strategy, which could be an alert or deceleration of the vehicle, and so on [21]. OpenCV represents open source computer vision. It is an open source BSD-authorized library that incorporates many propelled computer vision calculations that are upgraded to utilize equipment increasing speed. OpenCV is ordinarily utilized for AI, picture handling, picture control, and substantially more. OpenCV has a secluded structure. There are shared and static libraries and CV namespace. To put it plainly, OpenCV is utilized in our application to effortlessly stack bitmap records that contain arranging pictures and play out a mixed activity between two pictures with the goal that one picture can be found out of sight of another image. In this paper, the further sections were organized with a literature survey, specification of the proposed model which carries a detailed design of the work with a proposed algorithm, and further proceeded with a discussion on work, and the results of the work carried, and a detailed conclusion is attached at the end with a specification of the future scope.

2 Literature Survey

In this paper [6], a novel way to deal with basic pieces of face location issues is given, because of convolution neural networks (CNN) calculations. The proposed CNN calculations find and help to standardize human faces adequately while cause for most mishaps identified with the vehicle's accidents. Driver weariness their time necessity is a small amount of the recently utilized strategies [7]. The calculation begins with the identification of heads on shading pictures utilizing deviations in shading and structure of the human face and that of the foundation. By normalizing the separation and position of the reference focuses, all faces should be changed into a similar size and position. For standardization, eyes fill in as perspectives. Other CNN calculation finds the eyes on any grayscale picture via looking through trademark highlights of the eyes and eye attachments. Tests made on a standard database show that the calculation works exceptionally quickly, and it is solid.

In this paper [8], eye identification is required in numerous applications like the eye-stare following, iris recognition, video conferencing, auto-stereoscopic presentations, face identification, and face acknowledgment. This paper proposes a novel procedure for eye location utilizing shading and morphological picture preparation. It is watched that eye locales in a picture are described by low light, high-thickness edges, and high differentiation when contrasted with different pieces of the face. The technique proposed depends on the suspicion that a frontal face picture (full frontal) is accessible. Right off the bat, the skin area is recognized utilizing shading-based preparing calculation and a six-sigma strategy worked on RGB, HSV, and NTSC scales.

The further investigation includes morphological preparation utilizing limit locale location and recognition of light source reflection by an eye, normally known as an eye dab. This gives a limited number of eye competitors from which clamor is in this way expelled. This strategy is seen as exceptionally effective furthermore, exact for recognizing eyes in frontal face pictures. This paper [9] presents a vigorous eye location calculation for dark force pictures. The possibility of our strategy is to join the particular points of interest of two existing strategies, includes based technique furthermore, format-based technique, and to defeat their inadequacies. Initially, after the area of the face district is recognized, a component-based technique will be utilized to distinguish two unpleasant areas of the two eyes on the face. At that point, a precise discovery of iris focuses will be preceded by applying a format-based technique in these two unpleasant districts. Aftereffects of analyses to the countenances without displays show that the proposed approach is not just vigorous yet also very effective [22].

This paper [10] depicts our progressing chip away at continuous face discovery in dark-level pictures utilizing edge direction data. We will show that edge direction is a ground-breaking nearby picture includes showing objects like countenances for recognition purposes. We will introduce a straightforward and proficient strategy for format coordinating and object demonstrating dependent on anxious direction data. We too tell the best way to acquire an ideal face model in the edge direction space

from a lot of preparing pictures. Dissimilar to numerous methodologies that model the dim-level appearance of the face our methodology is computationally extremely quick.

3 Proposed Model

The existing system of driver drowsiness detection system requires a lot of hardware and maintenance. Mainly, using two cameras in the system one for monitoring the head movement and the other one for facial expressions. The other disadvantage is the aging of sensors, and all the sensors are attached to the driver's body which may affect the driver. Many non-intrusive approaches are developed, but the accuracy and pace at which they can alert the driver are not sufficient to prevent the accident. Drowsiness detection based on steering wheel movement measure; standard deviation of lane position proved effective, but these work only in limited circumstances' non-physiological methods like determining the inner state of drives by methods like electroencephalography and electrooculography by placing electrodes around the head and close to eyes. These methods give good results, but implementing them in a car is not a feasible solution to avoid accidents.

The concentration level of the driver decreases due to less rest, long consistent driving, or some other ailment like health issues and so forth. A few reviews on street mishaps conclude that almost 30% of mishaps are a result of the weariness of the driver. At the point when the driver goes on driving for sometimes greater than driver capacity, unnecessary exhaustion is caused and brings about fatigue which results in drowsy condition or loss of cognizance. Our proposed technique is to structure and build up a minimal effort framework, which depends on the installed stage for fatigue discovery that keeps the driver concentrated out and about. Numerous structures and models have been executed to stay away from such mishaps by keeping the entire spotlight and focus on precisely checking the open and shut conditions of the driver's eye continuously. These days, the driver's being in the vehicle is one of the most needed frameworks to maintain a strategic distance from mishaps. Our goal task is to guarantee the safety framework. For improving safety, we are recognizing the eye flickers of the driver and assessing the driver's status, and control the vehicle likewise. Utilizing a non-intrusive fatigue recognition framework is the most encouraging bearing to fabricate a genuine life-relevant arrangement. An answer that could be upgraded by exploiting various strategies.

As data science is a field that identifies some meaning within a data. ML algorithms help to work like that, whereas deep learning is a subfield of ML, and this work uses dlib methods in Python which are also in OpenCV; the live data collected from all the frames were analyzed to detect whether the driver is in fatigue or not from which an alert message can be sent. In the framework, we have utilized facial landmark expectation for eye identification. Facial landmarks are utilized to restrict and speak to remarkable districts of the face. Facial landmarks have been effectively applied to confront arrangement, head present estimation, flicker detection, and significantly

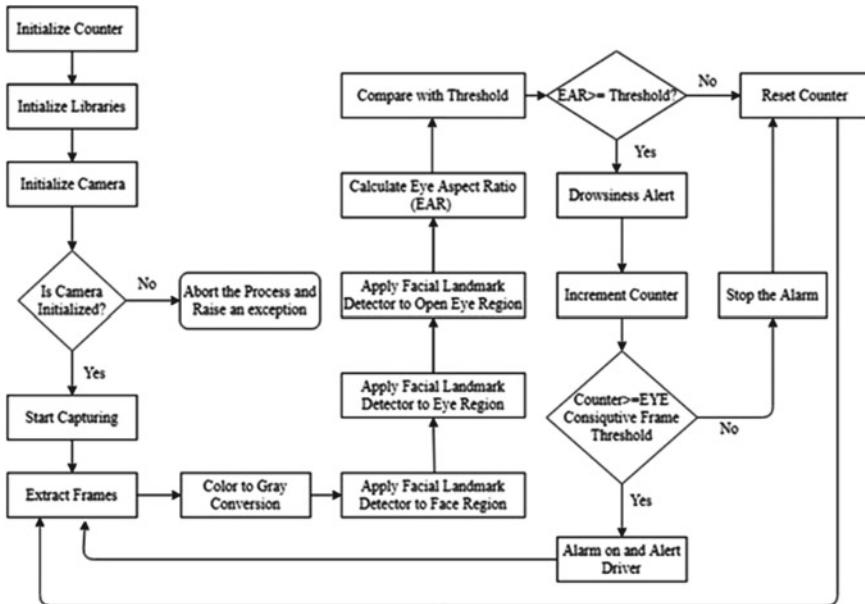


Fig. 1 System design

more. Concerning facial landmarks, our objective is distinguishing significant facial structures on the face utilizing shape forecast techniques. Figure 1 depicts how the paper works to produce the desired result. Whenever the system is started, it first initializes the frame counter to zero. The system then initializes all the libraries required to execute. Using open CV functions, the system starts taking the live feed from the camera. The frames are extracted from the video stream, and the dlib face landmark detection library is applied to detect the face landmarks from the frames obtained. The library employed here identifies 68 facial landmarks of the face present in the frame. The eye landmarks are identified to calculate the eye aspect ratio (EAR). Each eye has six facial landmarks. The EAR is calculated from the obtained data and will be compared to the threshold value. The EAR is the Euclidean distance of the eye landmarks. If the threshold value is greater than the obtained EAR, the system proceeds and extracts the next frame. If the EAR is greater than the threshold value assigned, counter is incremented by one. If the counter is greater than the consecutive frame threshold, the alarm is triggered. The drowsiness is detected, and an alarm sound will alert the driver immediately. After the driver is awake, the EAR will be less than the threshold, so the counter will be reset to zero again. So, the system continuously extracts frames and checks for drowsiness in every frame.

3.1 Proposed Algorithm

- **Input:** Video of the vehicle driver face
 - **Output:** Drowsiness of the driver
1. Start
 2. Whenever the system is started, it first initializes the frame counter to zero.
 3. The system then initializes all the libraries required to execute.
 4. Using OpenCV functions, the system starts taking the live feed from the camera.
 5. The frames are extracted from the video stream, and the dlib face landmark detection library is applied to detect the face landmarks from the frames obtained.
 6. The library employed here identifies 68 facial landmarks of the face present in the frame.
 7. The eye landmarks are identified to calculate the eye aspect ratio (EAR). Each eye has six facial landmarks.
 8. The EAR is calculated from the obtained data and will be compared to the threshold value. The EAR is the Euclidean distance of the eye landmarks.
 9. If the threshold value is greater than the obtained EAR, the system proceeds and extracts the next frame. If the EAR is greater than the threshold value assigned, counter is incremented by one.
 10. If the counter is greater than the consecutive frame threshold, the alarm is triggered. The drowsiness is detected, and an alarm sound will alert the driver immediately.
 11. After the driver is awake, the EAR will be less than the threshold, so the counter will be reset to zero again. So, the system continuously extracts frames and checks for drowsiness in every frame.

For each video outline, the eye landmarks are identified. The eye viewpoint proportion (EAR) among stature and the width of the eye is registered. Where $p1 \dots p6$ is the 2D milestone areas, portrayed in figure beneath. The EAR is for the most part consistent when an eye is open and is drawing near to zero while shutting an eye. It is somewhat individual and head present obtuse. The perspective proportion of the open eye has a little fluctuation among people, and it is completely invariant to a uniform scaling of the picture and in-plane pivot of the face. Since eye squinting is performed by the two eyes synchronously, the EAR of the two eyes has arrived at the midpoint; the proposed system is a non-intrusive approach for detecting the fatigue of the driver. This paper is developed to reduce the cost of implementation and provide safety to the driver by a warning. In this paper, dlib library is used to identify distinct face landmarks of the driver's face, and OpenCV is used as an image processing tool. Image processing with OpenCV proved to be efficient and quick in results. Eye aspect ratio (EAR) is calculated, and a threshold value is set to monitor whether the driver is drowsy or not. Since this approach is non-intrusive, it requires no intervention by the driver for it to work. This system continuously

compares the threshold value with the EAR of the driver. A camera with at least 15 fps is required to implement this system. The results obtained will be on par with many other approaches but at a very low cost for implementation. The goal of this paper is to warn as soon as possible if the system detects the driver is drowsy.

4 Results and Discussions

The choice for the eye state is made dependent on the EAR determined in the past advance. On the off chance that the separation is closed or is almost shut, the eye position is delegated “shut” in any case the eye state is distinguished as “open,” The last advance of the calculation is to decide the individual’s state dependent on a pre-set condition for drowsiness. The normal flicker span of an individual is 100–400 ms (e.g., 0.1–0.4 of a second). Subsequently, if an individual is languid, his eye conclusion must be passed this interim. We set 2 s. If the eyes stay shut for at least two seconds, drowsiness is recognized and alarm concerning this is triggered. We use the confusion matrix as a visual tool to show the performance of a binary classifier in Table 1. The instances in which the subject is detected to be drowsy are classified as positive, and the instance in which the subject is detected to be active is classified as negative. Once all the testing instances are classified, the output labels are compared against the target labels.

From the results obtained with Python and OpenCV which incorporates the accompanying advances: Successful run time capturing of video with sensors, the captured stream will be separated into frames, and every one of the edges will be broke down. Effective discovery of face followed by identification of eye. If the conclusion of the eye for progressive casings was distinguished, at that point, it is named tired state else it is viewed as an ordinary flicker and circle of capturing a picture, and checking the condition of the driver is completed over and over. In this execution during the drowsy state, the eye is not encompassed by a circle or it is not identified, and a relating message has appeared. Our model is intended for the location of the tired condition of the eye and gives a ready sign of cautioning as sound alert, and the experiment results were shown in Table 2. However, the reaction of the driver in the wake of being cautioned may not be sufficient to quit causing the mishap

Table 1 Confusion matrix in testing

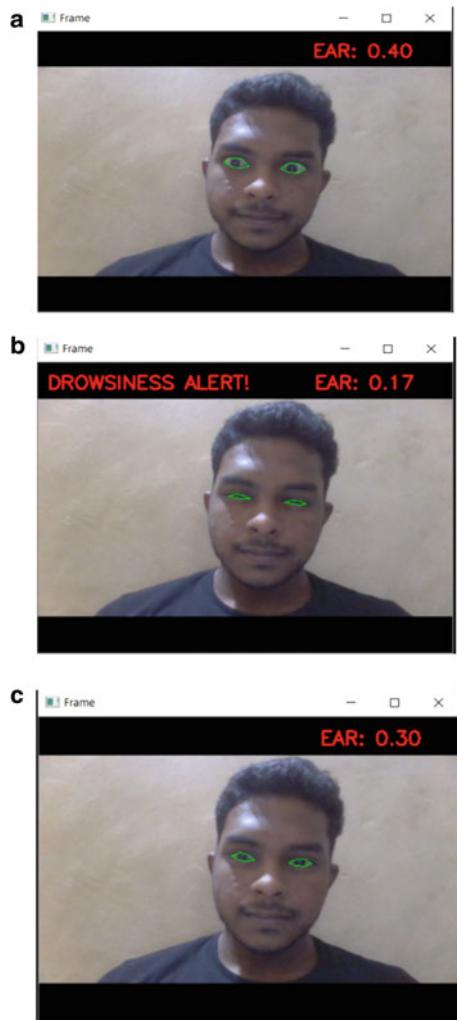
$N = 100$	Positive	Negative
True	TP = 40	TN = 38
False	FP = 4	FN = 18

Table 2 Experimental results of the proposed system

Functional accuracy	Error rate	Sensitivity	Specificity
0.85	0.15	0.90	0.67

implying that on the off chance that the driver is delayed in reacting toward the notice signal, at that point, mishap may happen. Henceforth to maintain a strategic distance from this, we can plan and fit an engine-related framework and synchronize it with the notice signal, so the vehicle will back off in the wake of getting the notice signal consequently. We can likewise furnish the client with an Android application that will give the data of his/her sluggishness level during any trip. The client will know the normal state, and drowsy state was shown in Fig. 2a–c, respectively.

Fig. 2 **a** When the subject eyes are wide open. **b** When a subject is drowsy, an alarm is triggered. **c** Subject is not in a drowsy state, and EAR is displayed



5 Conclusion

Continuous eye fatigue identification calculation was introduced. We quantitatively showed that dlib-based facial landmark identifiers are sufficiently exact to dependably appraise the positive pictures of face and a degree of eye receptiveness. This can be used to detect the drowsiness of the driver cost-effectively with low computational delay. Experimental results show that the real-time accuracy of the developed work is 84%. In most cases, it correctly detects if the driver is drowsy, and an alarm is triggered based on the EAR. As a future scope, the proposed work can be extended in order does not work at its best when the driver is not facing the camera by also sending alert messages about the attentiveness toward the drive as well as by facing the camera, such that it retrieves more accurate results also this system cannot detect the face landmarks accurately when the driver is turning his head left or right. Disambiguation also affects the performance of this system because the camera we used is not a night vision sensor. So, for better accuracy, it is to be improved by using effective cameras that also work on night vision. We took live streaming data, and 85% of it correctly classifies, and to make it more effective, it is suggested to choose more efficient cameras such that it should capture more valid data through live stream. So, the developed paper is 85% accurate. Our proposed system is working for more than 30 frames per second, so the paper works are smooth without any computational problems.

References

1. Hu X, Downie JS (2010, June) Improving mood classification in music digital libraries by combining lyrics and audio. In: Proceedings of the 10th annual joint conference on digital libraries, pp 159–168
2. Amiripalli SS, Bobba V (2018) Research on network design and analysis of TGO topology. *Int J Netw Virtual Organ* 19(1):72–86
3. Amiripalli SS, Bobba V (2019) Trimet graph optimization (TGO) based methodology for scalability and survivability in wireless networks. *Int J Adv Trends Comput Sci Eng* 8(6):3454–3460
4. Amiripalli SS, Bobba V (2019) An optimal TGO topology method for a scalable and survivable network in IOT communication technology. *Wireless Pers Commun* 107(2):1019–1040
5. Jitendra MSNV, Radhika Y (2021) Singer gender classification using feature-based and spectrograms with deep convolutional neural network. *Int J Adv Comput Sci Appl (IJACSA)* 12(2)
6. Sri Jayathi K, Vedachary M (2013) Implementation of the driver drowsiness detection system. *Int. J. Sci. Eng Technol Res (IJSETR)* 2(9):1751–1754
7. Amiripalli SS, Bobba V (2019) Impact of trimet graph optimization topology on scalable networks. *J Intell Fuzzy Syst* 36(3):2431–2442
8. Amiripalli SS, Bobba V (2020) A fibonacci based TGO methodology for survivability in ZigBee topologies. *Int J Sci Technol Res* 9(2):878–881
9. Amiripalli SS, Kumar AK, Tulasi B (2016, Feb) Introduction to TRIMET along with its properties and scope. *AIP Conf Proc* 1705(1):020032

10. Amiripalli SS, Kollu VVR, Jaidhan BJ, Srinivasa Chakravarthi L, Raju VA (2020) Performance improvement model for airlines connectivity system using network science. *Int J Adv Trends Comput Sci Eng* 9(1):789–792
11. Thota JR, Kothuru M, Shammuk Srinivas A, Jitendra MSNV (2020) Monitoring diabetes occurrence probability using classification technique with a UI. *Int J Sci Technol Res* 9(4):38–41
12. Jitendra MSNV, Radhika Y (2020) A review: Music feature extraction from an audio signal. *Int J Adv Trends Comput Sci Eng* 9(2):973–980
13. Ramiah Chowdary P, Challa Y, Jitendra MSNV (2019) Identification of MITM attack by utilizing artificial intelligence mechanism in cloud environments. *J Phys Conf Ser* 1228(1):012044
14. Jitendra MSNV, Naga Srinivasu P, Shanmuk Srinivas A, Nithya A, Kandulapati SK (2020) Crack detection on concrete images using classification techniques in machine learning. *J Crit Rev* 7(9):1236–1241
15. Srinivasu PN, Rao TS, Balas VE (2020) 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. <https://doi.org/10.4018/IJISMD.2020010105>
16. Naga Srinivasu P, Rao T, Dicu AM, Mnerie C, Olariu I (2020) A comparative review of optimisation techniques in segmentation of brain MR images. *J Intell Fuzzy Syst* 38:1–12. <https://doi.org/10.3233/JIFS-179688>
17. Naga Srinivasu P, Srinivasa Rao T, Srinivas G, Prasad Reddy PVGD (2020) A computationally efficient skull scraping approach for brain MR image. *Recent Adv Comput Sci Commun* 13:833. <https://doi.org/10.2174/221327591266190809111928>
18. Ji Q, Zhu Z, Lan P (2004) Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Trans Veh Technol* 53(4):1052–1068
19. Gupta I, Garg N, Aggarwal A, Nepalia N, Verma B (2018) Real-time driver's drowsiness monitoring based on dynamically varying threshold. In: 2018 eleventh international conference on contemporary computing (IC3). IEEE, pp 1–6
20. Manu BN (2016) Facial features monitoring for real time drowsiness detection. In: 2016 12th international conference on innovations in information technology (IIT). IEEE, pp 1–4
21. Rahman A, Sirshar M, Khan A (2015) Real time drowsiness detection using eye blink monitoring. In: 2015 national software engineering conference (NSEC). IEEE, pp 1–7
22. Chen JH, Asch SM (2017) Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med* 376(26):2507

Milestones in Autonomous Vehicle and Evaluation Using Computer Vision



J. Pavan Satish, Sai Harsha, T. Prem Jacob, A. Pravin, and G. Nagarajan

Abstract This research work depicts different improvements done on self-sufficient vehicles and assessment is done dependent on instruments accessible through PC vision. To improve the vehicle's reaction and driving calculations, it ought to be given a similar viewpoint as an individual, with the goal that they can abstain from hitting into objects or smashing. Current days self-driving vehicles have accomplished various achievements to improve driving wellbeing and moving. Individuals rely upon their optical vision to recognize any approaching snags. By giving different boundaries, for example, path checking identification, traffic signals, approaching vehicles, and furthermore, the person on foot framework can be improved to expand the productivity of the accessible framework. Different OpenCV libraries have been created to give designers to make or redesign a current framework.

Keywords Autonomous vehicles · Lane detection modules · Lane edge detection · Lane assist system · Convolutional neural network · Lane departure warning

1 Introduction

Car crashes happen because of individuals' error in misinterpreting traffic stream and their recklessness [1, 2]. The new study says that auto collisions have been expanding since the previous decade. It is because of advancement vehicles innovation and least staff assignment for street safety [3, 4]. Self-sufficient vehicles have been improving for as long as two decades [5, 6]. The pace of advancement of different innovation expects us to rely upon the autonomous framework to accomplish or achieve our day-by-day exercises without people indulging [7–9]. To acquire a superior self-driving vehicle, the proposed technique needs to defeat different difficulties including in improving security proportions of a vehicle. Among these achievements, lane recognition and path checking revelation assume a huge part in the path the executives framework in self-ruling vehicles [10]. Different global organizations are putting

J. Pavan Satish (✉) · S. Harsha · T. Prem Jacob · A. Pravin · G. Nagarajan
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

away cash, man and time in decreasing blunder rate and improving the limit of the framework to incorporate different boundaries like traffic signals, land-mark recognition, and furthermore walkers with the goal that the quantity of street mishaps can be diminished. In path recognition calculations, different enhancements are made to recognize even a bended lane [11]. They have created calculations to keep client vehicles in an assigned path utilizing lane departure warning (LDW) and lane keeping assist system (LKAS) [12]. Another angle remembered for the self-governing framework is to distinguish moving items around the vehicle, along these lines, expanding mindfulness by adding boundaries that influence vehicle way. Not many thoughts were additionally included, for example, a path change module [13–15]. The path change framework can be utilized to recognize the most secure path to change for a vehicle to keep away from any deterrent or improve time efficiency [16–18]. The initial lane discovery modules are created to distinguish a bended way and furthermore discover the ebb and flow of the bended path. The majority of the cameras these days have a restricted inclusion territory. A basic camera focal point of an Advanced Driver-Assistance System (ADAS) can just identify a bit of a profoundly bended lane [1]. A 360° View of the encompassing is given to the vehicle interface [19, 20]. The assessment of the present talked about perspectives are actualizes utilizing different apparatuses and modules accessible in computer vision [21]. We make utilization of Hough Transform to identify the path and furthermore path stamping to expand vehicle familiarity with the environmental factors. Present day innovation permits the vehicle to have a web connection [22, 23]. This encourages the vehicle to know thick traffic zones early and furthermore decline the time intricacy. A few organizations remember a voice partner for assisting the client with exploring or caution him approaching article behind the vehicle [24]. To get a superior self-driving vehicle, the proposed strategy needs to defeat different difficulties including in improving wellbeing proportions of a vehicle. Among these achievements, lane location and path checking revelation assumes a critical part in the path the board framework in self-sufficient vehicles [25, 26].

2 Related Work

The main module utilized for path location was gotten from a paper distributed in 2018 by G. Deng and Y. Wu named “Twofold Lane Edge Detection Method Based on Hough Transform.” In this paper, they utilize a monocular camera to get the picture as greyscale and utilize canny edge location alongside Hough Transform to discover a path and furthermore assist the framework with adjusting the vehicle along with the focal point of the way. This paper likewise acquainted different formulas with distinguish the curve of the bended street and to move it as needs be.

S. Lucashas presented “Optical character acknowledgment with Hough change based neural organizations” which was distributed in 1993, presented different PC vision-based modules to recognize numerous path stamping detection [2]. This

module is additionally actualized in another vision-based administration framework. Chan, Y.-C., Lin, Y.-C., and Chen, P.-C. Distributed paper named “Path Mark and Drivable Area Detection Using a Novel Instance Segmentation Scheme 2019” distributed in 2019. This paper presented the recognition of different path markings to build attention to self-ruling vehicles. This further builds the proficiency of the proposed system [14, 27]. T. Chang and Chen-Ju Chou, “Backside crash cautioning framework by virtue of a backside checking camera” distributed in 2009 acquainted a wellbeing system with alert the driver of approaching obstructions from the back of the vehicle. This module increases the wellbeing of the driver and furthermore builds the familiarity with the driver [22]. This module is incorporated into a proposed framework to 360° see familiarity with the vehicle. J. H. Jung, Y. Shin, and Y. Kwon, “Expansion of Convolutional Neural Network with General Image Processing Kernels,” in 2018. This paper presented expanded proficiency of the Lane the board framework by adding convolutional neural network (CNN) which refines the video contributions through rehashed outline max pool and along these lines expanding clearness of the picture which can be utilized to identify faint edges of the more modest object [15, 28].

3 Proposed System

In this section, the lane management algorithm we have used experimental pictures from a publicly available dataset. First, the image pre-processing module which turns the image into a greyscale and detects for edges using canny edge detection. For some video, processing we are going to use color gradient threshold value to remove the colors. Noise generated is reduced by blurring the image using the Gaussian blurring system. We are going to use color and gradient threshold value to optimize the image (Fig. 1).

On the second, we have the lane detection system used to detect lane and also helps in keeping the vehicle in the dedicated lane. But there is a small catch in lane detection, i.e., regular camera lens coverage cannot capture tightest of turns and curves. Thus, we are going to implement a 360° view of the surroundings. Some of the publishers have included a deep learning approach to implement Advanced Driving Assist System. The third module consists of the lane mark detection system using to improve the awareness of surrounding parameters to the vehicle. To detect the lane markings in the lane, we are going to use various wrapping techniques available in the open CV product documentation to convert the image to the required dimensions using scaling, transformation. The final module consists of a safety protocol, which is used to improve safety measures. They usually consist of halting the vehicle whenever a vehicle detects incoming objects (Fig. 2).

The final module contains object detection and security protocol to ensure the safety probability of the passenger. This module is going to detect various object that comes in the way of the vehicle. It is going to alert passengers and halt the car with various safety precautions. We can implement multiple more features into this

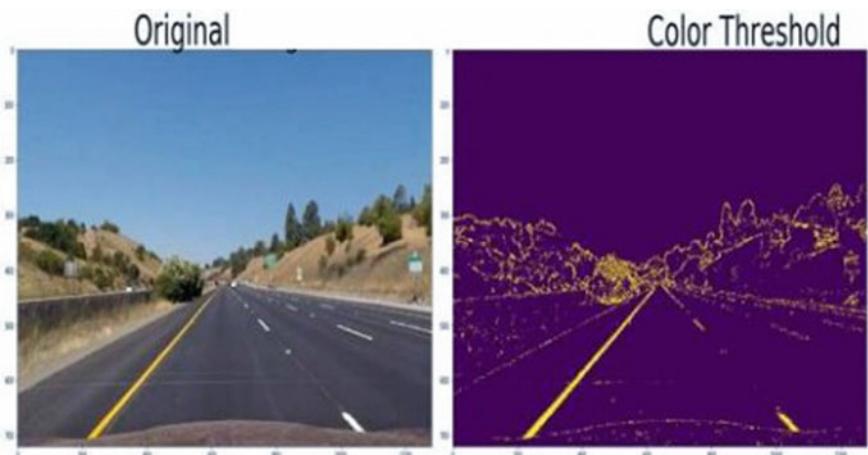


Fig. 1 Gradient thresholding

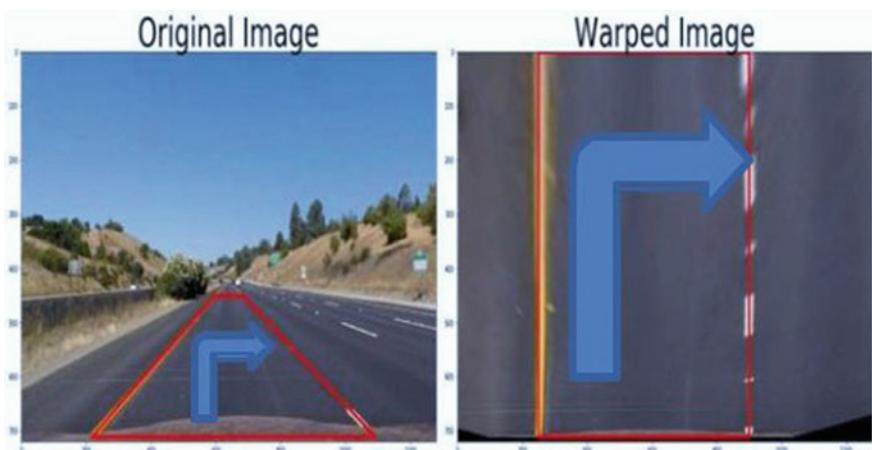


Fig. 2 Wrapping of image

product to make it more user-friendly and more aware of its surroundings. The above image roughly shows the pictorial representation of the 360° view, so that we can increase the awareness in the system (Fig. 3).

The above image roughly shows the pictorial representation of the 360° view, so that we can increase the awareness in the system. Recent technology advancement has increased various combinations of artificial intelligence into camera lensing, thus increasing the adaptive lensing mechanism to implement a robust and adaptive advanced driver-assist mechanism. This mechanism involves many deep learning methods providing less disturbing during the journey (Fig. 4).

Fig. 3 360° view mechanism



Fig. 4 Live object detection

Object detection in lane management system can be implemented using many algorithms like Keins Nearest Neighbor (KNN) that is going to detect the nearest pixel variation to identify an object. Other is real-time object detection, which is using the Gaussian block method to enclose various purposes in blocks and differentiating them according to the motion or any other parameter. Additional passenger security measures are added to ensure that the user can have a relieved journey, thereby decreasing the accident rate in various countries.

4 System Architecture

To start with, the image pre-preparing module which transforms the picture into a dim scale and distinguishes for edges utilizing vigilant edge discovery. On the second, we have the path recognition framework used to distinguish path and furthermore helps in keeping the vehicle in the devoted lane. We utilize canny edge discovery alongside

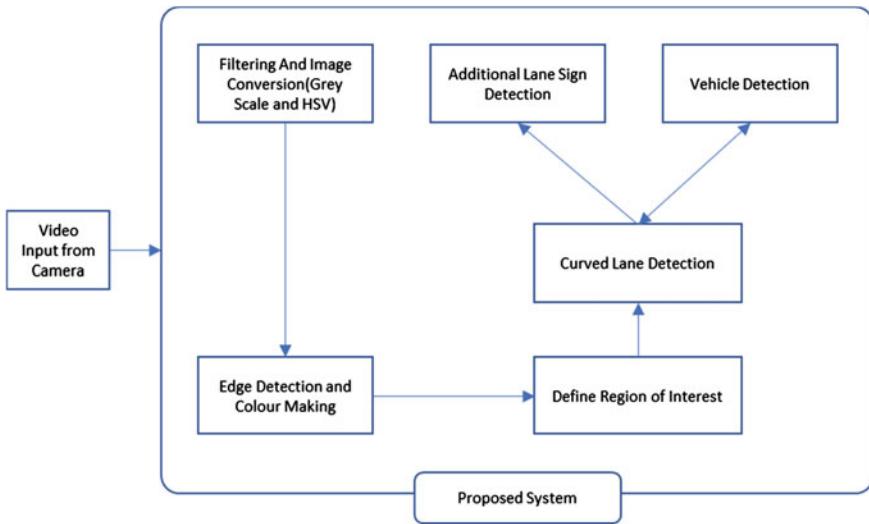


Fig. 5 Overall systems design

Hough Transform. The third module comprises of the path mark location framework utilizing to improve the familiarity with encompassing boundaries to the vehicle. To additional expansion mindfulness among the framework, we can add different path recognition requirements like discovery of traffic lights and approaching article location (Fig. 5).

5 Algorithms Used

- Step 1 The video input is taken from the camera on the top of the car. The data is blurred to remove noise in the image frames.
- Step 2 Then, the video input is converted into grayscale by each frame. Which is better than HSV when dealing with edge detection.
- Step 3 Video is transformed using the threshold method of removing noises less than the threshold value.
- Step 4 Using Erosion, which reduces the edge value, that can be used to remove left out errors, and by using Dilatation, we can thicken the border to former density.
- Step 5 The Remaining edges in the video input are used to find lanes and various lane Markings present on the road.
- Step 6 The second program to find live object detection is implemented to integrate the data into the first output.
- Step 7 The objects that are detected are isolated using contours in computer vision. And this is integrated to find pedestrians, cars, and obstacles on the road.

- Step 8 A safety protocol is implemented in the proposed algorithm to improve the safety of the passenger by alerting the system if any object approaches the vehicle.
- Step 9 All of the data is integrated into the single output panel and allows driving assistance in finding the best route possible.

6 Results and Discussion

From the results, we understand that a working individual after having a tiring day work, cannot concentrate on driving. In this case, the autonomous system in the vehicle can help guide the user in a safe passage or even use the lane keeping assist system to automate driving. Increases the efficiency of humans in daily life. It is advised to rely purely on the autonomous system, but this may lead to laziness. Current standard cameras are not adaptive in low-light areas. This may lead to unreliability during nightfall.

7 Conclusion

This paper contains a summed-up form of different angles associated with autonomous vehicles. This Proposed design of the program is utilized to assess path the executive's ideas utilizing PC vision. We utilize canny edge identification alongside Hough Transform. The proposed was at first acquainted with the self-governing vehicle where a framework is utilized to drive the vehicle; however, this can likewise be utilized as semi-self-sufficient, which can assist the driver with monitoring approaching vehicles from the side and back better compared to standard mirrors. To additional increment mindfulness among the framework, we can add different path discovery limitations like location of traffic lights and approaching article discoveries. For future works, we can incorporate an association with the web so the proposed framework can take the most recent updates of traffic and recommend different options for the objective. These models can be utilized in different vehicle the executive's framework and furthermore progressed advanced mechanics for the improvement of a superior AI. With regards to autonomous vehicles, there are different upsides and downsides.

References

1. Xuan H, Liu H, Yuan J, Li Q (2018) Robust lane-mark extraction for autonomous driving under complex real conditions. *IEEE Access* 6:5749–5765. <https://doi.org/10.1109/ACCESS.2017.2731804>
2. Chen T, Lu S (2015) Context-aware lane marking detection on urban roads. In: 2015 IEEE international conference on image processing (ICIP), Quebec City, QC, 2015, pp 2557–2561. <https://doi.org/10.1109/ICIP.2015.7351264>
3. Andrade DC, Bueno F, Franco FR, Silva RA, Neme JHZ, Margraf E, Amaral R, Dos S (2018) A novel strategy for road lane detection and tracking based on a vehicle's forward monocular camera. *IEEE Trans Intell Transp Syst*:1–11. <https://doi.org/10.1109/its.2018.2856361>
4. Jacob, T. P., & Ravi, T. (2013). Optimization of test cases by prioritization.
5. Sterlin CS, Refonaa J, Ramalavanya R (2006) Secure data offloading using auction based mechanism
6. Velmurugan A, Ravi T (2017) Optimal symptom diagnosis for efficient disease identification using Somars approach. *J Comput Theor Nanosci* 14(2):1157–1162
7. Srinivasan N, Lakshmi C (2016) A novel prediction based tree structured data using machine learning techniques. *Res J Pharm Biol Chem Sci* 7(5):527–531
8. Krishna RSB, Aramudhan M (2014, July) Feature selection based on information theory for pattern classification. In: 2014 international conference on control, instrumentation, communication and computational technologies (ICCICCT). IEEE, pp 1233–1236
9. Nithya G, Shabu SL (2016) A novel framework in reusing the ontological health record. *Res J Pharm Biol Chem Sci* 7(3):215–220
10. Chan Y-C, Lin Y-C, Chen P-C (2019) Lane mark and drivable area detection using a novel instance segmentation scheme. In: 2019 IEEE/SICE international symposium on system integration (SII). <https://doi.org/10.1109/sii.2019.8700359>
11. Deng G, Wu Y (2018) Double lane line edge detection method based on constraint conditions Hough transform. In: 2018 17th international symposium on distributed computing and applications for business engineering and science (DCABES), Wuxi, 2018, pp 107–110. <https://doi.org/10.1109/DCABES.2018.00037>
12. Zhang X, Zhao R (2006) Automatic video object segmentation using wavelet transform and moving edge detection. In: 2006 machine learning and cybernetics, Dalin, China, 2006, pp 3929–3933. <https://doi.org/10.1109/CMLC.2006.258748>
13. Chang T, Chou C-J (2009) Rear-end collision warning system on account of rear-end monitoring camera, 2009, pp 913–917. <https://doi.org/10.1109/IVS.2009.5164401>
14. Lucas S (1993) Optical character recognition with Hough transform-based neural networks. In: IEE colloquium on Hough transforms, London, UK, 1993, pp P7/1–P7/5
15. Rama Mohan Reddy G (2017) Internet of things: power controlling through in smart mobiles. *Int J Pure Appl Math (IJPAM)* 118(17):791–800
16. Brumancia E, Sabarinathan S, Mugesh R (2015) Distributed wormhole detection algorithm for wireless sensor network. *Int Rev Comput Softw* 10(3):307–314
17. Christy A, Praveena A, Shabu J (2019) A hybrid model for topic modeling using latent dirichlet allocation and feature selection method. *J Comput Theor Nanosci* 16(8):3367–3371
18. Vijeya Kaveri V, Maheswari V (2019) Mining social data to identifying user behavior in med help forum on health-related topics. *Int J Recent Technol Eng* 8(2 Special issue 3):1306–1310
19. Kanth RR, Jacob TP (2019) A survey on privacy preserving schemes and performance with multiple and individual data sets. *J Crit Rev* 6(4):57–64
20. Jany Shabu SL, Jayakumar C (2018) Multimodal image fusion using an evolutionary based algorithm for brain tumor detection
21. Minu RI, Nagarajan G, Pravin A (2019) BIP: a dimensionality reduction for image indexing. *ICT Express* 5(3):187–191
22. Pravin A, Jacob TP, Asha P (2018) Enhancement of plant monitoring using IoT. *Int J Eng Technol (UAE)* 7(3):53–55

23. Subhashini R, Jeevitha JK, Samhitha BK (2019) Application of data mining techniques to examine quality of water. *Int J Innov Technol Exploring Eng (IJITEE)* 8(5S). ISSN: 2278-3075
24. Ashokkumar K, Deepak CV, Chowdary DVR (2019, Oct) Sign board monitoring and vehicle accident detection system using IoT. *IOP Conf Ser Mater Sci Eng* 590(1):012015
25. Jung JH, Shin Y, Kwon Y (2018) Extension of convolutional neural network with general image processing Kernels, TENCON 2018. In: 2018 IEEE region 10 conference, Jeju, Korea (South), 2018, pp 1436–1439. <https://doi.org/10.1109/TENCON.2018.8650542>
26. JanyShabu SL, Jaya Kumar C (2018) Multimodal image fusion and bee colony optimization for brain tumor detection. *ARPN J Eng Appl Sci* 13:1819–6608
27. Durga GR, Iswariya R, Pravin A (2006) Enhanced security and immediate acknowledge of moving object in surveillance
28. Nagarajan G, Minu RI, Devi AJ (2020) Optimal nonparametric bayesian model-based multimodal BoVW creation using multilayer pLSA. *Circuits Syst Signal Process* 39(2):1123–1132

Sentinel-2 Images-Based Intelligent Crop Type Determination



L. Sujihelen, N. Naga Praveen Kumar, P. Pramod Sai, and G. Nagarajan

Abstract Agriculture is the primary source of each livelihood which forms the backbone of our country. Nowadays in agriculture, there is a low yield of crops due to the climatic changes, poor irrigation, and reduction in soil fertility. Some of the factors on which agriculture is dependent are soil, climate, flooding, fertilizers, temperature, precipitation, crops, insecticides and herb. The conditions of soil and crop production in farming change from every day as well as all through the developing season. This proposed work predicts which crop is grown in which soil. Also, it predicts which crop is grown in which season. The soil is detected by using the Sentinel-2 images. The proposed work performs better when compared with the existing system.

Keywords IoT · Crop type identification · Random forest (RF) · Sentinel-2A · Climatic changes

1 Introduction

Environmental change influences the worldwide agribusiness and nourishment security in complex manners. Currently, there are two primary strategies for crop arrangement utilizing satellite remote-detecting innovation: One strategy is to utilize the phantom highlights of high spatial goals information joined with multi-temporal attributes and some assistant highlights, (for example, height data and textural data) to improve characterization precision [1]. The other path is to utilize high worldly goals information by investigating crop phonological succession examples or time arrangement vegetation files, joined with some ghastly highlights [2]. The primary strategy is mostly founded on the distinctive unearthly qualities of various crops; however, the wonder of outside bodies with a similar range certainly affects characterization exactness. Albeit expanding height and textural and multi-temporal attributes can improve crop arrangement precision as per crop qualities, and there is still a lot of opportunity to get better. The other crop arrangement strategy includes concentrating

L. Sujihelen (✉) · N. Naga Praveen Kumar · P. Pramod Sai · G. Nagarajan
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

the development design as well as phenology attributes of land. The time arrangement information of huge fleeting goals pictures was utilized for recognizable proof. This strategy adequately improves the arrangement exactness yet in addition has more advanced prerequisites for information source. The obtaining of pictures with a similar condition over progressive timespans additionally constrains far reaching usage of the technique.

The fundamental purpose of horticultural crop the board in any nation is to protect nutrition assets for its populace [3]. Cultivating is today a monetarily, common and earth touchy segment. Satellite remote expose presently carries a goal, globally and exact way to deal with asset the controlling and has gotten urgent to agrarian inventories, crop forecast and checking crop wellbeing. Crop data incorporates both crop type as well as crop condition. Crop data is significant in showcasing of horticultural items in broadly and universally, in approving yield claims, implementing rural process, and in ground explicit yield the board. Obvious, infrared and the microwave observation are delicate to altogether different yield qualities. Information from optical as well as radar sensors is in this way reciprocal. Manufactured F-Stop Radar (SAR) reacts to the huge range crop structure insulator properties of the crop covering.

Reasonable forecasts are required by legislature to appraise well in front of reap time. Right now, strategy to enhance the estimation of corn crop in front of collect time to produce for the particular corn-developing conditions in the focal locale of nation. Complete measurable examinations have not been conceivable because of inadequate field information now and again, and crop condition evaluations have been subjectively settled through the visual correlations. These fundamental discoveries ought to be contrasted and crops in other land regions and under various creation frameworks. Specifically, examine is expected to evaluate the essentialness of immersion impacts and to research the collaboration between the optical as well as SAR symbolism [4].

Yield data incorporates crop type and crop position. Notwithstanding distinguishing the type of crop and assortment, recognizing crop development phase can be of worth. Crop situation is approximately characterized as the power and soundness of a crop. Crop situation is frequently identified with yield efficiency; however, this relationship was perplexing. Crop situation pointers can incorporate stature, leaf region, plant water substance, chlorophyll as well as nitrogen, among others. These markers should be connected to the trim development stage, as well as in this way it is important to screen those pointers beyond the whole developing season [5]. The target of the present examination is to explore the presentation and appropriateness of Sen2-Agri framework in the heterogeneous editing framework in India for operational crop type mapping at bundle goals.

2 Related Work

In past examinations, crop distinguishing proof was normally performed utilizing single-source, high spatial, or the high worldly goals fragile-detecting information

or complicated fusional information. Landscape information is generally utilized in yield order considers as a result of their qualities being accessible at no expense and having huge spatial and high fleeting goals. Be that as it may, the restricted highlights from single-date pictures can't meet the prerequisite of fine crop type grouping correctness. Along these lines, Landsat information is regularly utilized related to other information, for example, moderate resolution imaging spectro radiometer (MODIS) radiometer what's more, Sentinel, to make better worldly goals and get the phenology highlights. Multiple source information combination improves order precision yet additionally presents distinctive sensor adjustment mistakes and the tedious work of handling multi-source information.

MODIS information is likewise generally utilized in the paper of crop time arrangement information because of their high transient goals and accessibility [6, 7]. Be that as it may, because of the spatial goals of the examination zone, it is reasonable for enormous scale crop recognizable proof. In previous two years, the Sentinel information has been utilized in rural applications as well as information combination considers, for example, crop water request computations, crop phenology observing and debacle checking at no expense and with high spatial and fleeting goals. Additionally, the better worldly goals and special vegetation red edge band data in Sentinel information give more extravagant highlights to edit type fine grouping. AI is an approach to contemplate the PC's capacity to reproduce new human learning practices, gain new human abilities, and rearrange actual structures to constantly make better a PCs' exhibition [8, 9]. The use of AI has spread to different parts of man-made brainpower, for example, master frameworks, design acknowledgment, insightful robots, and numerous different fields.

AI for remote sensing reversal incorporates counterfeit neural systems, bolster support vector machines (SVMs), irregular woods and other assortment strategies, just as basic assumption thinking and the neuro-fluffy, hereditary calculations. Compare with customary order techniques, AI successfully uses more highlights and has the benefits of straightforward activity, brief timeframe utilization, and vigor in various information volumes and distinctive arrangement types. AI dependent on the SVMs, arbitrary timberland, and profound neural system techniques has numerous significant applications in crop ID dependent on remote-detecting pictures and has accomplished an elevated level of exactness contrasted and the strategies utilized in past investigations. Among these techniques, the SVM is demonstrated to be feebly touchy to highlight measurements as well as understand the vigor of the high impact, which was applied to multi-feature distinguishing proof issues. The arbitrary backwoods technique is pitifully influenced by information size affectability and commotion in the information and has a solid speculation capacity in multi-classification issues. Accordingly, SVMs and irregular backwoods have become significant research strategies for crop multi-classification issues.

Thinking about the present minimal effort, high-goals information source necessities for crop distinguishing proof and the extraction and utilization of more extravagant translation characters and ID highlights, high worldly as well as high spatial goals Sentinel-2A information is utilized as the information source right now. The SVMs and irregular woodland technique are contrasted and conventional

grouping strategies formulate feature crop recognizable proof. The irregular timberland technique is utilized for highlight determination right now, the best model for multi-featured crop grouping is developed.

3 Existing System

At present, there are two fundamental strategies for crop characterization utilizing satellite remote-detecting innovation: One strategy is to utilize the ghostly highlights of high spatial goals information joined with multi-temporal qualities and some helper highlights, (for example, height data and textural data) to improve order precision [6, 7]. The other path is to utilize high fleeting goals information by investigating crop phonological grouping examples or time arrangement vegetation records, joined with some otherworldly highlights. The principal strategy is for the most part dependent on the distinctive otherworldly qualities of various crops, yet the wonder of remote bodies with a similar range certainly affects arrangement exactness.

Albeit expanding height and textural and multi-temporal attributes can improve crop arrangement precision as indicated by crop qualities, there is still a lot of opportunity to get better. Thinking about the present minimal effort, high-goals information source necessities for crop distinguishing proof and the removal and utilizing more extravagant translation characters and ID highlights, huge worldly and large spatial goals Sentinel-2A information was utilized high-goals information source prerequisites for crop recognizable proof.

3.1 *Disadvantages of the Existing System*

- Overfitting issue in multi-dimensional component information characterization.
- Not reasonable for enormous districts.
- Time expending.

4 Proposed System

The proposed algorithm is suitable for large-scale crop identification is shown in Fig. 1. It also works well in small regions. It produces the result with high accuracy and also accurate forecast crop yield in a short time. Crop information includes crop type and crop condition. The condition of the crop has biomass, height, leaf area, plant water content, chlorophyll and nitrogen, among others.

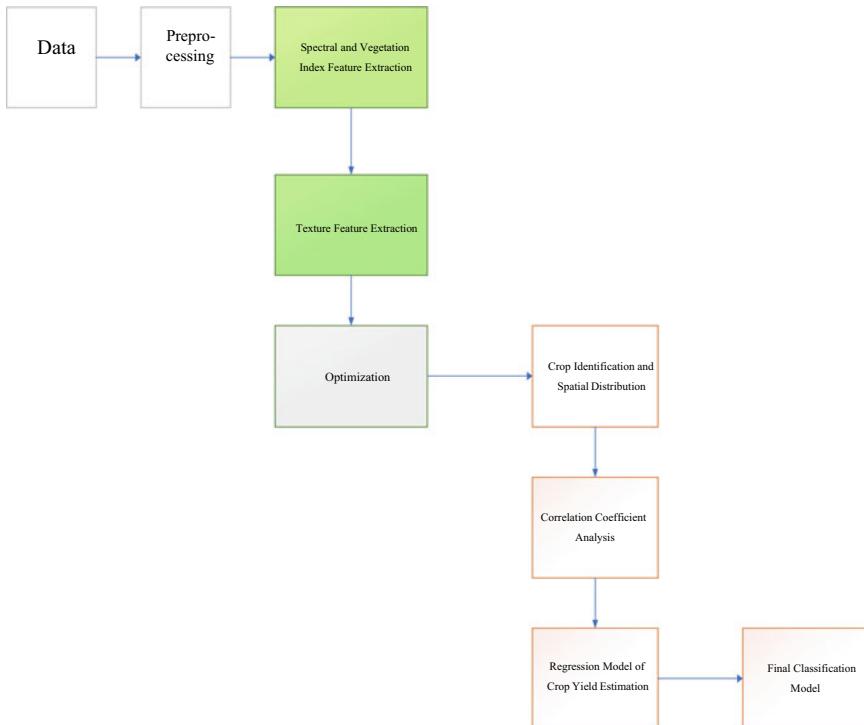


Fig. 1 Overview of the proposed system

4.1 Advantages of the Proposed System

- Short time utilization
- Suitable for huge scale crop distinction
- Work sensibly well in little areas
- High precision
- Accurate figure crop.

To process the proposed system an image enhancement algorithm is used for preprocessing steps convert color images to gray scale. Principal Component Analysis (PCA) technique is used in the proposed system for image whitening. The dataset has been trained by the algorithm based on the epoch value. The epoch value which says that how many times the data has to be trained and it helps to produce the result with high accuracy. The algorithm automatically takes some set of data to the training part and some set of data to the testing part. So that the data has been trained and tested various times and generates the result. In this project, we have been taken nearly 15,000 images of data to process shown in Fig. 2.

The algorithm used in the proposed method is Random Forest Method.

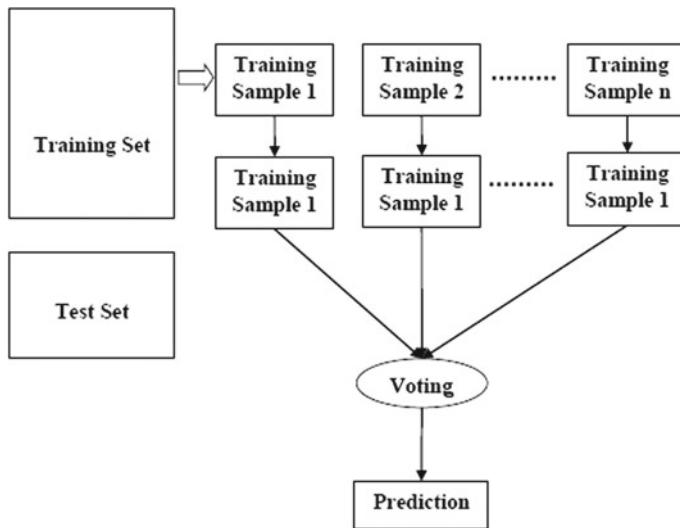


Fig. 2 Processing of training data

- 1 First, start with the selection of random samples from a given dataset.
- 2 Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- 3 In this step, voting will be performed for every predicted result.
- 4 At last, select the most voted prediction result as the final prediction result.

5 Results and Discussion

Some of the images that have been taken in our data set are soya bean, weed etc.

5.1 *Training and Validation Accuracy*

The given data has been trained based and the epoch value and validate the accuracy of the trained data is shown in Fig. 3.

5.2 *Training and Validation Accuracy*

The given data has been trained based on the epoch value and validate the loss of trained data is shown in Fig. 4.

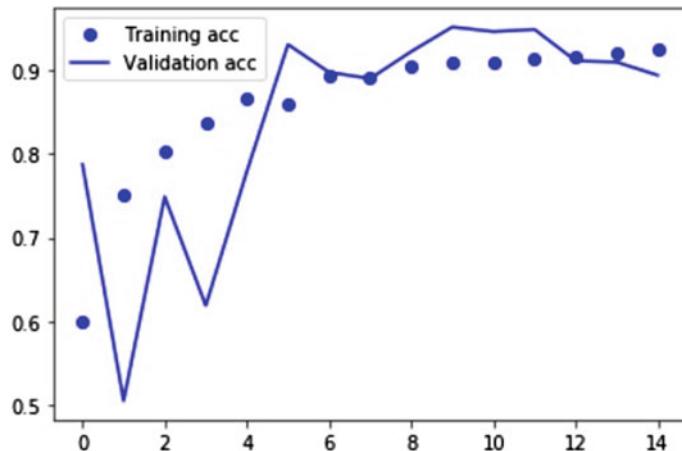


Fig. 3 Training and validation accuracy

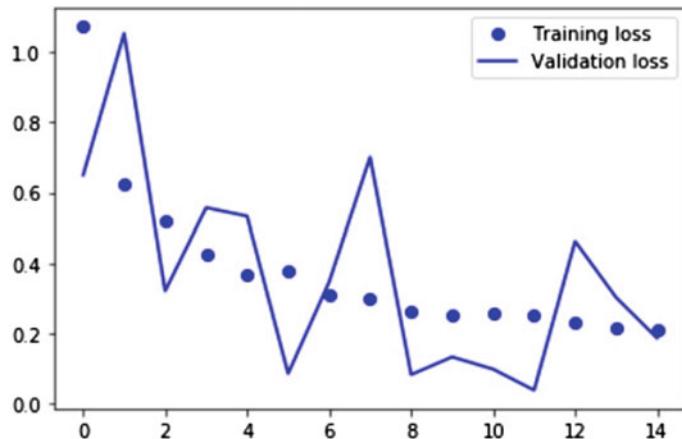
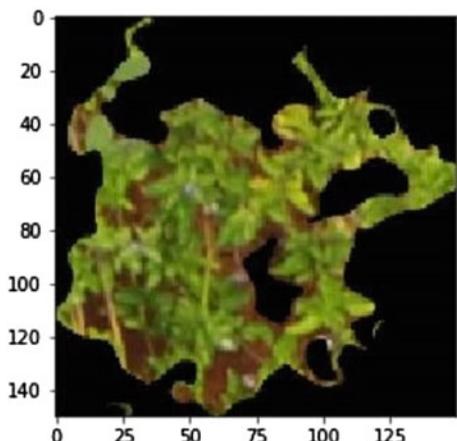


Fig. 4 Training and validation loss

5.3 Shape of the Input Image

After the training of the dataset has been finished the algorithm automatically generates the shape of the given input image and predict the output shown in Fig. 5

Fig. 5 Shape of the input image



6 Conclusions

A supervised classification approach based on Sentinel-2 backscatter time series was proposed, which exploited the temporal signatures of crops. The proposed work will predict which crop will grow in the soil. It predicts very fast when compared with the existing system. The image data has the ability to classify more crop types like wheat maize cotton and so on. The prediction increases the classification accuracy compared with other techniques.

References

1. Vorobiova N, Chernov A (2017) Curve fitting of MODIS NDVI time series in the task of early crops identification by satellite images. *Proc Eng* 201:184–195
2. Zhan Y et al (2018) The effect of EVI time series density on crop classification accuracy. *Optik* 157:1065–1072
3. Hansen MC, Loveland TR (2012) A review of large area monitoring of land cover change using Landsat data. *Remote Sens Environ* 122(1):66–74
4. Townshend J et al (2012) Global characterization and monitoring of forest cover using Landsat data: Opportunities and challenges. *Int J Digit Earth* 5(5):373–397
5. Gilbertson JK, Kemp J, Niekerk AV (2017) Effect of pan-sharpening multi-temporal Landsat 8 imagery for crop type differentiation using different classification techniques. *Comput Electron Agric* 134:151–159
6. Nagarajan G, Minu RI (2018) Wireless soil monitoring sensor for sprinkler irrigation automation system. *Wireless Pers Commun* 98(2):1835–1851
7. Hu Q et al (2017) How do temporal and spectral features matter in crop classification in Heilongjiang Province, China? *J Integr Agric* 16(2):324–336
8. Wang Q, Atkinson PM (2018) Spatio-temporal fusion for daily sentinel-2 images. *Remote Sens Environ* 204:31–42

9. Cherian S, Singh CS, Manikandan M (2014) Implementation of real time moving object detection using background subtraction in FPGA. In: 2014 international conference on communication and signal processing, Melmaruvathur, pp 867–871

An Efficient Modeling Based on XGBoost and SVM Algorithms to Predict Crop Yield



G. S. Mallikarjuna Rao, Sujani Dangeti, and Shanmuk Srinivas Amiripalli

Abstract India is an agricultural country; agriculture is known as the backbone of the Indian economy, where more than 60% of the Indian population depend on agriculture for their living. Crop yield prediction is an essential issue in agriculture. Crop yield analysis can be done by using machine learning techniques. Any person in the agriculture field will expect the yield that he is about to get. The things that need to consider for crop yield prediction include analyzing various parameters like pH value to determine soil fertility, location, etc. The number of nutrients is phosphorous (P), nitrogen (N), etc. The soil type, nutrients present in the soil, rainfall, soil composition, and all these things are considered, and all the data mentioned is analyzed. The data is trained by using suitable techniques and creating a model by using machine learning. The proposed system has the best accuracy in crop yield prediction, which shows precise results. The system provides suitable recommendations for the end user about the fertilizers that are fit for good crop yield depending upon the soil's climate and parameters. This benefits the farmer by increasing and support vector machine (SVM). The support vector machine is a supervised machine learning technique that is used to analyze the data and predict the crop yield in our system. XGBoost algorithm is a machine learning method that refers to “extreme Gradient Boosting.” It tunes the data and gives accurate, efficient, and scalable results. Crop yield production increases the revenue. The system uses two algorithms, known as the XGBoost algorithm.

Keywords Supervised machine learning technique · Predicting crop yield · Support vector machine · XGBoost algorithm introduction

G. S. Mallikarjuna Rao · S. Dangeti

Department of Computer Applications, Gayatri Vidya Parishad College of Engineering (Autonomous), Madhurawada, Visakhapatnam, Andhra Pradesh, India
e-mail: gsmrao_mcahod@gvpce.ac.in

S. S. Amiripalli (✉)

Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam, Andhra Pradesh, India

1 Introduction

India is an agricultural country, and also the population is very high. India has three climatic seasons, namely winter, rainy, and summer. We suggest to the farmers which crop is best in which season. This helps farmer in cropping time. We suggest the best crop for each season, like in the rainy season go for rice, maize. The type of crop plays an essential role in predicting crop yield. The main aim of the model is to predict the yield of the crop by using previous data. In this paper, we can compare the difference between the two algorithms. For better crop yield, we also depend on the type of soil, amount of fertilizers in that soil, and rainfall in that particular area; if it is rice crop better to go in the rainy season, if it is winter season better to go for tomato, or if it is summer better to go for watermelon. Crop yield prediction is an important one for both farmers and the government. The prediction of crop yield will give farmers high income, and the government secures the food resources. It also helps the Indian economy. Also, the Indian economy is depending upon agriculture. Generally, a machine learning algorithm will be predicting the best-fitting model to predict the crop yield. The challenge is to build an efficient model that predicts the crop yield output and to give our best accuracy. We need to compare different algorithms, which gives us less error rate fit to the model. The present paper compares algorithms like the support vector machine and XGBoost algorithm, and we predict the output by using the best-fit algorithm among the two. This paper's content below shows the proposed system, architecture, and experimental results in different paper sections.

2 Literature Survey

In [1–4], author used data mining methods to analyze the crop yield prediction. This paper mainly focused on creating a user-friendly model for the farmers that results in rice yield analysis depending upon available data. To maximize the production, different techniques were used from data mining; the techniques like k means algorithm show the pollution factor present in the atmosphere. In [5], the author presented a clear, detailed review dedicated to the agriculture production system using machine learning. In this paper, big data, high-performance computing, and other methods to analyze process in different agricultural sectors are used. The author used the support vector machine. Manjula et al. 2017 [6] used data mining techniques that focused on a model created using a prediction model that presented a data mining technique based on association rule mining for Tamil Nadu, India's selected region. Their experimental results showed that the proposed work efficiently predicted the crop yield production based on the parameters they used and predicted the crop yield. Ghosh et al. 2014 [7] proposed a system using machine learning technique and artificial neural networks that resulted in training the data that increased the accuracy of the model and the region and resulted in the form of a guide to recognizing the soil

properties that are relevant for plant growth and protection. Crossa and Zelenskiy et al. 2014 [8–12] used three experiments of genotypes in six different environments and showed how genotypes and environments interactions are dependent on each other. They experimented in many factors and analyzed the interactions. Ji, Sun, Yang 2017 [13–16] proposed a system and evaluated the model based on ANN to predict rice crop yield in a region of Fujian, China. They researched the mountainous region. Their work showed a comparison between the artificial neural networks and linear model. They trained the data by using models fed to the historical yield of data and field-specific rainfall data, and other weather records. They proved that the neural networks model, when performed with a linear model, resulted in high accuracy and accurate crop yield prediction, but they needed a massive amount of data. In [17, 18], the author produced crop yield by assembling machine learning models in this paper. Adaptive and AdaSVM are the proposed models. The implementation is done by using the techniques adaptive and AdaSVM. In [8], the author provided an approach to forecasting crop yield depending on various climatic parameters. The entire paper is implemented by using a decision tree. The algorithm produced the most climatic parameters on crop yield in particular selected districts. The author also developed a user-friendly web page to predict the climatic parameters that influence crop yield production. In [19], the author proposed a system for crop cultivation prediction. This paper produced a formula to match the crops with soil and fertilizer recommendations. In [20–24], the author used k -nearest neighbor and SVG to predict the crop yield at regional and other global parameters for its high accuracy and efficiency.

3 Proposed Method

The proposed system includes the prediction of crop yield by using various efficient algorithms and suggesting the number of fertilizers required for good crop yield. This is the sample dataset used in our system. Table 1 data predicts crop yield based on four factors like area, season, crop, and year. We create a machine learning model and predict the production and train the model. From Table 2, we can predict the production rate used to get the proper yield based on the parameters like nitrogen, phosphorous, etc. The output is the amount of fertilizer that is used respectively. In this, the input parameters 1, 2, 3, 4, and 5 represent high, very high, average, low, and deficient quantities present in the soil in that region [25].

3.1 Packages Used in the Implementation

1. Jupiter
2. NumPy
3. Pandas

Table 1 Sample dataset of crop data

State name	District name	Crop year	Season	Crop	Area	Production
Andaman and Nicobar	Nicobar	2000	Kharif	Aocurnut	1254	2000
Andaman and Nicobar	Nicobar	2000	Kharif	Pulses	2	1
Andaman and Nicobar	Nicobar	2000	Kharif	Rice	102	321
Andaman and Nicobar	Nicobar	2000	Whole year	Banana	176	641
Andaman and Nicobar	Nicobar	2000	Whole year	Cashew nut	720	165
Andaman and Nicobar	Nicobar	2000	Whole year	Coconut	18,168	65,100,000
Andaman and Nicobar	Nicobar	2000	Whole year	Dry ginger	36	100
Andaman and Nicobar	Nicobar	2000	Whole year	Sugarcane	1	2
Andaman and Nicobar	Nicobar	2000	Whole year	Sweet potato	5	15

Table 2 Sample fertilizer data

S. No.	N	P	K	Amt N	Amt P	Amt K
1	3	5	6	64	50	60
2	1	4	2	40	46	30
3	5	1	5	93	16	32
4	3	1	3	63	20	39
5	4	6	6	87	37	39
6	2	1	1	65	19	32
7	3	5	6	64	50	60
8	1	4	2	40	46	30
9	5	1	5	93	16	32

4. Matplotlib.pyplot
5. Scikit-learn
6. Tensor flow.

3.2 Architecture of Proposed System

Metadata is information about data. Metadata provides information on certain aspects. Means of creating the data time and date of creation Creator and author of data Location on a computer network. Data preprocessing means an empty variable is replaced with integer values. Data is primarily noisy. Sometimes it has missing values and also false values. Data in the real world is dirty, like incomplete means missing attribute value, noisy means containing errors in this system, and predicting production value is not less than zero. Data preprocessing helps improve accuracy. Duplicate or missing data may cause incorrect or even misleading results. Data preparation, cleaning, and transformations comprise most of the work in a data mining application [26] (Fig. 1).

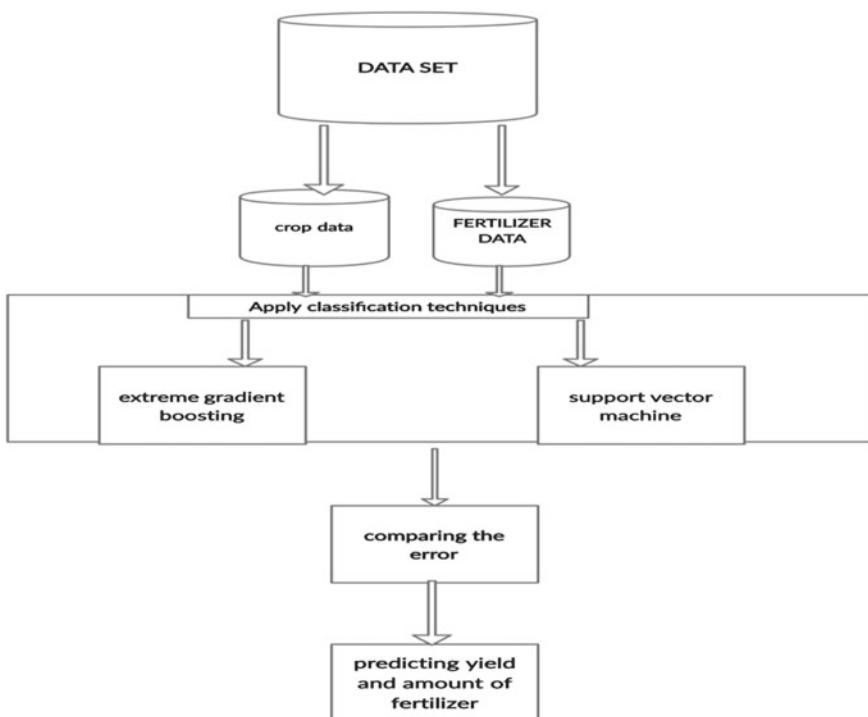


Fig. 1 System design

4 Proposed Algorithm

XGBoost is perfect when it comes to small-to-medium-sized datasets. It will perform well for both regression and classification tasks. XGBoost is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. The model performance of XGBoost is dominating the other datasets that are either well structured or in a tabular form. This entire process helps in predicting models based on regression and classification techniques.

4.1 XGBoost Algorithm

- **Input:** metadata
- **Output:** predicts crop yield
 1. Start
 2. Initial compute base model, which will give one output this output average of all these values.
 3. In this phase, the system will compute residuals, which are essentially errors. In this case, we will be utilising a simple loss function.
 4. Subtraction actual value with predicted value $h_1 = y - f_0$ (here “y” means actual value and “ f_0 ” means predicted value, h_1 means residual error).
 5. Now f_0 and h_1 give f_1 , the f_1 values are less than the f_0 .
 6. To improve performance, create a decision tree to reduce residual error.
 7. We are adding a decision tree sequentially with the help of residual values.
 8. This can be done until residual values will be minimized.
 9. Stop

4.2 Support Vector Machine

Support vector machine is known to be a critical classifier that is characterized by isolating a hyperplane. Basically, in two dimensions, the line that divides a plane into two parts includes one class on both sides. It creates a nonlinear class boundary and constructs a more complex linear classifier known as a hyperplane. Types of kernel functions are

1. linear kernel function: $K(u_i, u_j) = u_i \cdot u_j$
 2. polynomial kernel: $K(u_i, u_j) = (u_i \cdot u_j + c)^d$
- **Input:** metadata
 - **Output:** prediction result
1. Start

2. We are using a hyperplane and best right margin to differentiate our data.
3. The objective of the support vector machine algorithm is to find a hyperplane in n -dimensional space.
4. Let us take a simple example that is the linearly separable case in this case, take attributes $y = w_0 + w_1x_1 + w_2x_2 + w_3x_3$. Here y is the output, x_i indicates the attribute's value, and w_i , the learning function, has four weights to learn.
5. $y = b + \sum \alpha_i y_i(i)$. X , where y_i denotes the class value of training example $x(i)$, (dot), indicates the dot product between the input (x) and each support vector $x(i)$.
6. $y = b + \sum \alpha_i y_i k(x_i, x)$ here $k(x_i, x)$ known as kernel function, the main aim of kernel function is to convert two-dimensional data into high-dimensional data.
7. End

5 Results and Discussions

In this paper, we made an effort to show the crop production analysis processed by implementing both the algorithms—extreme gradient boosting algorithm (XGBoost) and support vector machine (SVM). These two models were experimented with various types of crops in India to predict the output result. Even the training of fertilizer data is possible and evaluated how much is needed for that particular land in that region. The two models were compared, and among them, XGBoost gave the highest accuracy. The comparison plotted in Fig. 2 shows the difference between

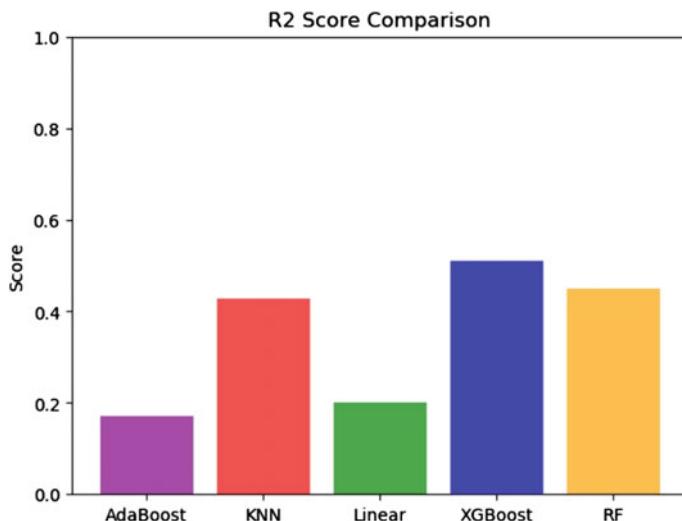


Fig. 2 Score comparison

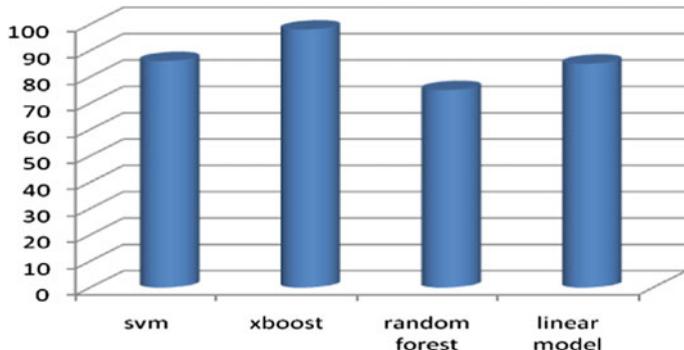


Fig. 3 Score comparison

```
In [63]: print("MAE =",mean_squared_error(b_test,b_pred))
print("MAE =",mean_absolute_error(b_test,b_pred))
print('R2 Score =', r2_score(b_pred, b_test))

MAE = 378758.9899470345
MAE = 83.72156521521087
R2 Score = 0.8674965138838631

In [64]: model.score(a_train,b_train)

Out[64]: 0.9999793170466165

In [65]: model.score(a_test,b_test)

Out[65]: 0.9258225341802667
```

Fig. 4 Result comparison

other models. The user needs to enter the data and details to predict the output, as shown in Figs. 3 and 4.

6 Conclusion

The prediction of crop yield and comparison between other models is successfully made and predicted. Along with that, an efficient algorithm is also obtained. In the future, various techniques help the user understand more clearly the production and prediction of crop yield. Also, developing a system helps the farmer grow crops that are more beneficial and gives better yield.

References

1. Mishra S, Mishra D, Santra GH (2016) Applications of machine learning techniques in agricultural crop production: a review paper. *Indian J Sci Technol* 9(38):1–14
2. Ramesh D, Vardhan BV (2015) Analysis of crop yield prediction using data mining techniques. *Int J Res Eng Technol* 4(1):47–473
3. Devika B, Ananthi B (2018) Analysis of crop yield prediction using data mining technique to predict annual yield of major crops. *Int Res J Eng Technol* 5(12):1460–1465
4. Amiripalli SS, Bobba V (2018) Research on network design and analysis of TGO topology. *Int J Netw Virtual Organ* 19(1):72–86
5. Amiripalli SS, Bobba V (2019) Trimet graph optimization (TGO) based methodology for scalability and survivability in wireless networks. *Int J Adv Trends Comput Sci Eng* 8(6):3454–3460
6. Amiripalli SS, Bobba V (2019) An optimal TGO topology method for a scalable and survivable network in IOT communication technology. *Wireless Pers Commun* 107(2):1019–1040
7. Crossa J, Cornelius PL (1997) Sites regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. *Crop Sci* 37(2):406–415
8. Jitendra MSNV, Radhika Y (2021) Singer gender classification using feature-based and spectrograms with deep convolutional neural network. *Int J Adv Comput Sci Appl (IJACSA)* 12(2)
9. Sri Jayathi K, Vedachary M (2013) Implementation of the driver drowsiness detection system. *Int J Sci Eng Technol Res (IJSETR)* 2(9):1751–1754
10. Ji B, Sun Y, Yang S, Wan J (2007) Artificial neural networks for rice yield prediction in mountainous regions. *J Agric Sci* 145(3):249
11. Amiripalli SS, Bobba V (2019) Impact of trimet graph optimization topology on scalable networks. *J Intell Fuzzy Syst* 36(3):2431–2442
12. Amiripalli SS, Bobba V (2020) A Fibonacci based TGO methodology for survivability in ZigBee topologies. *Int J Sci Technol Res* 9(2):878–881
13. Gandhi N, Armstrong LJ (2016) Rice crop yield forecasting of tropical wet and dry climatic zone of India using data mining techniques. In: 2016 IEEE international conference on advances in computer applications (ICACA), Oct 2016. IEEE, pp 357–363
14. Amiripalli SS, Kumar AK, Tulasi B (2016) Introduction to TRIMET along with its properties and scope. *AIP Conf Proc* 1705(1):020032
15. Amiripalli SS, Kollu VVR, Jaidhan BJ, Srinivasa Chakravarthi L, Raju VA (2020) Performance improvement model for airlines connectivity system using network science. *Int J Adv Trends Comput Sci Eng* 9(1):789–792
16. Eswari KE, Vinita L (2018) Crop yield prediction in Tamil Nadu using Bayesian network. *Int J Intellect Adv Res Eng Comput* 6(2)
17. Thota JR, Kothuru M, Shammuk Srinivas A, Jitendra MSNV (2020) Monitoring diabetes occurrence probability using classification technique with a UI. *Int J Sci Technol Res* 9(4):38–41
18. Jitendra MSNV, Radhika Y (2020) A review: music feature extraction from an audio signal. *Int J Adv Trends Comput Sci Eng* 9(2):973–980
19. Rice crop yield prediction using data mining techniques: an overview. *Int J Adv Res Comput Sci Softw Eng* 9
20. Ramiah Chowdary P, Challu Y, Jitendra MSNV (2019) Identification of MITM attack by utilizing artificial intelligence mechanism in cloud environments. *J Phys Conf Ser* 1228(1):012044
21. Jitendra MSNV, Naga Srinivasu P, Shammuk Srinivas A, Nithya A, Kandulapati SK (2020) Crack detection on concrete images using classification techniques in machine learning. *J Crit Rev* 7(9):1236–1241
22. Srinivasu PN, Rao TS, Balas VE (2020) 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. <https://doi.org/10.4018/IJISMD.2020010105>

23. Naga Srinivasu P, Rao T, Dicu AM, Mnerie C, Olariu I (2020) A comparative review of optimization techniques in segmentation of brain MR images. *J Intell Fuzzy Syst* 38:1–12. <https://doi.org/10.3233/JIFS-179688>
24. Naga Srinivasu P, Srinivasa Rao T, Srinivas G, Prasad Reddy PVGD (2020) A computationally efficient skull scraping approach for brain MR image. *Recent Adv Comput Sci Commun* 13:833. <https://doi.org/10.2174/2213275912666190809111928>
25. Veendhara S, Misra B, Singh CD (2014) Machine learning approach for forecasting crop yield based on climatic parameters. In: 2014 international conference on computer communication and informatics, Jan 2014. IEEE, pp 1–5
26. Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE et al (2016) Random forests for global and regional crop yield predictions. *PLoS One* 11(6):e0156571

Concentration Level of a Learner Using Facial Expressions on E-Learning Platform



S. Vijayakumar, Karlapudi Rahul Sai, and Bhumireddy Sohith Reddy

Abstract India is constantly considered as one of the biggest system of instructive establishments. Albeit a few imperatives are being related with our learning framework. We attempt to give a similar substance of educating to all understudies with various entomb individual aptitudes. The most significant factor is absence of understudy inspiration toward a subject, course, and so forth. Versatile learning is an instructive strategy that uses PCs as an intuitive educating gadget. In existing most instructive operators don't screen commitment expressly, yet rather accept commitment and adjust their connection dependent on the understudy's reactions to questions and undertakings. Consequently, unique understudy conduct investigation is an initial move toward a mechanized instructor input device for estimating understudy commitment. In our framework, we propose a crossover design framework summoning understudy facial feeling acknowledgment, eye stare checking, head developments distinguishing pieces of proof-based breaking down powerful understudy commitment/conduct in study hall and toward a particular course at e-learning stages. Our proposed design utilizes include extraction calculations like Principal Component Analysis (PCA) for facial feeling acknowledgment, Haar Cascade for student recognition and Local Binary Patterns for perceiving head developments. For AI approach and to give precise outcomes we propose Open CV. In this way dependent on the understudies input weightage is assigned, in the light of the last score, we do contrast and the edge esteem. On the off chance that the under studies consideration esteem is more noteworthy than the limited esteem, hypothesis-based redemption is suggested. On the off chance that the under studies consideration esteem is lesser than the limited esteem, video, brilliant class, persuasive video-based liberation is suggested. Experimental outcomes are been actualized utilizing Pycharm device.

S. Vijayakumar (✉)
Department of CSE, R.M.K. Engineering College, Chennai, India
e-mail: svk.cse@rmkec.ac.in

K. R. Sai · B. S. Reddy
Computer Science and Engineering, R.M.K. Engineering College, Chennai, India
e-mail: karl16207.cs@rmkec.ac.in

B. S. Reddy
e-mail: bhum16110.cs@rmkec.ac.in

Keywords Concentration · E-learning · Principle component analysis · Pycharm · Facial feelings

1 Introduction

In a virtual learning condition, students can lose inspiration and fixation effectively, particularly in a stage that isn't custom-made to their necessities. Our exploration depends on considering student's conduct on a web-based learning stage to make a framework ready to grouping students' dependent on their conduct, and adjusting instructive substance to their necessities.

As the expense of instruction (educational costs, charges and everyday costs) has soared in the course of recent decades, delayed graduation time has become a vital contributing element to the ever-growing understudy graduation. Truth be told, late investigations show that lone 50 of the in excess of 580 open four-year organizations in the United States have on-time graduation rates at or over 50% for their full-time understudies. To make school increasingly reasonable, it is along these lines essential to guarantee that a lot more understudies graduate on time through early mediations on understudies whose exhibition will be probably not going to meet the graduation measures of the degree program on schedule. A basic advance toward powerful intercession is to assemble a framework that can constantly monitor understudies' consideration level and precisely foresee their mind-set of tuning in and dependent on that information the educating can be conveyed.

Restriction is dynamic student conduct examination is absent in the current framework and biometric-based student conduct investigation is absent in the customary and e-learning stages likewise same substance of instructing is been conveyed to all sort of understudies. The superseding motivation behind this paper was to give an outline of dynamic understudy conduct examination is an initial move toward a computerized educator input instrument for estimating understudy commitment. Our proposed framework can be applied in both customary/e-learning frameworks. In our framework, we propose a half and half engineering framework conjuring understudy facial feeling acknowledgment, eye stare checking, head developments recognizable pieces of proof-based examining dynamic understudy commitment/conduct in study hall and toward a particular course at e-learning stages.

2 Literature Review

2.1 *Mediating the Expression of Emotion in Educational Collaborative Virtual Environments: An Experimental Study* Fabri, M., Moore, D. J., Hobbs, D. J. 2014

Web-based educating and e-learning philosophies have risen above higher than ever after the blast of data innovation age. Subsequently, the nature of instruction and number of online students has expanded considerably. All things considered, the modernized method of e-learning makes issue that influences an understudy's expectation to absorb information because of inaccessibility of any immediate management.

2.2 *Measuring the Impact of Emotion Awareness on E-Learning Situations* M. Feidakis, T. Daradoumis, S. Caballé and J. Conesa 2013

A teacher can give some knowledge into understudy's fulfillment during addresses, in this way understudy's contribution in class has direct relationship with the expert fitness of the educator. Direct management encourages learning as well as keeps the understudy synchronized with the course destinations because of moment correspondence with the educator whenever during the talk. Absence of correspondence has demonstrated that influenced understudies may encounter elevated levels of disappointment.

2.3 *An Infrastructure for Real-Time Interactive Distance E-Learning Environment* J. Yu 2010

Characteristic input on the substance being conveyed can be taken naturally from students by utilizing their outward appearances as an instrument to gauge intriguing quality of the substance and commitment of understudy in the online talk.

2.4 *Eye Tracking and E-Learning: Seeing Through Your Students* S. Al. Hend, G. K. Remya 2014

Hend et al. contended that the information gathered from eye GPS beacons demonstrates an individual's advantage level and the focal point of her consideration. From

eye position following and such aberrant measures as obsession numbers and length, look position, and flicker rate, data can be drawn about client levels of consideration, stress, unwinding, critical thinking, learning achievement, and exhaustion.

2.5 Learner Behavior Analysis Through Eye Tracking I. E. Haddiou, and M. Khaldi 2011

Ismail and Mohamed incorporated eye following innovation to gauge and examine student practices on an e-learning stage. They concentrated on the fascinating pieces of courses that reflect client feeling consideration, stress, unwinding, critical thinking, and weakness.

2.6 Facial Expression Recognition Utilizing Neural Network—An Overview Pushpaja V. Saudagare, D. S. Chaudhari 2012

Saudagare and Chaudhari [1] approached with a strategy to recognize articulation from feelings through neural systems. It audits the different strategies of articulation identification utilizing MATLAB (neural system tool compartment).

Mufti and Khanam [2] built up a fluffy principle-based feeling acknowledgment strategy utilizing outward appearance acknowledgment. In [3], Local Binary Pattern has been separated from static pictures to group outward appearance utilizing PCA. In [4], in view of the remaking blunder after the projection of each despite everything picture into symmetrical premise headings of various appearance subspaces, the outward appearance is perceived.

3 System Framework

This testing approach report is intended for Information and Technology Services' moves up to PeopleSoft. The report contains an outline of the testing exercises to be performed when an overhaul or improvement is made, or a module is added to a current application. The accentuation is on trying basic business forms, while limiting the time essential for testing while likewise relieving dangers. Note that diminishing the measure of testing done in an overhaul builds the potential for issues after go-live. The board should decide how much hazard is satisfactory on an update by overhaul premise.

Framework testing is basically trying the framework in general; it gets all the coordinated modules of the different segments from the combination testing stage

and consolidates all the various parts into a framework which is then tried. Testing is then done on the framework as all the parts are presently coordinated into one framework the testing stage will currently be done on the framework to check and expel any blunders or bugs.

It is the first and the most essential degree of Software Testing, in which a solitary unit (for example, a littlest testable piece of a product) is inspected in detachment from the rest of the source code. Unit testing is done to check whether a unit is working appropriately. As such, it checks the littlest units of code and demonstrates that the specific unit can work impeccably in segregation. In any case, one needs to ensure that when these units are consolidated, they work in a durable way. This guides us to different degrees of programming testing.

In PC programming, unit testing is a product testing technique by which singular units of source code, sets of at least one PC program modules along with related control information, utilization methodology, and working systems are tried to decide whether they are fit for use [3]. Intuitively, one can see a unit as the littlest testable piece of an application. In procedural programming, a unit could be a whole module, however, it is all the more normally an individual capacity or method. In object-arranged programming, a unit is regularly a whole interface, for example, a class, however, could be an individual strategy. Unit tests are short code parts made by software engineers or sporadically by white box analyzers during the advancement procedure. In a perfect world, each experiment is free from the others. Substitutes, for example, technique hits, mock articles, fakes, and test bridles can be utilized to help testing a module in seclusion. Unit tests are ordinarily composed and run by programming engineers to guarantee that code meets its structure and carries on as planned.

4 Proposed System

Along these lines, dynamic understudy conduct examination is an initial move toward a mechanized educator input apparatus for estimating understudy commitment. Our proposed framework can be applied in both conventional/e-learning frameworks. In our framework, we propose a half and half design framework summoning understudy facial feeling acknowledgment, eye stare checking, head developments recognizable pieces of proof-based dissecting dynamic understudy commitment/conduct in homeroom and toward a particular course at e-learning stages. Our proposed design utilizes highlight extraction calculations like Principal Component Analysis (PCA) for facial feeling acknowledgment, Haar Cascade for student discovery and Local Binary Patterns for perceiving head developments. For AI approach and to give exact outcomes, we propose Open CV. Trial results are been actualized utilizing Pycharm.

Framework implementation is the significant phase of venture when the hypothetical plan is fixed on useful framework. The primary stages in the usage are as per the following:

- Planning
- Training
- System testing and
- Changeover planning.

Arranging is the main undertaking in the framework execution. Arranging implies choosing the strategy and the time scale to be received. At the hour of execution of any framework, individuals from various offices and framework examination include. They are affirmed to handy issue of controlling different exercises of individuals outside their own information handling divisions. The line chiefs controlled through an execution organizing advisory group. The board thinks about thoughts, issues and objections of client division, it should likewise consider:

- The ramifications of framework condition;
- Self-choice and portion for usage errands;
- Consultation with associations and assets accessible;
- Standby offices and channels of correspondence.

5 Results and Discussion

A. Running the application

```

File Edit View Navigate Code Refactor Run Tools VCS Window Help I:\Learning\Code\face_eye.py face_eye.py preprocessor.py
E Learning face_eye.py
B E Learning C:\Users\Lenovo\Desktop\E_Learning\face_eye.py
3. Do any changes now?
    1.0.0
    377
    cv2.imshow("window_frame", bgr_image)
    if cv2.waitKey(1) & 0xFF == ord('q'):
        break
    return emotion

def eye():
    cap = None
    cap = cv2.VideoCapture(0)
    time.sleep(2)
    eye_list=None

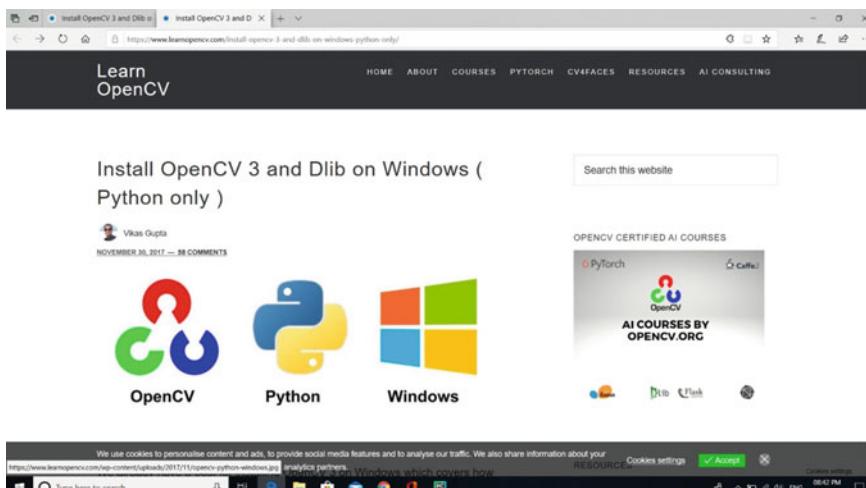
    numerator = 0
    denominator = 0
    while cap.isOpened():
        ret, frame = cap.read()
        if not ret:
            break
        eye_list=eye(frame)
    cap.release()

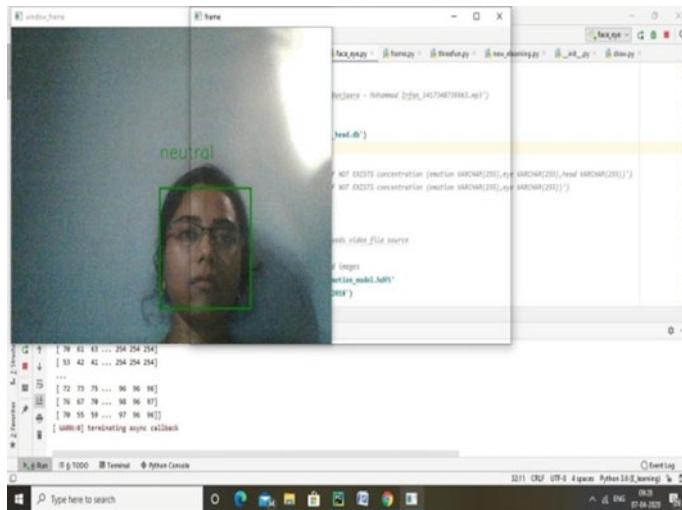
    np.int16 = np.dtype([('int16', np.int16, 1)])
    np.int32 = np.dtype([('int32', np.int32, 1)])
    np.int64 = np.dtype([('int64', np.int64, 1)])
    np.float32 = np.dtype([('float32', np.float32, 1)])
    np.float64 = np.dtype([('float64', np.float64, 1)])
    np.bool_ = np.dtype([('bool', np.bool_, 1)])
    np_resource = np.dtype([('resource', np.object_, 1)])

```

- B. Application is being executed and after verifying the facial expression and emotion if concentrated it will lead to a website (online education website) or it will display distracted in result.

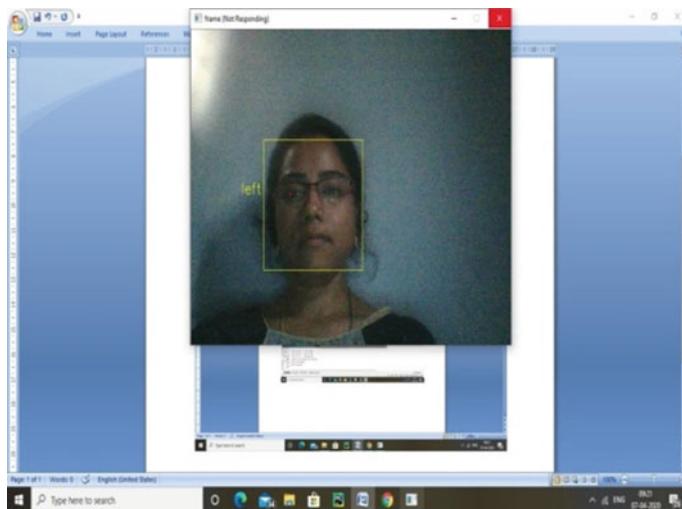
- C. The student is concentrated, so it led to the website opened in the browser to start their preparation.





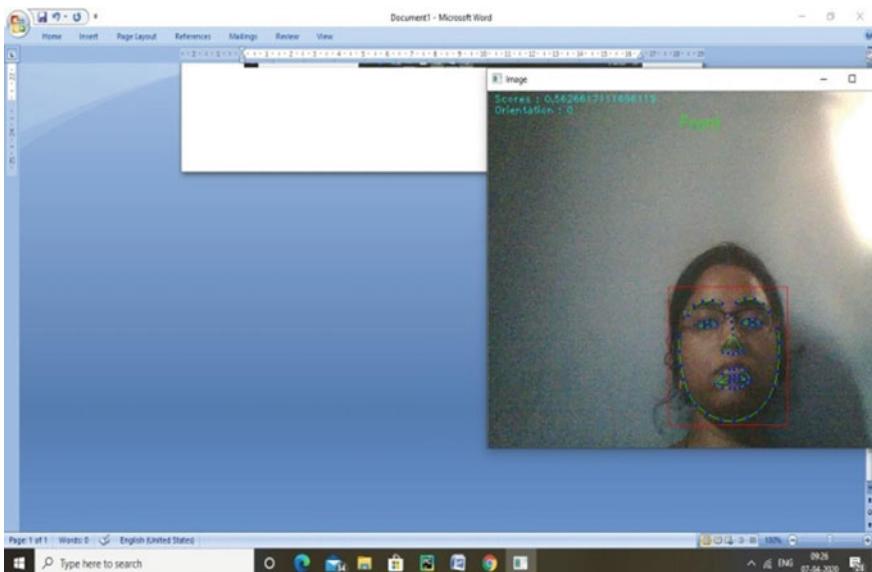
Real Time Facial Emotion Recognition

The real-time emotion detection from facial image is illustrated. Based on the feature extracted values and comparison with the trained values, the facial emotions like neutral, happy, sad, anger, surprise are detected.



Real Time Eye Gaze Detection and Recognition—Left

Eye movement detection from the input facial image based on the feature extraction algorithm and trained classifier model, eye movements toward top, bottom, left right and neutral is predicted.



Real Time Face Movement Detection—Front

Face movement detection from the input facial image based on the feature extraction algorithm and trained classifier model, head movements toward top, bottom, left right and neutral is predicted.

A concentration analysis can help confirm the presence of students in the online class room.

6 Conclusion

The hybrid biometric-based learner analysis does appear to be a promising new tool for evaluating learners' behavior dynamically. This technology can provide many benefits to e-learning, such as facilitating adaptive and personalized learning. Thus through this proposed system, the tutor can change the deliverance by dynamically analyzing the learner attention level. This would bring a revolution in the education sector.

In the future, the performance of proposed method will be improved and will be extended for detecting the students fatigue by measuring the degree of openness of the eye. Further, the real-time implementation using video camera will be conducted. Also this project can be extended in analyzing student head movements and students facial emotions.

References

1. Sheeson EC (2005) Computer anxiety and perception of task complexity in learning programming-related skills. *Comput Hum Behav* 21(5):713–728
2. Pattnaik M (2019) Infrastructure of data mining technique with big data analytics. *Int J MC Square Sci Res* 11(1):23–30
3. Korukonda AR, Finn S (2003) An investigation of framing and scaling as confounding variables in information outcomes: the case of technophobia. *Inf Sci*
4. Jay T (1981) Computerphobia: what to do about it. *Educ Technol*

A Non-negative Matrix Factorization for IVUS Image Classification Using Various Kernels of SVM



S. P. Vimal, M. Vadivel, V. Vijaya Baskar, and V. G. Sivakumar

Abstract Intravascular ultrasound (IVUS) is a medical methodology. It is a specially constructed catheter with a miniaturized Ultrasound Probe attached to the catheter's distal end is a medical imaging technique. An efficient method for IVUS image classification using non-negative matrix factorization (NNMF) and various support vector machine (SVM) kernels are presented in this study. The input IVUS images are given to NNMF for feature extraction and stored in the feature database. Finally, SVM kernels like linear, polynomial, quadratic and radial basis function (RBF) are used for prediction. The system produces a classification accuracy of 94% by using NNMF and different SVM kernels.

Keywords IVUS image classification · Non-negative matrix factorization · Support vector machine · Kernel function · Radial basis function

1 Introduction

A computerized ultrasound is connected to the proximal end of the catheter. IVUS—image-based atherosclerotic plaque characterization with the feature selection and SVM classification is presented [1]. The features are determined using multiple window sizes to change the different patterns in the region. Assessment of IVUS image identification technique is discussed [2]. Classification of diseases using the

S. P. Vimal (✉)

Department of Electronics and Communication Engineering, Sri Ramakrishna Engineering College, Coimbatore, Tamil Nadu 641022, India
e-mail: vimal.sp@srec.ac.in

M. Vadivel · V. G. Sivakumar

Department of Electronics and Communication Engineering, Vidya Jyothi Institute of Technology, Hyderabad, Telangana, India

V. Vijaya Baskar

Department of Electronics and Communication Engineering, School of EEE, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India

determination of artery cross-section layers is the adventiveness, media, and lumen layer.

Automatic classification and distinguish of IVUS and texture characteristics of atherosclerotic lesions in swine [3]. Texture steps have been used to minimize measurements, followed by the main component analysis. Two independent experts evaluated the research dataset and the findings were compared. Genetic, IVUS tissue characterization systems rule-based classification schemes [4]. Increase class discrimination, a rich array of textural features comes at various scales that include first-order statistics, co-occurrence matrices of gray rates, run lengths, waves, and local binary patterns for automatic identification of IVUS image using the Cascade Classification Stents [5]. Gentle Boost Cascade for identifying stent struts using structural features to code the details on the different sub-regions of struts. The IVUS images are fitted with a frost filter to eliminate the noise generated by ultrasound waves in the imaging technology [6, 7]. Automated coronary stent identification in IVUS pictures by using the classificatory cascade [8, 9]. Cascade of Gentle Boost classifiers for stent struts helps to identify the separate sub-regions of struts with structural features.

A non-negative matrix factorization for IVUS image classification using various kernels of SVM is described. Section 2 describes the methods and materials used for IVUS image classification. Section 3 describes the experimental result and discussion. Section 4 concludes the IVUS image classification.

2 Methods and Materials

Initially, the input IVUS images are given to NNMF for feature extraction. Then, different kernels in SVM like linear, polynomial, quadratic and RBF are for prediction. The workflow of the proposed system is shown in Fig. 1.

2.1 NNMF Feature Extraction

NNMF is a category of multivariate algorithms and linear algebra with a factor of matrix V in (usually) two matrices W and H , the property of which is the absence of all three matrixes. This negative effect promotes inspection of the resulting matrixes.

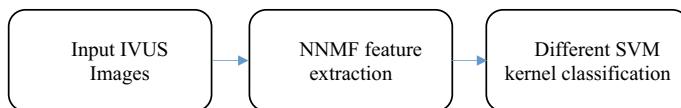
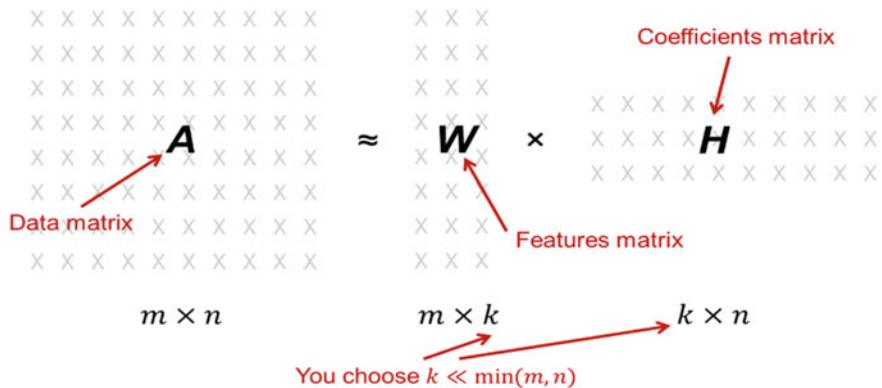


Fig. 1 Workflow of the proposed system

**Fig. 2** NNMF matrix order

Non-negativity is often fundamental to data study in applications such as audio spectrogram processing or muscular activity. As the problem is not necessarily resolvable precisely, it is normally numerically approximated. Figure 2 shows the matrix order of NNMF.

2.2 SVM Kernels Classification

SVM algorithms use a set of kernel-defined math functions. The kernel's job is to input data and convert it into the appropriate form. Different SVM algorithms are using different kernel functional forms. There can be various kinds of functions [9, 10]. The function of the kernel is to take data as input and transform it into the required form. Figure 3 shows the SVM kernel functions processed and mapped with the input data.

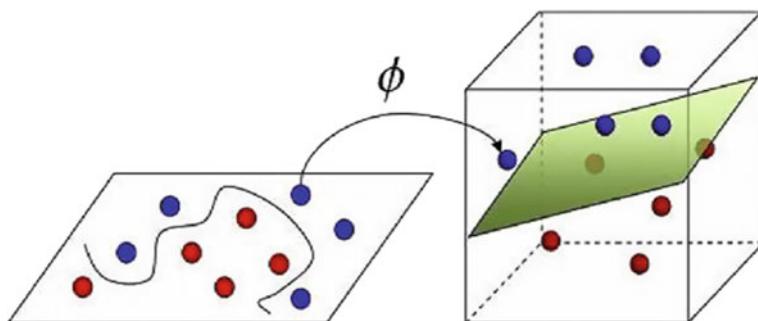
**Fig. 3** SVM kernel function

Table 1 The performance of the IVUS image classification system using NNMF and SVM kernels

<i>P</i>	NMFF computation time (s)	Classification accuracy (%)
16	5.14	77
30	8.01	83
44	5.04	88
58	5.35	91
72	5.07	94

3 Results and Discussions

The performance of IVUS image classification is measured in terms of accuracy, sensitivity and specificity. Table 1 shows the performance of the IVUS image classification system using NNMF and SVM kernel classifier. NMFF computation is calculated for time in seconds with the classification accuracy for various factor *P*-values.

From Table 1, it is inferred that the overall classification accuracy is 94% obtained by the SVM-RBF kernel by using NNMF factor *P*-value of 72 and its computation time is evaluated as 5.07 s.

4 Conclusion

An efficient method for IVUS image classification using NNMF and different SVM kernels are studied as described in the reviews. Initially, the NNMF is given for feature extraction with the *P*-factor values and the computation time is evaluated in seconds. At last, different SVM kernels like linear, RBF, quadratic and polynomial kernels are used for the prediction of IVUS images. The system yields an overall classification accuracy of 94% by using NNMF factor *P*-value of 72 and the computation time is evaluated as 5.07 s with different SVM kernels.

References

- Giannoglou VG, Stavrakoudis DG, Theocharis JB (2012) IVUS-based characterization of atherosclerotic plaques using feature selection and SVM classification. In: 2012 IEEE 12th international conference on bioinformatics & bioengineering (BIBE). IEEE, pp 715–720
- Sridevi S, Sundaresan M (2019) Evaluation of classification techniques for IVUS images. In: 2019 6th international conference on computing for sustainable global development (INDIACOM). IEEE, pp 998–1002
- Brathwaite P, Nagaraj A, Kane B, McPherson DD, Dove EL (2002) Automatic classification and differentiation of atherosclerotic lesions in swine using IVUS and texture features. In: Computers in cardiology. IEEE, pp 109–112

4. Giannoglou VG, Stavrakoudis DG, Theocharis JB, Petridis V (2012) Genetic fuzzy rule-based classification systems for tissue characterization of intravascular ultrasound images. In: 2012 IEEE international conference on fuzzy systems. IEEE, pp 1–8
5. Rajan A (2018) Classification of intravascular ultrasound images based on non-negative matrix factorization features and maximum likelihood classifier. *Int J Adv Signal Image Sci* 4(1):16–22
6. Narayanan KL, Ramesh GP (2017) Discrete wavelet transform based image compression using frequency band suppression and throughput enhancement. *Int J MC Square Sci Res* 9(2):176–182
7. Hemalatha RJ, Vijaybaskar V, Thamizhvani TR (2019) Automatic localization of anatomical regions in medical ultrasound images of rheumatoid arthritis using deep learning. *Proc Inst Mech Eng Part H J Eng Med*
8. Rotger D, Radeva P, Bruining N (2009) Automatic detection of bioabsorbable coronary stents in IVUS images using a cascade of classifiers. *IEEE Trans Inf Technol Biomed* 14(2):535–537
9. Rajan A, Ramesh GP (2015) Automated early detection of glaucoma in wavelet domain using optical coherence tomography images. *Biomed Pharmacol J* 8(2)
10. Kumarapandian S (2018) Melanoma classification using multiwavelet transform and support vector machine. *Int J MC Square Sci Res* 10(3):01–07

Tea Plant Leaf Disease Identification Using Hybrid Filter and Support Vector Machine Classifier Technique



S. Prabu, B. R. TapasBapu, S. Sridhar, and V. Nagaraju

Abstract In our country, Agriculture is the major occupation of the people. If a plant affected by any disease in long time, then there is a shortage in productivity in agriculture. Therefore, it is essential to diagnose and analyze the infection. Tea leaf cultivation is highly labor intensive and provides employment to about 2.0 million families engaged in tea cultivation, trade and trade across India. During cultivation, tea is most affected by the disease. In this research, various diseases in the tea plants are studied and also with help of image processing techniques and pattern recognition techniques, diseases are recognized at early infected stage. The method presented here is to arrange the leaf spot, rhizome rot, powdery mildew diseases and leaf blotch diseases which are infected in the tea leaf plantation. The color transformed images are sharply segmented using watershed transformation algorithm. Multiclass SVM classifier classifies the tea leaf diseases using gradient feature values of the tea leaf images. In this paper, we used hybrid filter which comprise of median filter and Gaussian filter for the purpose of edge detection and noise reduction. Finally, the performance evaluated in terms of accuracy, and it is found that the presented system is realizable and provides better classification than earlier techniques.

Keywords Image segmentation · Crop disease · Machine vision · GLCM · SVM

S. Prabu (✉)

Department of Electronics and Communication Engineering, Mahendra Institute of Technology, Namakkal, India

B. R. TapasBapu

Department of Electronics and Communication Engineering, S. A. Engineering College, Chennai, Tamil Nadu, India

S. Sridhar

Easwari Engineering College, Chennai, Tamil Nadu, India

V. Nagaraju

Department of Electronics and Communication Engineering, Rajalakshmi Institute of Technology, Chennai, Tamil Nadu, India

e-mail: Vankadarinagaraju@ritchennai.edu.in

1 Introduction

T Plant illness mainly on leaves is the main crucial regeneration of depletion in agriculture crops. The quantity and quality of productivity in agriculture are minimized due to pest's appearance in the leaves. Therefore, it increases in struggling and food fragility. Nowadays, more number of image processing mechanism are proposed to recognize the diseases.

The usual pests like fungus, aphids, caterpillars, flies, snails, etc. are generally considered in the plant disease. All the agriculturists are recognize pest's systematically over assessment by physical verification but time consumption is more in this approach.

The image processing techniques is applied in the agriculture sector to perform analysis on various agricultural applications. The crop decease detected by SVM classifier and hybrid filter with GLCM approach. The texture feature ID extracted using SVM classifier with GLCM. From the literature analysis, many researchers proposed methods which concentrate only on crop disease classification and not providing preventive measures. So the system presented in this paper provides preventive measures along with disease name. With the help of GLCM method along with first-order statistical moments, the textures are extracted.

In GLCM, pixel's spatial relationship is extracted to obtain texture classification. From the original image, GLCM is obtained and the differences obtained from the first non-singleton dimension of the input texture image. GLCM matrix indicates the pixel with grayscale intensity of some value i occurs with the nearest pixel with grayscale intensity j in the original image. GLCM matrix values considered for evaluation of the texture characteristics and evaluate the inequality in gray level of pixel of interest.

2 Literature Survey

Recently, many research works have done to detect the disease in leaf and the impact of diseases in leaf is major issue in agriculture domain.

Few research works in the same field are discussed here:

Rastogi et al. [1] in this paper explained about disease detection in cucumber plant by using machine vision technology and artificial neural network (ANN).

Pawar et al. [2] present a method to determine crop infection at earlier stage and this method is derived with help of digital image processing and ANN. This method is applied in cucumber crop disease to diagnose. The proposed algorithm gives 80.45% classification accuracy.

Zhang et al. [3] in this paper explained about feature selection of cotton disease leaves images by using fuzzy features (fuzzy curve and fuzzy surface) selection techniques without using nonlinear techniques.

Wu et al. [4] present leaf recognition algorithm for plant classification using probabilistic neural network (PNN). That classifies the 32 types of plants with help of leaf images.

Meunkaewjinda et al. [5] present study about detection of the grape leaf infection using back-propagation neural network (BPNN). BPNN effectively used to extracting color from leaf with complex backdrop.

Wang et al. [6] in this paper explained about the study of applications of neural networks which is used in the image recognition and based on the texture and shape features. Shen et al. [7] This paper presents the evaluation of leaf spot disease grading using image segmentation techniques.

Fujita et al. [8] in this paper explained about robust diagnostic of cucumber viral disease using convolution neural network. Pydipati and Phadikar et al. [9, 10]. This paper explained about the rice disease identification by using pattern recognition techniques. Images are classified using SOM neural network.

Satpathy and Zhang et al. [11, 12] in this paper explained about the image feature extraction of tobacco leaf using machine vision technique. Image extraction technique used for grading. Prabu et al. [13] proposed cognitive image filter for removing noise in medical images.

3 Proposed Method

Tea leaf plants are infected different types of diseases in the whole plantation without any forewarning of the diseases. The aim of this work is to detection and classification of leaf spot, leaf rot and powdery mildew diseases in the variety of tea leaf plants (leaves) in an earlier stage using digital image processing techniques. This chapter describes about processes in the current work. Block diagram of the proposed method is shown in Figs. 1 and 2.

4 Module Description

Proposed Method

The primary aim of the presented system is observation of leaf disease using SVM and hybrid filter. For experimental analysis, tea leaves are used as many types of diseases.

5 Image Processing Block Diagram

The modules in the current work are given below,

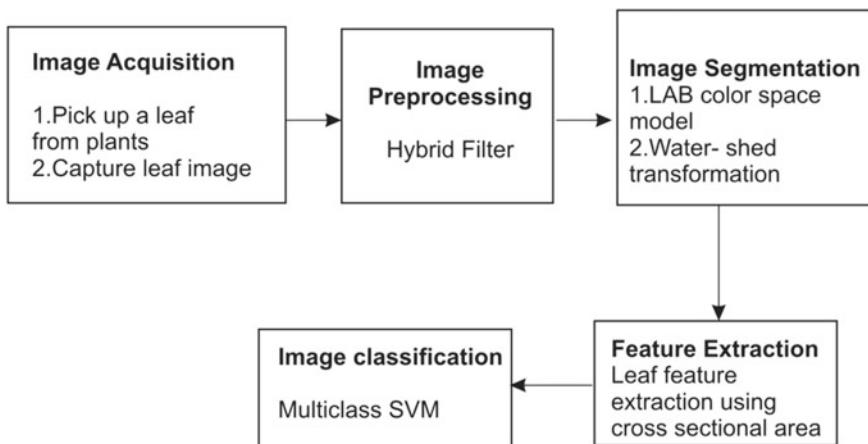


Fig. 1 Block diagram of the proposed system

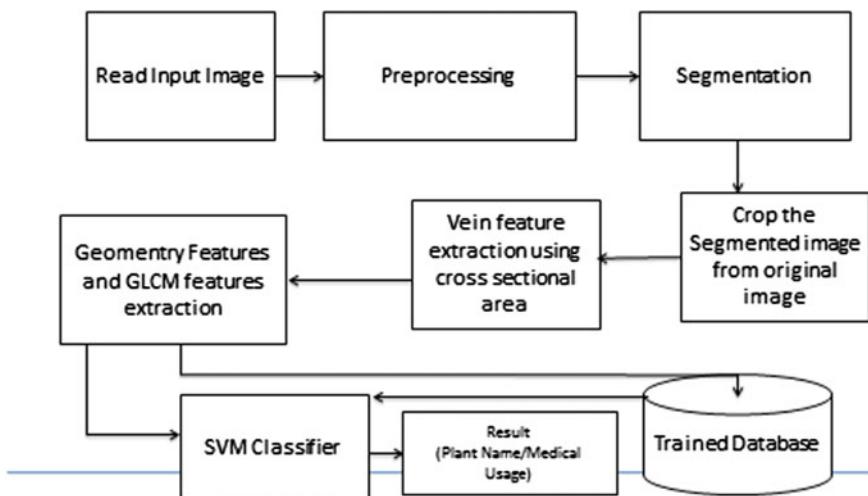


Fig. 2 Architecture of the proposed system

1. Image collection
2. Image preprocessing
3. Color transformation
4. Image segmentation
5. Feature extraction
6. Classification.

Image Preprocessing

Image preprocessing is the term for functioning on images at the subordinate level of abstraction. These functionings do not rise image data content, but they reduce if entropy is a data measure. The objective of preprocessing is development of the image input data which subdue unacceptable warp or raises relevant features for closer process and analysis task. The enrichment includes purifying which detach the noise and process the image accurately. The purification is done by using hybrid filter.

Hybrid Filter

1. Median filter.
2. Gaussian filter.

Median Filter

The median filter is controlled by arranging all the intensity levels in a sequential order, later it will be restored by middle grayscale value. Figures 3 and 4 show the original image, median filtered image.

In median filter, initially window is shifted then whole grayscale intensity values are classified. Later median of the values calculated. Then the median value set as middle pixel. In case of integer of constituent in window of $K \times K$ is odd, and then the center value of the window is set as middle or median value; otherwise, average of two median values is set as middle or median value.

Gaussian Filter

A Gaussian filter is a type of linear filter. Purpose of this filter is to blur the image or to reduce noise.

The Gaussian filter alone will fade edges and decrease contrast. The simple way to lessen noise in an image is nonlinear filter. Its declare to fame (over Gaussian for noise reduction) is that it extracts noise while keeping edges comparatively sharp.

Fig. 3 Original image



Fig. 4 Median filtered image



6 Classification

- (a) It incorporates K binary SVM classifiers, a single classifier for each class.
- (b) Each support vector machine is instructed to split one class from the remaining classes.
- (c) A hyperplane is calculated for each class, considering this class as positive (+1) class and the other classes as negative (-1).
- (d) Redo the process up to all classes are divided from the rest of classes.
- (e) A specimen is checked for each classifier and is set to the class that corresponds to SVM (Support vector machine) which have greatest output. The watershed segmentation of tea leaf is shown in Fig. 5.

Figure 6 shows multiclass classification in one versus all techniques.

Support vector machine is used for categorization or regression drawbacks. It uses an approach known as kernel trick to convert your information, and then based on these conversions, it notices an optimal margins between the feasible outputs. The merits of SVM and support vector regression cover that they can be used to remove the complexity of using linear duty in the high-dimensional characteristic expansion, and the optimization issue is converted into twin convex quadratic programs.

If the given values are close to the powdery mildew class, then the classifier recognized that untrained tea leaf belongs to powdery mildew diseases as shown in Fig. 7.

Multiclass SVM classifier is used to identify whether the tea leaf is affected or not and also to identify the type of diseases. Figure 8 shows classifier output for healthy leaves. Experimental result of the same in Fig. 9. Table 1 shows color space model.

If the given values are near to the leaf rot class, then the classifier recognized that in experienced tea leaf belongs to leaf rot diseases as shown in Fig. 10.

The outcomes are calculated in terms of Sensitivity and Specificity and demonstrated that the existing technique with 82.35% accuracy and the proposed method can improve the tea leaf disease detection with 95.85% accuracy.

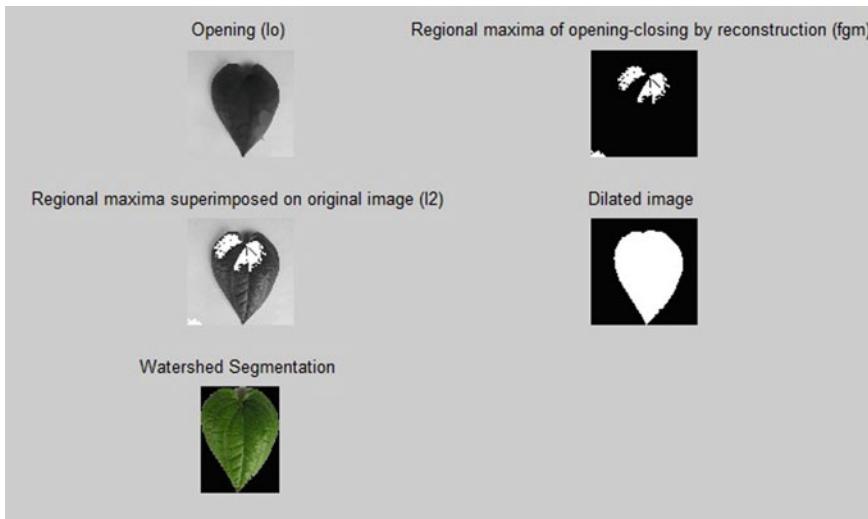
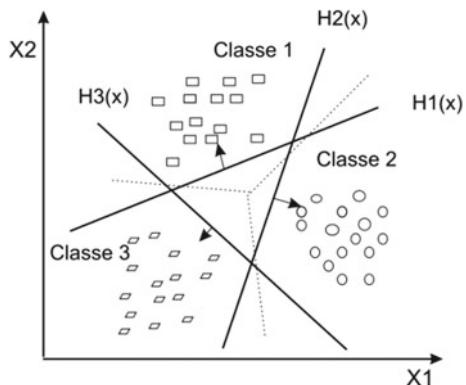


Fig. 5 Extracted tea leaf after watershed segmentation

Fig. 6 One versus all method



7 Performance Evaluation

In the evaluation metrics, confusion matrix is calculated to produce outcome. Table 2 shows confusion matrix in which

- TP—no. of healthy leaves are perfectly classified,
- FN—no. of infected leaves are misclassified as healthy leaves,
- FP—no. of healthy leaves misclassified as infected leaves,
- TN—no. of infected leaves are perfectly classified.

Accuracy Calculation

Fig. 7 Classifier output for powdery mildew diseases



Fig. 8 Classifier output for healthy leaves

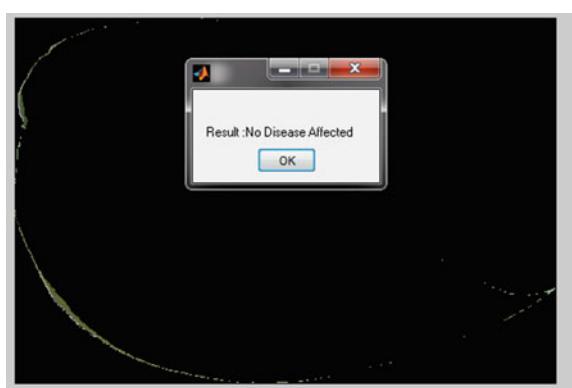
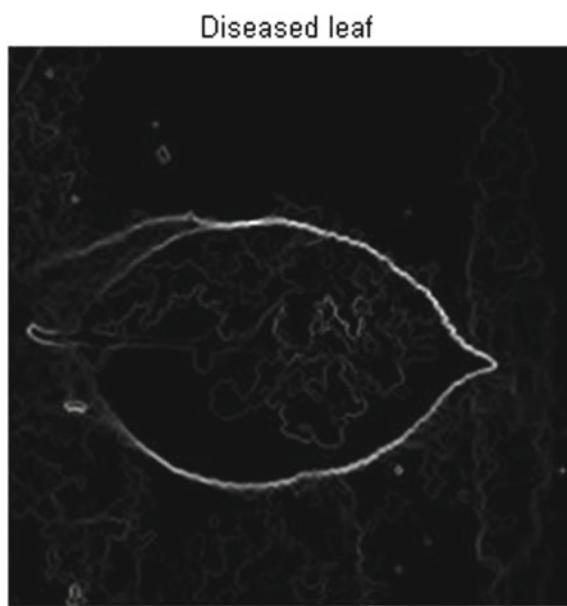


Fig. 9 Experimental result for healthy tea leaf



Table 1 Color space model

	Predicted	
	Healthy	Diseased
<i>Watershed transformation algorithm</i>		
True	Healthy	8
	Diseased	0
<i>Color space model</i>		
True	Healthy	3
	Diseased	11

Fig. 10 Classifier result for leaf rot disease**Table 2** Confusion matrix

		Predicted	
		Healthy	Diseased
True	Healthy	TP	FP
	Diseased	FN	TN

- TP = True Positive = sum(predicted value == True and ground truth == True)
- FN = False Negative = sum(predicted value == False and ground truth == True)
- FP = False Positive = sum(predicted value == True and ground truth == False)

- $TN = \text{True Negative} = \text{sum}(\text{predicted value} == \text{False and ground truth} == \text{False})$
- Specificity = $(TN/(FP + TN)) \times 100$
- Sensitivity = $(TP/(FN + TP)) \times 100$
- Accuracy = $((TP + TN)/(TP + FP + FN + TN)) \times 100$.

The confusion matrix for the existing algorithm (Color space model) and proposed algorithm (watershed transformation) are given below. Total number of tea leaf is 30, in that 20 for testing data and 10 for training data. For testing images, based on the true records and predicted records, the TP, TN, FP, FN values are calculated for all.

According to the confusion matrix, a set of metrics commonly evaluated by using the evaluation metrics are Sensitivity and Specificity.

8 Result Analysis

Based on the confusion matrix of existing and proposed work, the sensitivity and specificity and accuracy are calculated.

Figure 11 shows comparison table among color space model and watershed transformation algorithm in terms of sensitivity, specificity and accuracy. Compared with color space model, watershed transformation algorithm provide better sensitivity, specificity and accuracy.

The pooling layer output features are then finally applied into fully connected neural networks for reducing the error rate in the intermediate layers in Convolutional layers, as illustrated in Fig. 5. The output from this fully connected neural network is either stroke image or non-stroke image.

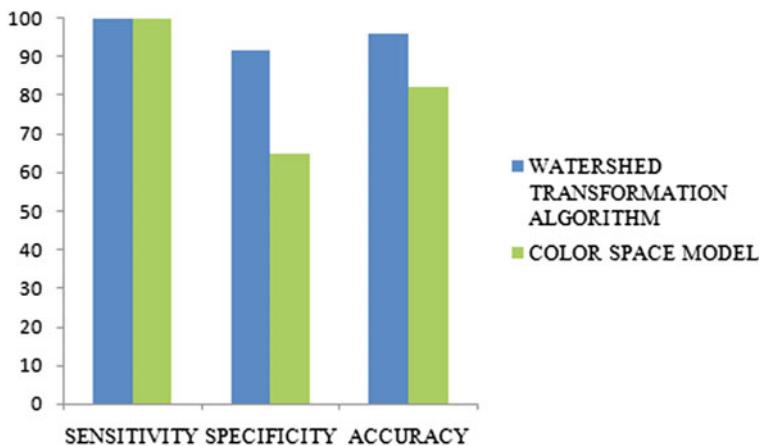


Fig. 11 Graphical results of sensitivity, specificity and accuracy

9 Conclusion

The system presented here is to classify the leaf spot, rhizome rot, powdery mildew diseases and leaf blotch diseases which are infected in the tea leaf plantation.

With the help of watershed transformation algorithm, the color transformed images are segmented sharply. After that, A channel is released from $1 * a * b$ color transformed images (RGB to $1 * a * b$). The gradient characteristics value of tea leaf images are obtained using HOG technique based on the shapes of the tea leaf. The watershed segmentation and Multiclass SVM classifier are the current techniques used in this paper. It is concluded that accuracy of watershed transformation technique detects the tea leaf diseases at 95.85% accuracy rate.

References

1. Rastogi A, Arora R, Sharma S (2015) Leaf disease detection and grading using computer vision technology & fuzzy logic. In: 2015 2nd international conference on signal processing and integrated networks (SPIN). IEEE, pp 500–505
2. Pawar P, Turkar V, Patil P. Algorithm for detecting crop disease early and exactly, this system is developed using image processing techniques and artificial neural network
3. Zhang Y-C, Mao H-P, Hu B, Li M-X (2007) Features selection of cotton disease leaves image based on fuzzy feature selection techniques. In: 2007 international conference on wavelet analysis and pattern recognition, vol 1. IEEE, pp 124–129
4. Wu SG, Bao FS, Xu EY, Wang Y-X, Chang Y-F (2007) A leaf recognition algorithm for plant classification using probabilistic neural network. In: IEEE 7th international symposium on signal processing and information technology
5. Meunkaewjinda A, Kumsawat P, Attakitmongkol K, Srikaew A (2008) Grape leaf disease detection from color imagery system using hybrid intelligent system. In: Proceedings of ECTICON. IEEE, pp 513–516
6. Wang HG, Li GL, Ma ZH, Li XL (2012) Application of neural networks to image recognition of plant diseases. In: International conference on systems and informatics
7. Shen W, Wu Y, Chen Z, Wei H (2008) Grading method of leaf spot disease based on image processing. In: Proceedings of 2008 international conference on computer science and software engineering, vol 06
8. Fujita E, Kawasaki Y, Uga H, Kagiwada S, Lyatomi H (2016) Basic investigation on a robust and practical plant diagnostic system. In: 15th IEEE international conference on machine learning and applications, pp 989–992
9. Pydipati R, Burks TF, Lee WS (2009) Identification of citrus disease using color texture features and discriminant analysis. *Comput Electron Agric* 52(2):49–59
10. Phadikar S, Sil J (2008) Rice disease identification using pattern recognition techniques. In: Proceedings of 11th international conference on computer and information technology, pp 25–27
11. Satpathy RB, Ramesh GP (2020) Advance approach for effective EEG artefacts removal. In: Balas V, Kumar R, Srivastava R (eds) Recent trends and advances in artificial intelligence and internet of things. Intelligent systems reference library, vol 172. Springer, Cham
12. Zhang X, Zhang F (2008) Images features extraction of tobacco leaves. In: Congress on image and signal processing. IEEE Computer Society
13. Prabu S, Balamurugan V, Vengatesan K (2019) Design of cognitive image filters for suppression of noise level in medical images. *Measurements* 141:296–301

A Multidimensional Data Mining Approach for Video Analysis and Ranking System



Anjan Dutta, Vaibhav Sinha, Punyasha Chatterjee, Narayan C. Debnath, and Soumya Sen

Abstract YouTube is the most popular comprehensive online video source where continuous real-time video uploading is done by users throughout the world at a tremendous rate. It has become the most popular streaming media where users' interaction is observed in terms of sharing, commenting, and rating the videos. Generally, these ratings represent the relevancy, validity, quality, and popularity of the videos. The comments made by the viewers are also a deciding factor for the new viewers to have an idea about the relevance and the quality of the videos. Hence the proper analysis of the sentiments exhibited in the comments can indicate the nature of the videos. Sometimes in the user's search result, a low quality, irrelevant, and inappropriate video content is often ranked higher due to the large number of views that seem untenable. In this work, a Natural Language Processing (NLP) framework based on sentiment analysis of the comments is presented to minimize this issue. Based on the positive and negative sentiments presented in the user comments videos are given a measure which is termed as polarity factor. Finally, the video impact measure is calculated based on the Polarity factor together with like-dislike and view counts. This study assists in mining the most appropriate and popular videos on YouTube based on the search key. A data-driven experiment has been done here to prove the effectiveness and accuracy of the proposed system.

Keywords Video ranking · Polarity · Video impact measure · Natural language processing · Like-dislike count

A. Dutta (✉) · V. Sinha
Techno International NewTown, Kolkata, India

P. Chatterjee
School of Mobile Computing and Communication, Jadavpur University, Kolkata, India

N. C. Debnath
School of Computing and Information Technology, Eastern International University,
Thu Dau Mot, Vietnam

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

1 Introduction

A challenging issue in large-scale retrieval of the multimedia data [1] is developing an efficient and effective framework to rank the search results [2]. In recent years social media sites like Facebook, Twitter, and online video sharing platforms like YouTube, Google+, etc. are gaining increasing popularity for becoming the medium of sharing real-time data in the form of videos and images uploaded by millions of users. YouTube is one of the largest platforms for multi-modal, multi-domain, multi-lingual, and multi-cultural video sharing and has become a comprehensive compilation of web-based video sources. Due to this large variation and attractiveness of the shared contents, it has drawn huge popularity and widespread attention. In Dec 2016, a web traffic analysis company Alexa Internet ranked YouTube as the second largest and most popular site. The user's interaction is increased by allowing them to express their opinion about the video in terms of likes, dislikes, and comments. These view counts and like, dislike counts indicate the global popularity of the particular video. Moreover, the users can find the relevant contents efficiently with the help of these metadata (view count, like, dislike count). When videos of a specific topic are searched then the videos appear in the search panel according to their popularity that is based on the number of likes, dislikes, and the number of views. The system sometimes suffers from implicit bias also. For example, a video is watched by a user simply because of its high rank not because the user liked it. As a result, the view count of this video is increased more. Due to this sometimes a video that exhibits derogatory and objectionable contents may become popular due to the high view count. The negative popularity earned by these videos is harmful to the viewers and these kinds of videos should be restricted for the wellbeing of society. Occasionally a popular video may have some negative comments but these negative sentiments are not considered for the video rankings. These may give rise to some adverse situations, e.g., a video tutorial that is having high popularity but demonstrating wrong concepts. Sometimes user's search query for videos related to a particular topic results in irrelevant and inconsistent outcomes. Some of the suggested videos often have a little link with the topic or domain the user is looking for. This kind of situation is commonplace and causes inconvenience to the user. It can be observed that a more number of positive comments are given to valid, relevant video in contrast to an inappropriate video. Hence, video quality, relevance, and popularity are decided by the nature of the comments it receives. However, due to unstructured characteristics user comments are difficult to analyze. In this study, a sentiment analysis [3, 4] approach based on Natural Language Processing (NLP) [5, 6] is used to analyze comments [7, 8] to find the relevant and popular videos. A unique ranking methodology is introduced here that will not only consider the common popularity indices like view count, like or dislike count but also include the sentiment analysis of the comments.

This paper is organized as follows. In Sect. 2 related works in this domain are discussed. The problem definition is given in Sect. 3. The methodology of the video

ranking process is explained in Sect. 4. Experiments and results are vividly discussed in Sect. 5. Conclusion and future scope of work are given in Sect. 6.

2 Related Work

With the advent of streaming media, significant research has been done on video processing [9–11] and search optimization and effectiveness. In [12] a novel ranking framework (SNDocRank) that considers the social network activities of the users are applied to video search. The authors have demonstrated that the SNDocRank framework offers more relevant search results to the user’s interest than any other traditional ranking methodologies. In [13] an aspect-based framework of YouTube video ranking is introduced. First comments are collected and pre-processing is done on them to remove spam and irrelevant comments. Thereafter aspect terms are extracted and their categories are determined by using SentiWordNet [14] module. Sentiment classification of the aspect terms is done after that. Finally, videos are ranked based on the aspects. A new comment classification approach to capture the salient aspects of the YouTube video comments is proposed in [15]. The authors demonstrated that this new classification method can assist in rapid semantic video analysis. Opinion Mining (OM) was carried out targeting YouTube video comments in [16]. The authors proposed a classification model that separates spam and irrelevant comments from meaningful ones. Thereafter the type classification of videos is done, e.g., whether a video is product-related or not. Finally, polarity was assigned to different meaningful comments that will help to distinguish comments related to the video and the target product. In [17] authors have presented an extensive study on near about 6 million comments related to 67,000 YouTube videos. They have analyzed dependencies among comments, comment ratings, views, and topic categories. Besides, the SentiWordNet module was used here to study the impact of sentiments expressed in the comments on the video ratings. Finally, different classifiers are built to predict the community acceptance of the unrated comments to estimate the ratings for these comments. As mentioned earlier YouTube videos have many comments those are irrelevant or contain advertisements and bad languages. Therefore these comments are not useful to the users who are interested to make valid opinions about the videos. To address this problem, in [18] authors have presented an application that acts as a relevant comment classifier by using a mix of external resources and open-source libraries. The retrieved comments are pre-processed by Weka [19] module by using a neural network classifier to separate the insignificant and irrelevant comments. The dynamics of view counts of YouTube videos and Internet market growth and the interdependency between them are studied in [20]. Three forms of varying market size are studied here; linear growth, exponential growth, and growth due to repeated viewership. An automated ranking system for domain-specific highlights is proposed in [21] by analyzing a large number of edited online videos. In [22] authors have done a feasibility study of the popularity

prediction models for online news rankings. The effectiveness of two ranking prediction methods—constant scaling model and linear model based on a logarithmic scale is compared here.

3 Problem Description

In this world of overwhelming and ever-expanding data, users are buried in the poll of information when searching for anything. Confusion may arise regarding which data is more important and relevant to the user. YouTube is one of the largest and most popular video information sources where real-time videos are uploaded continuously throughout the world. This is a medium where users can express their opinion about the videos in terms of likes, dislikes, and comments, and a particular video is ranked according to that. Sometimes an irrelevant and inappropriate video may get a higher ranking based on the number of view counts and likes. Hence users' search for the videos of a particular topic may result in the appearance of unwanted results in the search space. A Natural Language Processing (NLP) model-based sentiment analysis approach is presented here to minimize this issue. The proposed model not only relies on the like-dislike and view counts but also analyzes the sentiments of the user comments to rate a video and thus helps to find out the most relevant content on YouTube.

4 Methodology

In this work, Google YouTube API is used to extract the comments from the YouTube videos. Thereafter Natural Language Processing (NLP) is applied to do the sentiment analysis on the comments based on which the polarity of the comments are extracted. After that the like-dislike count of the videos is calculated. These two measures are collectively used to compute the video ranking which is termed as video impact measure (VIM). Thereafter the videos are sorted according to the descending order of this calculated ranking. The overall methodology is explained as below.

4.1 Polarity Calculation

The overall comment sentiments are analyzed by computing the polarity, i.e., degree of positivity or negativity exhibited by the comments. To calculate the positive polarity, the frequency of occurrence of the word appears in the positive dictionary is calculated, and thereafter, it is divided by the collective frequency of occurrence of the word in both the positive and negative dictionaries. Similarly the

negative polarity is calculated by dividing the total number of occurrence in the negative dictionary by the number of occurrences in both positive and negative dictionaries.

Video impact measure (VIM)

After calculating the polarity of the videos as mentioned in the above section, the view count, like and dislike counts of the comments are extracted. The VIM is calculated from these measures sequentially as mentioned below

Step 1: Polarity factor (PF) is calculated as below

$$\text{PF} = \frac{P - N + V}{V} \quad (1)$$

where P = positive polarity count, N = negative polarity count and V = view count.

Step 2: Relative like-dislike factor (LDF) is computed as

$$\text{LDF} = \frac{\text{LC} - \text{DLC} + V}{V} \quad (2)$$

where LC = like count, DLC = dislike count.

Step 3: VIM is taken as the average of PF and LDF

$$\text{VIM} = \frac{\text{PF} + \text{LDF}}{2} \quad (3)$$

Step 4: Videos are ranked according to the descending order of the individual VIM.

5 Experiment, Results and Discussion

5.1 Generate Video Title and ID

VideoID, title, and name are extracted based on search query. In Fig. 1, the searched keywords and videoIDs are shown in rectangular box.

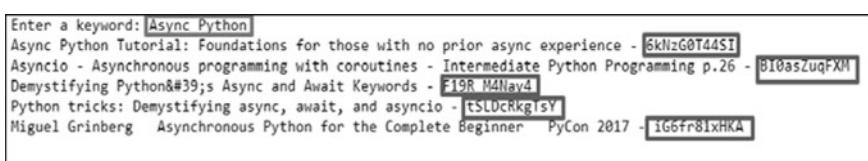


Fig. 1 Showing the search key and extracted video ids

5.2 Generate Videos Statistics

The project asks for Search Query (Async Python). The app does the web scrapping of the Top 15 videos and shows data like videoID, channelID, likes, dislikes, number of comments, etc. This is shown in Fig. 2.

5.3 Data Labeling and Polarity Score Calculation

The video comments are extracted based on the search query. Polarity score is calculated on this unlabeled data as discussed in Sect. 4.1 by applying the TextBlob module in Python. The comments and the corresponding polarity score are shown in Fig. 3.

5.4 Finding the Most Important Video

The positive and negative polarity values are represented as 1 and -1 and the neutral values are also appended as 0 values. This is shown if Fig. 3. The total number of positive and negative counts of a particular video ID is calculated and the final merged result is shown in Fig. 4. Polarity factor and like-dislike factor is

```
Please input your search query
Async Python
Search Completed...
Example output per item, snippet
dict_keys(['publishedAt', 'channelId', 'title', 'description', 'thumbnails', 'channelTitle', 'liveBroadcastContent', 'publishedTime'])
Top 3 results are:
Async Python Tutorial: Foundations for those with no prior async experience, (Live Python),
Asyncio - Asynchronous programming with coroutines - Intermediate Python Programming p.26, (sentdex),
```

Fig. 2 Showing the search key and different extracted parameters

	Video ID	Comment ID	comments	polarity	pol_cat
0	nYh-n7EOtMA	UgymFAuzR1M6wpgQbsJ4AaABAq	I can't hear this song now without pict...	0.0	-1
1	31crA53Dgu0	UgwGs6dQEPGLPWJAa8d4AaABAq	Anyone in November 2018	0.0	-1
2	nYh-n7EOtMA	UgnDssvFPTSKWMGj1B4AaABAq	There is English There is Urdu There...	0.0	-1
3	31crA53Dgu0	UgyTOlhw66ah2yNOBz54AaABAq	Just think those dude dancers will alw...	0.0	-1
4	6mqbAnrtWHO	UgwZx5t6j7r1cG7uYS14AaABAq	Sia S she I is A amazing	0.6	1

Fig. 3 Showing the polarity score of the comments

VideoID	Positive Polarity Count	Negative Polarity Count
tSLDcRkgTsy	114	6
BI0asZuqFXM	73	13
iG6fr81xHKA	31	2
Xbl7XjFYsN4	22	1
F19R_M4Nay4	19	1
Mj-Pyg4gsPs	19	1
L3RyxVOLjz8	18	3
6KNzG0T44SI	14	3
L061F07s7gw	6	2
StWIxBcvy10	4	1
c6uoXhaenHg	4	1
fI9QYp_ybfU	4	0
dJttCNhh850	3	0
NLaeCFr_FlI	3	0
h2IM-OPofqg	1	0

Fig. 4 Merged data frame of positive and negative polarity

viewCount	Polarity Factor	like_dislike result	VIM
4602.0	1.000652	1.060513	1.030582
6511.0	1.003225	1.041310	1.022268
18940.0	1.000581	1.027141	1.013861
2230.0	1.001345	1.025858	1.013602
7537.0	1.000531	1.024664	1.012597
49383.0	1.002187	1.020380	1.011284
78240.0	1.000371	1.021043	1.010707
25567.0	1.000704	1.019542	1.010123
16229.0	1.000185	1.019984	1.010084
17759.0	1.001014	1.018609	1.009811
1251.0	1.000799	1.016346	1.008573
31947.0	1.000470	1.016289	1.008379
72580.0	1.000827	1.014921	1.007874
5157.0	1.000582	1.013992	1.007287
NaN	NaN	1.002919	NaN
5567.0	NaN	NaN	NaN

Fig. 5 VIM calculation

computed from the positive and negative polarity counts by Eqs. 1 and 2. Thereafter VIM is calculated through Eq. 3. This is shown in Fig. 5. After this step videos are ranked according to the descending order of the VIM value. After the model training, the accuracy score on the training dataset is **99.0498%**, which

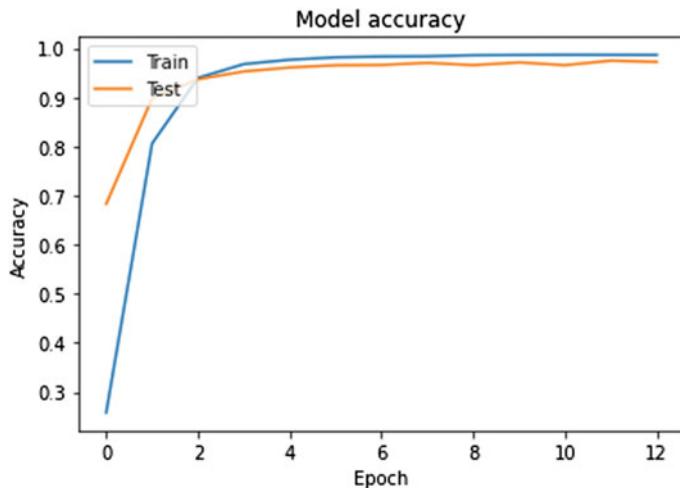


Fig. 6 Performance analysis of the proposed system

implies that the proposed model is predicting 99.0498% accurate results, which is quite satisfactory. In the next step, accuracy score on the test dataset is found. The accuracy of the model on the test dataset is **80.1886%**, which means the model is predicting 80.1886% accurate results on the unseen dataset. In Fig. 6 the performance of the suggested model is demonstrated.

6 Conclusion

A unique methodology for video ranking by analyzing their overall impact on the viewers' sentiment is proposed here. Video comments are analyzed and NLP is used to do the sentiment analysis of the comments and their polarities are found. A quantitative measure is given to the videos taking into account both the polarity and view count which is termed as polarity factor. Together with the polarity factor, the like-dislike count is considered to find the final video ranking which is called as Video Impact Measure. In this study, the relationship with the user sentiments expressed in the comments is analyzed to evaluate and rate a video in terms of its relevance, quality, and popularity. A test sample of nearly 15 YouTube comments is analyzed based on the search query. An in-depth study of video metadata using the NLP and TextBlob module is done here and the effectiveness of user comments in video ranking is demonstrated here. The efficiency of the proposed approach is revealed by the experimental result as 80.1886% in retrieving the appropriate video and it also encourages to do more future study and analysis on comments. The efficiency of the proposed model is proved hereby data-intensive experiment

w.r.t. accuracy, like/dislike results based on view counts and comment sentiments. In the future, an AI-based video recommendation system can be developed directly from this model by doing the sentiment analysis of the comments.

References

1. Karaa WBA, Dey N (2017) Mining multimedia documents. CRC Press
2. Hoi SC, Lyu MR (2008) A multimodal and multilevel ranking scheme for large-scale video retrieval. *IEEE Trans Multimed* 10(4):607–619
3. Ali MNY, Sarowar MG, Rahman ML, Chaki J, Dey N, Tavares JMR (2019) Adam deep learning with SOM for human sentiment classification. *Int J Ambient Comput Intell (IJACI)* 10(3):92–116
4. Baumgarten M, Mulvenna MD, Rooney N, Reid J (2013) Keyword-based sentiment mining using twitter. *Int J Ambient Comput Intell (IJACI)* 5(2):56–69
5. Bird S, Klein E, Loper E (2009) Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.
6. Manning C, Schütze H (1999) Foundations of statistical natural language processing. MIT Press
7. Dey N, Mishra R, Fong SJ, Santosh KC, Tan S, Crespo RG (2020) COVID-19: psychological and psychosocial impact, fear, and passion. *Digit Gov Res Pract* 2(1):1–4
8. Dey N, Babo R, Ashour AS, Bhatnagar V, Bouhlel MS (2018) Social networks science: design, implementation, security, and challenges
9. Dey N, Ashour A, Patra PK (eds) (2016) Feature detectors and motion detection in video processing. IGI Global
10. Kamble SD, Thakur NV, Bajaj PR (2018) Fractal coding based video compression using weighted finite automata. *Int J Ambient Comput Intell (IJACI)* 9(1):115–133
11. Pal G, Rudrapaul D, Acharjee S, Ray R, Chakraborty S, Dey N (2015) Video shot boundary detection: a review. In: Emerging ICT for bridging the future—proceedings of the 49th annual convention of the computer society of India CSI, vol 2. Springer, Cham, pp 119–127
12. Gou L, Chen HH, Kim JH, Zhang X, Giles CL (2010) Sndocrank: a social network-based video search ranking framework. In: Proceedings of the international conference on multimedia information retrieval, Mar 2010, pp 367–376
13. Chauhan GS, Meena YK (2019) YouTube video ranking by aspect-based sentiment analysis on user feedback. In: Soft computing and signal processing. Springer, Singapore, pp 63–71
14. Esuli A, Sebastiani F (2006) Sentiwordnet: a publicly available lexical resource for opinion mining. In: LREC, May 2006, vol 6, pp 417–422
15. Schultes P, Dorner V, Lehner F (2013) Leave a comment! An in-depth analysis of user comments on YouTube. *Wirtschaftsinformatik* 42:659–673
16. Severyn A, Uryupina O, Plank B, Moschitti A, Filippova K (2014) Opinion mining on YouTube
17. Siersdorfer S, Chelaru S, Nejdl W, San Pedro J (2010) How useful are your comments? Analyzing and predicting YouTube comments and comment ratings. In: Proceedings of the 19th international conference on world wide web, Apr 2010, pp 891–900
18. Serbanou A, Rebedea T (2013) Relevance-based ranking of video comments on YouTube. In: 2013 19th international conference on control systems and computer science, May 2013. IEEE, pp 225–231
19. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. *ACM SIGKDD Explor News* 11(1):10–18

20. Aggrawal N, Arora A, Anand A, Irshad MS (2018) View-count based modeling for YouTube videos and weighted criteria-based ranking. *Adv Math Tech Eng Sci* 149–160
21. Sun M, Farhadi A, Chen TH, Seitz S (2016) Ranking highlights in personal videos by analyzing edited videos. *IEEE Trans Image Process* 25(11):5145–5157
22. Tatar A, Antoniadis P, De Amorim MD, Fdida S (2014) From popularity prediction to ranking online news. *Soc Netw Anal Min* 4(1):174

Impulse Noise Restoration Using Combined Fuzzy Logic and Adaptive Trimmed Median Filter



Priyaranjan Kumar, Aswini K. Samantaray, Anshuman Kashyap, and Chinmayee Biswal

Abstract This work focuses on the development of algorithm for the removal of impulse noise from digital images in spatial domain in low density as well as high density of noises. Various issues related to the noisy problem are studied for grayscale as well as color images and suitable filtering methods are suggested. In this paper, fuzzy logic is used to detect the noisy pixels in an impulse noise corrupted image and an adaptive trimmed median filtering technique is proposed to remove the detected noisy pixel from gray and color images. When some of the pixels in a considered window of an image are noisy pixels, then median value of the noise-free pixels in that window replaces the processing detected noisy pixel. However, when all the pixels in a selected window are all noisy pixels, then in such cases the possible solution is to replace the processing pixel by the mean and standard deviation values of the elements in the selected window. The combined fuzzy logic and adaptive trimmed median filter approach is also used to preserve the edges and fine details of the images. To assess the performance of the proposed method, several standard grayscale and color test images are used in the experiments which have distinctly different features. The efficacy of the various filtering systems is evaluated both qualitatively and quantitatively in terms of PSNR and MSE for grayscale as well as color images.

P. Kumar

Gandhi Institute for Education and Technology, Bhubaneswar, India

A. K. Samantaray (✉)

National Institute of Technology Goa, Ponda, India

e-mail: asamantaray@nitgoa.ac.in

A. Kashyap

e-Infochips Private Limited, Ahmadabad, India

e-mail: anshuman.kashyap@einfochips.com

C. Biswal

C V Raman College of Engineering, Bhubaneswar, India

Keywords Impulse noise · Fuzzy logic · Membership function · Median filter

1 Introduction

Impulse noise which is also known as salt and pepper noise is generated by various factors such as faulty memory locations in hardware, malfunctioning pixels elements in camera sensors, bit errors in transmission of a noisy channel, or timing errors in the digitization process. The impulse noise corrupted image has dark spots in bright region and bright spots in dark region of the image. The corrupted pixels of the image are set to the minimum or maximum gray value depending upon the number of bits used to represent the gray value of a pixel. The values of the impulse noise are distributed over the image with equal probability [1]. The impulse noise degrades the performance of several image processing techniques such as data compression, image segmentation, and edge detection. Hence, it is required to remove the impulse noise from a corrupted image before applying any subsequent image processing technique to an image. If the impulse noise denoted by n distributed uniformly corrupts an original image f , then the image \hat{f} with impulse noise can be mathematically modeled as below.

$$\hat{f}_{i,j} = \begin{cases} n_{i,j} & \text{with probability } p \\ f_{i,j} & \text{with probability } (1 - p) \end{cases} \quad (1)$$

Last few decades have observed many attempts made to remove the impulse noise from noisy images. Standard median filter (SMF) [1] is one of the commonly used filter techniques for the removal of salt and pepper noise. In this filtering technique, the noisy center pixel of a filtering window is replaced by the median value of its neighborhood pixels of the same window. SMF is simple and efficient to remove the salt and pepper noise from an image. This technique works better when the noise density is low. However, in case of high-density noise, the filter fails to remove the noise completely from an image. Even at low noise density, SMF blurs the image and removes thin lines while removing the noise from the image. There are several derivatives of the SMF known as weighted median filter (WMF) [1] and adaptive median filter (WMF) [1]. These filtering techniques give different weights to different pixels in a particular filtering window. Though these techniques remove impulse noise better than SMF, they are unable to remove high-density noise from the image.

In order to overcome the drawbacks of median filters, several other methods have been proposed for the removal of salt and pepper noise. An improved fast peer group filtering (IFPGF) technique is proposed in [2] to remove the impulse noise by arithmetic mean filtering after detecting the corrupted pixels. The authors in [3] used a decision-based trimmed median filtering after considering all the pixels having 0 and 255 gray value as noisy pixels. However, pixels having 0 or 255 gray value are not necessarily noisy pixels. The method proposed in [4] uses fuzzy logic and improved

version of median filter to remove high-density salt and pepper noise from images. This technique removes all the noisy pixels from a filtering window and replaces the center noisy pixel by the median value of non-noisy pixels after using fuzzy logic to detect the noisy pixel. An efficient method is used in [5] to remove high-density impulse noise from images and videos. A decision-based technique is proposed in [6] for the removal of high-density salt and pepper noise from images. This technique first detects the noisy pixels and then removes it by using suitable filtering techniques. The methods used in [7, 8] use fuzzy filtering and median derivative filters to detect the noise and preserve the details of the image after filtering. A decision-based filtering technique depending on first-order neighborhood is proposed in [9] to restore the image details from impulse noise corrupted image. This paper gives emphasis on the detection of the noisy pixels using fuzzy logic. Next, it uses an adaptive technique to restore the original value of the image pixel using trimmed median filtering technique.

The rest part of the paper is organized as follows. An introduction to fuzzy logic is presented in Sect. 2. Section 3 presents the steps of proposed algorithm. Section 4 describes the results and discussions with peak signal-to-noise ratio (PSNR) graph of different test images followed by conclusion in Sect. 5.

2 Fuzzy Set Theory

Fuzzy set theory can generally be used as a classical set theory which has a membership value between zero and one [10]. This is a gradual transition between belonging to and not belonging to [6]. A fuzzy set S in the universe U is characterized by $U \rightarrow [0, 1]$ mapping a membership function μ_S . This membership function assigns every element x in U degree of membership $\mu_S(x) \in [0, 1]$ in the fuzzy set U . In this paper, the number of 0s and 255s in a particular window of the image defines the fuzzy membership function. The processing pixel in the selected window is defined by a function as follows.

$$S(x) = (S_0, S_{255}) \quad (2)$$

where S_0 and S_{255} are the number of 0s and number of 255s in the selected window. The membership function $\mu_S(x)$ of $S(x)$ is defined as follows.

$$\mu_S(x) = \begin{cases} \text{std}(U) & \text{if } S_0 \geq T_1 \\ \text{std}(U) \times \frac{S_0}{S_{255}} & \text{if } T_2 < S_0 < T_1 \\ \text{mean}(U) & \text{if } S_{255} \geq T_1 \\ \text{mean}(U) \times \frac{S_{255}}{S_0} & \text{if } T_2 < S_{255} < T_1 \end{cases} \quad (3)$$

where U is the number of neighborhood pixels, std is the standard deviation, and mean is the average of pixel values in the selected window. T_1 and T_2 are the predefined threshold values which depend on the number of 0s and number of 255s present in the selected window.

3 Proposed Method

It is observed that most of the filtering techniques fail to remove the impulse noise from the images at high density of noise. The SMF and its derivative filters give better result at low density of noise, i.e., at (20–30)%. However, when the noise density increases, these filters are not able to restore the original value of the corrupted pixel in the image. In addition to that, these filters produce blurring effect in the filtered image. Recent filtering techniques such as IFPGF and modified decision-based unsymmetric trimmed median filter (MDBUTMF) perform better up to 70% noise density. However, when noise density increases more than 70%, these filtering techniques unable to remove the noise completely from the image. At these high density of noise, these techniques also unable to restore the fine details of the image and produce patch like structure in the images. In this paper, fuzzy logic is used to detect the noisy pixels present in the image and a new adaptive trimmed median filtering technique is used to restore the original gray of the corrupted pixels. The proposed algorithm is described in details in Algorithm 1.

Algorithm 1: Algorithm of proposed method

- 1 Select a window of the corrupted image of size $W \times W$ (where initially W is 3). The center pixel of the window is the processing pixel which is denoted by $a_{i,j}$.
 - 2 If $0 < a_{i,j} < 255$, then $a_{i,j}$ is declared as a non-noisy pixel and the value of the pixel is kept unchanged.
 - 3 If $a_{i,j} = 0$ or $a_{i,j} = 255$ then $a_{i,j}$ is considered as a noisy pixel. Next, the noisy pixel is processed as follows.
 - (I) If all the neighborhood pixels in the selected window ($W \times W$) are not 0 and 255, then a number of pixels having values lying between 0 and 255 are counted. If the counted value is 60% or more of total pixels of window $W \times W$, the center pixel is replaced by the median value of the noise-free pixels. If the count is less than 60%, then the window size is increased and steps 1–3(I) are repeated. The window size is increased to a predefined size.
 - (II) If the value of all the pixels in the selected window are 0s and 255s, then there are four categories of noise density defined based on noise density using fuzzy rule. Those categories are very high, very low, low, and high. These four categories are defined based on two threshold values T_1 and T_2 . Then the corrupted center pixel is replaced by mean and standard deviation values as described in previous section.
 - 4 Steps 1–3 are repeated until the processing of all the pixels of the entire image is completed.
-

4 Result Analysis

The efficacy of proposed algorithm is validated by comparing the performance with other existing methods. The proposed algorithm is implemented with three different gray as well as color images namely Lena, Baboon, and Fish. The test images are of different size. The test images are shown in Fig. 1. Salt and pepper noise of 10–90% noise density is added to all the images before applying the proposed algorithm. The

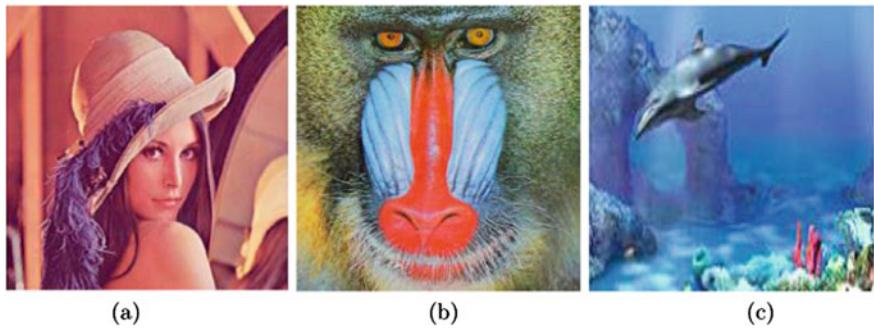


Fig. 1 a–c Represent the test images of Lena, Baboon, and Fish images, respectively

personal computer which is used for the experimentation purpose has the configuration of 2 GB RAM, 2.4 GHz, i3 processor, and 2 MB L2 cash. The performance of the proposed method is measured quantitatively by the peak signal-to-noise ratio (PSNR) which is defined in (2).

$$\text{PSNR}_{\text{db}} = 10 \log_{10} \frac{255^2}{\text{MSE}} \quad (4)$$

$$\text{MSE} = \frac{\sum_i \sum_j [f(i, j) - \hat{f}(i, j)]^2}{M \times N} \quad (5)$$

The PSNR of the filtered images using proposed algorithm is computed for all the gray and color test images, and the results are compared with the PSNR result of other existing methods such as SMF, IFPGF, and MDBUTMF for all noise densities ranging from 10 to 90%. The PSNR comparison result for gray and color images is shown in Tables 1 and 2, respectively. It is observed from Tables 1 and 2 that

Table 1 Comparison of PSNR of proposed method with different filtering techniques for gray test images

% of noise		10	20	30	40	50	60	70	80	90
Lena	SMF	53.28	51.15	48.23	44.91	41.14	37.69	34.99	31.81	29.92
	IFPGF	59.32	56.05	53.41	50.86	47.12	44.04	41.29	38.25	30.14
	MDBUTMF	59.86	56.52	54.04	50.99	47.73	44.56	41.85	38.88	31.39
	Proposed	66.52	62.38	58.84	54.59	52.04	49.93	46.29	42.88	41.56
Baboon	SMF	61.11	57.32	53.07	48.21	43.52	37.82	31.31	25.55	19.49
	IFPGF	64.09	61.52	58.22	54.14	50.56	45.54	39.33	30.81	20.34
	MDBUTMF	66.98	63.53	59.50	56.28	52.95	48.62	43.69	38.55	32.06
	Proposed	73.64	66.85	63.21	59.30	56.84	53.06	50.89	47.24	41.58
Fish	SMF	59.01	51.29	48.99	45.85	42.14	37.39	30.58	23.17	19.55
	IFPGF	63.05	56.03	50.58	48.08	44.15	40.53	34.73	27.19	20.70
	MDBUTMF	64.15	56.25	52.50	49.61	47.25	44.29	42.55	39.53	33.04
	Proposed	68.15	59.57	55.13	51.88	49.47	46.95	44.28	43.18	41.03

Table 2 Comparison of PSNR of proposed method with different filtering techniques for color test images

% of noise		10	20	30	40	50	60	70	80	90
Lena	SMF	47.82	44.51	43.03	40.99	38.14	35.69	33.63	30.81	28.52
	IFPGF	51.23	49.52	47.84	45.96	43.62	40.04	34.99	28.22	20.14
	MDBUTMF	51.63	50.20	48.64	46.96	45.33	43.56	41.48	38.38	32.39
	Proposed	68.62	61.33	57.14	53.97	51.41	49.13	46.97	44.82	42.26
Baboon	SMF	59.01	55.23	51.70	47.01	42.12	36.12	29.31	22.45	17.41
	IFPGF	64.90	60.25	56.25	52.47	48.06	42.74	35.38	28.01	19.43
	MDBUTMF	65.78	61.33	58.01	54.88	51.59	47.26	41.90	36.58	32.65
	Proposed	73.42	65.53	61.14	58.32	55.40	52.64	49.92	46.47	41.68
Fish	SMF	58.12	50.96	47.29	44.56	40.45	36.09	29.89	21.37	18.04
	IFPGF	61.35	54.03	49.88	46.88	43.52	39.13	33.37	26.99	19.27
	MDBUTMF	62.50	55.50	51.51	48.63	46.57	43.90	41.54	38.03	32.47
	Proposed	66.53	57.53	53.03	50.08	48.07	45.91	43.86	42.08	40.01

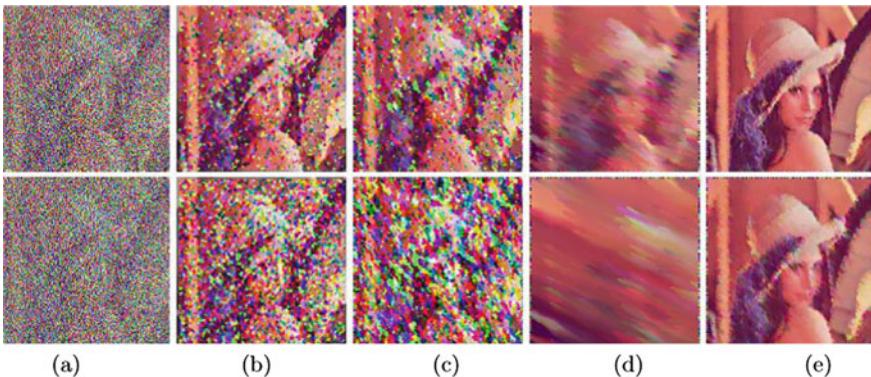


Fig. 2 **a** Lena image with 80 and 90% salt and pepper noise; **b–e** presents the outputs of SMF, IFPGF, MDBUTMF, and proposed method, respectively

though all the methods perform better in low density of noise, the proposed method outperforms all other existing methods in low as well as high density of impulse noise. The performance of the proposed method is also measured qualitatively by comparing the visual result of the proposed method with other methods. The comparison results are shown in Figs. 2, 3, and 4 for Lena, Baboon, and Fish images, respectively. Though the result is found for all density of noises, the visual comparison is shown for 80 and 90% noises. The first column of all the figures represents the noisy images corrupted by 80 and 90% impulse noise. Column 2, column 3, column 4, and column 5 represent the filtering images of SMF, IFPGF, MDBUTMF, and proposed method, respectively. It is observed from Figs. 2, 3, and 4 that the proposed method removes noise better than other existing methods. The proposed method is also able to preserve the fine details of all the test images.

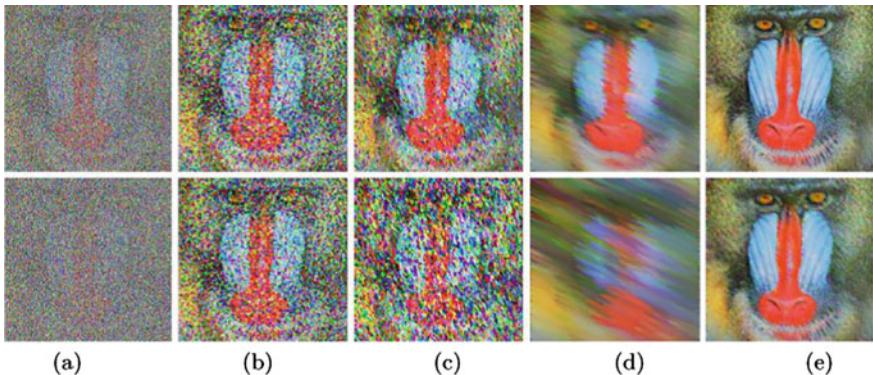


Fig. 3 **a** Baboon image with 80 and 90% salt and pepper noise: **b–e** presents the outputs of SMF, IFPGF, MDBUTMF, and proposed method, respectively

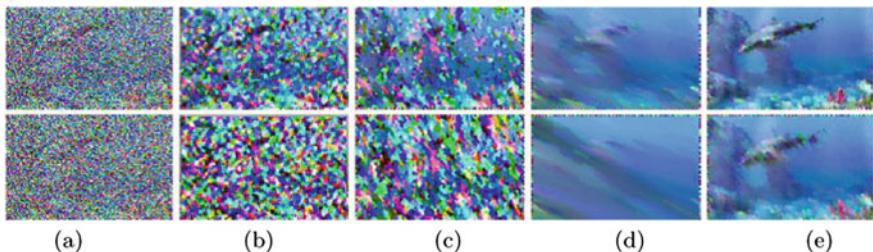


Fig. 4 **a** Fish image with 80 and 90% salt and pepper noise: **b–e** presents the outputs of SMF, IFPGF, MDBUTMF, and proposed method, respectively

5 Conclusion

In this paper, the main aim is to detect the impulse noise correctly in a noisy image and restore the original value of the corrupted pixel. This method overcomes the disadvantage of many existing filters that have been developed for this problem and performs better at low as well as high density of salt and pepper noise. In this work, an adaptive trimmed median filter with combination of fuzzy logic is proposed and developed for removal of impulse noise. First, the concept of fuzzy logic is applied to detect the extent of noise level in the corrupted image. In that, the proposed algorithm is conceptualized into two sections: (i) when all the pixels in the selected window are 0 or 255 and (ii) some pixels in the selected window are 0 or 255. It makes decision based on the number of the noise-free pixel in the selected window. In first case, the noisy pixel is replaced by mean and standard deviation of the pixels in the selected window. In later case, the noisy pixel is replaced by the median value of noise-free pixels in the selected window. The performance of the proposed method is tested in both gray-level and color images. It has been observed from the experimental results that the performance of the proposed filter gives better result than that of SMF, IFPGF, and MDBUTMF in terms of visual, PSNR, and MSE means.

References

1. Gonzalez RC, Woods RE (2013) Digital image processing, 3rd edn. Pearson Education
2. Camarena JG, Gregori V, Morillas S, Sapena A (2010) Some improvements for image filtering using peer group techniques. *Image Vis Comput* 28:88–201
3. Esakkirajan S, Veerakumar T, Subramanyam AN, Premchand CH (2011) Removal of high density salt and pepper noise through modified decision based unsymmetric trimmed median filter. *IEEE Signal Process Lett* 18(5):287–290
4. Veerakumar T, Esakkirajan S, Vennila I (2012) Combined fuzzy logic and unsymmetrical trimmed median filter for removal of high density impulse noise. *WSEAS Trans Signal Process* 8:32–42
5. Aiswarya K, Jayaraj V, Ebenezer D (2010) A new and efficient algorithm for removal of High density salt and pepper noise in images and videos. IEEE Computer Society, pp 409–413
6. Srinivasan KS, Ebenezar D (2007) A new fast and efficient decision based algorithm for removal of High density impulse noise. *IEEE Signal Process Lett* 15(3):189–192
7. Vellili DVD, Nachtegael M, Weken DVD, Kerre EE, Philips W (2003) Noise reduction by fuzzy image filtering. *IEEE Trans Fuzzy Syst* 11(4)
8. Chan RH, Ho CW, Nikolova M (2005) Salt and pepper noise removal by median type noise detectors and detail preserving regularization. *IEEE Trans Image Process* 14(10):1479–1485
9. Samantaray AK, Kanungo P, Mohanty B (2018) Neighbourhood decision based impulse noise filter. *IET Image Process* 12(7):1222–1227
10. Rajsekaran S, Pai GAV (2009) Neural networks, fuzzy logic and genetic algorithms synthesis and application. Prentice Hall India Learning Private Limited

Author Index

A

- Abja, Shaik Nakarikanth, 329
Acharya, Mousumi, 61
Ahmad, Tunku Salha Tunku, 283
Aishwarya, R., 83
Alugade, Vrushali, 273
Amiripalli, Shanmuk Srinivas, 565
Anandamurugan, S., 197
Anandhi, T., 265
Anil Kumar, D., 511
Anto Praveena, M. D., 335
Anusha, Diddi, 265
Ashokkumar, K., 127

B

- Bade, Kranti, 441
Baitharu, Tapas Ranjan, 419
Balabantray, Bunil Kumar, 61
Balasubramaniam, R., 371
Banerjee, Anasua, 363
Barik, Shiba Ch., 61
Behera, Anukampa, 497
Behera, Gopal, 295
Behera, Prafulla Kumar, 487
Bhamare, Mamta, 127
Bharathi, B., 335, 431
Bharti, P. K., 419
Bhatlu, Sridhar M., 223
Bhoi, Ashutosh, 295
Bhutia, Santosini, 349
Biswal, Chinmayee, 613
Borah, Samarjeet, 41, 253, 397
Brumancia, E., 265

C

- Chandel, Ashwani Kumar, 223
Chatterjee, Punyasha, 603
Clement, Nyior, 119

D

- Dangeti, Sujani, 565
Dansana, Debabrata, 487
Das, Smruti Rekha, 93
Das, Susanta Kumar, 511
Dash, Dillip Ku., 3
Dash, Ritesh, 3
Debnath, Narayan C, 603
Deenadhyalan, R., 197
Desai, Darshana, 211
Dhal, Sunil Kumar, 519
Drelichowski, Ludoslaw, 161
Dutta, Anjan, 603

G

- Gomathi, R. M., 265

H

- Harish, G. V., 431
Harsha, Sai, 545

I

- Ikponmwosa, Aimufua, 51
Indria, P., 265

J

- Jaglan, Vivek, 407
 Jaidhan, B. J., 533
 Jayanthi, R., 75
 Jitendra, Mukkamala S. N. V., 533
 Johnson, Sandra, 83

K

- Kadam, Ruchira, 147
 Kadu, Rajesh, 441
 Karuppasamy, K. M., 371
 Kashyap, Anshuman, 613
 Kaul, Adarsh, 147
 Kaul, Kapeesh, 343
 Kaviya, S. M., 75
 Kumar, Prankul, 189
 Kumar, Priyaranjan, 613
 Kumar, Rahul, 253

L

- Laha, Suprava Ranjan, 137
 Longe, Olumide, 51

M

- Madhumathi, J., 83
 Mallikarjuna Rao, G. S., 565
 Mary Posonia, A., 329
 Minu, R. I., 463
 Mishra, Anil Kumar, 11
 Mishra, Bibhuti Bhusan, 19
 Mishra, Brojo Kishor, 407
 Mishra, Jyoti Prakash, 11, 41, 93, 233
 Mishra, Sambit Kumar, 11, 41
 Mishra, Suchismita, 19
 Mishra, Sujogya, 283
 Mohapatra, Aparna, 137
 Mohapatra, Shakti Ranjan, 101
 Mohapatra, Soumya S., 61
 Mohile, Sanjana, 147
 Mohmaed, Aezeden, 283
 Muduli, Kamalakanta, 283

N

- Naga Praveen Kumar, N., 555
 Nagarajan, G., 463, 545, 555
 Nagaraju, V., 591
 Nagendram, S., 29
 Nagwanshi, Kapil Kumar, 397
 Nand, Parma, 473
 Nimmagadda, Sudesh, 303

P

- Panda, Binayak, 449
 Panda, Gopikrishna, 519
 Panda, Radhe Shyam, 3
 Panigrahi, Chhabi Rani, 497
 Panigrahi, Ranjit, 253
 Panigrahy, Satyajit, 223
 Pani, Rojalin, 241
 Pani, Subhendu Kumar, 419, 519
 Patel, Namrata, 441
 Pathak, Sunil, 397
 Pati, Bibudhendu, 497
 Patil, Rutuja, 273
 Patil, Varsha, 273, 441
 Patnaik, Preetishree, 407
 Patnaik, Srikanta, 19, 137
 Patra, Bichitrana, 349
 Patra, Sai Rashmi, 101
 Pattanaik, Bikash Chandra, 241
 Pattanayak, Binod Kumar, 137, 233
 Pattnaik, Pradyumna Kumar, 283
 Pattnaik, Saumendra, 137, 241
 Pavan Satish, J., 545
 Polkowski, Zdzislaw, 41, 161
 Prabu, S., 591
 Pradhan, Akanksha, 273
 Pramod Sai, P., 555
 Prasanth, G. S. R. P., 431
 Pravin, A., 545
 Prem Jacob, T., 545
 Priyanka, Diddi, 265

R

- Rajesh, S., 197
 Rakshit, Sandip, 51, 119
 Rangarajan, Lalitha, 177
 Reddy, Bhumireddy Sohith, 575
 Reddy, Veeram Deepak, 303
 Rishi, Tadavarthi, 303
 Rout, Minakhi, 363

S

- Sahoo, Bidush Kumar, 233, 241
 Sahoo, Ladu K., 61
 Sahoo, Manoj Kumar, 407, 511
 Sahoo, Smruti Rekha, 233
 Sai Jayanth, S., 329
 Sai, Karlapudi Rahul, 575
 Sakthivel, S., 197
 Samantaray, Aswini K., 613
 Saravanan, M., 75

Sasikala, T., 463
Satpathy, Rabinarayan, 519
Sen, Soumya, 603
Senthamarai Kannan, K., 371
Shanmuk Srinivas, A., 533
Sharma, Pankaj, 473
Sharma, Shubham, 473
Singh, Needhi Kumari, 253
Sinha, Vaibhav, 603
Sivakumar, V. G., 313, 321, 585
Soam, Anubhav, 343
Sohani, Mayank, 397
Sree Ram Kiran Nag, M., 29
Sridevi, T. N., 177
Sridhar, S., 591
Srinivas, G., 29
Sujihelen, L., 555
Switek, Slawomir, 161

T

TapasBapu, B. R., 591
Thilagavathy, A., 303
Thota, Jyotsna Rani, 533
Tripathy, Satya Narayan, 449

U

Upadhyaya, Shweta, 189
Ushasukhanya, S., 189, 343

V

Vadivel, M., 313, 321, 585
Vajjhala, Narasimha Rao, 51, 119
Vakkalanka, Sairam, 29
Vanitha, L., 385
Vedha Pavithra, V., 83
Venkata Praneel, A. S., 533
Venkata Rao, K., 29
Venkatesan, B., 197
Venmathi, A. R., 385
Verma, Upendra, 397
Vijaya Baskar, V., 313, 321, 585
Vijayakumar, S., 575
Vimal, S. P., 313, 321, 585

W

Wankhade, Sunil, 147

Y

Yolmo, Passang Uden, 253